MATHÉMATIQUES ET SCIENCES HUMAINES

A. GUENOCHE
B. MONJARDET

Méthodes ordinales et combinatoires en analyse des données

Mathématiques et sciences humaines, tome 100 (1987), p. 5-47

<http://www.numdam.org/item?id=MSH_1987__100__5_0>

Math. Sci. hum. (25^{e} année, n°100, *1987*, p.5-47)

METHODES ORDINALES ET COMBINATOIRES EN ANALYSE DES DONNEES

A. GUENOCHE * et B. MONJARDET **

In the foreseable future, discrete mathematics will be an increasingly useful tool in the attempt to understand the world.
P.R. Halmos, 1981.
(*Applied Mathematics is bad Mathematics,* Mathematics Tomorrow, Springer Verlag)

Quite probably, with the development of the modern computing technique, it will be clear that in very many cases it is reasonable to conduct the study of real phenomena avoiding the intermediary stage of stylising them in the spirit of ideas of mathematics of the infinite and the continuous, and passing directly to discrete models.
A.N. Kolmogorov, 1970.
(International Congress of Mathematicians in Nice)

Résumé
Après quelques considérations générales sur les relations entre les mathématiques discrètes, l'informatique et l'analyse des données, ce texte présente un ensemble de méthodes utilisant des techniques ordinales ou (et) combinatoires. A une description succinte de chaque méthode sont jointes quelques références relatives à ses aspects théoriques ainsi qu'à ses implémentations accessibles aux utilisateurs. Pour présenter ces méthodes nous les avons classées suivant la nature des tableaux de données qu'elles permettent de traiter.

Abstract
In this paper first we formulate some remarks on relations between discrete mathematics, computer sciences and data analysis. Then we present a set of methods using ordinal or combinatorial techniques. Boolean analysis and graph theory approach for binary data, tree representations (additive trees, n-trees), seriation methods for symmetric or not symmetric dissimilarity arrays, preferences aggregation procedures and consensus problems are studied. For each method we give a brief description and some bibliographic references concerning theoretical aspects and description of algorithms. Methods are classified according to the structures of the data processed: one or several arrays of type **I** x **J** or **K** x **K**, with binary, ordinal or quantitative values.

* G.R.T.C. - C.N.R.S., 31 Ch. J. Aiguier, 13402 Marseille Cedex 9
** Université Paris V et C.A.M.S., 54 Bd. Raspail, 75270 Paris Cedex 6

Sommaire

---

1 MATHEMATIQUES DISCRETES, INFORMATIQUE ET ANALYSE DES DONNEES

En se démarquant des modèles probabilistes et de la statistique inférentielle classique, l'analyse des données a d'abord fait appel à l'algèbre linéaire et à la géométrie (euclidienne ou ultramétrique) pour développer ses formalismes et ses méthodes. Parallèlement depuis quelques années apparaissent de nombreuses méthodes qui relèvent des mathématiques des structures finies, qu'il s'agisse de structures algébriques (treillis, monoïdes,..) ou combinatoires (ordres, graphes, hypergraphes,..), méthodes constituant ce qui a été appelé par Arabie (1982) la "combinatorial data analysis".

A cette double approche méthodologique on peut faire correspondre une bi-partition des méthodes un peu arbitraire; en effet ces deux classes ne sont pas sans intersection (méthodes d'optimisation, plans d'expériences, valuation des arbres taxonomiques,..). Elle a néanmoins l'avantage de souligner le rôle des mathématiques discrètes en analyse des données, donc l'un de ses rapports essentiels avec la recherche en informatique: la définition d'algorithmes performants sur des structures finies et donc l'étude de leurs propriétés.

Si l'on s'interroge sur les raisons de cette évolution, on peut d'abord remarquer que des structures discrètes et/ou des problèmes combinatoires découlent naturellement de certains modèles ou méthodes d'analyse des données, notamment dans:

. La formalisation de données obtenues ou de modèles cherchés (comparaisons par paires, questionnaires, réseaux sociaux, n-uplets de variables ordinales, modèles typologiques ou ordinaux tels les hiérarchies, les échelles de Guttman ou les modèles de sériation), ainsi que la construction de plans d'expériences pour le recueil de certaines données.

. Les traitements, qu'il s'agisse d'opérer sur des données ou avec des modèles combinatoires du type ci-dessus, ou d'utiliser des méthodes combinatoires en analyse des données, par exemple lorsque des données quantitatives sont discrétisées (graphes seuils, préordres de relèvement sur les marges d'un tableau), ou qu'une structure discrète est recherchée par optimisation d'un critère (problème d'ajustement et de résumé) ou par des "manipulations" combinatoires (permutation de lignes et/ou colonnes, chemins dans des graphes..).

De tels traitements ont été développés au niveau international par divers chercheurs dans des domaines variés (biologie, écologie, psychométrie, économétrie, sociométrie, archéologie, anthropologie, histoire, géographie). Souvent il s'agit de rester "proche" des données recueillies, en évitant des codages numériques ramenant à des techniques statistiques usuelles, au prix d'un arbitraire parfois difficilement justifiable, ou simplement de remarquer que certains problèmes classiques (recherche d'une typologie ou d'une sériation), typiquement discrets, peuvent être étudiés par des méthodes purement combinatoires.

Si actuellement les domaines d'application relèvent plus fréquement des sciences humaines ou sociales que des sciences de la nature ou de la vie, il faut sans doute attribuer ce fait aux difficultés plus grandes que ces disciplines ont à décrire numériquement leurs problèmes (par exemple, il n'y a pas de quantification "naturelle" d'un problème d'évolution des formes d'objets), et aux réticences qu'elles ont à les plonger dans un espace euclidien.

Sans vouloir faire un bilan au niveau international des apports de l'Analyse Combinatoire des Données, on peut signaler quelques unes de ses lignes directrices :

.Méthodes spécifiques : Par exemple les analyses booléennes ou galoisiennes d'une correspondance, les méthodes basées sur la théorie des graphes en taxonomie ou en sériation.

.Comparaisons de structures : Un certain nombre de travaux concourent à une théorie de la construction d'indicateurs de ressemblance entre données de nature combinatoire. La définition, la caractérisation axiomatique et la construction effective de métriques sur des ensembles ordonnés (ou des graphes) étant au cœur du sujet.

.Ajustements de structures discrètes (les "discrete models" des ouvrages de Harary et al., Roberts,..) à un certain type de données; comme cas typique, citons l'ajustement d'un ordre total à une relation de tournoi, valué ou non.

.Résumés de structures : Là aussi, on progresse vers une "théorie de l'agrégation" qui a notamment pour intérêt d'étudier les rapports entre les différentes méthodes d'obtention de consensus : construction algébrico-combinatoire (opérations généralisées de moyenne et de médiane), optimisation de critères (résumé à distance minimum), approche axiomatique.

.Représentations de structures : Les développements technologiques de l'informatique (écrans graphiques de meilleure définition) ont rendu réalisables les représentations graphiques des structures modélisatrices de données (arbres, treillis, graphes, hypergraphes..).

On terminera ce rapide tour d'horizon en signalant que l'Analyse Combinatoire des Données suscite des recherches nombreuses conduisant à des travaux de trois types difficilement dissociables:

.Mathématiques, parce que la formalisation de problèmes et de données conduit à des situations nouvelles, points de départ de recherches spécifiques (structure des ultramétriques ou des hiérarchies, propriétés des relations médianes, énumération de fonctions logiques ou de configurations particulières,..).

.Informatiques, parce que la nécessité de disposer pour ces méthodes d'implémentations utilisables, pour traiter des données souvent en nombre important, conduit à définir des algorithmes originaux et des programmes efficaces. De plus l'étude de la complexité de ces algorithmes joue un rôle important, qu'il s'agisse d'énumérer certaines structures discrètes, de construire le treillis de Galois d'une correspondance ou de chercher des relations à distance minimum de relations données. Bon nombre de ces problèmes sont reconnus comme NP-difficiles (au sens de la théorie de la complexité algorithmique), et dans la lignée de recherches récentes en théorie des graphes, il s'avère particulièrement intéressant d'étudier et de caractériser les cas particuliers dans lesquels cette complexité devient polynomiale.

.Méthodologiques, parce que la définition d'une méthode d'analyse de données, souvent issue d'un problème concret, ne peut éluder les questions de l'applicabilité, de la pertinence et de la généralisabilité de cette méthode pour traiter une classe spécifique de problèmes. On peut remarquer à ce propos que du point de vue méthodologique, les méthodes présentées ci-dessous peuvent être classées en trois grandes catégories :
Celles qui opèrent une transformation "réversible" sur les données, en ce sens que celles-ci peuvent être totalement reconstituées (par exemple, réordonnancement des lignes et des colonnes d'un tableau, hiérarchie des graphes seuils d'une dissimilarité, treillis de Galois étiqueté d'une correspondance binaire).
Celles qui réduisent les données à un modèle facilement interprétable par des procédures de résumés ou d'ajustements (par exemple, ordre médian de relations de préférence, partition centrale de plusieurs partitions).
Les méthodes mixtes combinant les deux démarches (par exemple, échelle de Guttman associée à une chaîne maximale du treillis de Galois, hiérarchie de préordres d'implications entre variables booléennes).

2 INTRODUCTION AUX METHODES ORDINALES ET COMBINATOIRES

Rappelons que nous entendons par méthodes ordinales et combinatoires les méthodes qui mettent en jeu des structures "essentiellement" non numériques que l'on peut définir sur un ensemble fini, à l'opposé des méthodes, souvent qualifiées de linéaires, qui s'appuient sur la structure linéaire d'un espace euclidien.

Ces méthodes se présentent à des niveaux variables allant d'une définition purement théorique dans un article, à une implémentation effective sur ordinateur. Dans l'inventaire nullement exhaustif que l'on trouvera ci-dessous, on a privilégié les méthodes dont on est sûr qu'il existe au moins une implémentation. Cette dernière assertion recouvre toutefois des réalités très variées, allant de méthodes largement répandues à celles où pratiquement l'auteur seul dispose du programme correspondant à la mise au point de son algorithme.

Font partie du premier cas quelques méthodes classiques en analyse des données, mais présentes ici parcequ'elles relèvent d'une approche combinatoire (par exemple, la méthode du lien simple en classification); on les trouve donc dans bon nombre de logiciels d'analyse de données. Un cas voisin est celui des méthodes qui utilisent essentiellement des notions classiques de théorie des graphes, telles que la recherche des cliques maximales. Il faut toutefois noter qu'il n'existe pratiquement pas, du moins en France, l'équivalent pour le traitement des graphes ou plus généralement des problèmes combinatoires de ce que sont les nombreux logiciels disponibles en statistique ou en analyse de données. Signalons toutefois deux exceptions : le projet L.E.G. à l'université de Bordeaux I (orienté vers la conception d'un langage spécifique), et les divers projets CABRI au L.S.D. - C.N.R.S de Grenoble, à l'Ecole des Mines de Saint Etienne et l'E.N.S.T. de Brest (projets orientés vers l'aide à la recherche et à l'enseignement). Signalons aussi à l'étranger le logiciel GRADAP au Technisch Centrum de l'Université d'Amsterdam, qui comprend de nombreuses procédures d'analyse d'un graphe ou d'un réseau. Quant à l'implémentation des méthodes ordinales et combinatoires, nous ne connaissons qu'un logiciel consacré à ces méthodes : A.B.C.D. (Analyses Booléennes et Combinatoires de Données) développé par l'un des auteurs. On trouvera en annexe un résumé descriptif des programmes disponibles.

Chaque méthode présentée ci-dessous est répertoriée par un titre suivi d'un bref texte explicatif et de références. On a privilégié celles qui, outre l'exposé de la méthode, contiennent la description précise d'une procédure pour la mettre en œuvre, voire la description d'une implémentation. On a toutefois signalé, lorsqu'ils existent, des articles ou ouvrages de référence ou de synthèse permettant d'avoir une vue plus générale sur une méthode ou un ensemble de méthodes et en particulier d'en connaitre les auteurs initiaux.

Pour faciliter le repérage des méthodes on les a classées suivant la structure des données qu'elles peuvent traiter. On a distingué trois catégories principales :
-Le tableau de données est de la forme **I** x **J** , où | **I** | "individus" sont "décrits" par | **J** | "variables".
-Le tableau de données est de la forme **K** x **K** c'est à dire correspond à une relation binaire (valuée ou non) sur un ensemble **K** d'objets.
-On a plusieurs tableaux de données de la forme **I** x **J** ou **K** x **K**.

Les cases des tableaux contiennent toujours des nombres et en ce sens ces tableaux sont tous numériques. Mais ces nombres peuvent avoir des significations différentes. Il peuvent être de simples codes désignant diverses modalités d'une variable qualitative; en particulier si la variable a deux modalités -on dit aussi une *variable binaire*- on utilise très souvent un codage en 0-1; le tableau de données correspondant sera alors appelé *tableau 0-1*. Ces derniers se présentent aussi bien sous la forme **I** x **J** (description d'individus par des attributs dichotomiques, réponses vraies ou fausses de sujets à des questions), que sous forme **K** x **K** (présence ou absence d'un couple (i,j)). Dans d'autres cas les valeurs numériques du tableau correspondent à des rangs qu'un individu peut occuper sur une échelle ordinale; on dira alors qu'on a une *variable ordinale*. Enfin les valeurs numériques peuvent correspondre à des mesures sur une échelle d'intervalles; nous dirons qu'il s'agit d'une *variable quantitative*.

Ces distinctions permettent de raffiner la taxonomie précédente des tableaux de données et d'obtenir ainsi une classification commode pour la présentation des

méthodes. Pour des classifications plus élaborées, on se reportera notamment à Coombs (1964), Shepard (1972) ou à Caroll et Arabie (1981).

3 METHODES TRAITANT UN TABLEAU **I** x **J**

Pour un tel tableau, noté aussi T, on appelle individu tout élément de **I** et variable tout élément de **J**. On distingue le cas où toutes les variables sont binaires, de celui où elles sont toutes ordinales ou quantitatives.

3.1 Variables binaires (tableau 0-1)

Par variable binaire rappelons qu'on entend toute variable à deux modalités. On notera **J** = {A, B, C, .. } l'ensemble des variables, chacune (par exemple A) pouvant prendre les valeurs a (codée 1) et a' (codée 0). Le tableau de données est donc un tableau 0-1. Pour décrire et traiter un tel tableau nous utiliserons une terminologie et des notions empruntées soit à l'analyse des questionnaires (en sociologie) soit à l'algèbre de Boole. Nous commençons donc par présenter ces deux terminologies ainsi que leurs relations.

Dans la terminologie de l'analyse des questionnaires, où T contient les réponses d'individus à des variables binaires (questions dichotomiques), la réponse de l'individu i, c'est à dire la i-ème ligne du tableau notée t_i, est appelée le patron (de réponse) de i. Plus généralement, un *patron* t_i est une suite t_{ij} indexée par **J**, de termes égaux à 0 ou à 1; il correspond à une ligne -réelle ou virtuelle- du tableau de données. L'ensemble de tous les patrons possibles est noté $2^{\mathbf{J}}$. Un *sous-patron* d'un patron est une sous-suite de celui-ci. Deux patrons t_i et t_k peuvent être ordonnés par le *préordre de dominance* (des lignes) : $t_i \leq t_k$ ssi $t_{ij} \leq t_{kj}$ pour tout j. L'*intersection* de deux patrons t_i et t_k est le patron défini par la suite (min (t_{ij}, t_{kj}), $j \in \mathbf{J}$); l'*union* de ces deux patrons est défini par la suite (max (t_{ij}, t_{kj}), $j \in \mathbf{J}$). On définit de même un préordre de dominance et des opérations d'intersection et d'union pour les colonnes du tableau T.

Un tableau est dit *réduit* si on ne considère que le sous-tableau de ses patrons distincts; en conservant l'effectif des patrons identiques on obtient un tableau pondéré. Dans la suite nous supposerons toujours le tableau réduit (pondéré ou non); dans ce cas, l'ensemble des patrons du tableau pouvant être identifié à l'ensemble **I** des individus, nous le désignerons également par **I**.

Dans la terminologie de l'algèbre de Boole, on pose J = {a, b, ..j, .. }, J' = {a', b', ..j', .. }, et $J^* = J \cup J'$ est l'ensemble des 2.| J | "attributs" associés aux | J | variables. On appelle *monôme de longueur* k tout sous-ensemble à k éléments de J^* où chaque variable apparaît au plus une fois (sous forme non accentuée, dite *forme directe*, ou sous forme accentuée, dite *forme complémentée*); un tel monôme est noté par la concaténation des éléments qui le constituent. Par exemple μ = ab'd est un monôme de longueur 3. Un monôme est *complet* s'il utilise toutes les variables, donc s'il est de longueur | **J** |. On remarque que l'ensemble des $2^{|\mathbf{J}|}$ monômes complets correspond à l'ensemble des atomes de "l'algèbre de Boole libre" engendrée par les variables; celles-ci sont les *générateurs* de cette algèbre (cf. par exemple les ouvrages de Birkhoff et Bartee, Flegg, ou Kuntzmann cités dans la bibliographie).

La relation entre patrons et monômes est la suivante : A tout monôme complet, correspond le patron obtenu en donnant la valeur 1 (resp. 0) à toute variable apparaissant sous forme directe (resp. complémentée). A un monôme non complet on fait correspondre le sous-patron obtenu comme ci-dessus mais en ne considérant que les variables présentes dans le monôme. Inversement à tout patron ou sous-patron on fait correspondre le monôme dont les variables accentuées (resp. non accentuées) correspondent aux 1 (resp. 0). Si dans un patron t_i on a $t_{ij} = 1$ (resp. 0) on dit aussi que le patron *possède l'attribut* j (resp. j').

Soit T un tableau et μ un monôme. On note $T(\mu)$ l'ensemble des patrons de T admettant comme sous-patron le sous-patron correspondant à μ; par exemple T(a'c) est l'ensemble des patrons de T ayant 0 en première et 1 en troisième position. On note $n_T(\mu)$ ou plus simplement $n(\mu)$, le nombre $| T(\mu) |$ de tels patrons. On dit que le monôme μ implique le monôme ν si $T(\mu) \subseteq T(\nu)$ et dans ce cas on note $\mu \rightarrow \nu$.

. Enumération des monômes vides

Un monôme μ est dit *T-vide*, ou simplement *vide*, si $T(\mu)$ est vide. Si μ est de longueur 2 on parle aussi de *case vide*. L'existence d'un monôme vide peut s'interpréter de différentes manières, équivalentes, sous forme d'implications entre les monômes qu'il contient. Par exemple ab' vide (qui signifie que pour aucune ligne du tableau on ne trouve la valeur 1 pour A et 0 pour B) peut s'interpréter comme $a \rightarrow b$ (en effet lorsqu'un patron de T possède l'attribut a, il possède également b, puisque l'on n'a jamais ab'), ou comme $b' \rightarrow a'$. De même ab'c vide peut s'interpréter comme $ac \rightarrow b$ ou $b'c \rightarrow a'$ ou encore $a \rightarrow b \vee c'$ (a implique b ou c') etc ..

Considérons l'ensemble des monômes vides de longueur 2 de type ab' (ou a'b), c'est à dire avec une seule forme complémentée. En associant à ces monômes les implications $a \rightarrow b$ (ou $b \rightarrow a$), on obtient une relation de préordre sur J. Ce préordre d'implications entre variables est le dual de leur préordre de dominance.

Plusieurs algorithmes ont été proposés pour l'énumération des monômes vides d'un tableau T; certains sont décrits par Flegg ou Flament. Le meilleur (car il se prête à de nombreuses adaptations) reste celui de Kuntzmann.

Une méthode de construction de familles minimales d'implications entre monômes, dans le cas où on ne s'intéresse qu'aux formes directes des variables, fortement dépendante du treillis de Galois (cf. infra) correspondant à T, sera décrite plus loin.

FLAMENT Cl., *L'analyse booléenne de questionnaire*, Mouton, Paris, 1976.
FLEGG H.G., *L'algèbre de Boole et son utilisation*, Dunod, Paris, 1967.
KUNTZMANN J., NASLIN P., *Algèbre de Boole et Machines Logiques*, Dunod, Paris, 1967.

. Graphes d'implications

A deux variables A et B, on associe les quatre monômes de longueur 2 : ab, ab', a'b, a'b'. Les effectifs $n_T(\mu)$ de ces quatre monômes forment le tableau croisé classiquement associé aux deux colonnes A et B du tableau T. Pour chacun de ces monômes, par exemple μ = ab, on calcule les quotients $n_T(ab) / n_T(a)$ et $n_T(ab) / n_T(b)$.

Ils représentent le pourcentage de cas où dans T on observe les implications a → b et b → a. A un seuil s fixé de ce pourcentage, on fait correspondre l'ensemble des implications pour lesquelles leur quotient est supérieur à ce seuil. On définit ainsi sur l'ensemble J^* de sommets, un graphe d'implications au seuil s, dont on extrait un graphe transitif maximal. En faisant varier le seuil s, on obtient une famille de préordres emboîtés.

GUENOCHE A., Fonctions booléennes sur un tableau en 0/1, *Data Analysis and Informatics 4*, DIDAY E. et al. Eds., North Holland, Amsterdam, 1986, p. 443-451.

. Couvertures minimales

Etant donné un tableau T, appelons T' le tableau formé de tous les patrons qui n'apparaissent pas dans T. Si un monôme μ est T'-vide, $T(\mu)$ est non vide et définit une partie de **I** que l'on dit *couverte* par μ. Un ensemble $\{\mu_1, .. \mu_k\}$ de monômes tels que :

$$\cup_{i=1,..k} T(\mu_i) = \mathbf{I}$$

est une *couverture* de **I**. Il existe différents critères permettant de dire qu'un tel ensemble est minimal; de plus pour le même critère il peut exister plusieurs ensembles minimaux couvrant **I**. La recherche de tels ensembles, pour l'expression booléenne obtenue en considérant tous les monômes T'-vides, correspond à ce qui est appelé en algèbre de Boole la recherche de formes normales disjonctives minimales (par rapport au critère choisi) d'une expression booléenne. Construire une couverture minimale est un problème NP-difficile.

Les méthodes de minimisation d'une expression (ou fonction) booléenne sont utilisées dans le cadre de l'analyse booléenne de questionnaire présentée ci-dessous, ainsi que dans les méthodes développées par Ledley.

Un cas particulier intéressant est celui où l'on cherche une couverture minimale d'un sous-tableau de T (par exemple T(a)). Si l'on choisit des monômes T(a')-vides pour réaliser cette couverture, leur réunion constitue une fonction discriminante entre T(a) et T(a'), c'est à dire que cette fonction prend la valeur 1 (resp. 0) ssi un patron appartient à T(a) (resp. T(a')).

BIRKHOFF G., BARTEE T., *Modern Applied Algebra*, Mc. Graw-Hill, New York, 1967.
GUENOCHE A., Propriétés caractéristiques d'une classe relativement à un contexte, *Actes des Journées "Symbolique numérique"*, Paris, Décembre 1987.
KAUFMANN A., PICHAT E., *Méthodes mathématiques non numériques et leurs algorithmes*, 2 tomes, Masson, Paris, 1977.
KUNTZMANN J., *Algèbre de Boole* , Dunod, Paris, 1968.
LEDLEY R., Digital electronic computers in biomedical sciences, *Science*, 130, 1959, p. 1225-1234.

. Analyse booléenne de questionnaire

Elle a été définie dans un contexte où **J** représente un ensemble de questions à réponses dichotomiques. On construit à partir de T et d'une valeur $\alpha \geq 1$, le tableau réduit $T_\alpha = \mathbf{I}_\alpha \times \mathbf{J}$ formé de tous les patrons présents au moins α fois dans T. On cherche ensuite un ensemble minimal de monômes T'_α-vides, couvrant $\mathbf{I}_\alpha$, le critère retenu pouvant être de minimiser la longueur de ces monômes. Un tel ensemble minimal est appelé ensemble de *projections canoniques ultimes* (p.c.u.) et peut

s'interpréter (de différentes façons) comme un ensemble d'implications entre des réponses à des questions ou groupes de questions. Le choix de α, laissé à l'utilisateur, peut être guidé par la facilité d'interpréter les p.c.u, facilité dépendant de leur longueur. En particulier si toutes les p.c.u. sont de longueur ≤ 2, elles s'interprètent comme un préordre d'implications entre les réponses (a ou a') aux questions. Dans son article Van Buggenhaut cherche le seuil minimal pour lequel on se trouve dans cette situation.

DEGENNE A., *Techniques ordinales en analyse des données: Statistique*, Hachette, Paris, 1972.
FLAMENT Cl., *L'analyse booléenne de questionnaire*, Mouton, Paris, 1976.
VAN BUGGENHAUT J., Questionnaires booléens : schéma d'implications et degré de cohésion, *Math. Sci. hum.*, 98, 1987, p. 9-20.

. Graphe médian

Soit T un tableau ne contenant pas deux colonnes *complémentées* i.e. où les 1 (resp. 0) d'une colonne correspondent aux 0 (resp. 1) de l'autre. On lui associe un graphe non orienté appelé *graphe médian* de T; ses sommets sont les monômes complets qui ne contiennent aucun monôme T-vide de longueur 2; deux sommets sont reliés par une arête s'ils ne diffèrent que par une variable (par exemple si $| \mathbf{J} | = 4$, ab'c'd et ab'cd sont liés). Les sommets de ce graphe correspondent aux patrons "réels" présents dans T ainsi qu'à des patrons *latents*, rajoutés dans la mesure où ils n'introduisent pas de combinaisons nouvelles des valeurs de deux variables. On montre que ces patrons latents sont ceux engendrés par itération de l'opération "médiane" (opération qui à 3 patrons associe l'union de leurs intersections deux à deux) sur les patrons réels, et que le graphe obtenu est connexe et médian. De plus, par construction, la distance géodésique (nombre d'arêtes d'une plus courte chaîne) entre deux sommets de ce graphe est égale à la distance de la différence symétrique entre les deux patrons correspondants.

BARTHELEMY J.P., From copair hypergraphs to median graphs with latent vertices, *Annals of discrete maths.*, 1988.
BARTHELEMY J.P., GUENOCHE A., *Les arbres et les représentations des proximités*, Masson, Paris, 1988.

. Graphe distributif

On associe au tableau T un graphe non orienté appelé graphe distributif de T. L'ensemble des sommets de ce graphe est l'ensemble des monômes complets ne contenant aucun monôme T-vide de la forme ab' (ou a'b); c'est un sur-ensemble des sommets du graphe médian. De même que dans ce dernier, deux sommets sont reliés par une arête s'ils ne diffèrent que sur une variable. Les sommets de ce graphe correspondent donc aux patrons effectivement présents dans T ainsi qu'à tous les patrons rajoutés dans la mesure où ils n'introduisent pas de combinaisons nouvelles de valeurs de variables du type 10 (ou 01). On montre que ces patrons rajoutés sont ceux obtenus à partir des patrons présents par itération de deux opérations union et intersection de patrons. Le graphe obtenu est connexe, contient comme sous graphe le graphe médian précédent (lorsque celui-ci est défini), et peut être orienté pour constituer le diagramme d'un treillis distributif.

Ce treillis distributif correspond au treillis distributif des "parties finissantes" de l'ensemble **J** ordonné par l'ordre de dominance entre variables; dans le contexte de l'analyse de questionnaire, il est aussi appelé *tresse de Guttman.*

MONJARDET B., Tresses, fuseaux, préordre et topologies, *Math. Sci. hum.*, 30, 1970, p. 11-22.

. Graphe de Buneman

Au tableau T = **I** x **J**, on associe un graphe dit *bunemanien* de T. Les sommets sont les éléments de **J** et deux colonnes sont reliées par une arête si le tableau croisé correspondant contient au moins une case vide. Si le bunemanien de T est un graphe complet, alors le graphe médian de T est un arbre et tous ses sommets sont des sommets réels. Sinon, soit **C** $\subseteq$ **J** une clique maximale de ce graphe (cf. § 4.1.1.); on lui associe un arbre dont les sommets sont les ensembles de patrons du sous-tableau **I** x **C**; les arêtes correspondent aux variables présentes dans **C** et sont définies comme pour le graphe médian. Cet arbre est tel que la suppression d'une arête produit une bipartition des individus, bipartition associée à la variable correspondante (individus ayant la valeur 0 dans une classe, ceux ayant la valeur 1 pour cette variable dans l'autre). Le recouvrement de **J** par une famille de cliques constitue une *décomposition arborée* de T.

BARTHELEMY J.P., GUENOCHE A., *Les arbres et les représentations des proximités*, Masson, Paris, 1988.
BUNEMAN P.,The recovery of trees from measures of dissimilarity, *Mathematics in Archaeological and Historical Sciences*, Hodson F.R. et al. Eds, Edinburgh University Press, 1971, p. 387-395.

. Segmentation dichotomique

Le choix d'une variable A permet de subdiviser le tableau en deux sous-tableaux formés des monômes complets (patrons) contenant soit a soit a'. L'application récursive de cette subdivision permet de construire un arbre binaire de classification, donc une classification hiérarchique sur l'ensemble des patrons. Les critères de choix des variables sont fondés sur un calcul de distance entre les classes résultantes. La distance du χ^2 est la plus utilisée.

Quand on cherche à construire une fonction discriminante entre classes d'individus décrits par des variables binaires, on peut également construire un tel arbre de subdivision, jusqu'à ce que chaque feuille de l'arbre soit incluse dans une des classes de la partition initiale. C'est un des problèmes classiques de l'apprentissage à partir d'exemples.

GUENOCHE A., Propriétés caractéristiques d'une classe relativement à un contexte, *Actes des Journées "Symbolique numérique"*, Paris, Décembre 1987.
WHALLON R., A new approach to pottery typology, *American Antiquity*, 37, 1, 1972, p. 13-33.

. Fonctions dilemmes

A trois variables A,B,C on associe les huit monômes de longueur 3 : abc, a'bc, ab'c, .. a'b'c', ainsi que la partition correspondante des patrons : T(abc), T(a'bc), .. T(a'b'c'). Si au moins deux de ces huit classes sont vides, on peut trouver une disjonction exclusive de monômes (dite *fonction dilemme*) équivalente à la disjonction des monômes de longueur 3 non vides. A cette disjonction exclusive est associée une partition des patrons. Par exemple si ab'c, a'bc, a'bc' et a'b'c' sont non vides on obtient la disjonction équivalente a $\vee$ b $\vee$ b'c', et la partition T(a), T(b) et T(b'c') des patrons.

Le choix de trois variables du tableau permet d'obtenir une (ou plusieurs) fonction(s) dilemme(s) et la (les) partition(s) correspondante(s) des patrons. Pour chaque classe

obtenue on peut réitérer la procédure en choisissant trois autres variables. Ces choix successifs sont faits par l'utilisateur (en fonction des cardinaux des classes résultantes ou des possibilités d'interprétation des résultats obtenus). On construit encore un arbre de subdivision qui n'est plus nécessairement binaire.

GUENOCHE A., Classification using dilemma functions, *Computational Statistics Quarterly*, 2, 1, 1985, p.103-108.

. Blocs réguliers et/ou homogènes

Un sous-tableau **I'** x **J'** de **I** x **J** est dit *régulier* si chacune de ses lignes contient au moins un 1; il est dit *homogène* s'il ne contient que des 0 ou des 1. Si de tels sous-tableaux sont engendrés par le croisement de deux partitions, l'une sur **I** l'autre sur **J**, on dit qu'on a une partition du tableau en blocs réguliers et (ou) homogènes. Dans l'algorithme développé par Arabie et al., on cherche de telles partitions "signifiantes" dont les blocs homogènes sont des blocs de 0. On notera que la technique adoptée utilise une analyse factorielle d'une matrice de corrélation, ce qui fournit un exemple de méthode où un modèle combinatoire est recherché par une technique d'algèbre linéaire. De même un algorithme itératif, dû à Govaert, basé sur les classifications simultanées de **I** et **J**, converge vers une partition localement optimale du tableau initial en blocs presque "homogènes" de 0 et de 1. Ces méthodes ont été développées dans le contexte de l'analyse de réseaux sociaux.

ARABIE Ph., BOORMAN A., LEVITT P., Constructing block models : how and why, *Journal of Mathematical Psychology*, 17, 1978, p. 21-63.
DEGENNE A., FLAMENT Cl., La notion de régularité dans l'analyse des réseaux sociaux, *Bull. de Méthodologie Sociologique*, 2, 1984, p. 3-16.
GOVAERT G., Classification simultanée de tableaux binaires, *Data Analysis and Informatics 3*, Diday E. et al. Eds., North-Holland, Amsterdam, 1984, p. 223-236.

. Treillis de Galois

Un sous-tableau de T ne contenant que des 1 est aussi appelé rectangle de T et noté A x B ($A \subset \mathbf{I}$, $B \subset \mathbf{J}$). Les rectangles maximaux (i.e. qui ne peuvent être agrandis ni en ligne ni en colonne) constituent les éléments du treillis de Galois de T. L'ordre de ce treillis est donné par la relation suivante entre deux rectangles maximaux :

$$A \times B \leq C \times D \text{ si et seulement si } A \subseteq C \text{ et (ou) } B \supseteq D.$$

Si le treillis de Galois de T est un préordre total (i.e. se réduit à une chaîne), on peut donner à T une forme d'*échelle de Guttman*, où tous les 1 du tableau sont séparés de tous ses 0 par une frontière en escalier. Une relation entre **I** et **J** qui possède cette propriété s'appelle aussi une *relation de Ferrers* ou un *biordre*.

D'un point de vue pratique, si l'on est seulement intéressé par la liste des rectangles maximaux, on utilisera soit l'algorithme de Ganter si leur nombre est trop important pour qu'ils soient mémorisés, ou que l'on ne cherche que ceux d'une certaine taille, soit l'algorithme de Norris, plus rapide mais qui nécessite de garder en mémoire la liste de tous les rectangles maximaux. Si l'on veut également les relations d'inclusion entre rectangles, on utilisera l'algorithme de Bordat.

Le treillis de Galois est isomorphe et anti-isomorphe aux deux treillis de parties de I et de J obtenus en prenant les intersections soit des colonnes soit des lignes de T (en rajoutant éventuellement une colonne ou une ligne de 1).

Dans la "Begriffsanalyse" de R. Wille et autres, le tableau T est appelé *contexte* et les rectangles maximaux de T sont appelés les *concepts* de ce contexte.

BARBUT M., MONJARDET B., *Ordre et Classification, Algèbre et Combinatoire,* Hachette, Paris, 1970.
BORDAT J.P., Calcul pratique du treillis de Galois d'une correspondance, *Math. Sci. hum*, 96, 1986, p. 31-47.
GANTER B., RINDFREY K., SKORSKY M., Software for concept analysis, *Classification as a tool of research*, GAUL W., SCHADER M. Eds, North Holland, 1986, p. 161-168.
NORRIS E.M., An algorithm for computing the maximal rectangles in a binary relation, *Rev. Roum. Math. Pures et Appl.*, 23, 2, 1978, p. 243-250.
WILLE R., Restructuring lattice theory : an approach based on hierarchies of concepts, in *Ordered Sets*, Rival I. Ed., Dordrecht, Boston, 1982, p. 445-470.

. Famille minimale d'implications informatives

Soient $T = \mathbf{I} \times \mathbf{J}$ un tableau 0-1, μ et ν deux sous-ensembles de J (qui correspondent donc à des monômes ne faisant intervenir que des variables non accentuées, par exemple acd). On dit que μ implique ν (étant donné le contexte T) si $T(\mu) = \{ i \in \mathbf{I}$ tels que $t_{ij} = 1$ pour tout j de $\mu \}$ est inclus dans $T(\nu) = \{ i \in \mathbf{I}$ tels que $t_{ij} = 1$ pour tout j de $\nu \}$. En d'autres termes, si un individu possède tous les attributs de μ, il possède aussi tous ceux de ν. On dit que l'implication $\mu \rightarrow \nu$ est *informative* si ν n'est pas inclus dans μ (auquel cas elle est trivialement vérifiée). Une *famille minimale d'implications* (informatives) est une famille de telles implications qui permet d'inférer toutes les autres implications par application des règles ci-dessous, et qui est minimale avec cette propriété.

$$\mu \rightarrow \nu \text{ et } \nu \rightarrow \pi \text{ entraînent } \mu \rightarrow \pi \quad \text{et} \quad \mu \rightarrow \nu \text{ entraîne } \mu \cup \pi \rightarrow \nu \cup \pi.$$

On montre que toutes les familles minimales d'implications sont des familles de même cardinalité que la *famille* dite *canonique* d'implications. Le calcul de cette famille canonique (d'où on déduit aisément les familles minimales) a donné lieu à divers implémentations.

DUQUENNE V., Contextual implications between attributes and some representation properties for finite lattices, in *Beiträge zur Begriffsanalyse*, GANTER B., WILLE R., WOLFF K.E. Eds., Wissenchaftverlag, Mannheim, 1987, p. 213-240.
GANTER B., Algorithmen zur Formalen Begriffsanalyse, *Beiträge zur Begriffsanalyse*, GANTER B., WILLE R., WOLFF K.E. Eds., Wissenchaftverlag, Mannheim, 1987, p. 241-254.
GUIGUES J.L., DUQUENNE V., Familles minimales d'implications informatives résultant d'un tableau de données binaires, *Math. Sci. hum*, 95, 1986, p. 5-18.

. Dimension Ferrers (bidimension)

Tout tableau $\mathbf{I} \times \mathbf{J}$ peut être obtenu comme intersection de relations de Ferrers (ou biordres) définies entre **I** et **J**. Le nombre de telles relations nécessaires s'appelle la *dimension Ferrers* (ou *bidimension*) de T. Elle est égale à la dimension ordinale du treillis de Galois de T. Le problème du calcul de cette dimension est NP-difficile.

La dimension Ferrers de T est aussi le nombre minimum d'échelles nécessaires dans le "modèle conjonctif" de Coombs-Kao (généralisation multidimensionnelle de l'échelle de Guttman). Tout tableau $\mathbf{I} \times \mathbf{J}$ peut être également obtenu comme union de relations de Ferrers, d'où une notion de dimension disjonctive (correspondant au modèle "disjonctif" de Coombs-Kao).

COGIS O., On the Ferrers dimension of a digraph, *Discrete Math.*, 38, 1982, p. 47-52.

COOMBS C.H., *Theory of data*, Wiley, New York, 1964.
DOIGNON J.P., DUCAMP A., FALMAGNE J.C., On realizable biorders and the biorder dimension of a relation, *Journal of Math. Psychol.*, 28, 1984, p. 73-109.
KOPPEN M.G.M., On finding the bidimension of a relation, *Journal of Math. Psychol.*, 31, 1987, p. 155-178.

. Sériation

Le problème de la sériation consiste à déterminer un ordre sur les lignes et un ordre sur les colonnes tels que le tableau réordonné présente le plus de 1 au voisinage de la diagonale. Pour ce problème on peut chercher à optimiser certains critères, par exemple trouver un ordre sur **I** tel que pour chaque variable ses valeurs 1 soient consécutives (Fulkerson et Gross ont trouvé un algorithme de complexité polynomiale pour déterminer si T a cette propriété et pour construire la permutation correspondante sur **I**); l'hypergraphe H associé à T (cf. rubrique suivante) est alors un *hypergraphe d'intervalles*. On peut également chercher à minimiser la "*largeur de bande*" (bandwith) des 1, i.e. la somme, pour toutes les lignes et toutes les colonnes, de la largeur de l'intervalle où sont situés tous les 1 du tableau. Ce problème est NP-difficile; on trouvera des méthodes exactes et une étude comparative avec des heuristiques dans l'article de Laporte.

Plusieurs auteurs ont proposé la même heuristique basée sur l'ordre croissant de la moyenne des rangs des 1 des lignes et des colonnes, celles-ci étant calculées alternativement. Cette procédure itérative est une heuristique rapide. Pour des références à d'autres modèle de sériation ou plus généralement de réorganisation d'un tableau 0-1, on consultera l'article de Caraux.

CARAUX G., Réorganisation et représentation visuelle d'une matrice de données numériques; un algorithme itératif, *Rev. de Stat. Appl.*, 32, 4, 1984, p. 3-23.
FULKERSON D.R., GROSS O.A., Incidence Matrices and Interval Graphs, *Pacific Journal of Math.*, 15, 1965, p. 835-855.
GOLDMANN K., Some Archaeological Criteria for Chronological Seriation, *Mathematics in the Archaeological and Historical Sciences*, F.R. Hodson et al. Eds., Edimburg University Press, 1971, p.202-208.
HUBERT L., Problems of seriation using a subject by item response matrix, *Psychological Bulletin*, 81, 12, 1974, p. 976-983.
LAPORTE G., Solving a family of permutation problems, *R.A.I.R.O.*, 21, 1, 1987, p. 65-85.

. Représentation arborée

A toute colonne du tableau T = **I** x **J**, on associe l'ensemble des individus de **I** ayant 1 comme valeur de cette colonne. On définit ainsi un ensemble de parties de **I**, ou encore (en suivant la terminologie de Cl. Berge) un hypergraphe dont l'ensemble des sommets est **I** et l'ensemble des hyperarêtes les parties sélectionnées de **I**. On dit que cet hypergraphe H est *arboré* s'il existe un arbre A d'ensemble de sommets **I** tel que toute hyperarête de H est une partie connexe de cet arbre; on dit aussi que H est *rigide* sur A.

On détermine si un hypergraphe est arboré et si oui, on représente tous les arbres sur lesquels il est rigide. Cette représentation, ou la représentation similaire de l'hypergraphe associé aux lignes du tableau, est utilisée en "analyse de similitude". Un cas particulier d'hypergraphe arboré est obtenu si l'arbre A est une chaîne. On retrouve alors les hypergraphes d'intervalles précédents; on parle aussi de *relation* (ou *matrice*) *de Petrie.*

FLAMENT Cl., Hypergraphes arborés, *Discrete Math.*, 21, 1978, p. 223-227.
LECLERC B., Arbres minimum communs et compatibilités de types variés, *Math. Sci. hum.*, 98, 1987, p. 41-67.
DUCHET P., Tree hypergraphs and their représentation trees, preprint 1987.

3.2 Variables ordinales ou quantitatives

Dans le tableau T croisant les individus et les variables, les valeurs t_{ij} sont numériques. Ces nombres peuvent cependant avoir des significations différentes selon que la variable considérée est ordinale ou quantitative (échelle d'intervalles). Nous décrirons d'abord deux méthodes spécifiques puis nous donnerons un principe général qui permet de ramener certains problèmes posés à propos du tableau, par exemple celui de la classification des individus (ou des variables), aux méthodes traitant plusieurs relations binaires, méthodes exposées au § 5.

. Réordonnancement

On cherche à réordonnancer les lignes et les colonnes du tableau, de façon à faire apparaître une structure. La nature de cette dernière dépend du problème posé, et peut être par exemple des sous-tableaux de fortes valeurs. On trouvera dans le texte de Caraux de nombreuses références sur ce type d'approche. Celle de Caraux lui même consiste à réordonner d'abord les lignes, puis les colonnes et à donner une représentation graphique du résultat. Pour réordonner par exemple les lignes, on commence par choisir un indice de distance entre celles-ci, puis on réordonne le tableau des distances en optimisant un critère global choisi de façon que les éléments proches (au sens de cette distance) soient voisins (dans le nouvel ordre) et les éléments distants soient éloignés. En fait on utilise une heuristique rapide qui donne un ordre, optimum local du critère.

Plusieurs logiciels de traitements graphiques de ce type de données ont été réalisés dans la lignée des travaux de Bertin sur la "graphique". Les permutations manuelles des matrices de Bertin, ont été implémentées, puis optimisées pour un certain critère par Leduc.

BERTIN J., *La graphique et le traitement graphique de l'information*, Flammarion, Paris, 1977.
CARAUX G., Réorganisation et représentation visuelle d'une matrice de données numériques; un algorithme itératif, *Rev. de Stat. Appl.*, 32, 4, 1984, p. 3-23.
GRONOFF J.D., *Heurista, logiciel d'aide à l'interprétation de données en sciences humaines*, Notice d'utilisation, E.H.E.S.S., Marseille, 1984.
LEDUC A., Chaînage automatique des matrices ordonnables, *Colloque de Micro-Info-Graphique*, Rouen, 1982, p. 1-38.

. Représentation arborée

Le tableau T est ici vu comme l'ensemble des valeurs d'une "dissimilarité" entre les éléments de **I** et de **J**. Une condition nécessaire et suffisante pour que ces valeurs soient celles d'une distance additive d'arbre a été énoncée par Brossier. Elle caractérise les valeurs de dissimilarités qui lient deux triplets d'éléments, l'un dans **I**, l'autre dans **J**. Il propose ensuite un algorithme, basé sur la décomposition d'une distance arborée en somme d'une distance "à centre" et d'une distance ultramétrique (cf. 4.2.1). Il commence par compléter T en calculant des valeurs sur **I** x **I** et **J** x **J** de façon que le tableau complet contienne les valeurs d'une distance additive d'arbre sur (**I** ∪ **J**) x (**I** ∪ **J**). Dès lors il suffit d'appliquer à ce nouveau tableau (symétrique) les méthodes de représentation arborée du § 4.2.1.

BROSSIER G., Etude des matrices de proximités rectangulaires en vue de la classification, *Revue de Statistique Appliquée*, 35, 4, 1986, p. 43-68.

. Réductions relationnelles

Le principe de ces méthodes est d'associer à une variable numérique J une relation binaire qu'elle "induit" sur **I**. Il existe en fait deux manières canoniques d'induire une telle relation binaire :
. Dans la *réduction nominale* de la variable J, on lui associe la partition (ou la relation d'équivalence) de **I** définie par : i et i' sont dans la même classe ssi $t_{ij} = t_{i'j}$.
. Dans la *réduction ordinale* de la variable J, on lui associe le préordre total sur **I** défini par : $i \leq_j i'$ ssi $t_{ij} \leq t_{i'j}$.

Au tableau **I** x **J** numérique on associe ainsi | **J** | partitions de **I** ou | **J** | préordres totaux sur **I**. Si l'on cherche à classer les individus en groupes séparés on cherchera à résumer ces | **J** | partitions en une seule, par exemple une partition centrale (Régnier); si on cherche à les ordonner, on cherchera un ordre ou un préordre total résumant les | **J** | préordres totaux, par exemple un ordre médian. On est donc ramené aux méthodes d'agrégation de relations binaires exposées au § 5.

Dans certains cas, on peut utiliser des méthodes de réduction duales en associant à chaque ligne du tableau les relations binaires qu'elles induisent sur l'ensemble **J** des variables.

MIRKIN B.G., *Group choice*, Wiley, New York, 1979.
REGNIER S., Sur quelques aspects mathématiques de la classification automatique, *I.C.C. Bull.*, 4, 1965, p. 175-191, repr. *Math. Sci. hum.*, 82, 1983, p. 13-29.
SIBSON R., Order invariant methods for data analysis, *J. Roy. Statist. Soc. B.*, 34, 1972, p. 311-349.

4 METHODES TRAITANT UN TABLEAU **K** x **K**

Dans ce paragraphe on considère un tableau de la forme **K** x **K** qui associe à tout couple (i,j) de $\mathbf{K}^2$ une valeur t_{ij} numérique, quantitative, ordinale ou binaire. Si les valeurs de T sont binaires, les éléments de **K** sont les sommets d'un graphe, orienté si T n'est pas symétrique. Si le tableau est symétrique et à valeurs quantitatives, on parle d'une dissimilarité et les éléments de **K** sont appelés *objets*. On pose n = | **K** |.

4.1 Valeurs binaires

4.1.1 Tableau symétrique

Ce tableau est alors un codage, dit matriciel, d'un *graphe simple* dont **K** est l'ensemble des sommets, et les valeurs 1 désignent les couples de sommets liés par une arête. Si les valeurs sur la diagonale sont toutes nulles, on a un graphe sans boucle. Les problèmes issus de la théorie des graphes ont souvent une signification en analyse des données et les algorithmes construits pour les résoudre sont au centre de méthodes d'analyse de données, par essence combinatoires. Nous citons les plus classiques et invitons le lecteur intéressé à consulter les ouvrages d'algorithmique sur les graphes (Read, Golombic, Gondran et Minoux, Reingold et al., ..).

. Composantes connexes, k-connexes

Dans un graphe G une *chaîne* est une succession d'arêtes reliant deux sommets. Soit A une partie de l'ensemble **K** des sommets; nous définissons ci-dessous quatre propriétés que peut vérifier A relativement aux liaisons entre sommets :

-*connexité* : pour tout i,j ∈ A il existe au moins une chaîne reliant i à j.
-*k-degré connexité* : A est connexe et tout sommet de A est adjacent (relié par une arête) à au moins k sommets de A.
-*k-arête connexité* : A est connexe et il faut supprimer au moins k arêtes dans A pour rendre A non connexe.
-*k-sommet connexité* : A est connexe, | A | ≥ k+1 et il faut supprimer au moins k de ses sommets pour rendre A non connexe.
De plus, si A vérifie une quelconque de ces propriétés, disons P, et est maximale pour cette propriété, nous dirons que c'est une P-classe.

Les classes connexes forment une partition de **K**; k étant fixé (entre 1 et n), il en est de même des classes k-degré connexes, k-arête connexes (mais pas des classes k-sommet connexes). De plus, pour k fixé, une classe k-sommet connexe est incluse dans une classe k-arête connexe qui est elle-même incluse dans une classe k-degré connexe qui est elle-même incluse dans une classe connexe. En faisant varier, soit le parametre k, soit la propriété considérée, on obtient diverses classifications hiérarchiques des sommets de **K**.

HUBERT L.J., Some applications of graph theory to clustering, *Psychometrika*, 39, 1974, p.283-309.
MATULA D.W., Graph theoretic techniques for cluster analysis algorithms, *Classification and Clustering*, Van Ryson J. Ed., Academic Press, New York, 1977, p. 96-129.

. Cliques maximales

Une *clique* d'un graphe est un sous-ensemble de sommets deux à deux adjacents. Une clique est *maximale* si elle n'est incluse dans aucune autre clique, c'est à dire si tout sommet extérieur à la clique n'est pas adjacent à au moins un de ses sommets. On a rencontré cette notion dans la construction du graphe de Buneman (§ 3.1) et nous la retrouverons aux différentes rubriques Hiérarchies (§ 4.2.1), puisqu'elle intervient dans la définition de diverses typologies. Dans les applications on utilise parfois la notion de *k-clique maximale*, i.e. d'ensemble de sommets dont deux quelconques sont toujours reliés par une chaîne de longueur ≤ k et maximales avec cette propriété, ou des variantes de cette notion.

ALBA R.D., A graph-theoretic definition of a sociometric clique, *Journal of Mathematical Sociology*, 3, 1973, p. 113-126.
PEAY E.R., Non metric grouping : Clusters and Cliques, *Psychometrika*, 40, 3, 1975, p.297-313.

. Centres, diamètre

Il existe de nombreuses définitions possibles des centres d'un graphe, certaines issues de la recherche opérationnelle. L'une des plus classiques est de définir un *centre* comme un sommet dont la plus longue (en nombre d'arêtes) des chaînes minimales à tous les autres sommets est de longueur minimum. Le *diamètre* d'un graphe est la longueur de la plus longue des chaînes minimales entre deux sommets quelconques du graphe.

Les notions des trois rubriques précédentes sont au coeur de l'étude des réseaux (notamment sociaux) et interviendront constamment dans les méthodes de classification du paragraphe 4.2.1.

. Equivalences à distance minimum d'un graphe symétrique

La distance d'une relation d'équivalence à une relation symétrique étant celle de la différence symétrique, on cherche les équivalences à distance minimum d'une relation donnée. Le problème a été montré NP-difficile.

KRIVANEK M., MORAVEK J., NP-hard problems in hierarchical-tree clustering, *Acta Informatica*, 23, 1986, p. 311-323.
ZAHN C.T. Jr., Approximating symmetric relations by equivalence relations, *J. SIAM Appl. Math.*, 12, 1964, p. 840-847.

4.1.2 Tableau non symétrique

A un tableau T de données binaires, non symétrique, correspond un *graphe orienté*, avec la convention que l'on a un arc i → j si et seulement si $t_{ij} = 1$. Là encore les notions usuelles de théorie des graphes (chemin, cycle, ..) et les algorithmes associés (plus long chemin, cycle de longueur fixée, ..) sont porteuses de méthodes combinatoires. Si ce graphe n'a pas de cycle, ou si l'on déclare "incomparables" les sommets d'un même cycle, on obtient, par fermeture transitive, un ordre partiel. Si ce graphe est antisymétrique et complet ($t_{ij} = 1$ ssi $t_{ji} = 0$), on parle de *tournoi*, très utilisé en agrégation des préférences. En effet si un nombre (impair) de votants exprime ses préférences sur des candidats sous forme d'un ordre total, on posera $t_{ij} = 1$ si i est préféré à j par une majorité de votants, sinon $t_{ij} = 0$. On retrouvera ces tournois valués au § 4.2.2.

. Ordre(s) à distance minimum d'un tournoi non valué

La distance entre un ordre total et un tournoi est ici mesurée par le nombre d'arcs qu'il faut "inverser" (i → j devient j → i, ou encore on échange les valeurs de t_{ij} et t_{ji}) pour que ce tournoi devienne identique à cet ordre total. Si un ordre total est à distance minimum d'un tournoi, cette distance est aussi le nombre minimum d'arcs du tournoi à supprimer pour que celui-ci devienne sans circuit. La construction d'un ordre à distance minimum d'un tournoi est un problème NP-difficile, dont on peut obtenir, si n est petit, la solution par une méthode "branch and bound". Dans le cas où le nombre de candidats est élevé (>30), on aura recours à des méthodes heuristiques. L'une des plus performantes est due à Smith & Payne, qui à chaque itération inversent l'un des arcs inclus dans un nombre maximum de circuits de longueur 3; toutefois, même en essayant tous les choix possibles, elle n'aboutit pas nécessairement à un ordre optimum.

BERMOND J.Cl., Ordres à distance minimum d'un tournoi et graphes partiels sans circuits maximaux, *Math. Sci. hum.*, 37, 1972, p. 5-25.
BERMOND J.Cl., KODRATOFF Y., Une heuristique pour le calcul de l'indice de transitivité d'un tournoi, *R.A.I.R.O.*, 10, 1976, p. 83-92.
SMITH A.F.M., PAYNE C.D., An algorithm for determining Slater's i and all nearest adjoining orders, *Br. J. Math. Statist. Psychol.*, 27, 1974, p. 49-52.

. Partie essentielle, diagramme (de Hasse)

On appelle *partie essentielle* ou *base* d'un ordre partiel sur **K** le sous-ensemble minimal des couples de l'ordre dont la fermeture transitive est égale à cet ordre. Ces couples sont nécessaires et suffisants pour caractériser l'ordre partiel; les autres arcs s'en déduisent par transitivité. Le graphe de sommets **K** et dont les arcs lient les seuls couples de la partie essentielle, s'appelle le *graphe de couverture*. Si l'on s'astreint à représenter ce graphe de façon qu'un élément inférieur à un autre soit situé en dessous, on obtient le *diagramme (de Hasse)* de l'ordre partiel. Les algorithmes de calcul de la partie essentielle d'un ordre partiel sont de complexité polynomiale (équivalente à celle du calcul de la fermeture transitive d'un graphe orienté). Les méthodes de tracé des diagrammes et plus généralement des graphes orientés essayent de minimiser le nombre de croisements d'arcs. Les algorithmes de tracés optimaux sont eux NP-difficiles et on a recours à des solutions approchées.

DELARCHE M., *Quelques outils infographiques pour l'analyse structurale de systèmes*, Thèse de docteur-ingénieur, Grenoble, 1979.
KNUTH D., *The Art of Computer Programming*, Addison-Wesley, 1973.
ROSTAM H., *Construction automatique et évaluation d'un graphe d'implication issu de données binaires dans le cadre de la didactique des mathématiques*, Rapport de recherche 150, I.R.I.S.A., Rennes, 1981.

. Extensions linéaires

Un ordre total est une *extension linéaire* d'un ordre partiel si tous les couples de cet ordre sont dans l'ordre total. On dit également que cet ordre total est *compatible* avec l'ordre partiel. On ne connait de formule de dénombrement des extensions linéaires d'un ordre partiel donné que dans des cas très particuliers. Les algorithmes d'énumération des ordres totaux compatibles avec un ordre partiel donné ont été étudiés par différents auteurs (sous le nom de "topological sorting") et en particulier par D. Knuth. Pour l'étude de la complexité de ces algorithmes, ou de ceux calculant différents paramètres associés aux extensions linéaires, on consultera Bouchitte et Habib.

AIGNER M., *Combinatorial Theory*, Springer Verlag, Berlin, 1979.
BOUCHITTE V., *Propriétés algorithmiques des extensions linéaires*, Thèse de Doctorat, Université de Montpellier, 1987.
BOUCHITTE V., HABIB M., *The calculation of invariance for ordered sets*, Rapport de recherche n° 150, E.N.S.T. Brest, 1987.
KNUTH D., *The Art of Computer Programming*, Addison-Wesley, Reading, 1973.

. Dimension d'un ordre partiel

Tout ordre partiel peut être obtenu comme l'intersection de plusieurs ordres totaux. Le nombre minimum d'ordres totaux nécessaires pour engendrer l'ordre partiel est appelé sa *dimension*. Alors que l'on peut tester si la dimension d'un ordre partiel est ≤ 2 par un algorithme polynomial, tester si cette dimension est $\leq k$ ($k \geq 3$) est un problème NP-complet. Dans le cas de dimension 2, l'algorithme polynomial est obtenu à partir de résultats sur le graphe de comparabilité d'un ordre (cf. Even et Golombic). Dans le cas général, des résultats de Bouchet ramènent le problème à un problème de recouvrement minimum. Un autre algorithme, dû à Ducamp, utilise des méthodes booléennes. On peut aussi se servir des réductions définies pour le calcul de la bi-dimension par Koppen (cf. § 3.1.).

BOUCHET A., Etude combinatoire des ordonnés finis, Thèse U.S.M.G., Grenoble, 1971.
DUCAMP A., Sur la dimension d'un ordre partiel,*Théorie des Graphes,* Rosenstiehl P. Ed., Dunod, Paris, 1967, p.103-112.
EVEN Sh., *Algorithmic Combinatorics*, Mac Millan, New York, 1973.
GOLOMBIC M.C., *Algorithmic graph theory and perfect graphs*, Academic Press, New York, 1980.

. Codage booléen d'un ordre partiel (2-dimension)

Tout élément d'un ensemble partiellement ordonné peut être codé par un vecteur 0-1, noté c, de façon que l'on ait $x \leq y$ ssi $c(x) \leq c(y)$. Le nombre minimum de composantes booléennes nécessaires pour un tel codage est la *dimension booléenne* (ou 2-dimension) de l'ordre partiel. Des résultats théoriques (Bouchet) montrent que cette 2-dimension est aussi le nombre minimum de générateurs - pour les opérations supremum et infimum - du treillis distributif des parties finissantes de l'ordre partiel (F est finissante si $x \in F$ et $x \leq y$ impliquent $y \in F$). En analyse des données ces résultats permettent de caractériser tous les tableaux 0-1 dont l'ordre de dominance des lignes (ou des colonnes) est fixé.

Plus généralement on peut définir la k-dimension d'un ordre partiel, étudiée notamment par Bouchet et Trotter.

BOUCHET A., *Etude combinatoire des ordonnés finis*, Thèse U.S.M.G., Grenoble, 1971.
MONJARDET B., NETCHINE-GRYNBERG G., Formalisation ordinale de modèles pluriels du développement psychologique, *Math. Sci. hum.*, 96, 1986, p. 65-94.
TROTTER W.T. Jr., A note on Dilworth's embedding theorem, *Proc. Am. Math. Soc.*, 52, 1975, p. 33-39.

4.2 Valeurs ordinales ou quantitatives

4.2.1 Tableau symétrique

Les valeurs du tableau sont souvent vues comme des mesures de ressemblance entre n objets, éléments de **K**. On parle alors d'un tableau de *dissimilarité* (resp. *similarité* ou *similitude*) si les valeurs faibles (resp. fortes) correspondent aux couples d'objets voisins. Il se représente naturellement par un graphe simple, complet, non orienté et valué, dont les sommets sont les éléments de **K**, chaque arête i - j ayant pour valeur t_{ij}. Si seul l'ordre entre les valeurs du tableau est pertinent, on considèrera celui-ci comme définissant une *préordonnance* (totale) c'est à dire un préordre total sur l'ensemble noté $\mathbf{K}^2$ des paires d'objets; $\{i,j\} \leq \{k,l\}$ ssi $t_{ij} \leq t_{kl}$ (les objets i et j sont plus proches que les objets k et l). Les méthodes décrites ci-dessous concernent principalement trois problèmes : Classification (recherche d'une structure taxonomique sur **K**), représentation (des éléments de **K** par une structure d'arbre "conservant" les dissimilarités), sériation (recherche d'un ordre total sur tout ou partie de **K**. Un certain nombre de ces méthodes n'utilisent que la préordonnance associée au tableau.

. Arbre minimum

Un *arbre* est un graphe simple connexe et sans cycle. Un *arbre d'un graphe* (spanning tree) est un arbre dont les sommets sont tous les sommets du graphe et les arêtes certaines de ses arêtes. Par définition, dans un arbre il existe une et une seule *chaîne* (chemin) joignant deux sommets; sa longueur est égale à la somme des longueurs

des arêtes empruntées. Par extension, la longueur d'un arbre est égale à la somme des longueurs de ses arêtes. Un arbre minimum d'un graphe donné est un arbre de longueur minimum. Un graphe peut admettre plusieurs arbres minimaux si plusieurs arêtes ont des longueurs égales.

La construction d'un arbre minimum est un des sujets les plus classiques de l'algorithmique combinatoire. Les méthodes les plus utilisées sont celles de Kruskal (de complexité $O(n^2)$ si les valeurs de **T** sont ordonnées) et de Prim (de complexité $O(n^2)$); on trouvera un inventaire plus exhaustif chez Rosensthiel.

KRUSKAL J., On the shortest spanning tree of a graph and the travelling salesman problem, *Proc. Amer. Math. Soc.*, 7, 1956, p.48-50.
PRIM R.C., Shortest connection network and some generalizations, *Bell System Tech. Jour.*, 26, 1957, p. 1389-1401.
ROSENSTIEHL P., L'arbre minimum d'un graphe, *Théorie des graphes*, Rosenstiehl P. Ed., Dunod, Paris, 1967.

. Distance inférieure maximum

Si le tableau T n'est pas un tableau de distance, on peut construire une distance D, dont toutes les valeurs sont inférieures ou égales à T et qui est maximum pour cette propriété. La valeur D(i,j) de cette distance est donnée par la longueur d'une plus courte chaîne entre les sommets i et j du graphe (la longueur d'une chaîne étant ici la somme des valuations de ses arêtes). On est donc amené à déterminer toutes les plus courtes chaînes d'un graphe complet valué; on peut utiliser un algorithme de complexité $O(n^2)$ dû à Moore et Dijkstra, détaillé dans la première référence ci-dessous, dans laquelle on trouvera également de nombreux algorithmes cités dans ce paragraphe. Pour d'autres types de distance approchant inférieurement la dissimilarité on consultera l'autre référence.

GONDRAN M., MINOUX M., *Graphes et Algorithmes,* Eyrolles, Paris, 1979.
VAN CUTSEM P., Ultramétriques, distances, ϕ-distances maximum dominées par une dissimilarité donnée, *Statistique et Annalyse des données*, 8, 2, 1983, p. 42-63.

. Hiérarchie des classes connexes et filtrant des cliques

A chaque valeur d'un seuil s pris parmi les valeurs de T correspond un *graphe seuil,* dons les sommets sont les éléments de **K** et les arêtes celles du graphe complet dont la valeur est inférieure ou égale à ce seuil; ces graphes sont tous distincts (et "emboîtés"). Pour chacun d'eux, on détermine ses classes connexes et ses cliques maximales. Pour un graphe seuil donné l'ensemble de ses classes connexes est une partition de **K** (réduite à la seule classe **K** si le seuil s est supérieur ou égal à la plus grande longueur d'arête d'un arbre minimum). L'ensemble de toutes ces partitions constitue la *hiérarchie des classes connexes* associée au tableau T.

On considère aussi l'ensemble de toutes les cliques maximales de tous les graphes seuils; cet ensemble, ordonné par inclusion, a été appelé *filtrant* des cliques en analyse de similitude. Plusieurs méthodes de réduction de ce filtrant ont été proposées (cf. la rubrique Hiérarchie de recouvrements).

AUGUSTSON J.G., MINKER J., An analysis of some graph theoretical cluster techniques, *Journal of A.C.M.*, 17, 1970, p. 571-588.

DEGENNE A., VERGES P., Introduction à l'analyse de similitude, *Revue Française de Sociologie*, 14, 1973, p. 471-512.
FLAMENT Cl., L'analyse de similitude, *Cahiers du C.E.R.O.*, 4, 2, 1962, p. 63-97.

. Hiérarchies de partitions

Parmi les méthodes qui associent une hiérarchie de partitions à un tableau de dissimilarité, celles qui utilisent un algorithme de classification ascendante hiérarchique sont les plus anciennes. Le principe général d'un tel algorithme consiste à réunir à chaque étape les classes les plus voisines, au sens d'une dissimilarité entre classes caractéristique d'une méthode particulière. Une vaste catégorie de telles méthodes, mettant en jeu des propriétés de théorie des graphes, est obtenue en définissant cette dissimilarité de la manière suivante (Hubert, 1974) : Soit P une propriété pouvant être vérifiée par un graphe; la dissimilarité entre deux classes **C** et **C'** est la plus petite valeur $s = t_{ij}$ ($i \in \mathbf{C}$, $j \in \mathbf{C}'$) telle que le sous-graphe au seuil s dont l'ensemble des sommets est $\mathbf{C} \cup \mathbf{C}'$ vérifie la propriété P.

Si on prend pour P la propriété de connexité, on retrouve la méthode dite du *lien unique* et la hiérarchie des classes connexes définie à la rubrique précédente.
Si on prend pour P la propriété d'être une clique, on obtient la méthode dite du *lien complet* qui construit des partitions formées de cliques des graphes seuils. On notera que dans le cas où il existe des dissimilarités égales le résultat de cette méthode dépend (éventuellement fortement) de l'ordre d'examen des classes; une solution à ce problème est proposée dans l'ouvrage de Barthélemy et Guénoche.
D'autres propriétés P intermédiaires entre la connexité et la "totalité" ont été considérées, notamment les propriétés vues au § 4.1.1. de k-degré, k-arête, k-sommet connexité ou de k-clique.

On notera que toutes les méthodes définies ci-dessus ne dépendent que de l'ordre sur les dissimilarités.

BARTHELEMY J.P., GUENOCHE A., *Les arbres et les représentations des proximités*, Masson, Paris, 1988.
BENZECRI J.P. et al., *L'analyse des données. 1. La taxinomie*, Dunod, Paris, 1973.
BOCK H.H., Automatische Klassifikation, *Theoretische und pratische Methoden zur Gruppierung und Strukturierung von Daten*, Vandenhoech und Ruprecht, Göttingen, 1974.
CHANDON J.L., PINSON S., *Les méthodes d'analyse typologique*, Masson, 1980.
HUBERT L.J., Some applications of graph theory to clustering, *Psychometrika*, 39, 1974, p.283-309.
MONJARDET B., Théorie des graphes et taxonomie mathématique, *Regards sur la théorie des graphes,* Hansen P. et al. Eds, Presses Polytechniques Romandes, 1980, p. 111-125.

. Hiérarchie de recouvrements

Le filtrant des graphes seuils (cf. supra) peut contenir des ensembles se recouvrant fortement. Beaucoup de méthodes combinatoires ont été proposées pour obtenir une hiérarchie de classes où, à chaque niveau, les classes obtenues ne peuvent se recouvrir que de façon limitée; c'est en particulier le cas des méthodes B_k de Jardine et Sibson.

HUBERT L.J., Some applications of graph theory to clustering, *Psychometrika*, 39, 1974, p.283-309.
JARDINE N., SIBSON R., *Mathematical Taxonomy*, Wiley, New York, 1971.

MATULA D.W., Graph theoretic techniques for cluster analysis algorithms, *Classification and Clustering*, Van Ryson J. Ed., Academic Press, New York, 1977, p.96-129.
PEAY E.R., Non metric grouping : Clusters and Cliques, *Psychometrika*, 40, 3, 1975, p.297-313.
ROHLF F.J., A new approach to the computation of the Jardine-Sibson B_k clusters, *The Computer Journal*, 18, 2, 1975, p. 164-168.

. Approximation ultramétrique

Il est bien connu que la construction d'une hiérarchie indicée de partitions à partir d'une dissimilarité revient à approcher cette dissimilarité par une distance *ultramétrique* (la méthode du lien unique construit l'ultramétrique inférieure maximale). Certains auteurs se sont posé directement ce problème d'approximation. Roux procède de manière itérative, en modifiant les dissimilarités jusqu'à ce qu'elles satisfassent à l'inégalité ultramétrique; chaque itération est en $O(n^3)$. Defays d'une part, Chandon et al. d'autre part cherchent la meilleure approximation ultramétrique au sens de la distance de la norme L^1 (pour le premier) ou des moindres carrés (pour les seconds). Le premier de ces problèmes a été montré NP-difficile.

CHANDON J.L., LEMAIRE J., POUGET J., Construction de l'ultramétrique la plus proche d'une dissimilarité au sens des moindres carrés, *R.A.I.R.O.*, 14, 2, 1980, p. 157-170.
DEFAYS D., Recherche des ultramétriques à distance minimum d'une similarité donnée, *Bull. Soc. Roy. Sc. Lg.*, 5-6, 1975, p. 330-343.
KRIVANEK M., MORAVEK J., NP-hard problems in hierarchical-tree clustering, *Acta Informatica*, 23, 1986, p. 311-323.
ROUX M., *Un algorithme pour trouver une hiérarchie particulière*, Thèse de troisième cycle, I.S.U.P., Paris, 1968.

. Ultrapréordonnance à distance minimum

Une *ultrapréordonnance* sur **K** est une préordonnance qui vérifie la condition : $\forall$ i, j, k $\in$ **K** , on a $\{i,j\} \leq \{i,k\}$ ou $\{i,j\} \leq \{j,k\}$. A une préordonnance à q classes est associée une hiérarchie stratifiée de q-1 partitions. L'ensemble des préordonnances étant un sup demi-treillis semi-modulaire supérieurement, il existe une distance latticielle "canonique" entre préordonnances. Schader cherche l'ultrapréordonnance la plus proche pour cette distance d'une préordonnance donnée. A cette fin, il propose un algorithme itératif qui partant de cette préordonnance construit en examinant tous les triplets d'éléments de **K** une préordonnance à q classes. Cette méthode est utilisée pour la préordonnance associée au tableau T donné.

SCHADER M., Hierarchcal analysis : Classification with ordinal object dissimilarities, *Metrika*, 27, 1980, p. 127-132.

Les méthodes précédentes recherchent des hiérarchies de partitions ou de recouvrements des éléments de **K**. Nous donnons maintenent des méthodes qui construisent des partitions de **K**. Le principe général de ces méthodes consiste à définir un critère de proximité d'une partition au tableau de dissimilarité, puis à chercher une partition optimale pour ce critère.

. Partition de séparation maximum

On appelle *séparation* d'une partition de **K** la valeur minimum de la dissimilarité des arêtes interclasses (i.e. joignant deux classes distinctes) de cette partition. On cherche

parmi toutes les partitions en p classes de **K** celles de séparation maximum. On montre qu'une telle partition est unique et égale à la partition donnée par les p classes connexes d'un graphe seuil (en faisant varier p, on obtient la hiérarchie de partitions correspondant à la méthode du lien simple). Le cas où l'on définit la séparation non plus par la valeur minimum, mais par la somme des dissimilarités des arêtes interclasses et où l'on impose des conditions, par exemple de taille sur les classes, a donné lieu à de nombreuses propositions pour lesquelles on pourra consulter la première et la dernière des références ci-dessous.

CHRISTOFIDES N., BROOKER P., The optimal partitioning of graphs, *SIAM Journal Appl. Math.*, 30, 1, 1976, p. 55-69.
HANSEN P., DELATTRE M., Bicriterion cluster analysis as an exploration tool, *Multiple Criterion Problem Solving*, Lecture Notes in Economic and Mathematics Systems, 155, Springer Verlag, Berlin, 1977, p. 249-273.
HUBERT J.L., Data analysis implications of some concepts related to the cuts of a graph, *J. of Math. Psychol.*, 15, 2, 1977, p. 199-208.
LECLERC B., An application of combinatorial theory to hierarchical classification, *Recent Developments in Statistics*, Barra J.R. et al. Eds, North Holland, 1977, p.783-786.
MILGRAM M., DUBUISSON B., Un algorithme heuristique de décomposition d'un graphe, *R.A.I.R.O.*, 11, 2, 1977, p. 175-199.

. Partition de diamètre minimum

Le *diamètre* d'une partition est la valeur maximum de dissimilarité entre deux éléments d'une même classe. On cherche parmi toutes les partitions en p classes de **K** celles qui sont de diamètre minimum. Le cas p=2 a été résolu par Rao; un algorithme pour le cas général, basé sur les propriétés de coloration des graphes est présenté par Hansen et Delattre qui montrent aussi que le problème général est NP-difficile.

La partition à p classes donnée par un algorithme de lien complet n'est pas nécessairement de diamètre minimum (parmi toutes les partitions à p classes). Deux partitions de diamètre minimum à p et q classes ne sont pas nécessairement emboîtées (comparables).

HANSEN P., DELATTRE M., Complete-link cluster analysis by graph coloring, *J. Amer. Stat. Assoc.*, 73, 362, 1978, p.397-403.

. Partition efficiente

Sur l'ensemble des partitions de **K** à moins de p classes, on définit un ordre partiel : une partition est "avant" une autre si sa séparation lui est supérieure et si son diamètre lui est inférieur. Une partition maximale dans cet ordre est appelée une *partition efficiente*. Un algorithme de construction des partitions efficientes est présenté dans la référence ci-dessous.

DELATTRE M., HANSEN P., Bicriterion cluster analysis, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2, 4, 1980.

. Partition à éloignement minimum

On définit l'éloignement d'une partition au tableau de dissimilarité comme la somme des valeurs des dissimilarités des arêtes interclasses. La recherche d'une partition minimisant cet éloignement est NP-difficile. L'algorithme utilisé par Wakabayashi, qui est basé sur des méthodes de coupes en programmation linéaire entière, a permis d'obtenir une solution exacte jusqu'à l'ordre n=150.

Une partition centrale d'un ensemble de partitions (cf. § 5) peut être définie comme une partition à éloignement minimum pour une certaine dissimilarité et peut donc être recherchée par les même méthodes.

WAKABAYASHI Y., *Aggregation of binary relations : algorithmic and polyhedral investigation*, Thesis, Augsburg, 1986.

. Partition la plus proche d'une préordonnance

Les partitions de **K** correspondent aux préordonnances (et même aux ultrapréordonnances) à deux classes sur **K** : classe des paires de sommets réunis < classe des paires de sommets séparés. Il existe de nombreux coefficients d'accord entre deux préordres totaux qui peuvent être utilisés pour mesurer la proximité d'une préordonnance à deux classes (donc d'une partition) à la préordonnance associée au tableau T (cf. § 5). Ayant choisi un tel coefficient, on cherche alors les partitions optimales pour ce critère; on consultera Chah pour le critère de la différence symétrique ainsi que Chandon & Boctor. Ce dernier article définit aussi une notion de partition efficiente de niveau k. En particulier une *partition efficiente de niveau 0* est une partition telle que toute paire d'objets d'une même classe ait une dissimilarité plus petite que toute paire d'objets appartenant à des classes séparées et telle que le nombre de couples ({i,j},{k,l}) avec i et j réunis, k et l séparés et $t_{ij} < t_{kl}$ soit maximal. Un algorithme d'énumération de toutes les partitions efficientes de niveau 0 est présenté.

La distance latticielle entre préordonnances (cf. supra) est un coefficient d'accord purement ordinal entre préordonnances. Dans son article Schader donne un algorithme permettant d'obtenir l'ultrapréordonnance à deux classes la plus proche pour cette distance de la préordonnance donnée.

CHAH S., Calcul des partitions optimales d'un critère d'adéquation à une préordonnance, *Publications de l'I.S.U.P.*, 29, 1, 1984.
CHANDON J.L., BOCTOR F.F., Approximation d'une préordonnance par une partition, *R.A.I.R.O.*, 19, 2, 1985, p. 159-184.
SCHADER M., Distance minimale entre partitions et préordonnance dans un ensemble fini, *Math. Sci. hum.*, 67, 1979, p. 39-47.

. Arbres à distances additives

Les méthodes de construction d'*arbres à distances additives* ont pour but de représenter les valeurs de dissimilarité comme les longueurs de chemins dans un ***K*-arbre**. Un **K**-arbre est un arbre valué dont l'ensemble des sommets contient **K**, les sommets qui ne correspondent pas aux éléments de **K** étant de degré au moins 3. Les méthodes de construction reviennent à approcher une dissimilarité par une distance dite *additive d'arbre* ou *arborée* (dont les valeurs sont les longueurs des chemins dans l'arbre). A partir de l'article initial de Sattah et Tversky, plusieurs auteurs ont mis au point des algorithmes pour construire séparément ou simultanément une structure de **K**-arbre et des longueurs d'arêtes. On peut citer:

La *méthode de décomposition* (Brossier) basée sur la propriété des distances additives d'arbre d'être la somme d'une distance ultramétrique et d'une distance à centre.

La *méthode des groupements* (Luong) basée sur le nombre de relations "quadrangulaires" satisfaites lorsque l'on réunit deux sous-arbres par un sommet commun.

BARTHELEMY J.P., LUONG X., Représentation arborée des mesures de dissimilarités, *Statistique et analyse de données*, 11, 1, 1986, p. 20-41.
BARTHELEMY J.P., LUONG X., Sur la topologie d'un arbre phylogénétique : aspects théoriques, algorithmes et applications à l'analyse de données textuelles, *Math. Sci. hum.*, 100, 1987.
BARTHELEMY J.P., GUENOCHE A., *Les arbres et les représentations des proximités*, Masson, Paris, 1988.
BROSSIER G., Approximation des dissimilarités par des arbres additifs, *Math. Sci. hum.*, 91, 1985, p. 5-21.
GUENOCHE A., Cinq algorithmes d'approximation d'une dissimilarité par des arbres à distances additives, *Math. Sci. hum.*, 98, 1987, p. 21-40.
SATTAH S., TVERSKY A., Additive similarity trees, *Psychometrika*, 3, 42, 1977, p. 319-345.

. Approximation arborée

De même que pour l'approximation ultramétrique (cf. supra), sans construire un **K**-arbre, on peut chercher à approcher au mieux les dissimilarités par une distance additive d'arbre. De Soete résout le problème par des méthodes classiques d'approximation sous contraintes. Roux renouvelle sa méthode itérative fondée sur les modifications des dissimilarités de façon à ce qu'elles vérifient les inégalités "quadrangulaires", mais cette fois-ci chaque itération est en $O(n^4)$.

Une autre approche, lorsqu'une structure d'arbre est donnée, qu'elle soit définie par une méthode de classification ou une une méthode de construction d'un arbre à distances additives, consiste à calculer des longueur d'arêtes. On sait calculer leurs valeurs de façon à minimiser l'écart quadratique entre les dissimilarités et les longueurs des chemins dans l'arbre. Il s'agit d'un problème d'approximation sous contraintes positives.

DE SOETE G., A least squares algorithm for fitting additive trees to proximity data, *Psychometrika*, 48, 1983, p. 621-626.
GUENOCHE A., Représentations arborées des classifications, *R.A.I.R.O.*, 20, 1986, p.341-354.
ROUX M., Techniques of approximation for building two tree structures, Proceeding of the Franco-Japanese scientific seminar, *Recent developments in clustering and data analysis*, Tokyo, 1987, p. 127-146.

. Tracés d'arbres valués

Les méthodes d'approximation d'une dissimilarité par des distances ultramétriques ou des distances additives d'arbres, les méthodes de classification hiérarchique ou de construction de **K**-arbres aboutissent toutes à un arbre valué. Depuis leur origine les arbres de classification sont représentés par des *dendrogrammes* qui font clairement apparaître que les distances de la racine de l'arbre à toutes ses feuilles sont égales. Ce n'est pas le cas des **K**-arbres où de plus aucun sommet ne joue le rôle de racine de façon naturelle. Des méthodes de tracé automatique d'arbres valués ont donc été proposées, parmi lesquelles le *tracé radial* qui répartit dans des directions uniformes les feuilles, et la *tracé axial* qui utilise comme axe un plus long chemin dans l'arbre.

BARTHELEMY J.P., GUENOCHE A., *Les arbres et les représentations des proximités*, Masson, Paris, 1988.

Le problème de la sériation que nous avons déjà rencontré avec les tableaux en 0-1 (§3.1) peut également être abordé à partir d'un tableau de dissimilarité. Il s'agit toujours de construire un ordre total sur **K** ou sur une partie de **K**, encore appelé une

série si l'ordre est total sur **K** et une *chaîne* si c'est un ordre total sur une partie de **K**. Plusieurs modèles de sériation ont été proposés qui ramènent la sériation à des problèmes NP-difficiles; très peu conduisent à des méthodes exactes de résolution. La plupart donnent lieu à des méthodes heuristiques.

. Série de longueur minimum

La *longueur* d'une série est égale à la somme des dissimilarités entre éléments consécutifs dans la série. Le problème de la construction de la série la plus courte est identique au célèbre problème NP-difficile du "Voyageur de commerce". On a donc pour le résoudre, si n n'est pas trop grand, des algorithmes exacts du type "branch and bound" ou des méthodes d'optimisation en nombres entiers (méthodes de *coupes*). Bien que des problèmes de voyageur de commerce aient été résolus jusqu'à n=500, on a recours en général, dès que n > 50, à des heuristiques efficaces qui construisent des séries de longueur faible. On en trouvera un inventaire intéressant dans divers articles des actes d'un colloque édité par Hodson et al..

GUENOCHE A., Méthodes combinatoires de sériation à partir d'une dissimilarité, Actes du colloque "Data Analysis and Informatics V", Versailles, 1987, p. 115-123.
HODSON F.R., KENDALL D.G., TAUTU P., *Mathematics and Archaeological and Historical Sciences,* Edinburgh University Press, 1971.
HUBERT L.J., Some applications of graph theory and related non metric technics to problems of approximate seriation: the case of symmetric proximity measures, *Br. J. of Math. and Stat. Psychol.*, 27, 2, 1974, p. 133-153.
LAWLER E.L., LENSTRA J.K., RINNOOY KAN A.H.G., SHMOYS D.B., *The travelling Salesman Problem. A Guided Tour of Combinatorial Optimisation,* Wiley, 1985.

. Chaînes et séries robinsoniennes

Une chaîne vérifie la *propriété de Robinson* si à partir de tous ses éléments les dissimilarités vont croissant le long de la chaîne. Une chaîne robinsonienne est maximale si elle n'est pas incluse dans une chaîne qui vérifie également cette propriété. Si un tableau de dissimilarités qui admet une *série robinsonienne* est réordonné suivant cette série on obtient un tableau (dit aussi matrice) de Robinson : les valeurs des dissimilarités croissent en lignes et en colonnes à partir de la diagonale. Dans ce cas on dit que la *dissimilarité* est *robinsonienne* et que la série lui est *compatible*. Tous les triplets ordonnés i<j<k de la série sont alors des triplets *compatibles* (avec la dissimilarité) i.e. vérifient $t_{ik} \geq \max (t_{ij}, t_{jk})$.

Lorsqu'une dissimilarité n'est pas robinsonienne, il existe différentes méthodes de réordonnancement de son tableau suivant une série "presque" robinsonienne; plusieurs de ces méthodes sont basées sur la remarque qu'une série robinsonienne correspond à un arbre minimum de la dissimilarité (cf. Hubert). Une autre méthode (cf. Guénoche) énumère les chaînes robinsoniennes maximales, en partant de tout triplet compatible. Une méthode complémentaire consiste à chercher les séries qui maximisent le nombre de triplets compatibles avec cette dissimilarité. A cette fin on utilise une heuristique qui part d'une chaîne robinsonienne maximale et ajoute les éléments manquants de manière à maximiser, à chaque insertion, le nombre de triplets compatibles.

Les méthodes de recherche d'une série, évoquées dans la rubrique précédente, donnent toutes une série robinsonienne si la dissimilarité en admet (et permettent donc de reconnaitre ce cas). Diverses généralisations de la notion ci-dessus de compatibilité entre un ordre et une dissimilarité ont été

proposées notamment par Diday.

DIDAY E., Croisements, Ordres et Ultramétriques, *Math. Sci. hum.*, 83, 1983, p. 31-54.
GUENOCHE A., *Modèles et méthodes de sériation à partir d'une dissimilarité*, preprint G.R.T.C., 1986.
HUBERT L.J., Some applications of graph theory and related non metric technics to problems of approximate seriation: the case of symmetric proximity measures, *Br. J. of Math. and Stat. Psychol.*, 27, 2, 1974, p. 133-153.
ROBINSON W.S., A method for chronologically ordering archaeological deposits, *American Antiquity*, 16, 1951, p.293-301.

. Ordonnancement des hiérarchies

Si H est une hiérarchie indicée sur **K**, il existe de nombreux ordres totaux sur **K** qui permettent de représenter "sans croisement" cette hiérarchie. On peut se donner différents critères pour choisir un de ces ordres et construire la hiérarchie correspondante. Un de ces critère consiste à maximiser son accord - au sens du τ de Kendall ou du ρ de Spearman - avec un ordre fixé a priori; des algorithmes ascendants ou descendants sont décrits par Degerman et Brossier. Un autre critère consiste à chercher un ordre le plus possible "en accord" avec la dissimilarité donnée; les définitions de cet accord peuvent être celles données dans les deux rubriques précédentes (longueur minimum, nombre de triplets compatibles). Un algorithme ascendant et un autre descendant sont proposés dans Gruvaeus & Wainer et Brossier. D'autres critères et algorithmes sont présentés par Brossier (1984).

BROSSIER G., Représentation ordonnée des classifications hiérarchiques, *Statistique et Analyse des Données*, 2, 1980, p. 31-44.
BROSSIER G., Classification hiérarchique à partir de matrices carrées non symétriques *Statistique et Analyse de Données*, 7, 2, 1982, p. 22-40.
BROSSIER G., Ordonnancement de hiérarchies, Algorithmes et propriétés, *Data Analysis and Informatics 3*, Diday E. et al. Eds, North Holland, 1984, p. 317-321.
DEGERMAN R., Ordered binary trees constructed through an application of Kendall's tau, *Psychometrika*, 47, 4, 1982.
GRUVAEUS G., WAINER H., Two additions to hierarchical cluster analysis, *Br. J. Math. Statist. Psychol.*, 25, 1972, p. 200-206.

. Pyramides

Une *Pyramide* sur **K** est une famille de parties de **K**, appelées classes (ou "paliers"), contenant **K** et tous les singletons, telle que l'intersection de deux classes est soit vide, soit une classe, famille pour laquelle il existe un ordre total dont ces classes soient des *intervalles* (on dit que cet ordre est compatible avec la pyramide). Les pyramides sont donc des hypergraphes d'intervalles particuliers; on remarque également que toute hiérarchie sur **K** est une pyramide. On définit une notion de pyramide indicée (analogue à celle de hiérarchie indicée); la notion d'*indice pyramidal* (analogue à celle d'indice ultramétrique) est équivalente à celle de dissimilarité robinsonienne définie supra. A une pyramide indicée corespond une suite emboîtée de recouvrements de **K** par les classes de la pyramide.

Il existe plusieurs méthodes de construction de pyramides indicées à partir d'un tableau de dissimilarité, par des algorithmes ascendants ou descendants, ou en recherchant l'indice pyramidal le "plus proche" d'une dissimilarité donnée. Si l'on se fixe l'ordre compatible avec la pyramide, il existe une solution "sous-dominante" et une solution "sur-dominée".

Une classe particulière de pyramides (dites pseudo-hiérarchies) et d'indices pyramidaux associés (dites dissimilarités fortement robinsoniennes) a été étudiée par Durand et Fichet.

BERTRAND P., DIDAY E., A visual representation of the compatibility between an order and a dissimilarity index: The pyramids, *Computational Statistics Quaterly*, 2, 1, 1985, p. 31-42.
BERTRAND P, *Etude de la représentation pyramidale*, Thèse de 3-ième cycle, Université Paris-Dauphine, 1986.
DIDAY E., Orders and overlapping clusters by pyramids, *Multidimentional Data Analysis*, De Leeuw J. et al. Eds, D.S.W.O. Press, Leiden, 1986, p. 201-234.
DURAND C., FICHET B., One to one correspondances in pyramidal représentations: an unified approach, *Classification and related methods of data analysis*, BOCK H.H. Ed., North-Holland, 1988.

4.2.2 Tableau non symétrique

On peut considérer ce tableau comme définissant un graphe *orienté* complet et valué dont les sommets sont les éléments de **K** et l'arc (i,j) a pour valuation t_{ij}. On associe à ce graphe un graphe valué (par T'), asymétrique appelé aussi *tournoi majoritaire*, en posant : $t'_{ij} = \sup (0, t_{ij} - t_{ji})$.

. Ordre à distance minimum d'un tournoi valué

On définit l'éloignement d'un ordre total à un graphe valué comme la somme des valuations des arcs du graphe qui ne sont pas dans l'ordre total. On cherche l'ordre à éloignement minimum du graphe valué défini par le tableau T, ce qui est équivalent à chercher l'ordre à éloignement minimum du tournoi majoritaire. Le problème de construction d'un ordre total à distance minimum d'un tournoi est un problème NP-difficile que l'on peut résoudre par une méthode "Branch and Bound" si n est "petit". Pratiquement, on obtient "au même prix" la totalité des ordres optimaux. Si n est "grand", on utilisera comme Arditti des méthodes de relaxation lagrangienne, ou de programmation linéaires (Michaud & Marcotorchino) ou en nombres entiers (Reinelt); on utilise aussi des méthodes heuristiques, par exemple en adaptant comme Barthélemy et al. celle de Smith et Payne au cas des tournois valués, ou en cherchant des optimums "locaux" i.e. ne pouvant être améliorés par transformation d'un type fixé de l'ordre total (par exemple les "doubles décalages" de Michaud & Marcotorchino).

Le même problème quand on recherche un préordre total donne lieu aux mêmes approches; citons l'heuristique de Schader & Tüshaus conduisant à un optimum local par des opérations de permutation, de regroupement ou scission (isolaht un élément) de classes du préordre total.

ARDITTI D., Un nouvel algorithme de recherche d'un ordre induit par des comparaisons par paires, *Data Analysis and Informatics 3*, Diday E. et al. Eds., North Holland, 1984, p. 323-343.
BARTHELEMY J.P., GUENOCHE A., HUDRY O., *Median linear orders : Heuristic and Branch and Bound Algorithms*, preprint E.N.S.T., Paris, 1987.
MICHAUD P., MARCOTORCHINO J.F., *Optimisation en Analyse ordinale des données*, Masson, Paris, 1979.
REINELT G., *The linear ordering problem : Algorithms and Application*, Heldermann Verlag, Berlin, 1985.
SCHADER M., TÜSHAUS U., An Heuristic for Finding a Complete Preordrer, *Classification and related methods of data analysis*, BOCK H.H. Ed., North-Holland, 1988.

. Ordre minimisant une fonction des rangs

Soit r_O la fonction rang associée à l'ordre O ($r_O(i)$ est le rang de i dans O). Dans cette

méthode, on cherche l'ordre O qui minimise la somme des valeurs t_{ji} [$r_O(j) - r_O(i)$] pour tous les couples d'éléments (i > j) placés dans cet ordre dans O, c'est à dire tels que $r_O(i) < r_O(j)$. Si T est tel que $t_{ij} + t_{ji}$ = Cste (cas obtenu si on part d'un profil d'ordres totaux), tout ordre total contenu dans le préordre défini par les scores (somme des valeurs d'une ligne) est solution du problème. Dans le cas général, le premier article cité décrit un algorithme de construction de l'ordre solution. Des méthodes analogues sont proposées dans la seconde référence.

FREY J.J., YEHIA ALCOUTLABI A., Comparaisons par paires : une interprétation et une généralisation de la méthode des scores, *R.A.I.R.O.*, 20, 3, 1986, p. 213-227.
KANO M., SAKAMOTO A., Ranking the vertices of a paired comparision digraph, *SIAM J. Alg. Discrete Math.*, 6, 1, 1985, p. 79-92.

5 METHODES TRAITANT PLUSIEURS TABLEAUX **I** x **J** ou **K** x **K**

Nous nous limiterons en fait à des méthodes traitant plusieurs tableaux **K** x **K** lorsque ceux-ci sont des tableaux 0-1, i.e. au cas où l'on dispose comme données de plusieurs relations binaires, non valuées, sur le même ensemble **K**. De telles données peuvent être obtenues lorsque par exemple plusieurs individus expriment leurs préférences ou leurs classements (partitions) sur le même ensemble d'objets (notamment dans les méthodes de comparaisons par paires), mais on les obtient aussi à partir d'un tableau **I** x **J** en procédant aux "réductions relationnelles" évoquées au §3.2.

Bien que les cas où l'on dispose de plusieurs relations binaires valuées ou de plusieurs tableaux **I** x **J** pourraient donner lieu à des méthodes analogues à celles décrites ci-dessous, de telles méthodes ont été jusqu'ici très peu développées et encore moins implémentées et nous ne les mentionnerons donc pas.

. <u>Comparaison de deux relations</u>

Il s'agit d'évaluer la proximité de deux relations généralement du même type (par exemple de deux ordres totaux exprimant les préférences de deux individus). Une méthode classique consiste à définir un *coefficient d'accord* en normalisant une mesure de dissimilarité - qui est souvent une distance - entre les deux relations. Si elle s'exprime parfois comme une distance euclidienne (ou plus généralement de type L^p) entre deux vecteurs codant ces relations, elle est souvent définie de manière combinatoire comme le nombre minimum de "transformations élémentaires" à opérer pour passer d'une de ces relations à l'autre. Suivant la nature des transformations élémentaires considérées, le problème de calcul d'une telle distance, et donc du coefficient d'accord associé, peut être soit très simple (τ de Kendall entre deux ordres totaux) soit très difficile (le calcul de la distance κ entre deux partitions est NP-difficile, cf. Day et Wells). Il est évidemment hors de question ici de présenter dans le détail les nombreuses distances et coefficients d'accord entre relations de types variés qui ont été proposés dans la littérature et nous nous contentons de renvoyer à quelques références qui en contiennent elles-mêmes de multiples autres.

DAY W.H.E., The complexity of computing metric distances between partitions, *Math. Soc. Sci.*, 1, 1981, p. 269-287.
DAY W.H.E., WELLS R.S., Extremes in the complexity of computing metric distances between partitions, *IEEE Trans. Pattern. Anal. Mach. Intel.*, Vol. PAMI-6, 1, 1984, p. 69-73.

GIAKOUMAKIS V., MONJARDET B., Coefficients d'accord entre deux préordres totaux, *Statistique et Analyse des Données*, 1987, 30 p.
HUBERT L., ARABIE P., Comparing partitions, *J. of Classification*, 2, 1985, p.193-218.
MIRKIN B.G., *Group choice*, Wiley, New York, 1979.
ROHLF F.J., Consensus indices for comparing classifications, *Math. Biosci.*, 59, 1982, p.131-144.

. Concordance de plusieurs relations

Pour évaluer la concordance de plus de 2 relations binaires, on peut recourir à la normalisation de l'inertie d'une famille de vecteurs codant ces relations; mais on peut aussi utiliser une moyenne de coefficients binaires d'accord (il peut arriver que ces deux approches coïncident comme dans le cas du W de Kendall pour des ordres totaux). Une autre approche consiste à évaluer la concordance par la proximité des relations données à une relation consensus les résumant (cf. infra). Comme dans la rubrique précédente, le calcul effectif d'un coefficient de concordance est souvent soit très simple, soit très difficile. Nous nous contenterons là aussi de donner quelques références permettant d'accéder à la littérature du sujet.

DAY W.H.E., Mc MORRIS F.R., A formalization of consensus index methods, *Bull. of Math. Biol.*, 47, 2, 1985, p. 215-229.
LECLERC B., CUCUMEL G., Consensus en classification : Une revue bibliographique, *Math. Sci. hum.*, 100, 1987.
HUBERT L.J., Generalized concordance, *Psychometrika*, 44, 2, 1979, p. 135-142.
MARCOTORCHINO J.F., MICHAUD P., *Optimisation en analyse des données*, Masson, Paris, 1979.
MONJARDET B., Concordance et consensus d'ordres totaux : les coefficients K et W, *Revue de Statistique Appliquée*, 33, 2, 1985, p. 55-87.

Les rubriques suivantes concernent le problème de résumer (d'agréger) plusieurs relations binaires d'un certain type en une relation binaire du même (ou d'un autre) type. Là encore, il existe une littérature volumineuse sur ce thème et nous nous contentons de donner les quelques méthodes qui ont fait l'objet du plus grand nombre d'investigations et d'implémentations.

. Partition centrale (ou équivalence médiane)

On cherche une partition à "éloignement minimum" de plusieurs partitions données, cet éloignement étant défini comme la somme des distances de la différence symétrique entre les équivalences induites par ces partitions et celle induite par la partition cherchée; le problème du calcul d'une telle partition est NP-difficile. Cette méthode, due à Régnier, a donné lieu à des implémentations soit d'heuristiques (rapides mais ne garantissant pas l'optimalité) soit d'algorithmes cherchant une solution optimale et basés sur des techniques de programmation linéaire en nombres entiers ou de relaxation lagrangienne.

MARCOTORCHINO J.F., MICHAUD P., Heuristic approach of the similarity aggrégation problem, *Methods of Oper. Research*, 43, 1981, p. 395-404.
REGNIER S., Sur quelques aspects mathématiques de la classification automatique, *I.C.C. Bull.*, 4, 1965, p. 175-191, repr. *Math. Sci. hum.*, 82, 1983, p. 13-29.
SCHADER M., TÜSHAUS U., Subgradient methods for analyzing qualitative data, in *Classification as a tool of research*, GAUL W., SCHADER M., Eds., North-Holland, 1986, p. 397-403.

WAKABAYASHI Y., Aggregation of binary relations : algorithmic and polyhedral investigation, Thesis, Augsburg, 1986.

. Ordre (total) médian

On cherche un ordre total à "éloignement minimum" d'ordres totaux donnés, cet éloignement étant défini comme la somme des distances de Kendall (nombre de désaccords) entre ces ordres totaux et l'ordre cherché; le problème du calcul des ordres médians est NP-difficile. Cette méthode due à Kemeny a donné lieu à des implémentations variées : pour un nombre faible d'objets (≤ 20) on énumère tous les ordres médians par des méthodes du type "séparation et évaluation progressive" (Branch and Bound), pour un nombre plus élevé (≤ 100) on recherche un ordre médian par des techniques de programmation linéaire en nombres entiers ou de ralaxation lagrangienne, au delà on construit des ordres "presque médians" par des heuristiques rapides.

ARDITTI D., Un nouvel algorithme de recherche d'un ordre induit par des comparaisons par paires, *Data Analysis and Informatics 3*, Diday E. et al. Eds., North Holland, 1984, p. 323-343.
KEMENY J.G., Mathematics without numbers, *Daedalus*, 88, 1959, p. 577-591.

. Préordre (total) médian

On recherche un préordre total à "éloignement minimum" de préordres totaux donnés, avec un éloignement défini comme dans les deux rubriques précédentes. Le problème est encore NP-difficile et les techniques de calcul utilisées sont les mêmes que celles décrites dans les deux rubriques précédentes.

SCHADER M., TÜSHAUS U., Subgradient methods for analyzing qualitative data, in *Classification as a tool of research*, GAUL W., SCHADER M., Eds., North-Holland, 1986, p. 397-403.
SCHADER M., TÜSHAUS U., An Heuristic for Finding a Complete Preorder, in *Classification and related methods of data analysis*, BOCK H.H. Ed., North-Holland, 1988.

Il est important de remarquer que dans les trois cas précédents le critère à optimiser ne fait intervenir que le tableau des comparaisons par paires que l'on peut associer aux relations données, tableau qui suivant la nature de ces relations peut vérifier certaines propriétés. Ces problèmes sont donc des cas particuliers des problèmes de recherche d'un (pré)ordre total ou d'une équivalence à "distance minimum" d'une relation valuée, problèmes coonsidérés au § 4.2.1.; les techniques de calcul utilisées sont les mêmes, mais on peut tirer avantage de la structure particulière des données. En particulier si, par exemple, pour des ordres totaux donnés leur relation majoritaire est un ordre total, celui-ci est l'ordre médian et il est unique. On trouvera ci-dessous quelques références générales sur la définition, les propriétés, le calcul et l'utilisation de relations médianes en analyse de données.

BARTHELEMY J.P., MONJARDET B., The median procedure in cluster analysis and social choice theory, *Math. Soc. Sci.*, 1, 1981, p. 235-268.
BARTHELEMY J.P., MONJARDET B., The median procedure in data analysis : new results and open problems, in *Classification and related methods of data analysis*, BOCK H.H. Ed., North-Holland, 1988.
MARCOTORCHINO J.F., MICHAUD P., *Optimisation en analyse des données*, Masson, Paris, 1979.
SCHADER M., *Scharfe und unscharfe Klassifikation qualitativer Daten*, Athenaum, Königstern, 1981.
TÜSHAUS U., *Aggregation binaren Relationen in der qualitativen Daten Analyse*, Athenaum, Königstern, 1983.

. Préordre (total) moyen

On cherche un préordre total le "plus proche" d'un ensemble de préordres totaux au sens suivant : les préordres sont codés par leurs vecteurs rangs (au sens de Kendall) et le préordre le plus proche est celui qui minimise la somme des carrés des distances euclidiennes entre son vecteur rang et les autres vecteurs; on l'appelle *préordre moyen* des préordres donnés. On sait (Kendall) que le même problème, où l'on cherche non plus un préordre mais un ordre total, a pour solution l'ensemble des ordres totaux contenus dans le "préordre de Borda-Kendall", i.e. celui obtenu par la somme des rangs des préordres donnés. Un résultat théorique (Lemaire) montre que le préordre moyen contient le préordre de Borda-Kendall (ses classes sont obtenues par union des classes de ce dernier) mais qu'il ne lui est pas en général égal. Dans l'article de Cook et Seiford, le préordre moyen est obtenu par un algorithme de type "branch and bound" alors que dans celui de Lemaire le problème est ramené à celui de la recherche d'un chemin de valuation minimum dans un certain graphe valué.

Signalons deux variantes dues à Armstrong et autres; dans la première on utilise le même critère mais avec la distance de la norme L^1 entre les rangs; dans la seconde, on utilise le même critère mais avec un codage différent des préordres totaux.

ARMSTRONG R.D., COOK W., SEIFORD L.M., Priority Ranking II : Consensus formation allowing incomplete ranking, *Management Science*, 28, 6, 1982, p. 639-645.
ARMSTRONG R.D., COOK W., KUNG M.T., SEIFORD L.M., Priority ranking and minimal disagreement : a weak ordering model, *R.A.I.R.O.*, 16, 4, 1982, p. 309-318.
COOK W., SEIFORD L.M., On the Borda-Kendall consensus method for priority ranking problems, *Management Science*, 28, 6, 1982, p. 621-637.
LEMAIRE J., Agrégation typologique de données de préférences, *Math. Sci. hum*, 58, 1977, p. 31-50.

. Agrégation typologique

Etant donné un ensemble de relations binaires d'un certain type, on cherche à les partitionner en classes homogènes de relations, chaque classe étant elle-même résumée par une relation. Ces méthodes utilisent à la fois une méthode de classification (à partir du choix d'une distance entre relations) et une des méthodes d'agrégation décrite ci-dessus; dans les références ci-dessous la méthode de classification est du type "nuées dynamiques".

CHANDON J.L., LEMAIRE J., Agrégation typologique de quasi-ordres : un nouvel algorithme, *Analyse des Données et Informatique*, DIDAY E. et al. Eds., I.N.R.I.A., 1977, p. 63-75.
LEMAIRE J., Agrégation typologique de données de préférences, *Math. Sci. hum*, 58, 1977, p. 31-50.

6 CONCLUSIONS

Parvenu au terme de cet inventaire, force est de constater l'étendue du domaine. Nous nous devons néanmoins de signaler qu'il ne peut être considéré comme exhaustif, tant les méthodes proposées et leurs variantes, l'on été dans des contextes extrêmement différents, sous des formes et un vocabulaire disparates. L'une des vocations de cet article est justement de proposer un langage uniforme, hérité des mathématiques discrètes, qui permet de mieux voir les rapports entre méthodes voisines. Cet effort de clarification devrait permettre de mieux cerner les vraies difficultés, mathématiques, informatiques et méthodologiques, inhérentes à ces méthodes et de progresser vers leurs solutions.

Il est évident que l'étude des conditions de validité dans les domaines d'applications, que nous n'avons nullement abordés, et qui l'a été de manière insuffisante, devrait être considérablement développée. Quant à l'applicabilité, sauf quand nous avons signalé certains problèmes comme NP-difficiles, elle n'est pratiquement qu'esquissée ici. Le sujet est d'ailleurs traité de façon très inégale par les auteurs eux-mêmes. Là encore de nombreux problèmes restent à être étudiés.

Dans ce cadre, nous nous proposons de compléter et de tenir à jour cette liste. Pour ce faire, nous souhaiterions être avertis des méthodes qui nous ont échappé, des développements récents de celles qui sont présentées ici et bien sûr de toute nouvelle méthode combinatoire ou ordinale en analyse de données. Ceci pourrait se matérialiser par une rubrique régulière de Mathématique et Sciences humaines. C'est l'occasion de rappeler que cette revue a joué un rôle non négligeable dans le développement de ces modèles et de ces méthodes; elle a en particulier consacré cinq numéros spéciaux et des bibliographies aux thèmes suivants : Modélisation des Préférences (62, 63, en 1978), Métriques et Relations (67, en 1979), Combinatoire et Analyse de Données (98, 100, en 1987).

Cet article a été ébauché dans le cadre du groupe de travail "Méthodes ordinales et combinatoires" du GRECO Analyse des Données et Informatique du C.N.R.S. (cf. le rapport d'activités du GRECO en 1986). Pour le texte actuel, les auteurs remercient B. LECLERC et V. DUQUENNE, ainsi qu'un référé anonyme pour leurs remarques constructives.

7 BIBLIOGRAPHIE

AIGNER M., *Combinatorial Theory*, Springer Verlag, Berlin, 1979.
ALBA R.D., A graph-theoretic definition of a sociometric clique, *Journal of Mathematical Sociology*, 3, 1973, p. 113-126.
ARABIE Ph., BOORMAN A., LEVITT P., Constructing block models : how and why, *Journal of Mathematical Psychology*, 17, 1978, p. 21-63.
ARABIE Ph., Review of "Group choice" by B.G. Mirkin, *Psychometrika*, 47, 3, 1982, p. 361-364.
ARDITTI D., Un nouvel algorithme de recherche d'un ordre induit par des comparaisons par paires, *Data Analysis and Informatics 3,* DIDAY E. et al. Eds., North Holland, 1984, p. 323-343.
ARMSTRONG R.D., COOK W., SEIFORD L.M., Priority Ranking II : Consensus formation allowing incomplete ranking, *Management Science*, 28, 6, 1982, p. 639-645.
ARMSTRONG R.D., COOK W., KUNG M.T., SEIFORD L.M., Priority ranking and minimal disagreement : a weak ordering model, *R.A.I.R.O.*, 16, 4, 1982, p. 309-318.
AUGUSTSON J.G., MINKER J., An analysis of some graph theoretical cluster techniques, *Journal of A.C.M.*, 17, 1970, p. 571-588.
BARBUT M., FREY L., *Techniques ordinales en analyse des données. Algèbre et Combinatoire,* Hachette, Paris, 1972.
BARBUT M., MONJARDET B., *Ordre et Classification, Algèbre et Combinatoire,* Hachette, Paris, 1970.
BARTHELEMY J.P., MONJARDET B., The median procedure in cluster analysis and social choice theory, *Math. Soc. Sci.,* 1, 3, 1981, p.235-267.
BARTHELEMY J.P., LECLERC B., MONJARDET B., On the use of Ordered Sets in Problems of Comparison and Consensus of Classifications, *J. of Classification*, 3, 1986, p. 187-224.
BARTHELEMY J.P., LUONG X., Représentation arborée des mesures de dissimilarités, *Statistique et analyse de données*, 11, 1, 1986, p. 20-41.
BARTHELEMY J.P., GUENOCHE A., HUDRY O., *Median linear orders : Heuristic and Branch and Bound Algorithms*, preprint E.N.S.T., Paris, 1987.
BARTHELEMY J.P., LUONG X., Sur la topologie d'un arbre phylogénétique : aspects théoriques, algorithmes et applications à l'analyse de données textuelles, *Math. Sci. hum.,* 100, 1987.

BARTHELEMY J.P., GUENOCHE A., *Les arbres et les représentations des proximités*, Masson, Paris, 1988.
BARTHELEMY J.P., MONJARDET B., The median procedure in data analysis : new results and open problems, *Classification and related methods of data analysis*, BOCK H.H. Ed., North-Holland, 1988.
BARTHELEMY J.P., From copair hypergraphs to median graphs with latent vertices, *Annals of discrete maths.*, 1988.
BATTEAU P., JACQUET-LAGREZE E., MONJARDET B. (édit), *Analyse et Agrégation des Préférences,* Economica, Paris, 1981.
BENZECRI J.P. et al., *L'analyse des données. 1. La taxinomie*, Dunod, Paris, 1973.
BERMOND J.Cl., Ordres à distance minimum d'un tournoi et graphes partiels sans circuits maximaux, *Math. Sci. hum.*, 37, 1972, p. 5-25.
BERMOND J.Cl., KODRATOFF Y., Une heuristique pour le calcul de l'indice de transitivité d'un tournoi, *R.A.I.R.O.*, 10, 1976, p. 83-92.
BERTIN J., *La graphique et le traitement graphique de l'information*, Flammarion, Paris, 1977.
BERTRAND P., DIDAY E., A visual representation of the compatibility between an order and a dissimilarity index: The pyramids, *Computational Statistics Quaterly*, 2, 1, 1985, p. 31-42.
BERTRAND P, *Etude de la représentation pyramidale*, Thèse de 3-ième cycle, Université Paris-Dauphine, 1986.
BIRKHOFF G., BARTEE T., *Modern Applied Algebra*, Mc. Graw-Hill, New York, 1967.
BOCK H.H., Automatische Klassifikation, *Theoretische und pratische Methoden zur Gruppierung und Strukturierung von Daten*, Vandenhoech und Ruprecht, Göttingen, 1974.
BORDAT J.P., Calcul pratique du treillis de Galois d'une correspondance, *Math. Sci. hum.*, 96, 1986, p. 31-47.
BOUCHET A., Etude combinatoire des ordonnés finis, Thèse U.S.M.G., Grenoble, 1971.
BOUCHITTE V., *Propriétés algorithmiques des extensions linéaires*, Thèse de Doctorat, Université de Montpellier, 1987.
BOUCHITTE V., HABIB M., *The calculation of invariance for ordered sets*, Rapport de recherche n° 150, E.N.S.T. Brest, 1987.
BROSSIER G., Représentation ordonnée des classifications hiérarchiques, *Statistique et Analyse des Données*, 2, 1980, p. 31-44.
BROSSIER G., Classification hiérarchique à partir de matrices carrées non symétriques *Statistique et Analyse de Données*, 7, 2, 1982, p. 22-40.
BROSSIER G., Ordonnancement de hiérarchies, Algorithmes et propriétés, *Data Analysis and Informatics 3*, Diday E. et al. Eds, North Holland, 1984, p. 317-321.
BROSSIER G., Approximation des dissimilarités par des arbres additifs, *Math. Sci. hum.*, 91, 1985, p. 5-21.
BROSSIER G., Etude des matrices de proximités rectangulaires en vue de la classification, *Rev. de Stat. Appl.*, 35, 4, 1986, p. 43-68.
BUNEMAN P.,The recovery of trees from measures of dissimilarity, *Mathematics in Archaeological and Historical Sciences*, Hodson F.R. et al. Eds, Edinburgh University Press, 1971, p. 387-395.
CARAUX G., Réorganisation et représentation visuelle d'une matrice de données numériques; un algorithme itératif, *Rev. de Stat. Appl.*, 32, 4, 1984, p. 5-23.
CARROL J.D., ARABIE Ph., Multidimensional scaling, *Annual Review of Psychology*, Rosenzweig M., Porter L. Eds, Palo Alto, 1981.
CHAH S., Calcul des partitions optimales d'un critère d'adéquation à une préordonnance, *Publications de l'I.S.U.P.*, 29, 1, 1984.
CHANDON J.L., LEMAIRE J., Agrégation typologique de quasi-ordres : un nouvel algorithme, *Analyse des Données et Informatique*, DIDAY E. et al. Eds., I.N.R.I.A., 1977, p. 63-75.
CHANDON J.L., LEMAIRE J., POUGET J., Construction de l'ultramétrique la plus proche d'une dissimilarité au sens des moindres carrés, *R.A.I.R.O.*, 14, 2, 1980, p.157-170.
CHANDON J.L., PINSON S., *Les méthodes d'analyse typologique*, Masson, Paris, 1980.
CHANDON J.L., DE SOETE G., Fitting least squares ultrametric to dissimilarity data: Approximation versus optimisation, *Data Analysis and Informatics 3*, DIDAY E. et al. Eds., North-holland, 1984, p. 213-221.
CHANDON J.L., BOCTOR F.F., Approximation d'une préordonnance par une partition, *R.A.I.R.O.*, 19, 2, 1985, p. 159-184.
CHRISTOFIDES N., BROOKER P., The optimal partitioning of graphs, *SIAM Journal Appl. Math.*, 30, 1, 1976, p. 55-69.
COGIS O., On the Ferrers dimension of a digraph, *Discrete Math.*, 38, 1982, p. 47-52.

COOMBS C.H., *Theory of data*, Wiley, New York, 1964.
COOK W., SEIFORD L.M., On the Borda-Kendall consensus method for priority ranking problems, *Management Science*, 28, 6, 1982, p. 621-637.
DAY W.H.E., The complexity of computing metric distances between partitions, *Math. Soc. Sci.*, 1, 1981, p. 269-287.
DAY W.H.E., WELLS R.S., Extremes in the complexity of computing metric distances between partitions, *IEEE Trans. Pattern. Anal. Mach. Intel.*, Vol. PAMI-6, 1, 1984, p. 69-73.
DAY W.H.E., Mc MORRIS F.R., A formalization of consensus index methods, *Bull. of Math. Biol.*, 47, 2, 1985, p. 215-229.
DEFAYS D., Recherche des ultramétriques à distance minimum d'une similarité donnée, *Bull. Soc. Roy. Sc. Lg.*, 5-6, 1975, p.330-343.
DEGENNE A., *Techniques ordinales en analyse des données: Statistique*, Hachette, Paris, 1972.
DEGENNE A., VERGES P., Introduction à l'analyse de similitude, *Revue Française de Sociologie*, 14, 1973, p. 471-512.
DEGENNE A., FLAMENT Cl., La notion de régularité dans l'analyse des réseaux sociaux, *Bull. de Méthodologie Sociologique*, 2, 1984, p. 3-16.
DEGENNE A., Présentation de l'Analyse de similitude, *Informatique et Sciences Humaines*, 67, 1986, p. 7-26.
DEGERMAN R., Ordered binary trees constructed through an application of Kendall's tau, *Psychometrika*, 47, 4, 1982.
DELARCHE M., *Quelques outils infographiques pour l'analyse structurale de systèmes*, Thèse de docteur-ingénieur, Grenoble, 1979.
DELATTRE M., HANSEN P., Bicriterion cluster analysis, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2, 4, 1980.
DE SOETE G., A least squares algorithm for fitting additive trees to proximity data, *Psychometrika*, 48, 1983, p. 621-626.
DIDAY E., Croisements, Ordres et Ultramétriques, *Math. Sci. hum.*, 83, 1983, p. 31-54.
DIDAY E., Orders and overlapping clusters by pyramids, *Multidimentional Data Analysis*, De Leeuw J. et al. Eds, D.S.W.O. Press, Leiden, 1986, p. 201-234.
DOIGNON J.P., DUCAMP A., FALMAGNE J.C., On realizable biorders and the biorder dimension of a relation, *Journal of Math. Psychol.*, 28, 1984, p. 73-109.
DUCAMP A., Sur la dimension d'un ordre partiel,*Théorie des Graphes*, Rosenstiehl P. Ed., Dunod, Paris, 1967, p. 103-112.
DUCHET P., Tree hypergraphs and their representation trees, Preprint 1987.
DUQUENNE V., *Quelques aspects algébriques du traitement des données planifiées*, Thèse de 3-ième cycle, Université R. Descartes, Paris, 1980.
DUQUENNE V., What can Lattices do for Experimental Designs ?, *Math. Social Sciences*, 11, 1986, p. 243-281.
DUQUENNE V., Contextual implications between attributes and some representation properties for finite lattices, in *Beiträge zur Begriffsanalyse*, GANTER B., WILLE R., WOLFF K.E. Eds., Wissenchaftverlag, Mannheim, 1987, p. 213-240.
DURAND C., FICHET B., One to one correspondances in pyramidal représentations: an unified approach, *Classification and related methods of data analysis*, BOCK H.H. Ed., North-Holland, 1988.
EVEN Sh., *Algorithmic Combinatorics*, Mac Millan, New York, 1973.
FLAMENT Cl., L'analyse de similitude, *Cahiers du C.E.R.O.*, 4, 2,1962, p.63-97.
FLAMENT Cl., *L'analyse booléenne de questionnaire*, Mouton, Paris, 1976.
FLAMENT Cl., Hypergraphes arborés, *Discrete Math.*, 21, 1978, p. 223-227.
FLAMENT Cl., LECLERC B., Arbres minimaux d'un graphe préordonné, *Discrete Math.*, 46, 1983, p. 854-866.
FLEGG H.G., *L'algèbre de Boole et son utilisation*, Dunod, Paris, 1967.
FREY J.J., YEHIA ALCOUTLABI A., Comparaisons par paires : une interprétation et une généralisation de la méthode des scores, *R.A.I.R.O.*, 20, 3, 1986, p. 213-227.
FULKERSON D.R., GROSS O.A., Incidence Matrices and Interval Graphs, *Pacific Journal of Math.*, 15, 1965, p. 835-855.
GANTER B., RINDFREY K., SKORSKY M., Software for concept analysis, *Classification as a tool of research*, GAUL W., SCHADER M. Eds, North Holland, 1986, p. 161-168.
GANTER B., Algorithmen zur Formalen Begriffsanalyse, *Beiträge zur Begriffsanalyse*, GANTER B., WILLE R., WOLFF K.E. Eds., Wissenchaftverlag, Mannheim, 1987, p. 241-254.
GIAKOUMAKIS V., MONJARDET B., Coefficients d'accord entre deux préordres totaux, *Statistique et Analyse des Données*, 1987, 30 p.

GOLDMANN K., Some Archaeological Criteria for Chronological Seriation, *Mathematics in the Archaeological and Historical Sciences*, Hodson F.R. et al. Eds., Edimburgh University Press, 1971, p.202-208.
GOLOMBIC M.C., *Algorithmic graph theory and perfect graphs*, Academic Press, New York, 1980.
GONDRAN M., MINOUX M., *Graphes et Algorithmes*, Eyrolles, Paris, 1979.
GOVAERT G., Classification simultanée de tableaux binaires, *Data Analysis and Informatics 3*, Diday E. et al. Eds., North-Holland, Amsterdam, 1984, p. 223-236.
GRONOFF J.D., *Heurista, logiciel d'aide à l'interprétation de données en sciences humaines*, Notice d'utilisation, E.H.E.S.S., Marseille, 1984.
GRUVAEUS G., WAINER H., Two additions to hierarchical cluster analysis, *Br. J. Math. Statist. Psychol.*, 25, 1972, p. 200-206.
GUENOCHE A., Classification using dilemma functions, *Computational Statistics Quarterly*, 2, 1, 1985, p.103-108.
GUENOCHE A., Fonctions booléennes sur un tableau en 0/1, *Data Analysis and Informatics 4*, Diday E. et al. Eds., North Holland, Amsterdam, 1986, p. 443-451.
GUENOCHE A., Représentations arborées des classifications, *R.A.I.R.O. Recherche opérationnelle*, 20, 1986, p. 341-354.
GUENOCHE A., Cinq algorithmes d'approximation d'une dissimilarité par des arbres à distances additives, *Math. Sci. hum.* 98, 1987, p. 21-40.
GUENOCHE A., Propriétés caractéristiques d'une classe relativement à un contexte, *Actes des Journées "Symbolique numérique"*, Paris, Décembre 1987.
GUENOCHE A., Méthodes combinatoires de sériation à partir d'une dissimilarité, Actes du colloque "Data Analysis and Informatics 5", Versailles, 1987, p. 115-123.
GUENOCHE A., Modèles et méthodes de sériation à partir d'une dissimilarité, preprint G.R.T.C., 1986.
GUIGUES J.L., DUQUENNE V., Familles minimales d'implications informatives résultant d'un tableau de données binaires, *Math. Sci. hum.*, 95, 1986, p. 5-18.
HANSEN P., DELATTRE M., Bicriterion cluster analysis as an exploration tool, *Multiple Criterion Problem Solving*, Lecture Notes in Economic and Mathematics Systems 155, Springer Verlag, Berlin, 1977, p. 249-273.
HANSEN P., DELATTRE M., Complete-link cluster analysis by graph coloring, *J. Amer. Stat. Assoc.*, 73, 362, 1978, p.397-403.
HARARY F., NORMAN R.Z., CARTWRIGHT D., *Structural Models, an Introduction to the theory of Directed Graphs*, Wiley, New York, 1965.
HODSON F.R., KENDALL D.G., TAUTU P., *Mathematics and Archaeological and Historical Sciences*, Edinburgh University Press, 1971.
HUBERT L.J., Some applications of graph theory and related non metric technics to problems of approximate seriation: the case of symmetric proximity measures, *Br. J. of Math. and Stat. Psychol.*, 27, 2, 1974, p. 133-153.
HUBERT L.J., Problems of seriation using a subject by item response matrix, *Psychological Bulletin*, 81, 12, 1974, p. 976-983.
HUBERT L.J., Some applications of graph theory to clustering, *Psychometrika*, 39, 1974, p.283-309.
HUBERT J.L., Data analysis implications of some concepts related to the cuts of a graph, *J. of Math. Psychol.*, 15, 2, 1977, p. 199-208.
HUBERT L.J., Generalized concordance, *Psychometrika*, 44, 2, 1979, p. 135-142.
HUBERT L.J., ARABIE P., Comparing partitions, *J. of Classification*, 2, 1985, p.193-218.
JACQUET-LAGREZE E., Analyse d'opinions valuées et graphes de préférences, *Math. Sci. hum.*, 33, 1971, p. 33-55.
JACQUET-LAGREZE E., Représentation de quasi ordres et de relations probabilistes transitives sous forme standard et méthodes d'approximation, *Math. Sci. hum.*, 63, 1978, p. 5-24
JARDINE N., SIBSON R., *Mathematical Taxonomy*, Wiley, New York, 1971.
KANO M., SAKAMOTO A., Ranking the vertices of a paired comparision digraph, *SIAM J. Alg. Discrete Math.*, 6, 1, 1985, p. 79-92.
KAUFMANN A., PICHAT E., *Méthodes mathématiques non numériques et leurs algorithmes*, 2 tomes, Masson, Paris, 1977.
KEMENY J.G., Mathematics without numbers, *Daedalus*, 88, 1959, p. 577-591.
KNUTH D., *The Art of Computer Programming*, Addison-Wesley, Reading, 1973.
KOPPEN M.G.M., On finding the bidimension of a relation, *J. of Math. Psychol.*, 31, 1987, p. 155-178.

KRIVANEK M., MORAVEK J., NP-hard problems in hierarchical-tree clustering, *Acta Informatica*, 23, 1986, p. 311-323.
KRUSKAL J., On the shortest spanning tree of a graph and the travelling salesman problem, *Proc. Amer. Math. Soc.*, 7, 1956, p.48-50.
KUNTZMANN J., NASLIN P., *Algèbre de Boole et Machines Logiques*, Dunod, Paris, 1967.
KUNTZMANN J., *Algèbre de Boole* , Dunod, Paris, 1968.
LAPORTE G., Solving a family of permutation problems, *R.A.I.R.O.*, 21, 1, 1987, p. 65-85.
LAWLER E.L., LENSTRA J.K., RINNOOY KAN A.H.G., SHMOYS D.B., *The travelling Salesman Problem. A Guided Tour of Combinatorial Optimisation,* Wiley, 1985.
LECLERC B., An application of combinatorial theory to hierarchical classification, *Recent Developments in Statistics*, Barra J.R. et al. Eds, North Holland, 1977, p.783-786.
LECLERC B., Description combinatoire des ultramétriques, *Math. Sci. hum.*, 73, 1981, p.5-37.
LECLERC B., Arbres minimum communs et compatibilités de types variés, *Math. Sci. hum.*, 98, 1987, p. 41-67.
LECLERC B., CUCUMEL G., Consensus en classification : Une revue bibliographique, *Math. Sci. hum.*, 100, 1987.
LEDLEY R., Digital electronic computers in biomedical sciences, *Science*, 130, 1959, p. 1225-1234.
LEDUC A., Chaînage automatique des matrices ordonnables, *Colloque de Micro-Info-Graphique*, Rouen, 1982, p. 1-38.
LEMAIRE J., Agrégation typologique de données de préférences, *Math. Sci. hum*, 58, 1977, p. 31-50.
LERMAN I.C., *Les bases de la classification automatique*, Gauthier-Villars, Paris, 1970.
LERMAN I.C., *Classification et analyse ordinale des données*, Dunod, Paris, 1981.
MARCOTORCHINO J.F., MICHAUD P., Heuristic approach of the similarity aggregation problem, *Methods of Oper. Research*, 43, 1981, p. 395-404.
MATULA D.W., Graph theoretic techniques for cluster analysis algorithms, *Classification and Clustering*, Van Ryson J. Ed., Academic Press, New York, 1977, p.96-129.
MICHAUD P., MARCOTORCHINO J.F., *Optimisation en Analyse ordinale des données*, Masson, Paris, 1979.
MILGRAM M., DUBUISSON B., Un algorithme heuristique de décomposition d'un graphe, *R.A.I.R.O.*, 11, 2, 1977, p. 175-199.
MIRKIN B.G., Geometrical conceptions in analysis of qualitative variables, *Quality and Quantity*, 9, 1975, p. 317-322.
MIRKIN B.G., *Qualitative attributes analysis*, (en russe), Moscou, 1976.
MIRKIN B.G., *Group choice*, Wiley, New York, 1979.
MONJARDET B., Tresses, fuseaux, préordre et topologies, *Math. Sci. hum.*, 30, 1970.
MONJARDET B., Axiomatiques et propriétés de quasi-ordres, *Math. Sci. hum.*, 63, 1978, p. 51-82.
MONJARDET B., Théorie des graphes et taxonomie mathématique, *Regards sur la théorie des graphes,* Hansen P. et al. Eds, Presses Polytechniques Romandes, 1980, p. 111-125.
MONJARDET B., Concordance et consensus d'ordres totaux : les coefficients K et W, *Revue de Statistique Appliquée*, 33, 2, 1985, p. 55-87.
MONJARDET B., NETCHINE-GRYNBERG G., Formalisation ordinale de modèles pluriels du développement psychologique, *Math. Sci. hum.*, 96, 1986, p. 65-94.
NORRIS E. M., An algorithm for computing the maximal rectangles in a binary relation, *Rev. Roum. Math. Pures et Appl.*, 23, 2, 1978, p. 243-250.
PEAY E.R., Non metric grouping : Clusters and Cliques, *Psychometrika*, 40, 3, 1975, p.297-313.
PRIM R.C., Shortest connection network and some generalizations, *Bell System Tech. Jour.*, 26, 1957, p. 1389-1401.
READ R.C. (Ed.), *Graph theory and computing,* Academic Press, New York, 1972.
REGNIER S., Sur quelques aspects mathématiques de la classification automatique, *I.C.C. Bull.*, 4, 1965, p. 175-191, repr. *Math. Sci. hum.*, 82, 1983, p. 13-29.
REINGOLD E.M., NIEVERGELT J., DEO N., *Combinatorial algorithms: Theory and Practice*, Prentice-Hall, Englewood Cliffs, 1977.
REINELT G., *The linear ordering problem : Algorithms and Application*, Heldermann Verlag, Berlin, 1985.
RIVAL I. Ed., *Ordered sets,* D. Reidel Publishing Company, Dordrecht, 1982.
ROBERTS F.S., *Discrete Mathematic models*, Prentice-Hall, Englewood Cliffs, 1976.
ROBINSON W.S., A method for chronologically ordering archaeological deposits, *American Antiquity,* 16, 1951, p.293-301.

ROHLF F.J., A new approach to the computation of the Jardine-Sibson B_k clusters, *The Computer Journal*, 18, 2, 1975, p. 164-168.
ROHLF F.J., Consensus indices for comparing classifications, *Math. Biosci.*, 59, 1982, p.131-144.
ROSENSTIEHL P., L'arbre minimum d'un graphe, in *Théorie des graphes*, Rosenstiehl P. Ed., Dunod, Paris, 1967.
ROSTAM H., *Construction automatique et évaluation d'un graphe d'implication issu de données binaires dans le cadre de la didactique des mathématiques*, Rapport de recherche 150, I.R.I.S.A., Rennes, 1981.
ROUX M., *Un algorithme pour trouver une hiérarchie particulière*, Thèse de troisième cycle, I.S.U.P., Paris, 1968.
ROUX M., Techniques of approximation for building two tree structures, Proceeding of the Franco-Japanese scientific seminar, *Recent developments in clustering and data analysis*, Tokyo, 1987, p. 127-146.
SATTAH S., TVERSKY A., Additive similarity trees, *Psychometrika*, 3, 42, 1977, p. 319-345.
SCHADER M., Distance minimale entre partitions et préordonnance dans un ensemble fini, *Math. Sci. hum.*, 67, 1979, p. 39-47.
SCHADER M., Hierarchcal analysis : Classification with ordinal object dissimilarities, *Metrika*, 27, 1980, p. 127-132.
SCHADER M., *Scharfe und unscharfe Klassifikation qualitativer Daten*, Athenaum, Königstern, 1981.
SCHADER M., TÜSHAUS U., Subgradient methods for analyzing qualitative data, in *Classification as a tool of research*, GAUL W., SCHADER M. Eds., North-Holland, 1986, p. 397-403.
SCHADER M., TÜSHAUS U., An Heuristic for Finding a Complete Preorder, *Classification and related methods of data analysis*, BOCK H.H. Ed., North-Holland, 1988.
SIBSON R., Order invariant methods for data analysis, *J. Roy. Statist. Soc. B.*, 34, 1972, p. 311-349.
SHEPARD R.N., A taxonomy of some principal types of data and of multidimentional methods for their analysis, *Multidimensional scaling: Theory and applications in the behavioral sciences, Vol. 1: Theory*, Shepard R. et al. Eds., Seminar Press, New York, 1972.
SMITH A.F.M., PAYNE C.D., An algorithm for determining Slater's i and all nearest adjoining orders, *Br. J. Math. Statist. Psychol.*, 27, 1974, p. 49-52.
TROTTER W.T. Jr., A note on Dilworth's embedding theorem, *Proc. Am. Math. Soc.*, 52, 1975, p. 33-39.
TÜSHAUS U., *Aggregation binaren Relationen in der qualitativen Daten Analyse*, Athenaum, Königstern, 1983.
VAN BUGGENHAUT J., Questionnaires booléens : schéma d'implications et degré de cohésion, *Math. Sci. hum.*, 98, 1987, p. 9-20.
VAN CUTSEM P., Ultramétriques, distances, ϕ-distances maximum dominées par une dissimilarité donnée, *Statistique et Annalyse des données*, 8, 2, 1983, p. 42-63.
WAKABAYASHI Y., *Aggregation of binary relations : algorithmic and polyhedral investigation*, Thesis, Augsburg, 1986.
WHALLON R., A new approach to pottery typology, *American Antiquity*, 37, 1, 1972, p. 13-33.
WILLE R., Restructuring lattice theory : an approach based on hierarchies of concepts, in *Ordered Sets*, Rival I. Ed., Dordrecht, Boston, 1982, p. 445-470.
ZAHN C.T. Jr., Approximating symmetric relations by equivalence relations, *J. SIAM Appl. Math.*, 12, 1964, p. 840-847.

ANNEXE : A.B.C.D: Logiciel d'Analyses Booléennes et Combinatoires de Données.

A.B.C.D. est un logiciel d'Analyses Booléennes et Combinatoires de Données composé de trois ensembles de programmes: A. SIMIL qui opère sur des tableaux de dissimilarité (**K** x **K**), A. BOOLE qui traite des tableaux en 0-1 (**I** x **J**) et A. PREF qui réalise des méthodes d'agrégation des préférences. A.B.C.D. est écrit en Basic Microsoft et implémenté sur MacIntosh. Une notice d'utilisation d'une trentaine de pages peut être demandée à l'auteur, ainsi que le logiciel dont les programmes sources sont diffusés sans restriction pour toute utilisation à but non lucratif.

A. SIMIL contient les implémentations des rubriques : Composantes connexes, Cliques maximales, Arbre minimum, Filtrant, Hiérarchies de partitions, Arbres à distances additives, Approximation arborée, Tracé d'arbres valués, Série minimum, Chaînes Robinsonniennes.
A. BOOLE contient les implémentations des rubriques : monômes vides, Graphe d'implication, couverture minimale, Graphe médian, Graphe distributif, Graphe de Buneman, Fonctions dilemmes, Treillis de Galois, Sériation.
A. PREF contient les implémentations des rubriques : Ordre à distance minimum d'un tournoi (valué ou non), Partie essentielle, Extensions linéaires, Préordre localement minimum, Ordre minimisant une fonction des rangs.

GUENOCHE A., *A.B.C.D: Logiciel d'Analyses Booléennes et Combinatoires de Données*, G.R.T.C., Marseille, 1984.

1. **Les méthodes d'Analyse de Similitude** (A. SIMIL)

A. SIMIL est un ensemble de programmes qui s'appliquent à une matrice de dissimilarité qui quantifie les écarts entre objets. Ces tableaux sont symétriques, de diagonale nulle et contiennent des valeurs positives ou nulles. Les algorithmes implémentés ont principalement deux fonctions:
- d'une part la représentation de ces écarts par des méthodes de classification ou la construction d'arbres additifs. Ces arbres sont valués et peuvent être dessinés à l'écran.
- d'autre part la sériation de ces mêmes objets, par la construction d'ordres partiels ou totaux.
Les différents programmes traitent des fichiers qui présentent l'extension .DIS. Présentons brièvement les fonctions disponibles par la liste des programmes:

- SAISIE DIS : Programme de Saisie de Mise à jour et d'Impression d'une matrice de dissimilarité ou de similitude, avec passage de l'une à l'autre par une inversion de l'ordre des valeurs qui conserve les écarts. On peut à partir de ces valeurs, calculer la distance des plus courts chemins entre objets et constituer le graphe des arêtes nécessaires pour réaliser ces plus courts chemins.

- SIMILITUDE : A partir des valeurs d'un tableau de dissimilarité considéré comme un graphe complet valué, on étudie les graphes correspondant à différents seuils, suivant les méthodes de l'analyse de similitude de A. Degenne et Cl. Flament. On construit leurs composantes connexes (ce qui revient à construire la hiérarchie de partitions du lien unique), l'arbre minimum du graphe complet et le filtrant, graphe dont les sommets sont les cliques maximales à différents seuils et les arêtes marquent les relations d'inclusion entre ces cliques.

Trois méthodes de construction d'une hiérarchie de partitions, et trois méthodes de construction d'un arbre à distances additives sont implémentées. Les arbres ainsi construits sont valués, mais on peut améliorer leurs longueurs d'arêtes de façon qu'elles approximent au mieux les dissimilarités. C'est ce que fait le programme LONGARET. De plus, tous ces arbres peuvent être tracés à l'aide du programme TRACE D'ARBRE. Pour utiliser l'un de ces programmes, il faut enregistrer l'arbre, sous un nom de fichier donné par l'utilisateur.

- C.A.H. : Programme de construction d'une hiérarchie indicée de partitions par les méthodes du lien unique, du lien moyen ou du lien complet. Les valeurs égales des dissimilarités sont prises en compte, si bien que l'arbre de classification n'est pas nécessairement binaire, et le résultat est indépendant de la numérotation des objets.

- DECOMPOSITION : Une distance additive d'arbre peut être décomposée en somme d'une distance à centre et d'une distance ultramétrique. Après avoir éventuellement modifié les

dissimilarités pour obtenir une telle distance (par une méthode itérative due à M. Roux) on soustrait une distance à centre (le centre choisi est la médiane) et on réalise une approximation des dissimilarités en construisant l'ultramétrique du lien moyen. En rajoutant cette distance à centre, on obtient un arbre à distances additives. L'algorithme utilisé est dû à G. Brossier.

- SCORES : Ce programme a pour but de construire un arbre à distances additives par évaluation des scores des paires de sommets. Pour tout quadruplet {i,j,k,l} d'une distance additive d'arbre D, il existe une bi-partition, ici {i,j} et {k,l}, telle que:

$$D(i,j) + D(k,l) \leq \text{Inf} \{ D(i,k) + D(j,l) , D(i,l) + D(j,k) \}.$$

Le score d'une paire est le nombre de quadruplets tels que cette paire est l'une des deux classes de la bi-partition. A chaque étape de l'algorithme on regroupe les sommets dont le score est maximum, ce qui crée un nouveau sommet dont la dissimilarité à un autre sommet est égale à la moyenne des dissimilarités entre ce sommet et les sommets regroupés. L'algorithme implémenté est dû à X. Luong.

- LONGARET : Si, étant donné un arbre, on considère la distance entre 2 sommets pendants comme la longueur du chemin qui va de l'un à l'autre, cette longueur est la somme des longueurs des arêtes qu'il faut emprunter. Mais la distance entre objets est donnée dans le tableau de dissimilarité. Ceci permet de calculer les longueurs des arêtes de l'arbre qui représente la hiérarchie de partitions ou la distance additive. La solution de ce problème n'est en général pas exacte, c'est à dire que les distances dans l'arbre ne peuvent être égales aux dissimilarités. On détermine alors la meilleure approximation, au sens des moindres carrés et sous contraintes positives, des longueurs des arêtes, de façon que la somme des carrés des écarts, entre les distances dans l'arbre et les dissimilarités, soit minimum.

- TRACE D'ARBRE : Ce programme permet de représenter des arbres valués, qui sont calculés par les programmes C.A.H., Décomposition, ou Scores éventuellement suivis d'une meilleure approximation des longueurs des arêtes. L'arbre peut être planté en tout sommet. Le programme réalise quatre algorithmes de tracé:
. Tracé radial, dans lequel à partir d'un sommet racine, on distribue circulairement et uniformément les sommets qui lui sont adjacents, à une distance égale à la longueur de l'arête;
. Tracé axial, dans lequel on commence par tracer un plus long chemin sur un axe horizontal, puis on fait pousser les sous-arbres dont les racines sont sur cet axe, perpendiculairement;
. Tracé arborescent, dans lequel on plante l'arbre en un sommet, ce qui définit une arborescence, et l'on place plus bas les sommets adjacents, à une distance verticale égale à la longueur de l'arête;
. Tracé planté, similaire au tracé radial, mais dans le demi-plan situé sous la racine.

- ROBINSON : Dans le modèle de sériation de Robinson, la ressemblance est une fonction décroissante de l'écart dans la série. Si l'on range le tableau de dissimilarité suivant l'ordre de cette série, à partir de tout élément en se déplaçant sur une ligne (vers la gauche ou la droite) et sur une colonne (vers le haut ou le bas), on doit rencontrer des valeurs de plus en plus grandes. C'est bien sûr une condition très forte et en général, aucun ordre total sur les valeurs d'un tableau de dissimilarité n'a la propriété de Robinson. On en est donc réduit à énumérer les ordres partiels, les chaînes, qui ont cette propriété. C'est la première fonction de ce programme. La seconde construit à partir d'une chaîne donnée un ordre qui minimise le nombre de "triplets mal ordonnés".

- SERIATION : La problème de la construction d'une série optimale, au sens d'un écart au modèle de Robinson, ou au sens d'une longueur minimum, est un problème d'optimisation sur un ensemble discret de complexité trop élevée pour ce type d'ordinateur. Aussi nous avons recours à des méthodes approchées. A partir d'une matrice de dissimilarité, on applique plusieurs heuristiques de sériation qui produisent des ordres totaux. Pour chacune des séries obtenues on calcule son écart au modèle de Robinson, la longueur de la série et son nombre de triplets mal ordonnés.

2. **Les fonctions d'analyses booléennes** (A.BOOLE)

A. BOOLE est en lui même un logiciel qui contient des programmes qui s'appliquent à un tableau de données en 0/1. Plusieurs formalisations de ces données sont utilisées, suivant le cadre méthodologique que l'on utilise. Ces données peuvent être vues comme un simple tableau de variables dichotomiques ou de présence/absence de caractéristiques, comme le codage d'un graphe

biparti (lignes et colonnes constituent les deux classes de sommets), comme un hypergraphe dont chaque colonne est une hyperarête, ou dans le modèle booléen. Dans ce cas, les colonnes du tableau sont les générateurs d'une algèbre de Boole, notés A, B, C, etc... Introduisons quelques points de vocabulaire.
Chaque générateur (par exemple A) peut figurer sous forme directe notée a correspondant aux 1 ou sous forme complémentée notée a' correspondant aux 0. a et a' désignent également les ensembles de lignes qui ont les valeurs 1 et 0 pour A. Les lignes distinctes sont appelées patrons ; ils peuvent être pondérés. Une conjonction de générateurs sous forme directe ou complémentée est un monôme; ce monôme est dit complet, si tous les générateurs y sont présents. A un monôme donné correspond un ensemble de patrons, ceux qui possèdent ces générateurs sous la même forme. On dit que ce monôme contient ces différents patrons. Un monôme est dit vide si aucun patron n'y est inclus.

Présentons brièvement les différents programmes:

- SAISIE 0/1 : Ce programme permet de tirer au hasard (avec une probabilité fixée de 1) et de saisir depuis le clavier ou dans un fichier "texte", sous forme de vecteurs ou sous forme de patrons pondérés (tous distincts), un tableau en 0/1 d'au plus 25 générateurs et de l'archiver. On peut mettre à jour ce tableau soit en inversant une colonne, soit en ajoutant ou supprimant une colonne ou un patron, soit en modifiant un patron. On peut également en extraire un nouveau tableau par une sélection des colonnes. Les nouveaux patrons résultant de ces mises à jour sont calculés avec leurs nouvelles pondérations. On peut également calculer soit entre les patrons, soit entre les générateurs la distance de la différence symétrique et archiver le tableau de dissimilarité ainsi constitué pour lui appliquer les méthodes de A.SIMIL.

- IMPLICATION : Une relation d'implication entre 2 variables A et B, par exemple a' --> b s'entend comme: "si la variable A prend la valeur 0, alors B prend la valeur 1". Cette relation doit être vraie sur tout le tableau de données. Cela veut dire que pour aucune ligne du tableau on ne trouve la combinaison 0 0 pour ces deux générateurs.
Dans ce programme on énumère d'abord toutes les implications portant sur deux variables par la recherche des combinaisons non attestées encore appelées cases vides. Puis nous construisons des implications non systématiques, mais seulement vraies dans un certain pourcentage de cas spécifié par l'utilisateur. Cette méthode amène à construire une famille de graphes d'implication.

- KUNTZMANN : La méthode de Kuntzmann mise en oeuvre ici, énumère tous les monômes vides du tableau, quel que soit leur nombre de générateurs. On les traduit aisément sous forme d'implications, mais si rien ne limite dans cette méthode le nombre de variables à considérer, au delà de trois, l'interprétation des implications n'a en général rien de naturel et par ailleurs le temps de calcul et les places mémoires nécessaires sont excessives. On peut n'obtenir que les monômes vides avec un nombre limité de générateurs, ainsi que ceux qui ne présentent au plus qu'un seul générateur sous forme conjuguée.

- RECOUVREMENT : Ce programme construit un recouvrement d'une classe de patrons du tableau 0/1, définie par un monôme (a ou a') construit sur un seul générateur (A). S'il s'agit de recouvrir a, on énumère les monômes vides du sous tableau correspondant à a'. Chaque monôme vide de a' qui n'est pas vide sur a couvre une partie de a. Une couverture de a est donc réalisée par une union de ces monômes. On peut, pour tout autre générateur que celui qui sert à définir la classe à recouvrir, indiquer si l'on accepte leur forme complémentée (on peut ainsi considérer des variables mutuellement exclusives). Un recouvrement de a par une union de monômes vides sur a' est une fonction caractéristique de a.

- DILEMME : Ce programme permet d'énumérer, à partir d'un tableau en 0/1, les fonctions dilemmes (disjonctions exclusives de conjonctions de générateurs sous une forme quelconque) construites sur au plus 3 générateurs. Au choix (par l'utilisateur) d'une de ces fonctions correspond une partition de l'ensemble des patrons dont chaque classe reçoit un numéro. Chaque classe peut ainsi être partitionnée ce qui permet d'obtenir un arbre de subdivision qui n'est pas nécessairement binaire.

- GALOIS : Un tableau en 0/1 définit une correspondance entre parties de l'ensemble des lignes et des colonnes. Cette correspondance peut se représenter par son treillis de Galois dont les sommets sont les produits cartésiens de 2 parties. A x B est un sommet si les ensembles A et B sont en

correspondance, c'est à dire que le sous-tableau A x B ne contient que des 1. De plus ces blocs sont maximaux, c'est à dire que si l'on ajoute une ligne ou une colonne à A ou B le sous tableau A x B ne contient plus que des 1. Ce programme énumère tous les blocs maximaux qui sont contenus dans le tableau 0/1 initial. De plus il construit les arêtes du treillis qui représentent les relations d'inclusion entre les parties A ou B. L'algorithme utilisé est celui de Bordat.

- GRAPHE MEDIAN : Pour obtenir une représentation des proximités entre lignes d'un tableau en 0/1 à m colonnes, on considère les lignes comme les sommets d'un graphe et une arête lie deux sommets s'ils ne diffèrent que par la valeur d'une seule colonne. Autrement dit ce graphe est une partie du simplexe $\{0,1\}^m$. La longueur du plus court chemin entre deux sommets est égale au nombre de colonnes à valeurs distinctes entre les deux lignes correspondantes. Si l'on s'en tient aux sommets du simplexe présents dans le tableau, ce graphe n'est en général pas connexe. Après l'avoir construit et au prix de l'ajout d'un certain nombre de sommets "virtuels" qui n'introduisent pas certaines combinaisons nouvelles des colonnes deux à deux, on obtient un graphe connexe (l'utilisateur choisit entre le graphe médian ou le graphe distributif). Celui ci constitue une représentation exacte des écarts entre patrons, dans la mesure ou la distance de la différence symétrique est exactement représentée.

- SCALOGRAMME : Ce programme détermine à partir d'un tableau en 0/1, un ordre sur les lignes et les colonnes tel que le tableau réordonné présente le plus de 1 au voisinage de la diagonale. C'est donc une méthode de sériation, aussi connue sous le nom de "méthode des moyennes réciproques". L'algorithme itératif implémenté ici est dû à K. Goldmann. A chaque étape, on ordonne lignes et colonnes suivant l'ordre croissant des moyennes des rangs des 1, jusqu'à ce que ces ordres soient stables. Cette méthode n'est applicable que s'il y a au moins un 1 par ligne et par colonne; sinon ils faudra supprimer ces vecteurs qui ne contiennent que des 0. Le nombre d'itérations est en général petit (de l'ordre de 5); c'est donc une méthode rapide. Elle produit l'ordre obtenu sur les lignes et les colonnes et réordonne le tableau qui peut être imprimé.

3. **Les méthodes d'agrégation des préférences** (A. PREF)

On adopte ici la terminologie classique de l'agrégation des préférences, à savoir un certain nombre de "votants" expriment leurs préférences sur des "candidats".

- SAISIE : Ce programme permet d'exprimer des préférences. Celles-ci peuvent être définies par des notes, des ordres ou des préordres, ou enfin par un tournoi valué ou non. Chacune de ces données est entrée par un éditeur de texte, suivant un format défini et peu contraignant. Suivant la méthode ultérieurement utilisée, ce programme permet de passer d'un type de données à l'autre, essentiellement des notes aux ordres ou préordres, des ordres ou préordres au tournoi ou au tournoi majoritaire, ou d'en supprimer la valuation. Chaque type de données est un fichier de type texte qui peut être mis à jour au moyen d'un éditeur. Chaque type est caractérisé par une extension (.not, .ord, .tour) au nom de fichier donné par l'utilisateur.

- HEURISTIQUES : Ce programme applique un certain nombre de méthodes heuristiques qui calculent autant d'ordres totaux, dont on espère que leur distance à un tournoi valué sera la plus petite possible. En dehors d'heuristiques basées sur le "bon sens", on applique celle de Smith et Payne au tournoi considéré comme non valué et celles de Barthélemy, Guénoche et Hudry, celle des doubles décalages de Michaud et Marcotorchino, ainsi que l'ordre minimisant une fonction des rangs, méthode due à Frey et Yehia Alcoutlabi. Pour chacune des solutions trouvées, on examine les ordres voisins sur le permutoèdre (par permutation de deux candidats consécutifs) jusqu'à atteindre un minimum local.

- TOURNOI : Ce programme énumère les ordres à distances minimum d'un tournoi valué ou non. Autrement dit, il construit tous les ordres médians d'un profil d'ordres (ou de préordres), totaux ou partiels. Dans une première étape, il calcule de "bons" ordres par des heuristiques basées sur l'élimination des cycles de longueur 3, ce qui permet d'obtenir une borne supérieure de la distance minimum d'un ordre total au tournoi. Ensuite, il met en œuvre une procédure "branch and bound" pour construire tous les ordres médians qui sont listés. Une arborescence jusqu'à 5000 sommets peut être développée.

-PREORDRE : Ce programme construit un préordre dont la distance aux ordres (ou préordres) initiaux est un minimum local de cette fonction de distance. De fait il part d'un préordre donné qui peut être tiré au hasard (dans ce cas, c'est un ordre total) et par permutation et réunion des classes d'équivalence du préordre, ou par création d'une nouvelle classe, il construit une séquence de préordres dont la distance au profil initial va décroissant. L'algorithme utilisé est dû à Schader et Tüshaus.

- EXTENSIONS : Ce programme construit toutes les extensions linéaires d'un ordre partiel donné. Celui-ci est défini par un ensemble de chaînes totalement ordonnées. Le programme calcule simultanément la partie essentielle de l'ordre partiel initial, puis liste les ordres totaux compatibles avec cet ordre partiel, s'il y en a; c'est dire qu'il détecte les cycles.