CLASSIFICATION DES DONNEES

(cas non supervisé)

I.C. LERMAN

A. PROPOSITION POUR UNE ORGANISATION GENERALE DES METHODES

Le non initié qui aborde l'étude de la Classification retire le sentiment d'une grande profusion de méthodes sans liens évidents entre elles.

En réalité - comme nous allons chercher à le montrer - la Classification offre un point de vue tout à fait original et fécond dans la représentation initiale puis synthétique des données. D'autre part, la classification fournit une riche méthodologie qui a un caractère parfaitement autonome pour l'analyse des données.

Relativement à une problématique supposant la collecte d'une grande masse de données, un problème de reconnaissance globale est posé. Pour ce problème on suppose pouvoir construire un tableau de données chiffrées ou codées, à double entrée, croisant un ensemble E d'objets ou individus avec un ensemble V de variables descriptives d'un seul type, quelconque, ou d'ailleurs, de plusieurs types.

L'objet de la classification automatique est de fournir la "meilleure" organisation en classes et sous-classes de "proximité" aussi bien de

l'ensemble E des objets que de celui V des variables de description. Intuitivement et très approximativement parlant, si C désigne l'un des deux ensembles (E ou V), une classification (i.e. partition) $\{C_1, C_2, \ldots, C_j, \ldots, C_k\}$ répond au problème si la densité des paires d'élements $\{c,d\}$ "proches" (resp. "éloignées") est nettement plus grande si c et d sont pris dans une même classe C_j (resp. dans deux classes distinctes C_j et C_h, $j \neq h$). En jouant sur le degré de proximité, on obtient un système de classes emboîtées ou empiétantes.

On peut vouloir dans la situation la plus classique - celle d'un tableau objets ou individus × variables numériques $\{x_i^j = v^j(o_i) / (i,j) \in I \times J\}$ où I (resp. J) indexe l'ensemble des objets (resp. variables) - présenter l'approche classificatoire par rapport à l'approche linéaire et géométrique, fournie dans ce cas par l'Analyse en Composantes Principales (A.C.P.). Alors que pour l'approche A.C.P., on propose une synthèse de l'ensemble des variables $\{v^j / j \in J\}$ au moyen de facteurs "indépendants" et de forte variance, dont chacun s'exprime au moyen d'une combinaison *linéaire* $\phi = \Sigma\{\phi_j v^j / j \in J\}$, la Classification quant à elle, offre une décomposition en classes et sous-classes de proximité qui, à un niveau de relation, sont "relativement indépendantes"; donc en "facteurs et sous facteurs". Mais, ce n'est là qu'un aspect de la comparaison.

Nous allons commencer par considérer quelques caractéristiques dichotomiques très générales qui permettent de situer de façon schématique une méthode ou un aspect d'une méthode de classification des données.

1. Caractère hiérarchique ou non hiérarchique de l'ensemble des classifications pouvant être obtenues par l'algorithme

Dans le premier cas, la suite $(P_0, P_1, \ldots, P_j, \ldots, P_k)$ des partitions qu'on obtient est totalement ordonnée par finesse décroissante :

$P_0 < P_1 < \ldots < P_j < \ldots < P_k$; en d'autres termes, il s'agit d'une suite "emboîtée" de partitions, où P_j se déduit de $P_{(j-1)}$ par réunion d'une ou de plusieurs paires de classes. D'ailleurs, un algorithme de classification hiérarchique procède en général par agrégations successives des classes les plus "proches".

Si un algorithme de classification hiérarchique produit en une étape de son exécution une suite ordonnée de partitions, le propre d'un algorithme de classification non-hiérarchique est de bâtir - en une étape de son exécution - une partition dont le nombre k de classes est, à peu de choses près, fixé.

Lorsque l'état final de l'algorithme de classification non-hiérarchique - avec un nombre k de classes présupposé - dépend de son état initial (e.g. méthode des "nuées dynamiques"), une suite de répétitions indépendantes de l'algorithme conduit à une suite de partitions dont on considère l'intersection (dont les classes sont appelées par Diday "formes fortes") et la réunion (dont les classes sont appelées "formes faibles").

Lorsque le résultat de l'algorithme (de classification non-hiérarchique) est par nature stable et ne dépend pas d'un état initial qu'on peut choisir a priori (e.g. méthode des "pôles d'attraction"), la répétition de l'algorithme pour k variant : $k = 2,3,\ldots,\ell$, conduit à une structure descendante mais *non hiérarchisée* de partitions $(Q_2,Q_3,\ldots,Q_\ell)$.

2. Représentation géométrique ou ensembliste des variables de description

La prise en compte de variables qualitatives est très souple dans le cadre d'une approche classificatoire où on peut considérer une représentation ensembliste définie relativement à l'ensemble E des objets ou individus décrits.

De façon générale la représentation géométrique d'une variable se conçoit comme la projection - conformément à une métrique choisie pour des raisons algébriques et statistiques - sur un axe ou, plus généralement, sur un sous-espace. De façon duale et relativement à $\mathbb{R}^n$ ($n = \text{card}(E)$), une même variable est représentée par un vecteur, voire un sous-espace de dimension 1 ou supérieure à 1 (dans le cas de variables qualitatives nominales). On ne voit pas comment représenter géométriquement une variable qualitative ordinale.

Pour ce qui est de la représentation ensembliste des variables, nous distinguons deux principaux types :

I. Variables d'incidence ou de contingence

II. Variables relationnelles.

Dans le cadre de I, on distingue

I.1. l'attribut descriptif ou variable logique de présence-absence :

$a : E \to \{0,1\}$

$x \to a(x)$

a est représenté par $E(a) = a^{-1}(1)$.

I.2. si on assimile chaque objet à une classe d'une partition donnée de l'ensemble E des individus, notée $P = \{E_1, E_2, \ldots, E_i, \ldots, E_{|I|}\}$, une même modalité j d'une variable qualitative, définit dans son croisement avec P, une partition $Q_j = \{F_{ji} = E_i \cap F_j / 1 \leq i \leq |I|\}$ de la classe F_j d'individus possédant la j-ème modalité, $1 \leq j \leq |J|$.

I.3. la variable quantitative v qui est regardée comme définissant une "pondération" sur E :

$v : E \to \mathbb{R}$

$x \to v(x)$.

Dans le cadre II, on distingue

II.1. la variable qualitative nominale qui se présente comme une application de E dans l'ensemble sans structure de codes $\{0,1,2,\ldots,(k-1)\}$:

$$c : E \to \{0,1,2,\ldots,(k-1)\}$$
$$x \to c(x).$$

La variable c - à k modalités - définit une partition $\{E_0,E_1,\ldots,E_{(k-1)}\}$ où $E_j = c^{-1}(j), 0 \leq j \leq (k-1)$. Cette partition peut être représentée au niveau de l'ensemble $P_2(E) = \{\{x,y\}/x,y \in E, x \neq y\}$ des paires ou parties à deux éléments de E, par le sous ensemble $R(\pi)$ (resp. $S(\pi)$) formé des paires réunies (resp. séparées) par la partition π :

$$R(\pi) = \Sigma\{P_2(E_j)/0 \leq j \leq (k-1)\}$$

(resp.

$$S(\pi) = \Sigma\{E_j * E_h/0 \leq j < h \leq (k-1)\}).$$

Avec une telle représentation, comparer 2 variables (qual. nom.) c_1 et c_2, revient à comparer $R(\pi_1)$ et $R(\pi_2)$ (resp. $S(\pi_1)$ et $S(\pi_2)$) où π_1 et π_2 sont respectivement associées à c_1 et c_2.

II.2. la variable qualitative ordinale qui se présente comme une application de E dans l'ensemble totalement ordonné des codes $\{0,1,2,\ldots,(k-1)\}$ $(0 < 1 < 2 < \ldots < (k-1))$. La variable à k modalités - que nous notons encore c - définit un préordre total à k classes $E_0,E_1,\ldots,E_{(k-1)}$, où $E_j = c^{-1}(j)$ et où $E_0 < E_1 < \ldots < E_{(k-1)}$ pour l'ordre quotient. Désignons par ω le préordre total sur E. Ce préordre total peut être représenté au niveau de l'ensemble $E \times E$ par

$$R(\omega) = \Sigma\{E_j \times E_h/1 \leq j < h \leq (k-1)\}.$$

On peut également considérer la variable qualitative partiellement ordinale

II.3. la variable "rang" est généralement associée à une variable "note" qui permet d'ordonner totalement et strictement l'ensemble E des individus ou objets. Cela suppose que la taille n de E est suffisamment petite et la finesse de la variable assez grande de telle sorte que l'application qui associe à chaque individu de E sa note, soit injective. La variable "rang" r se présente comme une application de E dans l'ensemble des n premiers entiers :

$$r : E \to \{1,2,\ldots,n\}$$
$$x \to r(x) = \text{card}\{y/y \in E \text{ et } y \leq x\}.$$

En désignant par o l'ordre total et strict défini sur E par la variable "rang", on peut représenter cette variable au niveau de E×E par

$$R(o) = \{(x,y)/(x,y) \in E \times E \text{ et } r(x) < r(y)\}.$$

Pour une même variable relationnelle, un codage naturel de la variable - au niveau de l'ensemble des couples d'objets - correspond à considérer la fonction indicatrice du sous-ensemble de E×E qui représente la variable. Toutefois, on peut a priori envisager d'autres codages; ainsi, la variable qualitative ordinale (cf. II.2) peut être codée comme suit :

$$(\forall\ (x,y) \in E \times E)\ ,\ \phi(x,y) = 1,\ 0 \text{ ou } -1$$

respectivement, selon qu'il existe $0 \leq j < h \leq (k-1)$ tels que

$$(x,y) \in E_j \times E_h\ ,\ (x,y) \in E_\ell \times E_\ell \quad (\text{pour } \ell = j \text{ ou } h) \quad \text{ou}$$

$$(x,y) \in E_h \times E_j\ .$$

II.4. la variable "pondération" sur E×E qui détermine un "graphe valué sur E" peut d'une certaine manière être considéré comme une généralisation

du précédent codage par couples. Cette variable est représentée par la matrice carrée $\{\mu_{xy}/(x,y) \in E \times E\}$.

A partir du codage au niveau de $E \times E$ d'une variable relationnelle, on peut toujours représenter une telle variable par un vecteur de $\mathbb{R}^{n^2}$. Mais la géométrie par trop riche de cet espace est démesurée par rapport au sens du codage, de sorte qu'il est difficile d'exploiter les propriétés géométriques de l'espace de représentation et de les interpréter dans une analyse de données qualitative .

Le mieux qu'on puisse faire géométriquement par rapport à E, est de représenter une variable par un vecteur ou sous-espace de $\mathbb{R}^n$. Nous avons déjà mentionné qu'une variable numérique v peut être représentée par le vecteur $(x_1,x_2,\ldots,x_i,\ldots,x_n)$, où x_i, $1 \leq i \leq n$, est la mesure de v sur l'objet codé i.

D'autre part, à une variable qualitative nominale on associe le sous-espace de $\mathbb{R}^n$ engendré par les indicatrices des modalités.

Enfin, il arrive qu'on exploite une interprétation géométrique de la représentation d'une variable "rang" au moyen du vecteur $(r_1,r_2,\ldots, r_i,\ldots,r_n)$ où r_i est le rang de l'individu codé i pour la variable. Cette exploitation s'entend lors de la comparaison de plusieurs variables définissant chacune un ordre total et strict sur E.

III. Représentation des objets

Reprenons la situation la plus classique n objets $\times$ m variables numériques où le tableau des données s'écrit

$$\{x_i^j = v^j(o_i)/(i,j) \in I \times J\}.$$

C'est relativement à la représentation essentielle d'une variable numérique par une forme linéaire coordonnée de $(\mathbb{R}^m)^*$ que se conçoit la représentation de l'ensemble E des objets par un nuage de points dans $\mathbb{R}^m$ muni d'une métrique définie positive :

$$N(I) = \{(o_i, \mu_i)/i \in I\}$$

où μ_i est un poids affecté - pour des raisons statistiques - à l'objet o_i , $1 \leq i \leq n$.

C'est ce type de représentation de l'ensemble des objets qui est fondamentalement considéré lorsqu'on fait admettre aux variables une représentation géométrique.

Dans le cas où les variables sont qualitatives d'une même type et dans la mesure où une même variable se présente comme une application dans une échelle discrète (nominale, ordinale (totalement ou partiellement)), l'ensemble des objets peut également être représenté par un nuage de points dans un espace discret de la forme $\prod_{1 \leq j \leq m} E_j$, où E_j est l'échelle des valeurs de la j-ème variable. On remarquera que cette représentation n'est pas possible dans le cas II.4.

Toutefois, dans le cas qualitatif, un objet peut être représenté de façon ensembliste à partir de l'ensemble des modalités qu'il possède des différentes variables.

3. Nature inertielle ou covariationnelle-corrélationnelle des indices et critères

Les notions d'indices de proximité et de critères de classification se réfèrent de façon essentielle à deux concepts fondamentaux; l'un d'eux est celui de "distance" et l'autre, celui de "corrélation".

Le propre de l'analyse des données est de passer de l'un des concepts à l'autre et ce passage est d'une grande richesse et fécondité méthodologiques. Mais de façon primitive et essentielle, la notion de "corrélation" s'adresse d'abord à la comparaison de variables et la notion de "distance" à la comparaison d'objets géométriquement représentés par des points d'un espace euclidien affin.

L'interprétation "linéaire" du coefficient de corrélation entre deux variables numériques v et w est bien connue. Mais il faut savoir que cet indice qui se met sous la forme

$$\rho(v,w) = \frac{\sum_i x_i y_i - n\,\bar{x}\,\bar{y}}{\sqrt{(\sum_i (x_i-\bar{x})^2)(\sum_i (y_i-\bar{y})^2}} ,$$

s'obtient à partir d'une approche parfaitement non-linéaire dans le cadre de la statistique non-paramétrique : en partant de l'indice $\Sigma\{x_i y_i / 1 \le i \le n\}$ (que nous appelons "brut") , on associe la v.a. $\Sigma\{x_{\sigma(i)} y_{\tau(i)} / 1 \le i \le n\}$ où (σ,τ) est un couple de permutations aléatoires indépendantes de l'ensemble G_n - muni d'une probabilité uniforme - des $n!$ permutations sur $(1,2,\ldots,i,\ldots,n)$ qui code E. Le calcul de la moyenne et de la variance de cette v.a., montre que l'indice brut centré et réduit n'est autre - au coefficient $\sqrt{(n-1)}$ près - que le coefficient $\rho(v,w)$.

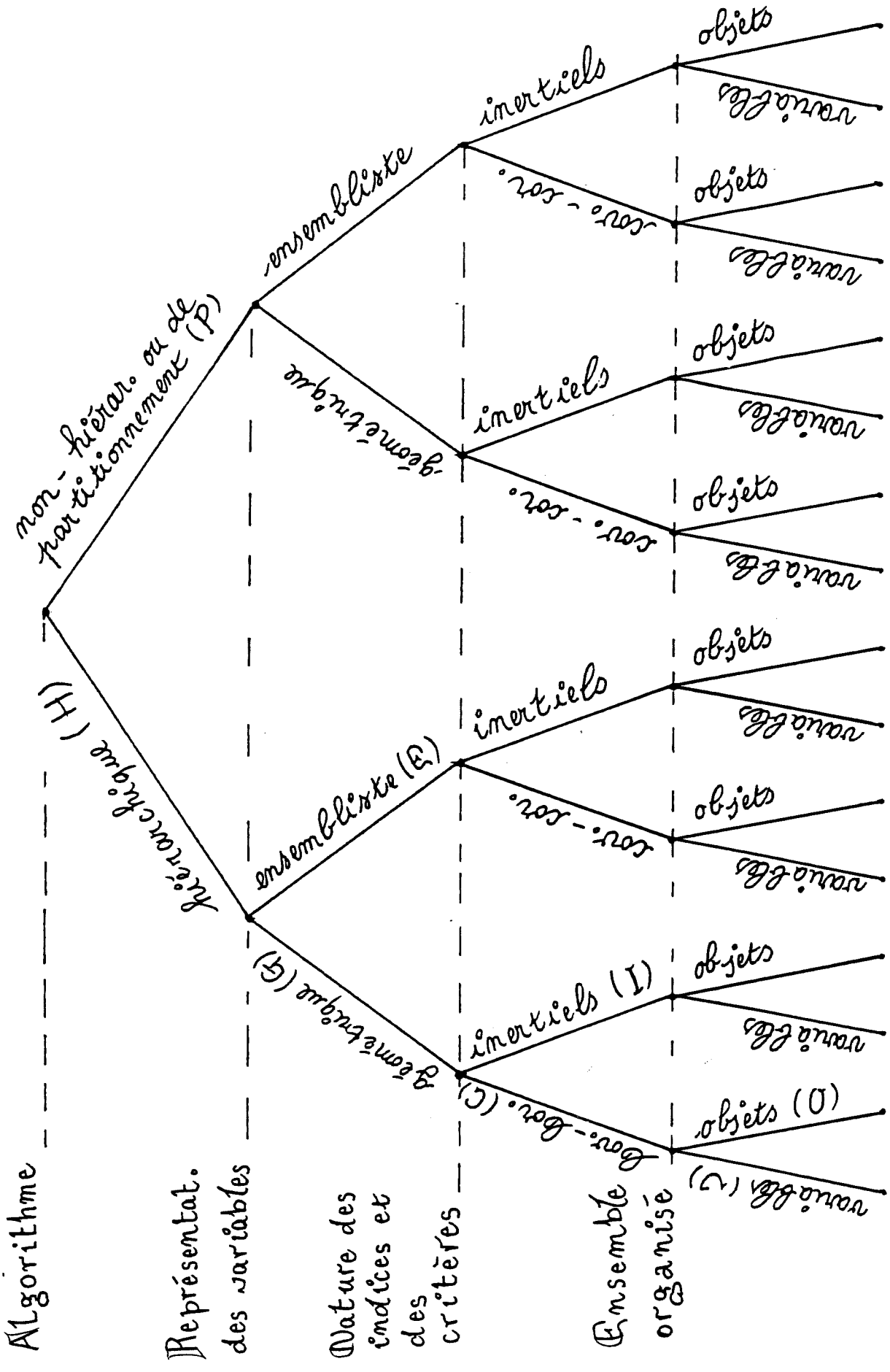
Cette approche s'exprime très bien dans le cadre de la représentation ensembliste des variables où elle mène - après un point de vue combinatoire - à des indices tout à fait originaux. Signalons - à titre d'illustration - ce que devient cette démarche dans le cas de la comparaison de deux attributs descriptifs a et b; cas le plus simple entre tous.

L'indice brut est alors $s = \text{card}[E(a) \cap E(b)]$. A $(E(a),E(b))$, on associe le couple (X,Y) de parties aléatoires indépendantes où X (resp. Y) "respecte" la cardinalité de $E(a)$ (resp. $E(b)$). A s se trouve ainsi associée la v.a. $\text{card}(X \cap Y)$ dont la moyenne et la variance permettent de "standardiser" s et d'obtenir au coefficient $\sqrt{n}$ près, le nouveau type de coefficient de corrélation ou d'association.

L'analyse factorielle a beaucoup développé les métriques des espaces euclidiens affins de représentation, qui permettent de traduire les dissemblances entre objets représentés par des points. Comme nous l'avons déjà mentionné, il pourra être question - pour une représentation vectorielle des variables - d'une notion de distance entre variables.

D'autre part, nous avons introduit pour la représentation ensembliste des variables, une notion de proximité entre objets qui a un sens corrélatif.

Dans ces conditions, nous aboutissons au schéma suivant où une même méthode - ou à tout le moins un de ses aspects - peut être décrite par une chaîne qui correspond en quelque sorte à sa signature. Dans le cadre de ce schéma, on suppose que toutes les variables induisent le même type de structure sur l'ensemble des objets.


Algorithme
Représentat. des variables
Nature des indices et des critères
Ensemble organisé
non-hiérar. ou de partitionnement (P)
hiérarchique (H)
ensembliste
géométrique
ensembliste (E)
géométrique (G)
inertiels
non-inertiels
inertiels
non-inertiels
inertiels
non-inertiels
inertiels (I)
non-inertiels (C)
objets
nuées
objets
nuées
objets
nuées
objets
nuées
objets
nuées
objets
nuées
objets
nuées
objets (O)
nuées (N)

Relativement à un type de méthodes, il y a toujours un ou plusieurs cadres très naturels pour la conception de la méthode (e.g. tableau de contingence pour la classification des moindres-carrés basée sur le χ^2, tableau d'incidence pour l'A.V.L., tableau de données numériques pour une méthode de reallocation-recentrage,...). Toutefois, un même type d'approche que traduit une méthode, est plus ou moins naturellement généralisable dans la structure des données qu'il peut prendre en compte et dans la forme des structures de condensation pouvant être mises en évidence. Cette extension est très féconde, aussi bien sur le plan théorique où elle conduit à des concepts nouveaux que sur le plan des applications où elle permet d'affiner les résultats.

Enfin, une approche peut être plus ou moins riche et soulever une problématique qui lui est propre.

Ce schéma permet d'entrevoir toute la variablilité méthodologique en Classification et il est exclu de présenter l'ensemble des méthodes. Nous allons nous contenter de présenter le principe de quelques méthodes classiques qui sont implémentées dans MODULAD.

Il faut savoir que dans le cadre d'une même méthode - signée par un concepteur - il y a d'une part l'algorithme qui utilise un type de critère pour l'émergence des classes, il y a d'autre part, toute une approche de nature statistiquo-algorithmique pour la préhension de l'interprétation significative des classifications.

Conformément au schéma, chaque méthode ou aspect d'une méthode sera accompagné de sa signature. Nous présenterons - en respectant l'ordre chronologique de leur apparition - deux méthodes de classification hiérarchique et trois méthodes de classification non-hiérarchique.

Les deux méthodes de classification hiérarchique sont :

H1. Classification hiérarchique basée sur le critère de l'"inertie expliquée" (HGIO) [WARD(1963)],
H2. Classification hiérarchique basée sur le critère général de la "vraisemblance du lien" (HEGV) [LERMAN(1970)].

Les trois méthodes de classification non-hiérarchique sont :

P1. "k-means" ou algorithme de "reallocation-recentrage" autour des centres mobiles de gravité" (PGIO) [HALL & BALL(1965)],
P2. "nuées dynamiques" (PGIO) [DIDAY(1972)],
P3. "pôles d'attraction" (PGIO) [LERMAN - LEREDDE(1977)].

B. PRESENTATION SUCCINTE DU PRINCIPE DE QUELQUES METHODES FONDAMENTALES

H1. Formation ascendante hiérarchique de l'arbre des classifications basée sur le critère de l'inertie expliquée (HGIO)

Il y a lieu ici de se référer à la représentation de l'ensemble O des objets au moyen d'un nuage de points

$$N(I) = \{(o_i, \mu_i)/i \in I\} \qquad (1)$$

de $\mathbb{R}^m$ muni d'une métrique définie positive. μ étant la masse totale du nuage $(\mu = \sum_{i \in I} \mu_i)$, le centre de gravité g du nuage est le point défini par l'écriture ponctuelle :

$$g = \frac{1}{\mu} \sum_i \mu_i o_i \quad . \qquad (2)$$

Le moment total d'inertie est défini par

$$\sum_{i\in I} \mu_i \, ||o_i - g||^2 \, . \qquad (3) \qquad (3)$$

Relativement à une partition $\{I_h / 1 \leq h \leq k\}$ de I qui indique une partition de $\mathcal{O}$, on a la décomposition suivante de (3) :

$$\sum_{1\leq h\leq k} \sum_{i\in I_h} \mu_i \, ||o_i - g_h||^2 + \sum_{1\leq h\leq k} \eta_h \, ||g_h - g||^2 \, , \qquad (4)$$

inertie perdue — inertie expliquée

où g_h désigne le centre de gravité de la h-ème classe :

$$g_h = \frac{1}{\eta_h} \sum_{i\in I_h} \mu_i o_i \, , \qquad (5)$$

avec $\eta_h = \sum_{i\in I_h} \mu_i$: masse de la h-ème classe.

Partant de la partition la plus fine où chaque classe contient un seul objet, on procède par fusions successives de paires de classes : un même pas de l'algorithme consiste à réunir la paire de classes qui rend minimale la variation du critère global que définit l'inertie expliquée ($\Sigma\{\eta_h \, ||g_h - g||^2 / 1\leq h\leq k\}$).

Si on considère la partition $\{I_h / 1\leq h\leq k\}$, la part locale - du critère global - affectée par la fusion de deux classes I_j et $I_{j'}$ est égale à

$$\Delta(j,j') = \frac{\eta_j \eta_{j'}}{(\eta_j + \eta_{j'})} \, ||g_j - g_{j'}||^2 \, . \qquad (6)$$

On remarquera avec intérêt que cette part se réduit à un critère purement local; c'est-à-dire, ne faisant intervenir que des caractéristiques liées aux deux classes codées I_j et $I_{j'}$.

Dans ces conditions, une même étape de l'algorithme représenté par un même niveau de l'arbre des classifications consiste en la fusion de la paire de classes (resp. des paires de classes en cas d'ex aequo) qui rend (resp. rendent) minimale (6).

La matrice des distances pondérées (6) entre classes est - après la réunion de deux classes d'indices j et j' - réactualisée au moyen de la formule :

$$\Delta(h,j\cup j') = \frac{\eta_h+\eta_j}{(\eta_h+\eta_j+\eta_{j'})}\ \Delta(h,j) + \frac{\eta_h+\eta_{j'}}{(\eta_h+\eta_j+\eta_{j'})}\ \Delta(h,j')$$

$$- \frac{\eta_h}{(\eta_h+\eta_j+\eta_{j'})}\ \Delta(j,j') \ . \qquad (7)$$

Lorsque plusieurs paires $\{j,j'\}$ de classes réalisent "en même temps" la valeur minimale de (6). Ces agrégations sont repérées par des noeuds d'un même niveau de l'arbre "détaillé" des classifications. Ainsi, on fusionne toutes les paires de telles classes $\{j,j'\}$ avant de consulter à nouveau la matrice des distances pondérées (6), laquelle étant toutefois réactualisée au fur et à mesure, après chaque agrégation de deux classes.

H2. Formation ascendante hiérarchique de l'arbre des classifications basée sur le critère général de la "vraisemblance du lien" (HEGV)

Nous allons considérer le cas le plus simple où il s'agit de découvrir une structure en classes et sous-classes de "proximité" sur un ensemble A d'attributs descriptifs. Insistons sur le fait que la méthode admet à son entrée n'importe quelle structure mathématique du tableau des données et permet aussi bien l'organisation des lignes que des colonnes du tableau.

On commence - conformément au paragraphe A.3 - par associer à chaque paire $\{a,b\}$ d'attributs l'indice

$$Q(a,b) = [s-E(S)]/\sqrt{var(S)} = \sqrt{n}\, r(a,b) \ , \tag{8}$$

où $s = \text{card}\,[E(a) \cap E(b)]$, $S = \text{card}(X \cap Y)$. On établit ainsi la table des indices "localement" réduits (ou normalisés) :

$$\{Q(a,b)/\{a,b\} \in P_2(A)\}, \tag{9}$$

où $P_2(A)$ désigne l'ensemble des paires (ou parties à deux éléments) de A.

On procède ensuite à une réduction "globale" des similarités rapportant chaque $Q(a,b)$ à la distribution observée (9). Plus précisément, on substitue à la table (9), celle :

$$\{Q_s(a,b)/\{a,b\} \in P_2(A)\} \ , \tag{10}$$

où

$$Q_s(a,b) = [Q(a,b) - moy_e(Q)]/\sqrt{var_e(Q)} \ , \tag{11}$$

où $moy_e(Q)$ et $var_e(Q)$ sont respectivement la moyenne et la variance de la distribution empirique (9).

En associant à A - selon un modèle aléatoire qui "respecte" la cardinalité de chaque $E(a)$, $a \in A$ - un ensemble A^* d'attributs aléatoires indépendants, nous démontrons que la suite des $m(m-1)/2$ v.a. $\{Q_s(a^*,b^*)/\{a^*,b^*\} \in P_2(A^*)\}$ est une loi multinormale.

L'indice de la "vraisemblance du lien" entre deux attributs a et b parmi A, est alors défini par

$$P(a,b) = Pr\{Q_s(a^*,b^*) < Q_s(a,b)\} = \Phi[Q_s(a,b)] \ , \tag{12}$$

où Φ est la fonction de répartition de la normale $N(0,1)$.

Pour cet indice et de façon intuitive, le degré d'association entre les deux attributs a et b est mesuré par le complément à l'unité du degré d'invraisemblance de la grandeur de s dans le cadre du modèle aléatoire de l'hypothèse d'absence de liaison $A \to A^*$.

Ainsi, à la table (10) des indices, on substitue

$$\{P(a,b)/\{a,b\} \in P_2(A)\} \ . \tag{13}$$

Si C et D sont deux classes (i.e. deux parties disjointes de A), l'indice d'association entre C et D repose sur la distribution - dans l'hypothèse d'absence de liaison - de

$$\max\{P(c^*,d^*)/(c^*,d^*) \in C^* \times D^*\} = p(C^*,D^*) \ . \tag{14}$$

Cet indice s'exprime par

$$P(C,D) = Pr\{p(C^*,D^*)<p(C,D)\}$$

$$= [p(C,D)]^{card(C)\times card(D)} \ , \tag{15}$$

où $p(C,D) = \max\{P(c,d)/c,d) \in C \times D\}$.

Pour la formation ascendante hiérarchique de l'arbre de classifications sur A, il s'agit à chaque pas de réunir la paire de classes (resp. les paires de classes en cas d'ex aequo) pour laquelle (resp. pour lesquelles) l'indice (15) est maximal.

A la différence de la situation précédente, il s'agit d'un critère directement local (cf. H1).

Comme seul l'ordre des valeurs de (15) importe, nous considérons -

pour des raisons de précision calcul - la fonction croissante [-Log(-Log)], de sorte que l'indice effectivement calculé est

$$S(C,D) = -Log[-LogP(C,D)] = -Log[card(C)] - Log[card(D)] - Log[-Log(p(C,D))] . \quad (16)$$

Dans ces conditions, la formule de réactualisation devient :

$$S(B,C\cup D) = -Log[card(C) + card(D)] + \max\{S(B,C) + Log[card(C)], S(B,D) + Log[card(D)]\}. \quad (17)$$

Dans le cas où plusieurs paires de classes réalisent "en même temps" la valeur maximale de (16), on procède comme dans le cas H1 précédent.

H3. Contrainte de contiguïté ou de discontiguïté

Dans le cas de la classification d'un ensemble d'unités géographiques, l'administrateur impose aux classes formées d'être connexes (i.e. d'un seul tenant). Dans le cas de la classification d'un ensemble d'attributs descriptifs, on ne veut pas admettre l'association de deux classes d'attributs, s'il n'existe pas deux attributs - appartenant respectivement aux deux classes - dont l'indice brut (cf. § A3) est supérieur à un seuil donné.

Dans ces conditions, on associe à chaque élément de l'ensemble à classifier, sa table de contiguïté (resp. discontiguïté) dans le premier (resp. second) cas.

Dans la table des indices de proximité entre classes, on procède à une recherche monotone croissante (pour la valeur de l'indice) des paires de classes agrégeables et on retient la paire (resp. les paires) de

classes qui réalise (resp. réalisent) la valeur maximale de l'indice et qui satisfont la contrainte de contiguïté dans le premier cas et celle de discontiguïté dans le second cas.

H4. Noeuds et niveaux significatifs d'un arbre de classification

H4.1. Niveaux significatifs

Trois critères globaux de jugement de la partition produite à un niveau donné de l'arbre des classifications peuvent être considérés et mis en oeuvre.

Le plus classique est celui de l'inertie expliquée qui sert d'ailleurs directement dans la formation de l'arbre H1; si $(P_0, P_1, \ldots, P_\ell, \ldots, P_m)$ est la suite des partitions produites aux différents niveaux de l'arbre $0, 1, \ldots, \ell, \ldots, m$, on associera à la partition P_ℓ, indiquée par $\{I_h^\ell / 1 \leq h \leq k_\ell\}$, la valeur du critère

$$\sum_{1 \leq h \leq k_\ell} \eta_h \, ||g_h - g||^2 \, , \tag{18}$$

avec des notations que l'on comprendra en se reportant au paragraphe H1 ci-dessus.

Plusieurs remarques importantes s'imposent :

- ce critère ne peut être utilisé que dans le cas où l'ensemble à classifier peut être représenté de façon naturelle par un nuage de points dans un espace euclidien;
- il y a de bons critères pour l'émergence des classes et il y a de bons critères pour l'évaluation des partitions obtenues et ce ne sont pas nécessairement les mêmes. D'ailleurs, dans le cas H1 ci-dessus, ce critère ne fait que décroître sur la suite des niveaux de l'arbre des classifications qu'il a servi à bâtir. En fait, il importe que le critère d'évaluation soit relativement indépendant de la méthode de formation de l'arbre et ait un caractère très général, ce qui n'exclut pas - lorsque cela est possible - d'utiliser (18) dans le cas de H2.

Un critère très général est celui basé sur la "préordonnance". Si nous désignons pa K l'ensemble à classifier qui peut correspondre, soit à l'ensemble O des objets, soit à l'ensemble V des variables, la préordonnance sur K est un préordre total sur l'ensemble $L = P_2(K)$ des paires d'éléments de K. Pour ce préordre que nous supposons ici - pour simplifier - un ordre total et strict, le rang d'une paire est une fonction croissante de la ressemblance entre ses composantes, mesurée par l'indice Q de proximité choisi :

$$(\forall\ (p,q) \in L \times L)\ ,\ p < q \iff Q(p) < Q(q)\ . \qquad (19)$$

Nous représentons dans L×L la préordonnance $\omega(K)$ par son graphe :

$$gr(\omega) = \{(p,q)/(p,q) \in L\times L\ ,\ p < q \text{ et non } q < p \text{ pour } \omega\} \qquad (20)$$

Une même partition π qui ensuite jouera le rôle de P_ℓ $(0 \leq \ell \leq m)$, sera représentée dans L×L par le "rectangle" $R(\pi) \times S(\pi)$ où $R(\pi)$ (resp. $S(\pi)$) est l'ensemble des paires réunies (resp. séparées) par la partition π. $R(\pi) < S(\pi)$ pour l'ordre quotient.

L'indice brut entre la préordonnance $\omega(K)$ et la partition π est alors

$$s(\omega,\pi) = \text{card}[gr(\omega) \cap (S(\pi) \times R(\pi))]\ . \qquad (20')$$

Nous opérons une normalisation de cet indice en associant à la partition π, une partition aléatoire π^* dans l'ensemble - muni d'une probabilité uniformément répartie - $P(n;t)$ de toutes les partitions de même type cardinal que π (i.e. dont la suite des cardinaux des classes est la même que celle de π).

La forme la plus simple de l'indice normalisé qui prend le nom de "statistique globale" est la suivante :

$$[s(\omega,\pi)-(r(\pi)s(\pi)/2]/\sqrt{r(\pi)s(\pi)[r(\pi)+s(\pi)+1]/12}, \qquad (21)$$

où $r(\pi)$ (resp. $s(\pi)$) est le cardinal de $R(\pi)$ (resp. $S(\pi)$).

On peut se rendre compte que l'indice (20) peut en fait se mettre sous la forme de la somme des rangs (calculés conformément à $\omega(K)$) des paires réunies par la partition π.

On introduit dans ces conditions un autre critère d'adéquation - normalisé de la même façon - et basé sur la somme des similarités des paires réunies par la partition. Une forme simplifiée de ce critère est la suivante :

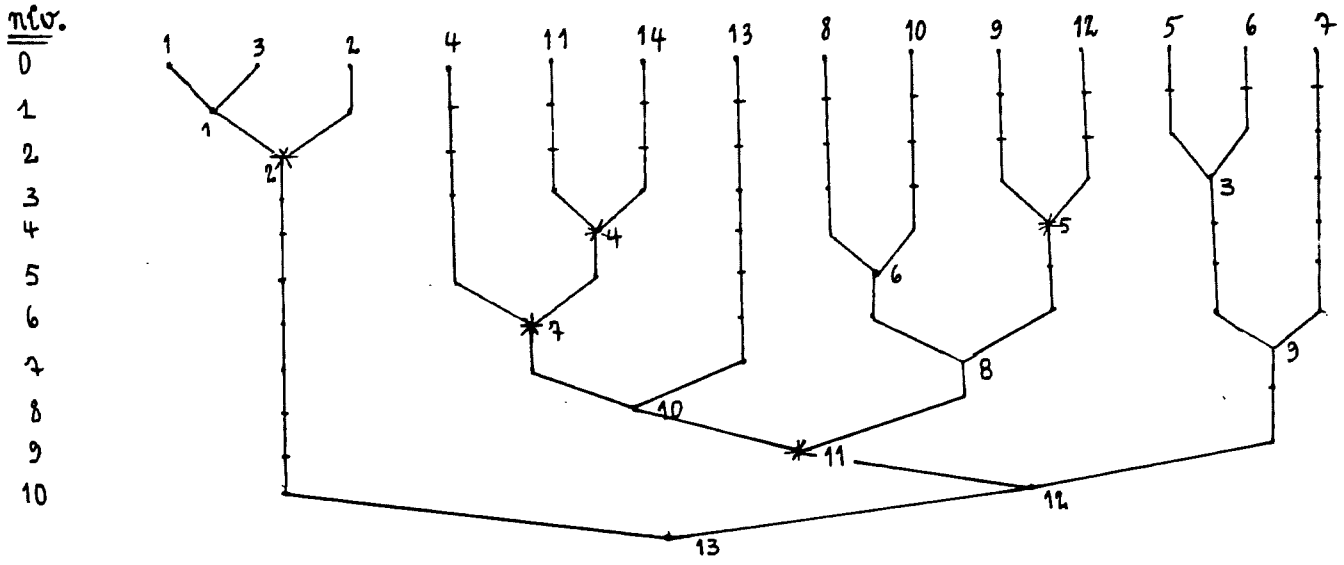
$$\frac{1}{\sqrt{r(\pi)s(\pi)/[r(\pi)+s(\pi)-1]}}\ \Sigma\{\varepsilon(p)Q_s(p)/p \in L\}\ , \qquad (22)$$

où Q_s est un indice globalement réduit (cf. formule (11) du paragraphe H2) et où $\{\varepsilon(p)/p \in L\}$ est la fonction indicatrice de $R(\pi)$.

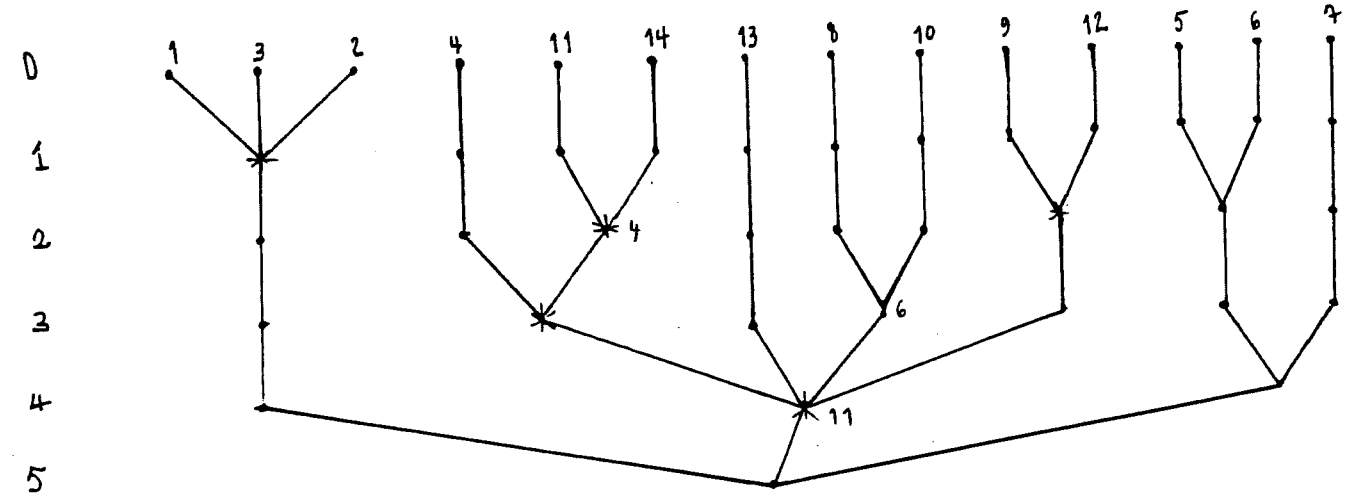
La suite des valeurs d'un critère global d'adéquation tel que (21), (22) ou (18) permet de reconnaître quels sont les principaux états d'équilibre dans la synthèse automatique, fournie niveau après niveau dans l'arbre détaillé des classifications emboîtées.

H4.2. Noeuds significatifs

Si Σ est un critère global d'adéquation ((21),(22) ou (18), nous avons ci-dessus associé la suite des valeurs observées $\{\Sigma_\ell = \Sigma(P_\ell)/0 \leq \ell \leq m\}$ de Σ sur la suite des partitions d'un arbre hiérarchique de classification. En attachant maintenant à chaque niveau ℓ, le taux d'accroissement $\theta_\ell = [\Sigma_\ell - \Sigma_{(\ell-1)}],1 \leq \ell \leq m$, on peut déterminer les noeuds "significatifs" d'un arbre de classification; un tel noeud apparaît



Arbre détaillé ; les nœuds significatifs ont été marqués par une étoile *



Arbre condensé aux niveaux où apparaît un nœud « significatif ».

(Les nœuds formés entre deux nœuds « significatifs » sont tassés vers le niveau du dernier nœud « significatif ».)

correspondre à un maximum local de la distribution observée de θ le long de la suite des niveaux de l'arbre. L'expérience montre en effet que la valeur de θ augmente lorsqu'une classe en cours de formation se confirme et que cette valeur tombe devant l'arrêt de l'enrichissement d'une classe ayant quelque consistance, au profit de la création de l'embryon d'une autre classe. Ces maxima locaux correspondent ainsi à des niveaux d'achèvement de classes. Nous proposons dans ces conditions une condensation de l'arbre aux niveaux où apparaît un noeud "significatif".

L'examen conjoint des deux distributions observées $\{\Sigma_\ell / 1 \leq \ell \leq (m-1)\}$ et $\{\theta_\ell / 1 \leq \ell \leq (m-1)\}$ permet une interprétation dynamique de l'arbre des classifications.

H5. Degré de neutralité des éléments

A chaque élément y de l'ensemble K à classifier, nous associons un indicateur de variance résultant de la comparaison de y avec la suite des éléments de $(K-\{y\})$.

Si on travaille avec des distances, cet indicateur peut être ($k = \text{card}(K)$) :

$$\frac{1}{(k-1)} \Sigma\{\delta^2(x,y)/x \in K-\{y\}\} \tag{23}$$

ou encore

$$\frac{1}{(\mu-\mu_y)} \Sigma\{\mu_x \, \delta^2(x,y)/x \in K-\{y\}\} \tag{23'}$$

avec $\delta(x,y) = [d(x,y)-\bar{d}(y)]$, où, respectivement dans le premier et second cas :

$$\bar{d}(y) = \frac{1}{(k-1)} \Sigma \{d(x,y)/x \in K-\{y\}\} \quad , \tag{24}$$

$$= \frac{1}{(\mu-\mu_y)} \Sigma \{\mu_x \, d(x,y)/x \in K-\{y\}\} \, . \tag{24'}$$

Si on travaille avec des indices d'association de la forme (10) ci-dessus, cet indicateur se met sous la forme :

$$\frac{1}{(k-1)} \Sigma \{\chi_s^2(x,y)/x \in K-\{y\}\} \quad , \tag{25}$$

avec $\chi_s(x,y) = [Q_s(x,y) - \bar{Q}_s(y)]$, où

$$\bar{Q}_s(y) = \frac{1}{(k-1)} \Sigma \{Q_s(x,y)/x \in K-\{y\}\} \quad . \tag{26}$$

Le rangement des éléments d'une même classe conformément à la suite des valeurs décroissantes de ce type d'indice de dispersion, permet de reconnaître les éléments moteurs ou d'entraînement de la classe (pour lesquels l'indice est le plus élevé) et ceux les plus neutres (pour lesquels l'indice est le plus bas).

P1. "k-means ou algorithme de "reallocation-recentrage" autour des centres mobiles de gravité (PGIO)

Le contexte est le même que celui défini dans le cas H1 ci-dessus : même représentation et même critère.

L'entier k étant fixé, le but de cet algorithme est la construction d'une partition de E en k classes au plus, optimisant au mieux un critère de cohésion des classes formées. Le nombre de classes obtenu résulte en fait de l'aléa du déroulement de l'algorithme, il reste dans la pratique très près de k.

Plus précisément, il s'agit - pour $h \leq k$ - de minimiser :

$$\sum_{1\leq j\leq h} \sum_{x\in E_j} \mu_x \, ||x-g_j||^2 \quad , \tag{1}$$

où $P = \{E_j/1\leq j\leq h\}$ est la partition en h classes et où $\{g_j/1\leq j\leq h\}$ est le système des centres de gravité des classes E_j (g_j : centre de gravité de $E_j/1\leq j\leq h$).

Pour le démarrage de l'algorithme, nous supposons la donnée a priori d'un ensemble quelconque de k points $\{o_1^{(0)}, o_2^{(0)},\ldots,o_k^{(0)}\}$ qui définit le système initial de "centres d'attraction". Le pas de l'algorithme consiste à définir un "zonage" de E en attachant chacun des sommets du nuage au centre d'attraction le plus proche. En cas d'ex aequo dans la proximité, l'usage est d'attribuer le sommet au centre d'indice le plus petit. On définit ainsi une partition de E en k(1) classes $\{E_1^{(1)},E_2^{(1)},\ldots,E_{k(1)}^{(1)}\}$ où $k(1) \leq k$; en effet, certains centres ne correspondant pas à des sommets du nuage peuvent demeurer seuls.

De façon plus précise, l'hyperplan médiateur d'un même couple de sommets $(o_j^{(0)},o_h^{(0)})$ tel que $j < h$, découpe le nuage N(E) en deux morceaux :

$$E_{jh} = \{x/x \in E, d(x,o_j^{(0)}) \leq d(x,o_h^{(0)})\}$$

et

$$E_{hj} = \{x/x \in E, d(x,o_j^{(0)}) > d(x,o_h^{(0)})\} , \qquad (2)$$

de sorte que la "zone" attribuée à $o_j^{(0)}$, se trouve définie par l'intersection

$$Z(o_j^{(0)}) = \left(\bigcap_{h>j} E_{jh}\right) \cap \left(\bigcap_{\ell<j} E_{j\ell}\right) . \qquad (3)$$

$Z(o_j^{(0)})$ - s'il n'est pas vide - définit l'une des classes $E_{j'}^{(1)}, 1 \leq j' \leq k(1)$.

A chacune des classes $E_{j'}^{(1)}$, on associe son *centre de gravité* formant ainsi une nouvelle suite de centres d'attraction : $(o_1^{(1)},o_2^{(1)},\ldots,o_{k(1)}^{(1)})$. L'ensemble des centres ainsi obtenus est de cardinal inférieur ou égal à k(1); en effet, il peut se faire qu'un même centre de gravité corresponde à plus d'une seule classe.

D'où une nouvelle décomposition de E en classes dont les centres de gravité définiront les nouveaux centres d'attraction et ainsi de suite... Nous allons voir que dans le cadre défini ci-dessus, le processus - dont chaque pas diminue le critère (I) - converge nécessairement et *on espère* que l'ensemble des derniers centres viendra occuper les régions à forte "densité" du nuage N(I).

Ne serait-ce que pour le problème de la classification du nuage N(I), cet algorithme a entraîné toute une famille de méthodes basées sur le principe de "reallocation-recentrage". Ces méthodes diffèrent aux deux niveaux suivants :

- choix ou détermination du système initial de centres d'attraction ou noyaux,
- affectation des éléments aux classes en cours de formation.

Ces variantes peuvent correspondre à des aspects algorithmiques ou statistiques (définition de critères). Signalons par exemple la procédure de Mc Queen (1967) qui a introduit la dénomination des "k-means" : après chaque affectation d'un sommet à la classe - en cours de formation - dont le centre de gravité est le plus proche, on substitue au précédent centre d'attraction de la classe d'accueil, le centre de gravité de la classe enrichie d'un élément.

P2. "Nuées dynamiques" (PGIO)

Sous l'appellation "Algorithme des nuées dynamiques (A.N.D.)", E. Diday (1972) a introduit en France ce type d'algorithme en lui apportant une certaine formalisation et en cherchant à généraliser au maximum sa portée. La généralisation porte d'abord sur la notion de représentation d'une classe qui ne se fait plus nécessairement au moyen d'un centre de gravité, mais à partir de la définition d'un "noyau" de la classe; lequel peut par exemple - dans la situation décrite ci-dessus - correspondre à un

sous-ensemble de faible cardinal de la classe. Ce dernier définissant un "squelette" de la classe, "épouserait mieux sa forme" que ne le ferait un centre de gravité. Le deuxième aspect de l'extension de l'algorithme porte sur la nature de l'espace de représentation des données et d'ailleurs, la notion de "noyau" dépend de cet espace et du problème de reconnaissance posé. Enfin, un problème important de métriques et critères se pose, pour alimenter de façon cohérente ces développements de l'algorithme.

Pour simplifier la description de l'algorithme nous allons considérer la représentation d'une même classe par un de ses points (version d'ailleurs souvent utilisée dans l'A.N.D.).

On introduit deux fonctions qui correspondent aux deux articulations fondamentales de l'algorithme. La première que nous notons $\pi^{(*)}$, pour rappeler que son objet est de former une partition, est de "reallocation" de l'ensemble des sommets du nuage $N(E)$ autour des centres d'attraction formant le système des noyaux. Dans notre cadre, L_k désignant l'ensemble des parties de E, de cardinal inférieur ou égal à k et P_k l'ensemble des partitions de E en au plus k classes, π est une application de L_k dans P_k dont le sens a été précisé ci-dessus (cf. (3)).

La deuxième fonction que nous notons $\nu^{(*)}$, pour rappeler que son objet est de former un système de noyaux, est de "recentrage" des classes. Il s'agit d'une application de P_k dans L_k, associant à chaque classe celui de ses points par rapport auquel est minimal le moment d'inertie de la classe.

Si (L,P) est un élément de $L_k \times P_k$, on désigne par $w(L,P)$ la valeur du critère (1) pour (L,P); plus précisément,

$$w(L,P) = \sum_{1 \le j \le h} \min\{M_2(E_j, o_j)/o_j \in L\} , \qquad (4)$$

(*) Ces fonctions sont notées - par l'auteur de la méthode - f pour π et g pour ν.

où nous avons noté $P = \{E_j / 1 \leq j \leq h\}$ $(h \leq k)$ et où

$$M_2(E_j, o_j) = \sum_{x \in E_j} \mu_x \, ||x - o_j||^2 \qquad (5)$$

est le moment d'inertie de E_j par rapport à o_j, élément de E_j.

Considérons la partition $\pi(L)$ que nous pouvons noter $\pi(L) = \{G(o)/o \in L\}$ où $G(o)$ est la classe attribuée au centre d'attraction o. On a

$$w[L, \pi(L)] \leq w(L,P) \; . \qquad (6)$$

D'autre part, de par la propriété caractéristique d'un centre d'attraction, on a de façon immédiate

$$w[\nu(P), P] \leq w(L,P) \; . \qquad (7)$$

Ainsi, chacune des deux articulations d'un même pas de l'algorithme - reallocation définie par π et recentrage définie par ν - fait diminuer la quantité critère qui est l'inertie perdue par la classification.

Bien qu'il s'agit de déterminer une "bonne" partition de E, minimisant au mieux l'inertie perdue, il faut regarder l'algorithme de "reallocation-recentrage" comme procédant à la recherche d'un "bon" système de noyaux - centres d'attraction. $L^{(0)}$ désignant le système initial de noyaux, lequel pouvant être choisi de façon quelconque, on a

$$L^{(1)} = \nu.\pi[L^{(0)}], \; L^{(2)} = \nu.\pi[L^{(1)}], \; \ldots, \; L^{(r)} = \nu.\pi[L^{(r-1)}], \; \ldots \qquad (8)$$

Chaque étape faisant diminuer la quantité critère w qui reste positive, le processus converge nécessairement et s'arrête après le t-ème pas si

$$\nu.\pi[L^{(t)}] = L^{(t)} \quad : \tag{9}$$

système de noyaux qu'il n'est plus possible d'améliorer, point fixe de l'application $\nu.\pi$. Le couple $[L^{(t)},\pi(L^{(t)})]$ est le résultat de l'algorithme et on peut écrire

$$L^{(t)} = (\nu.\pi)^t(L^{(0)}) \quad , \tag{10}$$

t étant défini comme le plus petit entier pour lequel

$$(\nu.\pi)^{t+1}(L^{(0)}) = (\nu.\pi)^t(L^{(0)}) \quad . \tag{11}$$

Dans la pratique et en tenant compte de la précision calcul de l'ordinateur, on arrête le processus dès que la différence entre le premier et le second membre de (11) devient inférieure à un certain seuil.

Il est d'usage de procéder à un certain nombre de répétitions (de l'ordre de 5) de l'algorithme à partir de différents systèmes initiaux de noyaux. Si r est le nombre de répétitions, on obtient de la sorte une suite $(P_1,P_2,\ldots,P_r)$ de r partitions de E.

Les classes du croisement de ces r partitions définissent ce que Diday appelle les "formes fortes".

P3. "Pôles d'attraction" (PGIO)

Comme c'est le cas pour les méthodes précédentes, nous présentons une seule variante de cette dernière. Comme pour la définition des éléments plus ou moins "neutres" (cf. H5), l'idée de cette méthode résulte du souci d'exploiter les n (= card(E)) distributions des distances ou proximités de chacun des objets aux $(n-1)$ autres. Dans la variante qui nous concerne ici, nous travaillerons avec les distances et il

s'agit donc de mettre à profit la famille de distributions

$$\{\mathcal{D}(x)/x \in E\} \tag{12}$$

où

$$\mathcal{D}(x) = \{d(x,y)/y \in E-\{x\}\} , \tag{13}$$

où $d(x,y)$ est la distance définie sur l'ensemble E des objets.

La méthode opère de façon divisive, par segmentation autour de "pôles" d'attraction de classes. Ces pôles sont déterminés par une analyse mettant en oeuvre les distances mutuelles entre pôles et les moments absolus d'ordre 2 des distributions $\mathcal{D}(x)$ où x appartient à l'ensemble des pôles.

Plus précisément, on commence par déterminer un ensemble $P_2 = \{p_1,p_2\}$ formé de deux pôles, où p_1 réalise

$$\max\{M_2(x)/x \in E\} , \tag{14}$$

où $M_2(x)$ est le moment absolu d'ordre 2 de $\mathcal{D}(x)$:

$$M_2(x) = \frac{1}{(n-1)} \Sigma \{d^2(y,x)/y \in E-\{x\}\} .$$

p_2 est choisi - dans $E-\{p_1\}$ - de façon à réaliser un compromis entre les deux exigences suivantes :

(a) $M_2(p_2)$ le plus grand possible,

(b) $d(p_1,p_2)$ le plus grand possible.

Nous considérons le critère suivant :

$$\max_{y \in E-\{p_1\}} \{M_2(y)d^2(y,p_1)\} . \tag{15}$$

Plus généralement si P^* est l'ensemble des pôles déjà extraits, on peut considérer pour la détermination du nouveau pôle, une règle de type max min; en clair, le nouveau pôle réalise

$$\max_{y \in (E-P^*)} \{\min_{p \in P^*} M_2(y) d^2(y,p)\} . \qquad (16)$$

La première partition formée est en deux classes, résultant de l'affectation de la suite - comme elle se présente - des éléments de E, à l'un ou à l'autre des deux premiers pôles p_1 et p_2.

A chaque création d'un nouveau pôle, une nouvelle partition est formée en reaffectant la suite des éléments de E. Si à une étape donnée de l'algorithme, P est l'ensemble des pôles et si $Cl^*(q)$ désigne la classe déjà formée autour du pôle q appartenant à P, on affectera l'objet x de $(E-C)$ - où $C = \cup\{Cl^*(q)/q \in P\}$ - au pôle p de P, si le couple (x,p) réalise

$$\min\{\frac{1}{card(Cl^*(q))} \sum_{x \in Cl^*(q)} d^2(x,y)/(y,q) \in (E-C) \times P\} ; \qquad (17)$$

en d'autres termes, l'objet x est affecté à la classe $Cl^*(p)$ si le moment d'inertie de $Cl^*(p)$ par rapport à x est le plus petit.

On détermine ainsi une suite de partitions : la première en deux classes autour des deux premiers pôles p_1 et p_2, la deuxième en trois classes autour des trois premiers pôles p_1, p_2 et p_3,..., la $(k-1)$-ème en k classes autour des k premièrs pôles $p_1, p_2, \ldots, p_{(k-1)}$ et p_k.

Il est important de remarquer que cette suite "descendante" de partitions n'est pas hiérarchique (i.e. ordonnée par finesse croissante).

A chacune des partitions de la suite définie par l'algorithme, on affecte l'un ou l'autre des critères (18) ou (22) (§ H41) ce qui permet une règle d'arrêt et la reconnaissance des partitions les plus "significatives".

Dans le cas où l'ensemble des unités de données est par trop hétérogène, pour éviter l'influence des éléments aberrants pour lesquels $M_2(y)$ est "grand" parce que y se trouve "éloigné de tout", on commencera par tenir à l'écart les points y pour lesquels $[M^2(y)/M_2(y)]$ - où $M(y)$ est la moyenne des distances à y - est "trop voisin" de 1 (on a en fait $[M^2(y)/M_2(y)]$ compris entre 0 et 1).

Une fois obtenue la partition en k classes $\{Cl(1),Cl(2),\ldots,Cl(k)\}$ autour de la suite des k premiers pôles $p_1,p_2,\ldots,p_k$, on peut reformer par recentrage et reallocation une - éventuellement "meilleure" - partition en k classes. Pour cela, on associera à chacune des classes $Cl(j)$, $1 \leq j \leq k$, son représentant o_j le plus central, lequel étant l'élément de $Cl(j)$ dont la somme des carrés des distances aux autres éléments de la classe, est minimale.

C. CONCLUSION : autres méthodes

Comme nous l'avons déjà souligné ci-dessus, chacune des méthodes présentée l'a été dans un cadre très naturel, en général le plus simple. Mais, il y a pour une méthode, d'autres cadres tout aussi naturels de présentation. D'autre part, comme nous l'avons également mentionné ci-dessus, il y a des possibilités naturelles de plus ou moins grande extension d'une approche méthodologique. On peut par exemple signaler que l'algorithme des nuées dynamiques a été adapté pour le problème de la séparation d'un mélange de lois Gaussiennes multidimensionnelles de probabilité. On peut également signaler que l'algorithme de la vraisemblance du lien permet la classification de l'ensemble des objets ou individus décrits par des variables qualitatives

nominales, ordinales, ou encore beaucoup plus générales quant à la structure de ressemblance que chacune détermine sur l'ensemble des objets.

Il y a certainement dans le choix des méthodes ci-dessus présentées un minimum inéluctable de partialité. Toutefois, il suffit parfois d'ajouter peu de mots pour comprendre le principe de telle ou de telle autre méthode implémentée dans MODULAD.

Ainsi "Boules Optimisées" diffère de l'algorithme des "nuées dynamiques" (resp. "k-means") en ce que tout point affecté doit être à une distance inférieure à une distance R - fixée a priori - du centre d'attraction auquel le point est attribué.

Comme "Boules optimisées" et par rapport aux "k-means", ISODATA pose des contraintes d'un autre type :
- éclater une classe (à partir d'un seuil de dispersion maximale de chacune des classes),
- fusionner deux classes (à partir d'un seuil de distance minimale entre centres de deux classes),
- supprimer une classe (à partir d'un seuil quant au nombre minimum d'éléments par classe).

CROKI2 et CROMUL correspondent - à partir des nuées dynamiques - à la recherche aussi conjointe que se peut - d'un couple de partitions sur l'ensemble des lignes et sur l'ensemble des colonnes d'un tableau de données; CROKI2 s'adresse à un tableau de contingence et CROMUL à un tableau disjonctif complet. En fait, partant d'un couple de partitions (de l'ensemble des lignes et de l'ensemble des colonnes), l'algorithme a un caractère alternatif : optimisant au mieux la partition de l'un des côtés (du tableau des données), on passe à celle de l'autre côté pour revenir au premier, jusqu'à ce que l'inertie - conformément à une métrique adéquate - expliquée par le couple de partitions, ne puisse plus augmenter.

L'algorithme des "transferts" a été pour la première fois proposé par S. Regnier (1965) dans le cadre de sa méthode des "partitions centrales" sur l'ensemble des objets qui est muni d'une famille de partitions, définie par une suite de variables descriptives qualitatives nominales. Les développements les plus récents et les plus opérationnels de l'approche classificatoire qui en a résulté sont dûs à J.F. Marcotorchino et F. Michaud qui utilisent - à partir d'un codage sur l'ensemble des paires - la programmation linéaire en nombres entiers (1979).

L'algorithme des transferts implémenté dans MODULAD travaille dans le cas où les variables sont numériques et se propose la classification de l'ensemble des objets représentable par un nuage de points dans un espace euclidien. Partant d'une partition initiale en k classes, un même pas de l'algorithme consiste à opérer tous les transferts possibles d'un élément d'une classe dans une autre, à accompagner chacun des transferts de la variation de l'inertie expliquée et à retenir celui des transferts (d'un objet d'une classe dans une autre) qui rend maximum cette variation.

Nous n'avons pas pu dans les limites de cette présentation générale mentionner les très intéressantes idées algorithmiques de la classification hiérarchique - rapide et de faible encombrement mémoire - de "gros ensembles": "voisinages réductibles", "voisins réciproques", "classification en parallèle", "classification séquentielle" (pour le cas non hiérarchique).

Enfin, une fois dégagées les principales classifications de l'ensemble des variables (de description) et de l'ensemble des objets, il y a des méthodes qui empruntent une approche corrélative ou une approche inertielle qui conduisent à l'élaboration d'indices qui permettent :

- le croisement entre une classification sur l'ensemble des variables et une classification sur l'ensemble des objets ou individus; cette dernière pouvant être exogène et définie par rapport à un caractère extérieur,

- l'évaluation du rôle d'un individu ou d'une classe d'individus dans la formation d'une classe de variables,
- l'"explication" d'une classe d'individus par une variable ou classes de variables,

et ce, pour différentes structures du tableau de données.

A travers INTERP de MODULAD certains indices relatifs à la troisième rubrique existent. Ils concernent le cas où les variables sont numériques et sont de nature inertielle.

BIBLIOGRAPHIE TRES SUCCINTE

[1] ANDERBERG M.R. (1973), "Cluster analysis for applications", Academic Press, New York.

[2] DIDAY E. et coll. (1980), "Optimisation en classification automatique", IRIA.

[3] JAMBU M. & LEBEAUX N.O. (1983), "Cluster analysis and data analysis", North Holland.

[4] LERMAN I.C. (1981), "Classification et analyse ordinale des données", Dunod, Paris.

[5] MARCOTORCHINO J.F. & MICHAUD P. (1979), "Optimisation en analyse ordinale des données", Masson, Paris.