JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

I. C. LERMAN

La classification : concepts et caractéristiques d'une méthodologie d'analyse des données

Journal de la société statistique de Paris, tome 122, n° 2 (1981), p. 70-90

<http://www.numdam.org/item?id=JSFS_1981__122_2_70_0>

LA CLASSIFICATION : CONCEPTS ET CARACTÉRISTIQUES D'UNE MÉTHODOLOGIE D'ANALYSE DES DONNÉES

(Communication faite devant les Sociétés de statistique de Paris et de France le 13 novembre 1980)

I.C. LERMAN

professeur de statistique, laboratoire de statistique
Université de Rennes I, I.R.I.S.A.

Partant des premières approches dues aux naturalistes et de leurs prolongements actuels, nous insistons sur la manière dont une méthode d'analyse des données (quelle qu'elle soit d'ailleurs) se trouve déterminée par la structure mathématique des données dans le cadre desquelles elle a été conçue.

A partir de là nous développons notre propre contribution dans le domaine, laquelle repose sur quelques idées fondamentales qui ne permettent que le minimum d'arbitraire dans la construction de nos méthodes qui ont prouvé leur fécondité dans le cadre de nombreuses et riches applications, empruntées aux diverses disciplines des sciences de l'observation.

Après un aperçu conséquent de cas réels, l'article se termine par une bibliographie importante groupant 66 références.

Starting from the first naturalist approaches and their current continuations, we point out how a method of data analysis (no matter which one it is) is in fact determined by the mathematical structure of the data in whose frame it has been conceived.

From there on, we develop our own contribution to the field, which rests on a few fundamental notions which allow only a minimum arbitrariness in the construction of our methods which have proved fruitful through numerous and rich applications borrowed to the various branches of observation sciences.

PRÉAMBULE

Cet article est le reflet direct de l'exposé de la conférence débat du 13 novembre 1980 de la Société de Statistique de Paris et de France. Suite à une demande que nous a faite le regretté R. Gibrat, il s'agit pour nous de présenter nos méthodes de classification et d'analyse ordinale

Journal de la Société de statistique de Paris, tome 122, n° 2, 1981.

des données en les situant dans une dimension historique et par rapport aux autres méthodes d'analyse des données.

Par conséquent, nous allons tenter dans le cadre de cet article :

a) de préciser les bases intuitives sur lesquelles s'est bâtie la démarche classificatoire dans l'analyse des données,

b) de fournir les principales idées de nos méthodes de classification et d'analyse ordinale en insistant plus particulièrement sur celle, très générale, de classification hiérarchique,

c) de donner un aperçu sur les nombreux et très riches exemples que nous avons eu à traiter.

I — Les premières approches et leurs prolongements

Notre premier point d'ancrage peut être la vaste classification que l'homme a édifié du règne animal et du règne végétal (espèce, genre, tribu, famille, ...) et la notion non explicite qui en découle de classe « naturelle ». Dans ce contexte nous allons nous référer à deux ancêtres de la taxinomie.

Le premier est Adanson qui en 1757 (cf. [1]) donne l'idée sous une forme primitive d'un algorithme de classification hiérarchique utilisant de façon intuitive la notion de ressemblance entre objets : « Je me contenterai de rapprocher les objets suivant le plus grand nombre de degrés de leurs rapports et de leurs ressemblances... Les objets ainsi réunis formeront plusieurs petites familles que je réunirai encore ensemble, afin d'en faire un tout dont les parties soient unies et liées intimement ».

Dans la pratique, associant à chaque caractère descriptif la partition de l'ensemble des individus où une même classe est caractérisée par la possession d'une modalité du caractère, Adanson ne retient une séparation que si plusieurs caractères se retrouvent pour la spécifier. La technique était approximative et pleine d'arbitraires; mais on est en droit de se demander si elle ne contient pas le germe de l'idée de base de la méthode des « partitions centrales » que S. Régnier avait en 1965 découverte et développée tout à fait indépendamment. (cf. [58]). En effet, comme Adanson, S. Régnier considérait la famille de partitions définie par la suite des caractères et posait le problème de la recherche des partitions centrales au sens d'une métrique adéquate sur l'ensemble des partitions d'un ensemble fini.

Le deuxième patriarche de la classification que nous citerons est Vicq d'Azyr qui en 1792 (cf. [63]) constate que la définition d'une classe par une propriété caractéristique (classe monothétique), est trop rigide pour les sciences de la nature et de l'homme. Il faut attendre Beckner (1959) (cf. [3]) pour une formulation intuitive (qui laisse donc encore une part d'arbitraire), d'une classe « naturelle ». Une telle classe se référerait à un ensemble fini d'attributs tel que chaque élément de la classe possède une proportion importante, mais non spécifiée, de ces attributs et réciproquement chaque attribut est possédé par une proportion importante des éléments de la classe. On appelle une telle classe polythétique (par opposition à monothétique).

Les récents travaux de deux chercheurs de l'Université de Lille B. Lefebvre et J. Losfeld (1979) (cf. [27]) correspondent justement à une formalisation plus précise et constructive de la notion de classe polythétique.

Comme d'ailleurs on se rend compte dans ces derniers travaux, il n'est pas possible de définir dans l'absolu une classe « naturelle ». On peut le faire relativement à un niveau de synthèse donné en admettant qu'il existe implicitement différents niveaux significatifs de synthèse dans l'organisation en classes et sous-classes de l'ensemble des lignes (resp. colonnes) du tableau

d'incidence des données formé de zéros et de uns, qu'on dit encore de présence-absence où un 1 (resp. 0) indique la présence (resp. absence) d'un attribut chez un individu.

Nos « nœuds significatifs » de l'arbre des classifications (cf. ci-après) fournissent précisément ces différents niveaux d'achèvement des classes.

La portée opérationnelle de la notion de classe polythétique reste limitée; en effet, elle ne concerne que l'organisation de l'ensemble des lignes (resp. colonnes) d'un tableau d'incidence. On se demande ce que devient cette notion dans le cas où les variables de description sont d'une autre nature que des traits, dans le cas par exemple où les variables sont numériques ou encore dans le cas plus difficile où les variables sont qualitatives ordinales.

D'autre part, la notion de classe « naturelle » de variables est aussi sinon, comme nous le montrons dans nos méthodes, peut être plus importante que celle d'individus et on ne voit pas comment préciser et étendre la définition de Beckner à de telles situations jusqu'à lui donner valeur opérationnelle.

Au lieu de cela il sera beaucoup plus fructueux de reprendre la démarche constructive d'Adanson. Nous allons le faire en nous référant dans un premier temps à deux structures du tableau des données : le tableau d'incidence où les variables sont des attributs et celui de mesures numériques où les variables sont à valeurs dans $\mathbb{R}$.

II — Aspects de la classification prenant comme argument premier le tableau d'incidence

Relativement à un tableau d'incidence, c'est pour des raisons historiques évidentes que c'est le problème de la classification de l'ensemble E des individus qui a été considéré comme majeur. Par rapport à ce problème, une multitude de techniques a été proposée par l'école anglo-saxonne dans les années 60. Les représentants les plus notoires de cette école sont P.H.A. Sneath et R. Sokal (cf. [60] et [59]).

Le point de départ de chacune des méthodes proposées était le choix d'un indice de similarité, fonction numérique sur $E \times E$, associant à chaque couple d'individus (x, y), un nombre $S(x, y)$ sensé mesurer la ressemblance entre les deux individus et établi à partir des vecteurs logiques formés de zéros et de uns de description. Nous n'insisterons pas sur l'expression axiomatique à laquelle nous avons pu aboutir pour la définition de cet indice (cf. [31]). Bien qu'elle permette de mener à bien la recherche que nous allons bientôt mentionner, cette définition ne résout rien quant à l'arbitraire du choix, d'un indice de similarité, dans la situation générale.

Or un des moteurs principaux de notre recherche est l'élimination de l'arbitraire de choix. Dans cette situation, compte tenu de la très grande incertitude où on se trouve, on peut songer à retenir comme information quant aux ressemblances entre éléments une structure plus générale qu'une pondération numérique sur $E \times E$ obtenue au moyen d'un indice particulier. S'inspirant des travaux de Shepard (1962) (cf. [61]), Benzecri et de la Véga (1965) retenaient comme donnée de base l'ordre total sur l'ensemble $F = P_2(E)$ des paires, induit par un indice de similarité, où par exemple le rang d'une paire est une fonction croissante de la dissemblance entre ses composantes. Cet ordre a été appelé « ordonnance ». Conformément à cette appellation, nous avons préféré distinguer ce cas de celui, plus général, où la donnée est une *préordonnance :* préordre total sur F. Cette distinction est d'autant plus judicieuse que nous montrons aisément que la donnée d'une chaîne totalement ordonnée par finesse de partitions est équivalente à celle d'une préordonnance satisfaisant des conditions particulières et dite *ultramétrique* (cf. [31]).

Bien qu'associée au choix d'une similarité, on peut espérer qu'une telle donnée soit plus stable. Nous entreprîmes donc d'étudier l'influence du choix de la similarité sur la préordonnance

associée (1966) (cf. [31]). Nous montrions que si le nombre d'attributs possédés par un même objet de E, est invariable, la préordonnance associée est la même quel que soit l'indice choisi; plus généralement, si ce nombre d'attributs a une variance faible, la préordonnance associée, lorsqu'on remplace un indice par un autre, varie peu au sens d'une métrique naturelle sur l'ensemble des préordres totaux sur F. Ces résultats sont importants dans la pratique pour la classification d'un échantillon décrit au moyen d'un ensemble A d'attributs établi à partir d'un questionnaire ou d'un code descriptif pour lesquels le nombre d'attributs possédés par un même objet est, sinon invariable, du moins de faible variance.

Cependant (et cela a conditionné de façon définitive notre recherche) dans la première expérience réelle que nous avions eu à traiter (recherche des modèles des personnages-enfants à travers la littérature enfantine (cf. [12], [35], [53])), il s'avérait plus pertinent de découvrir ces modèles par une organisation en classes et sous-classes de proximité, directement de l'ensemble des attributs de description.

Il est clair que les résultats précédents peuvent également être énoncés relativement au choix d'un indice de proximité sur l'ensemble A des attributs descriptifs. Mais dans cette situation duale de la précédente, la variance de la variable card (E_a) (i.e. nombre d'individus possédant un même attribut a) est le plus souvent loin d'être négligeable. Or dans ce cas, la préordonnance associée n'est plus stable par rapport au choix de l'indice et le problème de la définition d'un indice restait posé. Cette recherche a constitué une des motivations principales de celle qui a conduit à la forme actuelle de notre méthode de classification hiérarchique basée sur la vraisemblance des liens.

III — Aspects de la classification prenant comme argument premier le tableau de mesures numériques

Commençons par souligner ici qu'une méthodologie d'analyse des données se trouve très fortement déterminée par la structure de référence du tableau des données. A ce sujet, nous ne pensons pas qu'il soit exagéré d'affirmer que l'analyse des correspondances n'aurait pas été si le Professeur J.P. Benzecri ne s'était trouvé confronté, pour l'analyse de données linguistiques, à des tables de contingence comme support de l'information.

Considérons à présent le cas d'un tableau de mesures $E \times V$ où $V = \{ v^j / 1 \leqslant j \leqslant p \}$ désigne l'ensemble des variables numériques de description. Dans ce cas également, c'est le problème de la classification de l'ensemble E des individus ou objets qui a dominé la préoccupation et le langage des taxinomistes. La représentation géométrique d'un tel tableau, donnée dans le cadre de l'analyse en composantes principales, n'est pas étrangère à cette appréhension. Tout le monde sait en effet que dans cette représentation on considère l'espace géométrique $\mathbb{R}^p$ et on associe à la variable v^j la jième forme linéaire coordonnée e^*_j, $1 \leqslant j \leqslant p$; de la sorte, l'objet x sera représenté par le point de $\mathbb{R}^p$ dont la suite des coordonnées est la suite des valeurs des différentes variables sur l'objet x.

La conception de la méthode des nuées dynamiques de E. Diday (1972) (cf. [14], [15]) se situe dans ce contexte, elle prend appui sur l'idée de la représentation d'une classe par un centre, sur la nécessité de définir un critère et surtout sur l'algorithme de réallocation. Ce qu'elle introduit de nouveau est la représentation d'une classe par un noyau dont la structure peut dépendre du problème posé et de la structure des données.

Dans cette méthode la classe s'organise autour d'un noyau qui se veut central, qui commence de façon aléatoire ou a priori et qui est réajusté au fur et à mesure de l'algorithme de réallocation.

Avec H. Leredde nous venons (1978-1979), après quelques années d'effort de mettre au point une nouvelle et très riche famille d'algorithmes de classification non-hiérarchique où une même classe est formée à partir de l'un de ses points « extrêmes » se situant dans une région de forte densité. La méthode est basée sur l'extraction de « pôles d'attraction » qui sont déterminés par une analyse simultanée de la variance et de la moyenne des proximités de chacun des éléments aux autres (cf. [29] et [39]).

Par rapport à la méthode des nuées dynamiques, si le système initial de noyaux est fourni par un système de « pôles d'attraction », cela enlève un degré notable dans l'arbitraire de choix et permet à l'algorithme de converger en une ou, tout au plus, deux étapes (résultat expérimental toujours constaté).

Toutefois, nous avons choisi de présenter surtout ici notre méthode de classification hiérarchique. Revenons par conséquent à la démarche Adansonienne dans le cas d'un tableau de données : Individus × Variables numériques. En vertu de ce que nous avons ci-dessus mentionné, c'est surtout la notion de distance qui a été mise en avant pour évaluer la ressemblance entre individus et pour comparer les proximités entre classes. Seulement, plusieurs types de distance ont pu être proposés sans qu'on puisse contrôler la nature du résultat par rapport au choix de la distance. Le plus élaboré en la matière est le critère de l'inertie ou variance expliquée (J.H. Ward (1963) cf. [66]), celui-là même qui est décomposé axialement dans la méthode de l'analyse en composantes. La procédure consiste alors à réunir à chaque pas la paire de classes telle que la variation de l'inertie expliquée (i.e. variance du nuage des centres de gravité pondérés) est minimale.

La méthode de J.P. Benzecri et M. Jambu de Classification ascendante hiérarchique (cf. [4], [19]) est exactement cette dernière mais elle concerne plutôt le tableau de contingence des cardinaux des classes du croisement des deux partitions ou ceux qu'on peut assimiler à ce dernier. Cette adaptation du critère de l'inertie expliquée passe nécessairement par la représentation euclidienne d'un tableau de contingence telle qu'elle est fournie dans l'Analyse factorielle des correspondances. Précisons que s'il s'agit de classifier l'ensemble des lignes d'un tel tableau, on considère une description des lignes à travers les colonnes au moyen de leurs profils; les colonnes joueront le rôle de variables et les lignes celui des points-individus. Au contraire, dans notre approche l'ensemble à classifier (celui des lignes ou des colonnes) est assimilé à celui des variables moyennant une représentation euclidienne duale (cf. [48]).

Relativement à un tableau de données Individus × Variables, c'est, nous venons de l'exprimer ci-dessus, l'optique de synthèse conduisant à l'organisation en classes et sous-classes de proximité de l'ensemble des individus, qui a polarisé la recherche des taxinomistes. Nous avons au contraire insisté sur le très grand intérêt d'une telle organisation, d'abord de l'ensemble des variables descriptives, surtout lorsque ce dernier ensemble est de cardinal élevé, comme c'est le cas des enquêtes par questionnaire. On obtient ainsi une décomposition du comportement de la population (étudiée à travers un échantillon) en tendances et sous tendances ou, si on veut utiliser le langage de l'Analyse factorielle, en facteurs et sous facteurs; mais où chacun apparaît comme un agrégat de variables, par rapport auxquelles on a un comportement plus ou moins semblable des sujets de l'échantillon. En Analyse factorielle par contre, on cherche une décomposition en facteurs « indépendants » où chacun se présente nécessairement sous la forme d'une combinaison linéaire des variables initiales :

$$\varphi = \Sigma\{\alpha_j v^j / 1 \leqslant j \leqslant p\}$$

Dans ces conditions, on peut saisir pourquoi il arrive qu'une opposition entre deux parties d'un plan factoriel puisse refléter une opposition entre deux nœuds d'une même classe apparue dans l'arbre des classifications. Indépendamment des nombreux exemples concrets traités, nous

avons effectué une analyse théorique systématique entre une approche factorielle et une approche classificatoire, qu'elle soit non-hiérarchique ou hiérarchique (1978) (cf. [43]).

IV — Introduction a notre méthode très générale de classification hiérarchique

1. *Les idées de base*

Résumons nous. Nous venons de faire apparaître que dans le cas où les variables sont numériques ainsi que dans le cas d'un tableau de contingence ou assimilé, on passe par une représentation géométrique et la plupart des taxinomistes mettent en avant le problème de la classification d'un nuage de points dans un espace euclidien où on utilise les concepts de distance et d'inertie ou de variance.

Or, les variables descriptives telles qu'elles se manifestent dans les sciences de l'homme et de la nature sont rarement numériques et il importe que la représentation mathématique des données soit fidèle. C'est pour cette raison que nous proposons une représentation ensembliste ou ordinale des données qui a un caractère mathématique fini.

D'autre part, avant la classification de l'ensemble des individus, nous favorisons le point de vue : classification des variables. De toute façon, nous proposons dans tous les cas de figure un couple d'arbres condensés des classifications, le premier sur l'ensemble des variables et le second sur celui des sujets. Le croisement de ces deux arbres permet de situer l'une par rapport à l'autre les deux synthèses automatiques; « expliquant de la sorte une même classe d'individus par rapport aux différentes classes de variables ou vice et versa (voir figure 2, § 4).

Quel que l'ensemble à organiser (celui des variables ou des individus), c'est une notion très générale de la corrélation entre structures statistiques de même type au sens mathématique du terme, qui est à la base de notre méthode de classification hiérarchique. Cette notion intervient à tous les niveaux de la recherche d'une structure en classes et sous-classes de proximité :

— indices de proximité entre variables (resp. entre individus),
— indice de proximité entre parties disjointes de l'ensemble à organiser,
— reconnaissance des niveaux et des nœuds « significatifs » de l'arbre des classifications,
— croisement de classifications,
— etc.

Cette corrélation sera en fait, pour les deux premières situations, un intermédiaire pour atteindre la vraisemblance du lien entre les deux structures à comparer. Nous verrons bientôt plus concrètement à quoi cela correspond.

2. *Typologie des tableaux de données*

Nous supposons que, pour un tableau de données $E \times V$, l'ensemble V des variables de description est formé de variables d'un même type (mathématique). Il en résulte que le type d'un tableau des données est celui commun des variables qui le définissent. Nous avons abouti à la typologie suivante qui n'est nullement étudiée pour elle-même, mais qui s'avère la plus cohérente conformément au développement combinatoire et statistique qui va suivre.

Nous distinguons deux principaux types d'une variable descriptive : celui, pouvant être représenté par une partie de l'ensemble E des objets ou une pondération sur E; et celui, dont la représentation est une partie de $E \times E$ ou une pondération sur $E \times E$.

Dans la première catégorie on peut classer l'attribut de description et la variable numérique. En effet, un attribut descriptif a peut être représenté par la partie $E(a)$ de E, formée des

objets qui le possèdent. D'autre part, une variable numérique peut être regardée comme définissant une pondération sur E en attachant à chaque objet x, le nombre $v(x)$, valeur de la mesure de la variable v sur l'objet x.

Dans la deuxième catégorie on peut classer toute variable qui définit une relation binaire sur E ou une relation binaire pondérée (on dit encore valuée) sur E. Les variables les plus classiques sont :

• la variable « rang » qui définit un ordre total σ sur E que nous représentons par son graphe

$$R(\sigma) = \{(x, y) \in E \times E / x < y \text{ pour } \sigma\}$$

• la variable dite « qualitative ordinale » qui définit un préordre total ω sur E que nous représentons par

$$R(\omega) = \Sigma\{E_i \times E_j / i < j\} \text{ (somme ensembliste)}$$

où E_i est la ième classe du préordre formée des objets possédant la ième modalité du caractère.

• le caractère descriptif à l'ensemble sans structure des modalités qui définit une partition π sur E que nous représentons dans l'ensemble plus réduit $F = P_2(E)$ des parties à deux éléments de E, par l'ensemble $R(\pi)$ des paires réunies par π ou bien par l'ensemble $S(\pi)$ des paires séparées par la partition :

$$R(\pi) = \Sigma\{P_2(E_i) / 1 \leqslant i \leqslant k\} \quad \text{et} \quad S(\pi) = \Sigma\{E_i \times E_j / 1 \leqslant i < j \leqslant k\}$$

Ces deux représentations sont équivalentes pour notre indice de proximité.

Entre les deux dernières structures d'une variable descriptive où d'une part l'ensemble des modalités du caractère est totalement ordonné et où d'autre part, cet ensemble n'est muni d'aucune structure, on peut imaginer différents types de variables discrètes où l'ensemble des modalités est muni d'une structure d'ordre partiel. On peut ici admettre que les différentes variables ne correspondent pas à la même structure d'ordre.

• la variable « pondération » sur $E \times E$ peut être représentée par une matrice carrée $\{\mu_{xy} / (x, y) \in E \times E\}$ où μ_{xy} est la pondération affectée au couple (x, y).

La représentation de l'ensemble des individus tient étroitement compte de celle de l'ensemble des variables descriptives. Nous n'allons pas détailler dans chacun des cas la nature de cette représentation. Signalons toutefois et à titre d'illustration que si, dans le cas où les variables sont numériques, cette représentation est celle classique au moyen d'un nuage de points dans un espace euclidien, cette représentation devient dans le cas où les variables sont qualitatives ordinales, au moyen d'un nuage de points dans le treillis distributif produit de p ordres totaux où un même sommet du treillis sera pondéré par la fréquence des sujets qu'il représente.

Arrivés à ce point, on comprendra pourquoi nous considérons un tableau de données représentant le croisement Individus × Variables comme fondamentalement dissymétrique. Formellement c'est clair; d'autre part, concrètement parlant, il y a à la limite autant de différence entre une variable et un individu qu'entre un appareil de mesure et l'objet sur lequel est effectué la mesure.

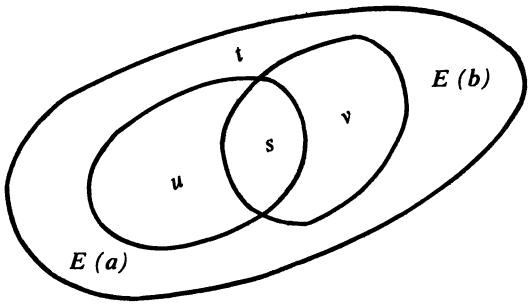
Il y a pourtant le cas très important d'un tableau de données parfaitement symétrique, il s'agit du tableau de contingence de croisement de deux variables partitions, dont l'analyse n'a d'intérêt que si le nombre de modalités par variable est « grand ». Pour ce type de tableau ainsi que pour ceux qui peuvent être assimilés à ce dernier et qui souvent résultent d'une juxtaposition de tables de contingence, il s'agira en fait d'analyser le comportement des lignes à travers les colonnes et vice et versa. Bien que les deux analyses se correspondent agréablement, pour

effectuer une même analyse *on peut* passer par une représentation euclidienne qui introduit une dissymétrie formelle entre les rôles respectivement joués par une ligne et une colonne.

Dans un article qui vient de paraître (cf. [48]) nous avons développé systématiquement notre point de vue et proposé notre solution pour le traitement de ce type de tableau de données en la comparant à d'autres plus classiques, sur la base dè différents exemples réels.

3. *Indice de proximité entre variables (resp. individus)*

Nous allons à présent introduire notre notion d'indice de proximité entre structures finies de même type et ce dans le cas le plus simple de la comparaison d'un couple $(E(a), E(b))$ de parties de E, représentant en l'occurrence un couple (a, b) d'attributs descriptifs.



Le point de départ de la construction de l'indice est $s = \text{card}\,[E(a) \cap E(b)]$ que nous appelons indice « brut » de proximité. Mais il est clair que la valeur de cet indice est un indicateur biaisé de la perception de la ressemblance; il suffit en effet que deux attributs soient fréquents (resp. rares) pour que la valeur de s soit relativement grande (resp. petite) et ce, indépendamment de la position relative de $E(a)$ et de $E(b)$. D'où l'idée d'introduire une hypothèse N d'absence de lien et de se demander : Quelle aurait été la valeur « attendue » du cardinal de l'intersection entre les deux parties si au lieu de $E(a)$ (resp. $E(b)$), on avait une partie aléatoire X (resp. Y) qui respecte d'une « certaine façon » le cardinal de $E(a)$ (resp. $E(b)$). Qui dit « valeur attendue » dit distribution de la v.a. card $(X \cap Y)$.

Les deux attributs ou les deux parties qui les représentent seront considérés d'autant plus proches que s apparaît plus invraisemblablement grand; c'est à dire, que $Pr\{\text{card}\,(X \cap Y) \geqslant s/N\}$ est plus petite, soit que $Pr\{\text{card}\,(X \cap Y) < s/N\}$ est plus grand. L'indice que nous adoptons est précisément ce dernier :

$$P(a, b) = Pr\{\text{card}\,(X \cap Y) < s/N\} \tag{1}$$

Il y a exactement trois formes N_1, N_2 et N_3 de l'h.a.l. qui peuvent s'exprimer au niveau du choix d'un élément aléatoire dans l'ensemble des parties d'un ensemble muni d'une mesure de probabilité adéquate (cf. [49]). La plus simple à exprimer est celle N_1 qui peut avoir la forme unilatérale suivante : on fixe $E(a)$ et on associe à $E(b)$ une partie aléatoire Y dans l'ensemble, muni d'une probabilité uniformément répartie, des parties de E de cardinal $n(b) = \text{card}\,(E(b))$. Soulignons que la distribution de la v.a. $S_a = \text{card}\,(E(a) \cap Y)$ est la même que celle de $S_b = \text{card}\,(X \cap E(b))$, où on fixe $E(b)$ et on associe à $E(a)$ une partie aléatoire X dans l'ensemble, muni d'une probabilité uniforme, des parties de E de même cardinal $n(a) = \text{card}\,(E(a))$.

La distribution commune est en fait hypergéométrique de moyenne $\mu = n\,(a)\,n\,(b)/n$ et de variance $\sigma_1^2 = n\,(a)\,n\,(\bar{a})\,n\,(b)\,n\,(\bar{b})/n^2\,(n-1)$.

Le calcul de $P\,(a, b)$ suppose l'approximation normale (en général très bonne) de la loi de la v.a. associée à s. Compte tenu de cette référence à une loi symétrique, on peut adopter un indice qui consiste à centrer et à réduire s par rapport à l'h.a.l. :

$$Q\,(a, b) = (s - \mu)/\sigma_1 \qquad (2)$$

et

$$P\,(a, b) = \phi\,[Q\,(a, b)]$$

où ϕ est la f.r. de la loi normale réduite.

Il s'avère que $Q\,(a, b)$ n'est autre que le coefficient d'association de K. Pearson dont le carré est le x^2 attaché au tableau de contingence de croisement des deux partitions en deux classes chacune $\{E\,(a), E\,(\bar{a})\}$ et $\{E\,(b), E\,(\bar{b})\}$.

On obtient des indices distincts quant à l'expression de la variance dans chacune des deux h.a.l. N_2 et N_3.

D'autre part, avec R. Gras, nous avons à partir d'une approche analogue défini un indice, non plus d'association, mais d'implication qui s'est avéré très fécond dans l'analyse des données didactiques où on se pose constamment la question : « Dans quelle mesure tel comportement à un stimulus donné a implique tel comportement à un autre stimulus b. » Cette recherche est développée dans [49].

Ce qui nous incite à dire que notre méthode est basée sur une notion très générale de la corrélation, c'est que, dans le cadre de l'h.a.l. N_1, l'indice centré réduit est celui d'association de K. Pearson si on a à comparer un couple d'attributs et, à un coefficient multiplicatif près, celui de corrélation de Bravais-Pearson (cf. [65]) si on a à comparer un couple de variables numériques. Dans ce dernier cas on se rend compte que l'indice brut doit nécessairement prendre la forme

$$s = \sum_{1 \leqslant i \leqslant n} u\,(i)\,v\,(i) \qquad (3)$$

où (u, v) désigne le couple de variables et où $I = \{1, 2, ..., i, ..., n\}$ indexe l'ensemble E des individus. Les deux v.a. duales se présentent alors nécessairement sous la forme :

$$S_u = \sum_{1 \leqslant i \leqslant n} u\,(i)\,v\,[\sigma\,(i)] \quad \text{et} \quad S_v = \sum_{1 \leqslant i \leqslant n} u\,[\sigma\,(i)]\,v\,(i), \qquad (4)$$

où σ est une permutation aléatoire.

Pour comparer deux variables d'un même type de la deuxième catégorie, l'approche devient techniquement plus complexe. Dans le cas discret, l'indice brut est le cardinal de l'intersection de deux sous ensemble particuliers de $E \times E$ et pour construire l'indice définitif, au lieu d'une partie libre aléatoire, on aura à considérer selon les cas, une partition aléatoire, un préordre total ou partiel aléatoire, etc.

Avec G. Lecalvé (cf. [26], [36]) nous traitons le cas de la comparaison de deux pondérations (on dit encore valuations) sur $E \times E$. On obtient également dans ce cadre des résultats tout à fait nouveaux et décisifs aussi bien du point de vue théorique que pratique.

L'approche s'adapte également à la comparaison des individus décrits par des variables d'un même type où on tient parfaitement compte de la dissymétrie de nature entre une variable et un individu. Avec J.Y. Lafaye, nous obtenons des résultats indiscutables (cf. [22]).

4. *Indice de proximité entre classes. Algorithme de la Vraisemblance du Lien*

Désignons par B l'ensemble à classifier. Il peut s'agir de l'ensemble V des variables de description ou de l'ensemble E des objets ou individus décrits, quel que soit le type algébrique

commun des variables formant V. A ce point de notre présentation on dispose de la table des proximités entre éléments de B

$$\{ P(c, d)/\{ c, d \} \in P_2(B) \} \tag{1}$$

où $P_2(B)$ est l'ensemble des paires ou parties à deux éléments de B et où $P(c, d)$ se réfère à une échelle de probabilité définie par la vraisemblance du lien, pour tout $\{ c, d \}$ de $P_2(B)$.

Soient C et D deux parties disjointes de B. Le point de départ de l'élaboration de l'indice de comparaison entre C et D est fourni, à partir de considérations topologiques, par la formule

$$p(C, D) = \max \{ P(c, d)/(c, d) \in C \times D \} \tag{2}$$

Mais si on se limitait à cette forme, en se reportant au schéma suivant où la densité des hachures reflète celle des points, la paire de classes $\{ C_1, D_1 \}$ est à fusionner avant celle $\{ C_2, D_2 \}$. En effet, la valeur du « saut minimal » pour passer de C_1 à D_1 est inférieure à celle pour passer de C_2 à D_2. Or la valeur de la plus grande proximité observée entre un élément de C_i et un élément de D_i est, compte tenu de la densité des points, sensiblement plus exceptionnelle pour $i = 2$ que pour $i = 1$. De sorte que l'indice $p(C, D)$ ci-dessus jouera le rôle de l'indice brut de proximité.

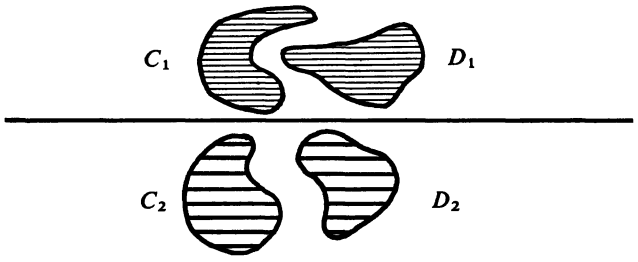


Fig. 1

L'indice final que nous retenons résulte de la distribution de la v.a. $p(C', D')$ associée à (2) où C' (resp. D') est une classe aléatoire associée à C (resp. D) dans l'h.a.l.. L'indice auquel on aboutit prend la forme très simple suivante :

$$P(C, D) = [p(C, D)]^{l \cdot m} \tag{3}$$

où $l = \text{card}(C)$ et $m = \text{card}(D)$.

La définition d'un indice de proximité entre parties disjointes de B permet la première étape de condensation sous la forme d'un arbre « détaillé » des classifications. Ce dernier s'obtient pas à pas, où à chaque pas on réunit la paire de classes (resp. les paires de classes s'il y en a plus d'une) qui réalise la plus grande valeur de l'indice de proximité (3). Partant de la partition la plus fine où chaque classe contient un élément, on aboutit à celle, la moins fine, à une seule classe, conformément au schéma suivant où on représente l'état initial et celui final.

L'état initial est fourni par un tableau de données croisant un ensemble E d'individus ou objets et un ensemble V de variables de description. On a illustré ci-dessous ce tableau dans le cas où $E = \{ e_i/1 \leqslant i \leqslant 7 \}$ et où $V = \{ v^j/1 \leqslant j \leqslant 5 \}$. L'état final est défini par un couple d'arbres de classifications, chacun construit séparément, le premier sur l'ensemble des variables et le second sur l'ensemble des individus. Dans le schéma précédent, la partition du niveau 2 de l'arbre sur V est $\{ \{ v^1, v^5 \}, \{ v^2, v^4 \}, \{ v^3 \} \}$, celle du niveau 3 de l'arbre sur E est $\{ \{ e_1, e_5, e_7 \}, \{ e_2, e_3, e_5, e_6 \})$. Nous avons déjà signalé ci-dessus que la détermination de ce couple d'arbres des classifications permet la réorganisation des lignes et colonnes du tableau des données conformément à la structure de proximité apparue.

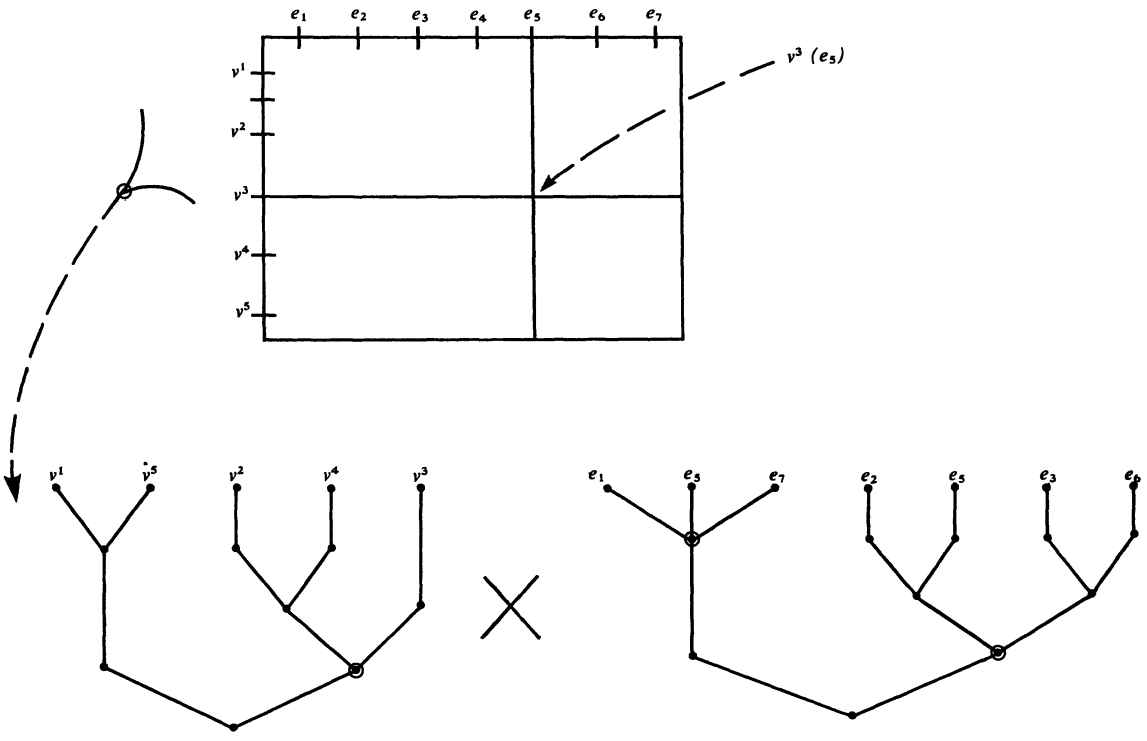


Fig. 2

L'algorithme qui construit la représentation polonaise de l'arbre des classifications à partir de (3) a été appelé « Algorithme de la Vraisemblance du Lien » (A.V.L.). Son comportement a été analysé des points de vue informatique et statistique par M^me^ Nicolaü (cf. [53] et [55]).

F. Nicolaü (cf. [54]) a imaginé dans sa thèse d'autres indices de même type : vraisemblance d'un lien; mais où le lien est mesuré par une autre fonction que le maximum de l'ensemble (2) des valeurs; par exemple à partir de la loi de la v.a., associée dans l'h.a.l. à la moyenne

$$\frac{1}{l \cdot m} \sum \{ P(c, d) / (c, d) \in C \times D \} \tag{4}$$

Toutefois, quel que soit l'intérêt de l'étude, les résultats comparatifs obtenus par la classification hiérarchique sur des données concrètes, montrent que c'est toujours l'indice (3) qui donne les résultats les plus raffinés et les plus cohérents dans leurs nuances.

Terminons ce paragraphe en soulignant qu'un bon critère pour l'émergence des classes n'est pas nécessairement le meilleur pour affecter un nouvel élément à une classe d'une classification déjà formée; ce n'est pas non plus celui qui doit être utilisé pour évaluer la qualité du résultat de la classification. Alors que la plupart des taxinomistes sinon tous utilisent le même critère pour les différents aspects.

5. *Nœuds significatifs; condensation de l'arbre*

Une étape décisive de la méthode consiste à condenser l'arbre aux niveaux où se produit un nœud « significatif » détecté à partir du comportement d'une statistique de proximité entre

une certaine forme de l'information quant aux ressemblances entre éléments de l'ensemble B à classifier et l'association entre deux classes correspondante au nœud.

Le principe d'élaboration de cette statistique de proximité est toujours le même (cf. § 3 ci-dessus). Mais pour se ramener à la comparaison de deux structures de même type, on ne retient de l'indice de proximité Q sur B que le préordre total associé sur l'ensemble $F = P_2(B)$ des paires d'éléments distincts de B, c'est à dire la préordonnance sur B, où

$$(\forall (p, q) \in F \times F), p < q \Leftrightarrow Q(p) > Q(q) \tag{5}$$

En effet, la donnée d'une partition π, éventuellement produite à un niveau de l'arbre, peut être regardée comme définissant un préordre total sur F à deux classes $R(\pi)$ et $S(\pi)$ 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.

Nous représentons dans $F \times F$ la préordonnance $\omega(B)$ par son graphe :

$$gr(\omega) = \{ (p, q) / (p, q) \in F \times F, p < q \text{ et non } q < p \text{ pour } \omega \} \tag{6}$$

et la partition π par le rectangle $R(\pi) \times S(\pi)$.

L'indice brut entre la préordonnance et la partition sera dans ces conditions

$$s(\omega, \pi) = \text{card}\,[gr(\omega) \cap (R(\pi) \times S(\pi))] \tag{7}$$

Ce cardinal a été introduit comme critère de classification par J.P. Benzecri (1965) sous la forme encore par trop métrique du « nombre d'inégalités entre les distances spécifiées par la partition et compatibles avec l'ordonnance ». Nous l'avons repris sous la forme (7) et surtout, nous avons étudié sa distribution lorsque π décrit l'ensemble des partitions d'un même type (i.e. dont les cardinaux des classes sont fixés) (1968, 1972) (cf. [31], [34]). Nous démontrons que cette distribution est asymptotiquement normale et nous caractérisons l'expression de chacun de ses moments.

La statistique notée Σ et obtenue en centrant et en réduisant (7) définit la « mesure » d'adéquation globale de la partition. La suite des valeurs observées de Σ sur la suite des niveaux de l'arbre des classifications permet une interprétation dynamique de ce dernier et sa condensation aux niveaux où se produit un nœud « significatif ». En attachant à chaque niveau i le taux d'accroissement $\theta_i = (\Sigma_i - \Sigma_{(i-1)})$, un tel nœud apparaît comme un maximum local de la distribution observée de θ le long de la suite des niveaux de l'arbre. L'examen conjoint de cette dernière distribution et de celle correspondante de Σ, permet de reconnaître les principaux états d'équilibre dans la synthèse.

6. *Méthodes environnantes*

Autour de l'armature principale que nous avons schématiquement décrite ci-dessus, nous avons développé un ensemble de méthodes plus locales relatives à des questions qu'on peut se poser sur la formation des classes et leur analyse dans tel ou tel type de données. Certaines méthodes précèdent la classification, d'autres la suivent.

Parmi les méthodes qui précèdent la classification signalons :

— l'affectation à chaque élément de l'ensemble à classifier, de la valeur d'un indice de « neutralité » par rapport à une visée classificatoire. Cet indice se présente sous la forme d'un rapport de deux variances; celle du numérateur est la variance des proximités (de la forme Q (cf. § 3)) de l'élément en question avec chacun des autres éléments.

Cet indice se calcule avant la classification mais s'exploite dans l'interprétation des résultats; sa valeur est d'autant plus grande que l'élément concerné intervient plus intimement dans la formation de la classe (cf. [32], [33]).

— mesure de la « classificabilité » ou aptitude de l'ensemble à être organisé en classes et sous-classes de proximité. Cet indice est conçu au niveau de la préordonnance, il caractérise la distorsion de sa structure par rapport à une préordonnance ultramétrique. Soulignons qu'une « bonne » classificabilité n'implique pas nécessairement une « bonne » cohésion qui se traduirait par une « forte » densité pour les classes formées par un algorithme de classification (cf. [31]).

Parmi les méthodes qui suivent la pratique de la classification signalons.

— croisement entre une classification sur l'ensemble des variables et une classification sur l'ensemble des individus; cette dernière pouvant être exogène et définie par rapport à un caractère extérieur (cf. [11], [44], [45].

— rôle d'un individu ou d'une classe d'individus dans la formation d'un profil de comportement défini par une classe d'attributs (cf. [11], [45]).

— « explication » d'une classe d'individus par une variable ou classe de variables pour différentes structures du tableau des données.

D'autre part, nous avons mis au point des algorithmes de recherche d'échelles d'attitude ou sériels qui permettent de dégager des structures d'enchaînement, réalisant ainsi une sorte d'analyse factorielle *ordinale* des données (cf. [29], [33], [47], [49]).

V — Applications

Les différents aspects méthodologiques que nous avons évoqué ci-dessus sont sous-tendus par un ensemble important de programmes auxquels ont travaillé de nombreux chercheurs. De jour en jour ces programmes se développent avec l'apparition de situations nouvelles et s'organisent de manière de plus en plus cohérente.

Les domaines d'application sont naturellement liés au cadre dans lequel nous avons évolué. Nous avons commencé nos travaux à la Maison des Sciences de l'Homme où nous étions surtout confrontés à des données en psychosociologie et en psychologie. En venant à Rennes, le champ d'application de nos méthodes s'est, au fur et à mesure, beaucoup étendu : didactique des mathématiques, médecine clinique, sociologie médicale, épidémiologie, économie rurale, urbanisme, gestion et marketing, préhistoire archéologique, fiabilité, phonétique, linguistique, etc.

En dehors du très grand intérêt qu'a porté l'utilisateur chercheur pour le résultat, chaque traitement nous a donné l'occasion d'une expérimentation méthodologique entraînant une meilleure connaissance et un apport nouveau.

Nous ne pouvons répondre dans l'absolu à la question qui revient souvent sur la bouche de l'utilisateur : « quelle est la taille nécessaire de l'échantillon pour dégager clairement les tendances comportementales? ». En effet, nos études ont montré que deux facteurs interviennent; le premier est lié aux variances des distributions marginales des variables descriptives et le second à la « force » des tendances sous-jacentes. Si ces tendances sont « fortes », un « petit » échantillon permettra de les extraire; si au contraire elles sont « faibles «, il faut un plus « grand » échantillon; la vérité générale étant que certaines tendances peuvent être fortes et d'autres plus faibles.

Comme nous l'annoncions déjà dans le préambule, nous allons chercher à donner ici un aperçu des résultats de différents traitements empruntés à différentes disciplines.

1. *Psychosociologie.* « *Perception des conditions de vie et image de la misère en Europe et en France* » (cf. [11], [45])

Les données proviennent d'une enquête dont le maître d'œuvre est l'I.F.O.P. et qui nous ont été communiquées par J.R. Rabier (Conseiller spécial à la commission des communautés Européennes). Le but de cette enquête à l'échelle Européenne est la préhension de la perception qu'a l'opinion publique des conditions de vie de la pauvreté et de la misère.

Un questionnaire d'une centaine de questions a été soumis en 1976 à des échantillons représentatifs des populations des neuf pays de la Communauté, âgées de quinze ans et plus, totalisant 8 622 personnes. Outre les questions habituelles d'identification des personnes interrogées, y compris quant à leurs attitudes religieuses et politiques, on a retenu des questions sur

— des observations d'ordre économique,

— les conditions de vie et les niveaux de satisfaction,

— la perception et image de la misère.

Chacune de ces questions définit à partir d'une échelle, une variable qualitative ordinale.

Pour chaque cas étudié (analyse au niveau de toute l'Europe pour définir une médiane Européenne et par pays), on a commencé par la classification hiérarchique de l'ensemble des échelles. Une telle organisation permet de dégager les thèmes majeurs ou principales tendances du comportement de la population étudiée.

On a ensuite éclaté chaque variable ou caractère de description en l'ensemble de ses modalités. La classification de l'ensemble des attributs-modalités, montre des profils d'attitude ayant un caractère local, qui « expliquent » l'apparition des tendances globales du comportement et qui sont plus directement exploitables.

Relativement à chacun de ces derniers profils, apparus dans l'étude au niveau de toute l'Europe, on a situé les différents pays selon une mesure de leur degré de responsabilité dans la formation du profil en question.

L'étude sur la France a été la plus complète. On a croisé chacune des classifications respectivement définies par chacune des variables d'identification avec chacun des profils d'attitude, encore une fois défini comme une classe d'attributs-modalités. L'indice d'association entre une telle classe et une même modalité d'une variable d'identification obéit au principe général exprimé au paragraphe IV. 3. On peut ainsi situer les différentes catégories socio-professionnelles, les différents votes politiques, les différentes religions, etc. par rapport aux différentes formes de la perception des conditions de vie.

2. *Pédagogie mathématique.* « *Formes d'aptitude et taxinomie d'objectifs cognitifs en mathématiques* » (cf. [17], [40], [49])

Les données proviennent d'un test élaboré par R. Gras, formé de 95 items recouvrant différentes notions à caractère mathématique et subi par un échantillon, choisi au niveau national et comprenant 1 621 élèves des 4e et 3e du C.E.S. Dans cette situation, chaque item définit un attribut ou variable logique 0-1, où le code 0 est attribué en cas d'échec et celui, 1 en cas de succès.

La séparation en étapes pour l'appropriation d'un concept est l'un des objets les plus importants de la recherche en didactique des mathématiques. Ces différentes étapes correspondraient à différents « objectifs cognitifs » tels qu'ils peuvent être définis conformément à la théorie

du développement de Piaget et Brüner. Il s'agit ensuite d'enchaîner ces objectifs conformément à une échelle d'appropriation d'un même concept.

La classification des items à partir du comportement réel des enfants a permis de faire apparaître à travers chacune des grandes classes un même thème mathématique : « géométrie des surfaces planes », « géométrie des lignes » et « nombre ». D'autre part et c'est essentiel, chacune des sous classes d'une même classe a précisément correspondu à un même « objectif cognitif ».

Ce résultat très encourageant a conduit R. Gras à construire un test autour du concept de symétrie centrale pour valider, à partir d'un indice d'implication les enchaînements entre comportements de réponse en s'appuyant sur une taxinomie d'objectifs cognitifs qu'il propose (cf. [17], [49]).

3. *Médecine.* « *Exploration fonctionnelle hépatique et recherche de profils biologiques* » (cf. [22], [23], [24])

Les données sont fournies par l' « I.N.S.E.R.M. (U 49) » et représentent un échantillon de malades de la pathologie hépato-biliaire, décrits à partir de 16 paramètres biologiques dont 4 représentaient d'ailleurs de nouvelles mesures à tester par rapport aux 12 autres plus classiques.

A partir d'un algorithme, optimal en un certain sens, de découpage de l'intervalle de variation d'un même paramètre en sous intervalles, on obtient deux codages « appauvris » des données. Le premier en remplaçant chacune des variables quantitatives numériques par une variable qualitative ordinale, dont une même modalité correspond à un sous intervalle. L'algorithme a permis d'obtenir un codage qui s'est avéré tout à fait pertinent par rapport à l'aspect « diagnostic médical ». Le second codage est encore plus pauvre puisqu'il correspond, comme ci-dessus, à associer à chaque modalité un attribut, ne tenant ainsi aucun compte de la structure d'ordre derrière l'ensemble des modalités d'un même caractère.

On a analysé le sens médical des différents types de résultats qu'on obtient pour les différents codages pouvant être adoptés. Le résultat le plus frappant dans l'analyse des différents types de classes obtenues est la stabilité des résultats lorsqu'on remplace le codage numérique par celui « qualitatif ordinal » et ce, aussi bien pour la classification des variables que des sujets. La classification des 16 variables a fait apparaître les principaux syndromes de la maladie et celle des sujets, les différents états maladifs à travers les grandes classes et les divers degrés de gravité d'un même état, à travers les sous classes.

On a d'autre part cherché à saisir la notion de « profil biologique » à partir de deux codages : celui « numérique » et celui en « attributs descriptifs » où le croisement de deux arbres condensés des classifications (l'un sur l'ensemble des attributs et l'autre sur celui des individus) a permis un examen plus analytique des comportements.

4. *Archéologie préhistorique.* « *Typologie d'un ensemble de bifaces du « Moustérien de tradition acheuléenne* » (cf. [16], [51], [52]).

A partir de quelques mensurations unidimensionnelles et rectilignes, la classification a mis en évidence des groupes de formes pouvant être décrites globalement. L'étude d'abord menée uniquement sur les contours de face, a été précisée par l'adjonction de mesures concernant le profil. Les formes classiques (ovalaires, cordiformes, limandes, discoïdes) se séparent bien; il apparaît en outre des catégories originales (hexagonaux, pentagonaux, losangiques, piriformes, etc.) basées sur l'angulosité des contours ainsi que sur l'opposition base/pointe. Une véritable réflexion de reconnaissance des formes a été conduite dans ce travail par l'étude des rapports

et liaisons entre mesures réalisées, par l'intermédiaire de classifications sur l'ensemble des paramètres et sur l'ensemble des objets décrits. On s'est notamment attaqué au problème fréquent de la réduction du nombre de mesures : choix des paramètres donnant une description suffisante pour une reconnaissance des formes. Cette reconnaissance s'est trouvée beaucoup simplifiée par l'association, à partir d'un critère adéquat, à chaque classe d'objets, d'un élément type.

Ici encore, l'analyse a pu bénéficier du croisement de classifications, mais sur le même ensemble des objets, où les deux classifications sont respectivement associées à deux ensembles disjoints de paramètres (par exemple : mesures de face et mesures de profil).

5. *Économie rurale.* « *La viticulture Girondine* » (cf. [57]).

La structure des données est ici celle d'une juxtaposition horizontale de tables de contingences, dont l'ensemble des lignes est indexé par l'ensemble des communes viticoles de la Gironde et dont l'ensemble des colonnes est indexé par l'ensemble des modalités de variables-partitions correspondantes à des tranches de surfaces cultivées ou de volume produit par catégorie de vin.

La classification des communes a été effectuée de deux façons; la première, dite « libre » où on respecte au mieux les proximités statistiques et la seconde sous contrainte de contiguïté, conduisant nécessairement à la formation de classes connexes d'un point de vue spatial.

Diverses zones ont été déterminées, se définissant autour de quatre grandes notions qui ont pu être évaluées par commune : le statut des exploitants, la surface viticole, le volume de la récolte et la taille des exploitations. Chacune des zones a pu être caractérisée, expliquée (à partir d'indices d' « explication ») et ses problèmes soulevées. Par la même, il a pu être remis fermement en question les régions économiques utilisées par les services publiques régionaux.

6. *Gestion.* « *L'impact de l'informatique sur les organisations* » (cf. [9])

L'objectif de cette recherche était de tenter de « mesurer » l'impact de l'informatique sur les organisations, impact sur la structure, sur les processus, sur le leadership et le comportement individuel ou de groupe. Nous nous contenterons de citer quelques unes des lois qui se dégagent à partir de l'analyse classificatoire.

Comme on pouvait s'y attendre, la dépendance vis à vis de l'informatique est plus grande dans le secteur tertiaire que dans le secteur secondaire où sa nature est tout à fait différente.

La dépendance à l'égard de l'ordinateur est davantage ressentie par les responsables que par les exécutants, cette dépendance est plus liée au pourcentage de documents émis vers l'ordinateur qu'au pourcentage de documents reçus de l'ordinateur qui est pourtant supérieur.

L'utilité d'un document est liée à la transformation possible des informations qu'il contient. Si cette transformation n'est plus à faire et si d'autres documents ne figurent pas en entrée du poste de travail concerné, ce dernier disparaîtra.

Le plus souvent, on n'a nullement repensé les principes d'organisation pour tenir compte de l'informatisation, se contentant de décalquer le passé, l'existant, sans toujours s'interroger sur le pourquoi de cet existant. C'est ainsi que si dans une organisation les documents manuels sont perçus comme peu utiles, il en est de même pour les documents informatisés.

On en arrive en fait à une véritable méthode de diagnostic du fonctionnement des entreprises et ce, à partir de 50 répondants seulement!

7. *Marketing.* « *Contribution à l'étude du commerce spécialisé* » (cf. [6])

On a pu mettre en évidence, en termes de structure, les magasins qui paraissent les plus fragiles. Trois situations se dégagent à ce sujet :

— les magasins qui ne sont pas assez spécialisés et qui restent sur le même « créneau » que les « grandes surfaces »; un signe qui ne trompe pas est la décroissance du chiffre d'affaires.

— les magasins qui sont par trop spécialisés où la spécialité peut correspondre à une mode passagère.

— les magasins dont la structure reste figée et transmise par filiation familiale.

On a aussi caractérisé ceux qui se développent. Par le croisement d'une classification sur l'ensemble des variables de structure et d'une classification sur l'ensemble des variables de comportement, on a pu décrire et analyser les comportements de gestion. C'est ainsi qu'on a observé des comportements insoutenables dans le cadre d'un certain type de structure.

Les liens entre structure, comportement et spécialités du magasin ont été dégagés pour différentes classes de magasins; notamment « accessoires », « prêt-à-porter féminin », « loisirs-hommes » et « luxe ».

8. *Fiabilité.* « *Étude des mesures statiques sur circuits intégrés logiques* » (cf. [50])

L'analyse a pu rejoindre certains aspects technologiques de la construction des composants. La classification des variables a mis en évidence les principaux groupes de mesures électriques et a permis de ne plus retenir que celles, essentielles. D'autre part, la classification des composants de différentes marques a permis de repérer les constructeurs qui maîtrisent le mieux la fabrication des composants.

Dans le cadre de cette étude, une véritable analyse du codage numérique des données a dûe être menée. En effet, l'expression des différentes mesures dépendait de choix arbitraires; d'autre part l'ordre de grandeur de la variance observée d'une même variable physique, était très différent d'un type de mesure à l'autre.

9. *Reconnaissance de la Parole.* « *Analyse de données phonétiques; codage et reconnaissance* » (cf. [28])

Les données phonétiques utilisées sont fournies par un vocodeur à 14 canaux. Chacun des phonèmes prononcé, élément de l'ensemble à classifier est ainsi représenté par un tableau de nombres à 14 colonnes et dont le nombre de lignes est variable. Chaque ligne correspond à ce qu'on appelle un « échantillon ». 13 ms séparent la production de deux « échantillons » consécutifs.

Les problèmes de segmentation et de codage sont ici essentiels. Pour pouvoir être comparés au moyen d'un indice de proximité, les différents phonèmes doivent avoir la même structure de représentation : tableau de nombres à 14 colonnes et à l lignes où l est fixé une fois pour toutes et toujours le même.

Une étude a déjà été effectuée au niveau d'un seul locuteur, une autre, sensiblement plus riche est en cours et concerne plusieurs locuteurs.

Relativement à un aspect de la première étude concernant un ensemble de voyelles, on peut signaler que l'avant dernier niveau de l'arbre des classifications distingue parfaitement les voyelles orales d'une part et nasales d'autre part. A un niveau plus bas, on sépare tout aussi parfaitement les voyelles fermées de celles ouvertes et ce, dans les deux cas. Le taux de reconnaissance reste acceptable au niveau le plus élémentaire de la reconnaissance d'une voyelle. Toutefois, il ne faut pas s'attendre à ce que ce dernier taux soit très élevé; en effet, l'élément traité (le phonème) a un caractère très local et les méthodes de segmentation arrivent difficilement à le distinguer et à le mettre en évidence.

Pour terminer, signalons que d'autres études de cas réels que ceux mentionnés dans le texte peuvent être consultés dans [2], [5], [7], [8], [10], [13], [25], [26], [37], [38], [48], [57] et [64].

BIBLIOGRAPHIE

[1] M. ADANSON. — Histoire naturelle du Sénégal. Coquillages, etc. Bauche; Paris (1757).

[2] R. BASTIDE, F. MORIN, F. RAVEAU. — Les Haïtiens en France. Mouton et Co, 1974, Paris.

[3] M. BECKNER. — The biological way of thought. Columbia University Press, New York, (1959).

[4] J.-P. BENZECRI. — L'Analyse des Données. 1. La Taxinomie, Dunod, Paris, 1974.

[5] A. BERGE, G. DENJEAN. — Comportement Digestif et Fonctionnement Intellectuel. *Revue de Neuro-psychiatrie Infantile*, 1974, 22 (6), pp. 355-370.

[6] R. BLOCH, J.-P. COURTELLE, M.-C. GAUTREAU (I.A.E. Bordeaux), L. BRETON, L. GUENNEGUEZ, A. PROD'HOMME (I.R.I.S.A. Rennes). — Rapport sur le commerce spécialisé. Secrétariat d'état au commerce, 1980.

[7] F. BONNIEUX, P. RAINELLI, T. CHANTREL, et I.C. LERMAN. — Construction d'indicateurs socio-économiques liés à la qualité de l'eau. Colloque international, I.R.I.A., « Analyse des Données et Informatique », Versailles, septembre 1977.

[8] P. BOUTIN, A. CHOLLET et B. TALLUR. — Essai d'application de techniques de l'Analyse des Données aux pointes à dos des niveaux aziliens de Rochereil. *Bulletin de la Société Préhistorique Française*, Études et Travaux, 1976.

[9] L. BRETON, A. PROD'HOMME, J. VILLARD. — Une contribution à l'étude de l'impact de l'informatique sur les organisations. Rapport final de l'A.T.P. C.N.R.S.-I.R.I.A. n° 75/2442, I.R.I.S.A., Rennes, novembre (1980).

[10] J.-L. BUARD. — Gestion statistique des demandes d'actes biologiques. Typologie des unités fonctionnelles. Thèse de 3e cycle, Université de Rennes I, décembre (1980).

[11] T. CHANTREL. — Nouvelle approche dans la classification et représentation d'un vaste ensemble d'échelles et profils d'attitudes. Application à des données en économie rurale et en psychosociologie. Thèse de 3e cycle, Université de Rennes I, mai 1979.

[12] M.-J. CHOMBARD DE LAUWE, Cl. BELLAN. — Enfants de l'image. Payot, Paris (1979).

[13] I. COHEN. — Classification d'une famille d'échelles au moyen d'un nouvel indice. Comparaison avec le traitement par l'Analyse des correspondances. Application à des Données en psycho-pédagogie et en sociologie rurale. Thèse de 3e cycle, Université de Paris VI (I.S.U.P.), février (1977).

[14] E. DIDAY. — Optimisation en classification automatique et reconnaissance des formes. R.A.I.R.O. série verte, 1973.

[15] E. DIDAY et collaborateurs. — Optimisation en classification automatique, Publications de l'I.N.R.I.A tomes 1 et 2, Rocquencourt, octobre 1979.

[16] J.-P. GEFFRAULT. — Reconnaissance des formes en archéologie préhistorique à partir de différents corpus de données par des méthodes de classification. Rapport de D.E.A. (sept. 79) et thèse de 3e cycle en préparation (Université de Rennes I) à soutenir en 1981.

[17] R. GRAS. — Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en Mathématiques. Thèse d'État, Université de Rennes I, octobre 1979.

[18] J.A. HARTIGAN. — Clustering Algorithms, John Wiley, New York, 1975.

[19] M. JAMBU. — Classification automatique pour l'analyse des données. Tome I : Méthodes et algorithmes, Dunod, Paris (1978).

[20] N. JARDINE, R. SIBSON. — Mathematical Taxonomy. John Wiley, New York, 1971.

[21] M.G. KENDALL. — Rank Correlation Methods. Charles Griffin, London (fourth edition, 1970).

[22] J.-Y. LAFAYE. — Les différentes formes de l'appréhension des données dans l'exploration fonctionnelle hépatique; discrétisation de variables numériques. Recherche de profils biologiques par une méthode de classification hiérarchique. Thèse de 3e cycle soutenue le 21/9/78, Université de Rennes I.

[23] J.-Y. LAFAYE. — Une méthode de discrétisation de variables continues. *Rev. Stat. Appl.*, 1979, n° 2.

[24] J.-Y. LAFAYE. — Une méthode automatique de discrétisation de variables numériques représentées par de petits échantillons. Actes du Congrès AFCET : « Reconnaissance des formes et intelligence artificielle », Toulouse 12-14 septembre 1979.

[25] F. LEBLANC-MARIDOR. — Le rôle du conditionnement dans la commercialisation des produits alimentaires. Thèse de 3e cycle en sciences de Gestion, Faculté des Sciences économiques de Rennes, novembre 1979.

[26] G. LECALVE. — Problèmes d'analyse des données, 2e partie d'une thèse d'état, Université de Rennes I, novembre 1976.

[27] B. LEFEBVRE, J. LOSFELD. — Formalisation constructive de la notion de classe polythétique, dans « Data Analysis and Informatics » E. Diday et al. (eds), North-Holland Publishing Company, 1980.

[28] A. LELIEVRE. — Étude théorique de l'appréhension de données en vue de leur codage optimal pour un traitement par une méthode de classification hiérarchique de « grands » tableaux. Application à la reconnaissance des phonèmes de la parole. Rapport CNET-Lannion et thèse de 3e cycle en préparation (Université de Rennes I) à soutenir en 1981.

[29] H. LEREDDE. — La méthode des pôles d'attraction, la méthode des pôles d'agrégation; deux nouvelles familles d'algorithmes en classification automatique et sériation. Thèse de 3e cycle, Université de Paris VI, octobre 1979.

[30] I.-C. LERMAN. — Analyse Hiérarchique, *Revue Mathématiques et Sciences humaines*, n° 17, Paris 1967; repris et complété dans [47].

[31] I.-C. LERMAN. — Les Bases de la Classification Automatique. Gauthier-Villars, collection Programmation, Paris, 1970.

[32] I.-C. LERMAN. — Sur l'analyse des données préalable à une classification automatique (proposition d'une nouvelle mesure de similarité), *Revue Mathématique et Sciences humaines*, 8e année, n° 32, 1970.

[33] I.-C. LERMAN. — Analyse du phénomène de la « sériation », *Revue Mathématiques et Sciences humaines*, n° 38, Paris 1972.

[34] I.-C. LERMAN. — Étude Distributionnelle de Statistiques de proximité entre structures finies de même type; application à la classification automatique, Cahiers du B.U.R.O., n° 19, Paris, 1973.

[35] I.-C. LERMAN. — Introduction à une méthode de classification automatique, illustrée par la recherche d'une typologie des personnages enfants à travers la littérature enfantine, *Revue de Statistique Appliquée*, vol. XXI n° 3, pp. 23-49, Paris, 1973.

[36] I.-C. LERMAN. — Formal Analysis of a General Notion of Proximity between Variables in Proceed of European Congress of Statisticians, 1976.

[37] I.-C. LERMAN, S. MORA OBREQUE, J. PAGES et R. ROBERT. — Contribution de deux méthodes d'analyse des données dans l'étude de la dynamique d'une population trispécifique de pucerons de la pomme de terre. Annales de l'E.N.S.A., année 1976.

[38] I.-C. LERMAN, M. BLANCARD, J.-Y. LAFAYE et M. MOREL. — Implémentation et évaluation d'une méthode de classification hiérarchique. Compte rendu contrat D.G.R.S.T., n° 757, 1459, janvier 1977.

[39] I.-C. LERMAN et H. LEREDDE. — La méthode des pôles d'attraction. Colloque international I.R.I.A., « Analyse des données en Informatique », Versailles, septembre 1977.

[40] I.-C. LERMAN. — Formes d'aptitudes et taxinomie d'objectifs cognitifs en Mathématiques d'après les travaux de R. Gras, *Revue Française de Pédagogie*, n° 44, Paris 1978.

[41] I.-C. LERMAN. — Méthodes combinatoires et statistiques dans le traitement des données du comportement. Bulletin de l'A.S.U., vol. 2, 1978, pp. 45-65.

[42] I.-C. LERMAN. — Étude formelle et statistique de la notion de ressemblance, Publ. I.R.I.S.A. nº 107, Rennes, décembre 1978.

[43] I.-C. LERMAN. — Les présentations factorielles de la classification. I dans RAIRO vol. 13, nº 2, pp. 107-128 et II dans RAIRO vol. 13, nº 3, pp. 227-251, 1979.

[44] I.-C. LERMAN. — Croisement de classifications floues, Publ. Inst. Stat. Univ. Paris, XXIV, fasc. 1-2, 13-46, 1979.

[45] I.-C. LERMAN, M. HARDOUIN et T. CHANTREL. — Analyse de la situation relative entre deux classifications floues, dans « Data Analysis and Informatics » E. Diday et al. (eds), North-Holland Publishing Company, 1980.

[46] I.-C. LERMAN. — Combinatorial analysis in the statistical treatment of behavioral data, Quality and Quantity, 14 (1980) 431-469.

[47] I.-C. LERMAN. — Analyse ordinale d'une classe d'échelles ou analyse hiérarchique dans « Analyse des données » tome I, Publication de l'A.P.M.E.P. nº 28, pp. 133-159, 1980.

[48] I.-C. LERMAN, B. TALLUR. — Classification des éléments constitutifs d'une jusxtaposition de tableaux de contingence, Publ. I.R.I.S.A., nº 127 et *Rev. Stat. Appl.*, 1980, nº 28, 3.

[49] I.-C. LERMAN, R. GRAS, H. ROSTAM. — Élaboration et évaluation d'un graphe d'implication pour des données binaires, Publ. I.R.I.S.A., nº 136 et à paraître sous la forme d'une suite de deux articles dans *Revue de Mathématique et Sciences humaines* en 1981.

[50] J.-R. MASSE. — Classes de tableaux équivalents en analyse descriptive des données; application à l'étude de mesures statiques sur circuits intégrés logiques. Thèse de 3e cycle, Université de Rennes I, octobre 1978. Ce travail a donné par ailleurs lieu à un rapport de recherche interne au C.N.E.T. (Dépt. « Fiabilité »).

[51] J.-L. MONNIER et R. ETIENNE. — Application des méthodes de classification hiérarchique de I.-C. LERMAN à deux séries de bifaces du Moustérien de tradition acheuléenne provenant des gisements de Kervouster (Finistère) et Bois-du-Rocher (Côtes-du-Nord), *Bulletin de la Société Préhistorique Française 1978*, tome 75/10.

[52] J.-L. MONNIER. — Le paléolithique de la Bretagne dans son cadre géologique. (Travaux du Laboratoire Anthropologie-Préhistoire-Protohistoire et Quaternaire Armoricain). Thèse d'État, Université de Rennes I, 1980.

[53] F. and M.H. NICOLAÜ. — Analyse d'un algorithme de classification et « Contributions au Traitement automatique de données », 2 thèses de 3e cycle. Université Paris VI, I.S.U.P., novembre 1972.

[54] F. NICOLAÜ. — Critérios de análise classificatória hierárquica baseados na função de distribuição, Faculté des Sciences de Lisbonne, Laboratoire de Statistique, Lisbonne 1980. Thèse de doctorat soutenue en février 1981.

[55] M.-H. NICOLAÜ. — Contribuições ao estudo dos cœficientes de comparação em análise classificatória, Faculté des Sciences de Lisbonne, Laboratoire de Statistique, Lisbonne 1980. Thèse de doctorat soutenue en février 1981.

[56] G. PIERAUT, LE BONNIEC and K. van METER. — Étude génétique de la construction d'une propriété relationnelle : la Relation de Passage, Monographies françaises de psychologie, nº 35, C.N.R.S., Paris 1976.

[57] A. PROD'HOMME. — Indices d'explication des classes obtenues par une méthode de classification hiérarchique respectant la contrainte de contiguïté spatiale. Application à la viticulture Girondine et à la construction de logements dans les Bouches-du-Rhône. Thèse de 3e cycle, Université de Rennes I, décembre (1980).

[58] S. REGNIER. — Sur quelques aspects mathématiques des problèmes de la classification automatique, I.C.C. Bull., vol. 4, 1965.

[59] P.H.A. SNEATH, R. SOKAL. — Numerical Taxonomy, W.H. Freeman and Co., San Francisco, 1971.

[60] R.-R. SOKAL, P.H.A. SNEATH. — Principles of numerical taxonomy. San Francisco and London, Freeman and Cº, 1963.

[61] R.-N. SHEPARD. — The analysis of proximities : scaling with an unknown distance function I et II. Psychometrika, 1962.

[62] W.-F. DE LA VEGA. — Techniques de classification automatique utilisant un indice de ressemblance. *Revue française de sociologie*, Paris, 1967.

[63] F. VICQ D'AZYR. — Quadripèdes, Discours préliminaire, dans « Encyclopédie méthodique », vol. 2. Panckoucke, Paris (1792).

[64] Ph. VILLOING. — Classification ascendante hiérarchique et indices de similarité sur données qualitatives nominales selon l'algorithme de la vraisemblance du lien. Thèse de 3e cycle, Université de Rennes I, décembre (1980).

[65] A. WALD, J. WOLFOWITZ. — Statistical Tests Based on Permutations of the Observations, *Ann. Math. Stat.*, vol. 15, 1944.

[66] J.H., Jr WARD. — Hierarchical grouping to optimize an objective function, J.A.S.A., 58, 236-244.