Statistique et analyse des données

I. C. Lerman

Méthodes combinatoires et statistiques dans le traitement des données du comportement

Statistique et analyse des données, tome 3, n° 2 (1978), p. 45-65

<http://www.numdam.org/item?id=SAD_1978__3_2_45_0>

Statistique et Analyse des Données
2 - 1978

METHODES COMBINATOIRES ET STATISTIQUES DANS LE TRAITEMENT DES DONNEES DU COMPORTEMENT

par I.C.LERMAN

Laboratoire de Statistique - I.R.I.S.A. - Université de Rennes -

1-INTRODUCTION

Cet article résulte de la mise en forme actualisée et complétée d'un exposé de synthèse aux premières journées nationales sur la Classification (VANNES, Mai 1977). Nous tenterons de donner un aperçu sur l'ensemble des méthodes et réalisations que nous avons pu, au bout de longues années, élaboré avec l'aide et la collaboration de nombreux chercheurs en Statistique et Informatique Appliquées.

Ces travaux ont un triple caractère

. Combinatoire et Statistique pour l'élaboration des méthodes.

. Informatique pour leur mise en oeuvre.

. Participation à la recherche dans diverses disciplines des Sciences Humaines, Economiques et de la Nature, pour la validité des méthodes, leur progrès et leur promotion.

La plupart des méthodes couramment utilisées en Analyse des Données se réfèrent à une représentation euclidienne. Nos méthodes qui permettent le traitement rapide des grands tableaux, se réfèrent à une représentation mathématique finie. S'il s'agit de situer ces méthodes par rapport à celles de J.P. BENZECRI ; on peut signaler rapidement que ces dernières utilisent une représentation géométrique et analysent métriquement la dépendance mesurée par le χ^2. Par contre, nos méthodes utilisent une représentation algébrique et combinatoire et procèdent d'une notion très générale de la corrélation entre structures. Cela permet de s'adapter avec souplesse à n'importe quel type de tableau de données et de rester ainsi plus près du langage posé par l'utilisateur.

La clarté de la conception statistique et mathématique de l'outil repose sur quelques idées fondamentales qu'on cherchera brièvement à exprimer ci-dessous et qui n'autorisent que le 'minimum d'arbitraire possible" dans la construction des méthodes. Mais la qualité de l'outil ne se trouve finalement établie que par la richesse et la cohérence dans les nuances des résultats concrets obtenus.

La première idée concerne un codage fidèle des données obtenu au moyen d'une représentation mathématique finie ayant un caractère combinatoire et algébrique. En effet les variables descriptives, telles qu'elles se manifestent dans les Sciences de l'Homme et la Nature, sont rarement numériques et il importe que la méthode de synthèse de l'information respecte une représentation fidèle des données. Ainsi, par exemple, une variable "qualitative ordinale" qui définit un préordre total sur l'ensemble E des objets, sera représentée par la partie suivante de E x E

$$\sum_{i<j} E_i \times E_j \qquad \text{(somme ensembliste)} \qquad (1)$$

où on a noté E_i , i=1,...,k; la i-ème classe du préordre. Dualement, un objet décrit par une suite de telles variables $(\omega_1, \omega_2, \ldots, \omega_m)$, sera représenté par un point du produit direct de m ordres totaux dont le j-ème est défini sur l'ensemble des modalités de la j-ème variable ou question.

La deuxième idée est relative à la structure de condensation la plus adaptée pour le problème posé par l'utilisateur. Cette structure a le plus souvent un caractère fini ; partition, chaîne de partitions, ordre,... et il s'agit de chercher à la découvrir directement et non espérer la retrouver plongée dans une structure plus riche tel par exemple, un plan euclidien. Pour cela, le critère de synthèse doit tenir compte de la représentation fidèle des données et du type de structure recherché tout en ayant un fondement statistique clair.

L'idée la plus importante qui est à la base de la richesse et de l'intérêt des résultats concrets que nous avons pu obtenir réside dans la manière de définir le critère de synthèse. Il s'agit d'une notion très générale de corrélation entre structures finies de même type à partir de laquelle on introduit une notion de "vraisemblance" dans celle de "ressemblance" entre structures de même type. La démarche que nous allons illustrer dans le cas de la comparaison d'un couple de partitions (Π_1, Π_2) sur E, peut être schématisée comme suit

$$(\alpha,\beta) \in A \times B \rightarrow (R(\alpha), R(\beta)) \in \Omega \times \Omega$$

$$\downarrow$$

$$s = \text{card}\,(R(\alpha) \cap R(\beta))$$

$$\downarrow$$

$$\Pr\{S = \text{card}\,(R(\alpha') \cap R(\beta')) < s\}$$

Dans l'ensemble cité $(\alpha,\beta) = (\Pi_1, \Pi_2)$, A (resp. B) est l'ensemble des partitions sur E de même type que Π_1 (resp. Π_2), $\Omega = P_2(E)$ est l'ensemble des paires d'objets distincts de E, $R(\Pi_1)$ (resp. $R(\Pi_2)$) est l'ensemble des paires réunies par Π_1 (resp. Π_2) et Π'_1 (resp. Π'_2) est un élément aléatoire de A (resp. B) muni d'une probabilité uniformément répartie.

Cette notion intervient à toutes les étapes de la recherche d'une classification ou d'une suite de classifications (ordonnée ou non, par finesse décroissante) :

- Indice de proximité entre variables ou bien entre objets
- Indice de proximité entre classes d'éléments de l'ensemble à classifier
- Détermination des noeuds ou bien des niveaux "significatifs"
- Contrôle de l'interprétation des résultats.

A partir de là, deux types d'algorithmes ont été développés pour la formation des classes. Le premier opère par agrégations successives et conduit à un arbre des classifications qu'on condensera à partir de la reconnaissance statistique des noeuds significatifs. La seconde stratégie opère par séparations. On établit une suite de classifications, la première en deux classes, la deuxième en trois classes, ..., la k-ème en (k+1) classes définie par aglutination autour d'un système de (k+1) "pôles d'attractions". Ces pôles d'attraction qu'on sait devoir entraîner des classes de bonne cohésion, sont déterminés par une analyse simultanée de la variance et de la moyenne des proximités (ou distances) de chacun des éléments à la suite des autres de l'ensemble à organiser (cf. [23] et chapitre 7 de [21]). Dans cette deuxième classe de méthodes nous utilisons davantage pour la formation et la signification des classes des critères d'inspiration géométrique ; par exemple, le moment d'inertie défini relativement à une métrique conforme à la classe de nos indices de proximité. Certaines représentations euclidiennes sont d'ailleurs proposées ; lesquelles permettent de dégager des "sériations" ou des "échelles d'attitude".

Si la classification des variables permet la découverte des principales tendances de comportement ou des principaux profils d'attitude de la population étudiée ; celle des objets ou individus permet la mise en évidence de groupements d'éléments proches ; une même classe d'objets s'"expliquant" par la dominance chez ses éléments de certains profils d'attitude.

La première étape consiste en la construction d'une table carrée des valeurs des proximités entre éléments de l'ensemble à classifier. Il peut s'agir de l'ensemble V des variables descriptives ou bien de l'ensemble E des objets.

II-TYPOLOGIE DES VARIABLES ; LEUR REPRESENTATION

Nous représentons une variable descriptive par une partie d'un ensemble ou une pondération sur ce dernier. Nous commençons par distinguer deux principaux types d'une variable ; 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 x E ou une pondération sur E x 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 définit 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

- La variable "rang" qui définit un ordre total o sur E que nous représentons par son graphe $R(o) \subset E \times E$:

$$R(o) = \{(x,y) \in E \times E / x < y \text{ pour } o\} \quad (1).$$

- Le caractère descriptif à l'ensemble totalement ordonné des modalités qui définit un préordre total ω sur E que nous représentons par la partie $R(\omega)$ de E x E définie par

$$R(\omega) = \{(x,y) \in E \times E / x < y \text{ et non } y < x \text{ pour } \omega\} \quad ; \quad (2)$$

on a en fait

$$R(\omega) = \sum_{i<j} E_i \times E_j \quad \text{(somme ensembliste)} \quad (3)$$

où E_i est la i-ème classe du préordre formée des objets possédant la i-ème 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. On a à priori les deux représentations

$$R(\pi)=\{ \{x,y\} \in F/\exists j, 1\leq j\leq k, x \in E_j \text{ et } y \in E_j\}, \quad (4)$$

ensemble des paires réunies par la partition $\pi=\{E_1,E_2,\ldots,E_k\}$

et $S(\pi)= \{ \{x,y\} \in F/ \exists i\neq j, x \in E_i \text{ et } y \in E_j\}$:

ensemble des paires séparées par la partition π.

Ces deux représentations sont équivalentes pour notre indice de proximité.

- La variable "pondération" sur ExE peut être représentée par une matrice carrée $\{^{\mu}xy/(x,y) \in E \times E\}$ indexée par E x E où $^{\mu}xy$ est la pondération affectée ,au couple (x,y). Ce type de variable est plus rare que ceux qui précèdent ; il se présente par exemple dans la définition d'une relation de communication, d'une relation de dominance, d'une relation "floue",...

III-TYPE D'UN TABLEAU DES DONNEES ; REPRESENTATION DES INDIVIDUS

On définit le type d'un tableau de données Variables X Individus comme étant celui, supposé commun, des différentes variables descriptives.

Nous allons à présent préciser la représentation des individus par rapport à l'ensemble des variables descriptives pour chaque type de tableau de données.

1- Cas où les variables sont des attributs descriptifs

Le tableau des données est un tableau d'incidence de zéros et de uns. Il y a analogie formelle entre d'une part le rôle joué par l'ensemble des attributs par rapport à celui des individus et d'autre part, le rôle joué par l'ensemble des individus par rapport à celui des attributs. On représentera un individu x par l'ensemble A_x des attributs qu'il possède.

2- Cas où les variables sont des caractères où l'ensemble des modalités est sans structure

En attachant à chaque modalité un attribut de description ; on se ramène conceptuellement à la situation précédente, mais avec une structure particulière de la représentation ; la particularité étant dûe à ce que les modalités d'un même caractère sont mutuellement exclusives.

3- Cas où les variables sont des caractères descriptifs où l'ensemble des modalités est totalement ordonné.

Nous avons déjà signalé dans l'introduction quel doit être dans ce cas la représentation d'un même individu. Si m est le nombre de variables, il s'agit d'un point du produit direct de m ordres totaux dont le j-ème est défini sur l'ensemble des modalités de la j-ème variable ou question. Ce point a pour coordonnées la suite des codes des modalités de réponse de l'individu aux différentes questions. L'ensemble des individus sera ainsi représenté par un "nuage" de points dans le treillis distributif produit direct de m ordres totaux ; un même sommet de ce treillis sera pondéré par la fréquence des sujets qu'il représente.

4- Cas où les variables descriptives sont des variables "rang".

Ce cas où chaque variable définit un ordre strict et total sur l'ensemble des individus, est formellement équivalent au précédent mais, plus particulier. Par rapport à ci-dessus, tout se passe comme si chaque variable avait n modalités de réponse (n = card (E)), également réparties. La pondération affectée à un même sommet du treillis est dans ce cas égale à 0 ou 1.

5- Cas où les variables descriptives sont numériques.

Dans ce cas, le plus classique, on associe à chaque variable une forme linéaire coordonnée de R^m (m= card (V)) et on représente un individu par le point de R^m dont la suite des coordonnées est la suite des valeurs des différentes variables sur l'individu. L'ensemble E se trouve ainsi représenté par un "nuage" de points dans R^m.

6- Cas d'un tableau de contingence.

Arrivés à ce point de notre analyse, on comprendra pourquoi nous considérons un tableau de données représentant le croisement Variables x Individus comme fondamentalement dissymétrique. Même dans le cas où V est formé d'attributs de description où pourtant la représentation de E par rapport à l'ensemble A des attributs peut être regardée comme formellement analogue à celle de A par rapport à E (en effet, de la même façon que A est représenté par un échantillon de points dans l'ensemble P(E) des parties de E, E peut être représenté par un échantillon de points dans P(A) ; on obtient des résultats plus significatifs pour la comparaison des éléments de E, en construisant une hypothèse d'absence de lien qui tient compte de la fréquence relative des différents attributs (cf.§I et §5 ci-dessous).

Il y a pourtant le cas d'un tableau de données parfaitement symétrique ; il s'agit du tableau de contingence

$$\{k_{ij}/(i,j) \in I \times J\} \quad , (1)$$

représentant les cardinaux des classes du croisement de deux partitions ; la première à n = card(I) classes et la deuxième à m = card(J) classes. Les k_{ij} sont donc des nombres entiers positifs ou nuls.

La représentation de l'ensemble I des lignes du tableau à travers l'ensemble J des colonnes du tableau est fournie dans le cadre de l'Analyse des Correspondances.

En notant, pour tout $(i,j) \in I \times J$,

$$k_{i.} = \Sigma\{k_{ij}/j \in J\} \quad , \quad k_{.j} = \Sigma \{k_{ij}/i \in I\} ,$$

$$k.. = \Sigma \{k_{ij}/(i,j) \in I \times J\}$$

$$f_{ij} = k_{ij}/k.. \quad , \quad p_{i.} = k_{i.}/k.. \quad , \quad p_{.j} = k_{.j}/k.. \quad , \quad (2)$$

$$f^i_J = (f^i_j/j \in J) \text{ où } f^i_j = f_{ij}/p_{i.} : \text{"profil de i à travers J"} \quad ; \quad (3)$$

on associe à I le nuage $\mathcal{N}(I)$ dans R^m muni de la métrique diagonale $(q_{jj}=1/p_{.j} \mid j \in J)$:

$$\mathcal{N}(I) = \{(f^i_J, p_{i.})/i \in I\} \quad , (4)$$

où $p_{i.}$ est la masse affectée au point f^i_J pour tout $i \in I$.

IV-INDICE DE PROXIMITE ENTRE VARIABLES DESCRIPTIVES

1-Expression générale de l'indice dans le cas "discret".

Nous allons reprendre ici, de façon un peu plus détaillée, le shéma de définition de la proximité.

Si (α,β) est un couple de variables d'un même type (i.e. définissant le même type de structure sur E), représenté par un couple de parties $(R(\alpha),R(\beta))$ de E, s'il s'agit d'attributs descriptifs et de E x E s'il s'agit de variables de la deuxième catégorie ; on introduit l'indice "brut" de proximité

$$s = \text{card}\,(R(\alpha) \cap R(\beta)) \quad , \quad (1)$$

Nous associons à α (resp. β) l'ensemble A (resp. B) des structures sur E de même type et ayant les mêmes caractéristiques cardinales que α (resp.β). Ainsi par exemple si α est un préordre total sur E de composition $u=(n_1,n_2,\ldots,n_k)$ $(n=n_1+n_2+\ldots+n_k)$; A peut être défini comme l'ensemble de tous les préordres totaux sur E de même composition u.

A s nous associons les deux variables aléatoires duales

$S_\alpha = \text{card}\,[R(\alpha) \cap R(\beta')]$ et $S_\beta = \text{card}\,[R(\alpha') \cap R(\beta)]$ où α'(resp.β') est un élément aléatoire dans l'ensemble A (resp. B) muni d'une probabilité uniformément répartie. Nous démontrons de façon synthétique, pour les différents cas ci-dessus envisagés (cf. § II), que la distribution de la v.a. S_α est la même que celle, S_β. On précise l'expression des moments d'une telle distribution et son caractère asymptotiquement normal sous certaines conditions, assez générales.

En désignant par $\mu_{\alpha\beta}$ et $\sigma^2_{\alpha\beta}$ la moyenne et la variance de la v.a. S_α (resp. S_β) ; l'indice "centré réduit" que nous adoptons se met sous la forme

$$Q(\alpha,\beta) = (s-\mu_{\alpha\beta})/\sigma_{\alpha\beta} \qquad (2)$$

N désignant l'hypothèse d'absence de lien ci-dessus exprimée, l'indice définitif que nous considérons et qui se réfère à une échelle de probabilité s'écrit

$$P(\alpha,\beta) = \Pr^N\{S < s\} \qquad (3)$$

où S est l'une des deux v.a. de même loi S_α et S_β. En d'autres termes, les deux variables descriptives α et β ont un degré de ressemblance d'autant plus grand que la valeur de s est invraisemblablement grande, par rapport à N.

Le passage de la formule (2) à celle, (3) se fait, avec une bonne approximation, au moyen de la relation

$$P(\alpha,\beta) = \Phi\,[Q(\alpha,\beta)] \qquad (4)$$

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

Les résultats obtenus dans ce cadre sont les suivants

- Dans le cas où (α,β) est un couple (a,b) d'attributs descriptifs, Q(a,b) n'est autre que le coefficient d'association de K. Pearson

- Dans le cas où (α,β) est un couple (o,o') de variables "rang" (i.e. totalement et strictement ordinales), Q(o,o') n'est autre que l'indice τ de M.G. Kendall ; mais où on aurait remplacé le dénominateur (qui n'est, dans cet indice, que le maximum de la valeur absolue du numérateur), par l'écart-type $\sigma_{oo'}$

- Dans le cas de la comparaison d'un couple (ω,ω') de préordres totaux, on obtient un indice tout à fait nouveau et on démontre que celui, proposé dans ce cas par M.G. Kendall, est "biaisé" dans le sens suivant : l'espérance mathématique de la variable aléatoire associée dans l'hypothèse N d'absence de lien, est différente de zéro. En effet, pour définir son indice, M.G. Kendall, n'a fait que retenir l'algorithme de calcul de τ en affectant, non sans arbitraire, la valeur d'une fonction ordinale aux différents objets d'une même classe du préordre, (traitement des ex quos ("tiés")).

- De façon tout à fait parallèle à ce qui vient de précéder, on obtient un indice tout à fait nouveau dans le cas de la comparaison d'un couple (π,π') de partitions sur E, défini par un couple de caractères descriptifs. L'indice obtenu est essentiellement différent de celui du χ^2 attaché au tableau de contingence de croisement des deux partitions. Des travaux à caractère expérimental de comparaison des deux statistiques sont déjà entrepris du point de vue de la simulation.

2- Généralisation de la comparaison d'un couple d'attributs à celle de la comparaison d'un couple de pondérations sur E.

Ayant codé E par l'ensemble des indices $I=\{1,2,\ldots,i,\ldots,n\}$, on peut voir aisément que d'une part, l'indice brut de proximité entre deux attributs descriptifs a et b se met sous la forme $s=\sum_{i\in I} \alpha_i \beta_i$ où α (resp. β) est la fonction indicatrice de la partie E_a (resp. E_b) et que d'autre part, les deux v.a. S_a et S_b sont équivalentes à celles

$$\sum_{i\in I} \alpha_i \beta_{\sigma(i)} \quad \text{et} \quad \sum_{i\in I} \alpha_{\sigma(i)} \beta_i, \quad \text{où } \sigma \text{ est un élément aléatoire dans l'ensemble, muni}$$

d'une probabilité uniformément répartie, de toutes les permutations sur I.

Par conséquent, la comparaison d'un couple de pondérations sur E qui généralise celle d'un couple d'attributs conduit à prendre comme indice brut $s = \sum_{i\in I} \alpha_i \beta_i$ où cette fois-ci α_i (resp. β_i) est le nombre affecté à l'objet codé i par la première (resp. seconde) variable. Les deux v.a. associées sont $\sum_{i\in I} \alpha_i\beta_{\sigma(i)}$ et $\sum_{i\in I} \alpha_{\sigma(i)}\beta_i$ où σ est un élément aléatoire dans l'ensemble, muni d'une probabilité uniforme, de toutes les permutations sur I. On sait que l'étude du comportement asymptotique d'une telle distribution est l'objet du célèbre théorème de Wald, Wolfowitz et Noether.

3- Généralisation à la comparaison d'un couple de pondérations sur ExE.

De la même façon, il s'agit d'étendre la notion de proximité entre deux variables discrètes de même type, que nous avons représentées par un couple de parties de E x E, telles que celles définissant des ordres totaux, préordres totaux ou partitions sur E, à la comparaison d'un couple de pondérations sur E x E, de la forme

$$\{\mu_{ij}/(i,j) \in I^{(2)}\} \quad \text{et} \quad \{\nu_{ij}/(i,j) \in I^{(2)}\}$$

avec $I^{(2)} = (I\times I-\Delta)$, où Δ est la diagonale et où pour tout (i,j) de $I^{(2)}$, $\mu_{ij} \in R^+$ (resp. $\nu_{ij} \in R^+$).

Il est naturel de considérer comme indice brut

$$s = \sum_{(i,j) \in I^{(2)}} \mu_{ij} \nu_{ij} \quad , \quad (5)$$

et d'associer les deux v.a. duales qu'on démontre être de même loi

$$S = \sum_{(i,j) \in I^{(2)}} \mu_{ij} \nu_{\sigma(i)\,\sigma(j)} \text{ et } T = \sum_{(i,j) \in I^{(2)}} \mu_{\sigma(i)\sigma(j)} \nu_{ij}, \quad (6) \ ;$$

où σ est un élément aléatoire dans l'ensemble, muni d'une probabilité uniforme, de toutes les permutations sur I.

En effet, nous établissons que les différents cas de comparaison d'un couple de variables discrètes, ci-dessus envisagé, peuvent formellement apparaître comme particuliers de la situation considérée ici.

S'inspirant d'un vieux papier de H.E.DANIELS, G. LECALVE a eu l'idée de cette extension qui a conduit à un nouvel indice. Nous avons toutefois repris cette approche de façon plus précise relativement notamment au calcul des moments ce qui permet une meilleure justification de la tendance vers la loi normale de la distribution de S (resp. T) (cf. [15] et [22]).

4- Sur les différents codages des données.

Nous avons fait transparaître au paragraphe II précédent le souci que nous avions d'une représentation mathématique d'une variable qui soit simple, traduisible dans le langage naturel et qui respecte exactement la "pauvreté" de l'échelle descriptive de la variable. Toutefois, une même variable peut admettre plus d'une représentation "fidèle" ainsi un attribut de description a peut être regardé comme définissant une partition dont la seule classe qui contient plus d'un seul élément est E_a : ensemble des objets possédant l'attribut. a peut également être considéré comme un préordre total à deux classes E_a^c (complémentaire de E_a) et E_a où $E_a^c < E_a$ pour l'ordre quotient. Il en résulte dans ces conditions pour a une représentation par une partie de E x E (cf.§ II). Cependant, nous avons montré d'un point de vue théorique et vérifié expérimentalement que la valeur de la proximité sous la forme d'un indice de vraisemblance (cf. formule (3) § IV) est, dans la plupart des cas, invariante quel que soit l'un des trois codages ci-dessus mentionnés pour un attribut de description. La vérification expérimentale s'est faite au niveau des résultats globaux de la classification des attributs à partir de l'"Algorithme de la Vraisemblance du Lien" dont il sera question au paragraphe VI. Ces résultats montrent l'uniformité de la démarche.

L'invariance des résultats de la classification après "appauvrissement" des échelles des différentes variables où les échelles respectives retenues sont choisies statistiquement les plus "significatives", est une préoccupation très actuelle de notre recherche. Elle est au centre des travaux de J.Y. LAFAYE (cf. (14)) qui, relativement à des données médicales, a analysé l'effet du remplacement de variables numériques définies en l'occurrence par des dosages cliniques, au moyen de variables préordinales. Ces dernières sont obtenues à partir d'un algorithme de découpage de l'intervalle de variation d'un même paramètre en sous intervalles ; chacun définissant une modalité du caractère descriptif associé au paramètre. Relativement à l'algorithme en question

qui assure le caractère statistiquement "significatif" de la subdivision ; on a pu observer une quasi-invariance des résultats de la classification des variables aussi bien que celle des sujets pour un indice de proximité entre individus dont le principe sera exprimé au paragraphe suivant. L'algorithme de formation des classes est celui du paragraphe VI ci-dessous.

Une autre préoccupation de la recherche concerne la caractérisation du type de résultats qu'on obtient à partir de la prise en compte d'une structure donnée de l'information. Ainsi, relativement à différentes études et notamment à une importante enquête psycho-pédagogique : 4 000 sujets décrits au moyen d'une centaine d'échelles d'attitude ; on a pu se rendre compte de la différence de nature dans la classification des variables lorsqu'on remplace chacune des échelles par un ensemble d'attributs descriptifs dont chacun est associé à une position de l'échelle (travaux de I.COHEN, cf. [5]).

V-INDICE DE PROXIMITE ENTRE INDIVIDUS ; CAS D'UN TABLEAU DE CONTINGENCE

Il y a lieu de commencer par considérer le cas symétrique d'un tableau de contingence où, pour se fixer les idées, on se pose le problème de la définition d'un indice de proximité entre éléments de J (cf.§ III.6 ci-dessus). Pour une description de I à travers J codée au moyen du nuage $\mathcal{N}(I)$ (cf. (4) § III.6), un tel indice correspondra à celui entre variables. Il doit toutefois remplir deux conditions :

(i) être conforme à la classe de nos autres indices de proximité entre variables ; c'est-à-dire, correspondre à une notion de corrélation avant référence à un indice de probabilité.

(ii) dériver de la métrique du X^2.

Cet indice a été mis au point par B.TALLUR (cf. [34]) qui l'a généralisé au cas de la juxtaposition selon un seul côté de tableaux de contingence et l'a appliqué avec des résultats pleins d'intérêt à l'étude de la structure de l'agriculture régionale Française (cf. [34]).

Considérons à présent le problème de la définition d'un indice de proximité entre individus dans le cas d'un tableau de données Variables x Individus, qui a donc un caractère fondamentalement dissymétrique. Le principe de la constitution d'un tel indice dans le cas discret (i.e. variables représentées chacune par une partie de E ou de E x E) est le suivant :

A la suite des valeurs $(x_1,x_2,\ldots,x_m)$ des différentes variables descriptives de même type sur un sujet x ; l'hypothèse d'absence de lien associe une suite $(X_1,X_2,\ldots,X_m)$ de v.a. indépendantes ; la loi de X_j étant définie par la distribution observée de la j-ème variable descriptive.

A un couple d'objets (x,y) dont le couple des vecteurs des valeurs des différentes variables descriptives est $((x_1,x_2,\ldots,x_m),(y_1,y_2,\ldots,y_m))$, on associe le couple de vecteurs aléatoires indépendantes $((X_1,X_2,\ldots,X_m), (Y_1,Y_2,\ldots,Y_m))$ et de même loi. A l'indice "brut"

$$s = \sum_{1\leq j\leq m} x_j y_j \quad , \quad (1)$$

on associe la v.a.

$$S = \sum_{1\leq j\leq m} X_j Y_j \quad , \quad (2)$$

dont la loi peut être approchée par la loi normale pourvu que m ne soit pas "trop petit".

D'où l'indice "centré réduit"

$$\left[s - \mathcal{E}(S)\right]/\sigma(S) \quad , \quad (3)$$

qui permet d'accéder à celui qui se réfère à une échelle de probabilité.

Cet indice a donné les résultats les plus raffinés dans le cas où les variables descriptives sont d'un type discret. Mais dans le cas où les variables sont numériques et où les données sont hétérogènes (ordre de grandeur de l'intervalle de variation ou de la variance d'un même paramètre, très variable d'un paramètre à l'autre), l'effet de "taille" des objets devient envahissant avec un tel indice. Cependant les meilleurs résultats ont été obtenus en appliquant le shéma précédent de la formation d'un indice (cf. formules (1),(2) et (3) ci-dessus) non pas à partir des mesures brutes ($v(x)/v \in V$) sur l'objet x, mais à partir de mesures transformées ($w(x)=\varphi\left[v(x)\right]/v \in V$) assurant au niveau de l'indice "brut" s certaines conditions d'invariance géométrique. Ainsi, on peut prendre pour w(x)

$$w(x) = \left[v(x)-m(x)\right] \Big/ \sqrt{\sum_{v\in V} \left[v(x)-m(x)\right]^2} \quad , \quad (4)$$

où

$$m(x) = \frac{1}{m} \sum_{v\in V} v(x) \quad ;$$

le coefficient s est alors celui de corrélation entre objets. On peut également prendre pour w(x)

$$w(x) = v(x)/\sqrt{\sum_{v\in V} \left[v(x)\right]^2} \quad , \quad (5)$$

l'indice "brut" s entre les deux objets x et y représente le cosinus de l'angle des deux vecteurs d'origine commune l'origine et d'extrémités respectives les points $(x_1,x_2,\ldots,x_m)$ et $(y_1,y_2,\ldots,y_m)$ représentent les deux objets x et y dans R^m.

VI-INDICE DE PROXIMITE ENTRE CLASSES ; "ALGORITHME DE LA VRAISEMBLANCE DU LIEN".

Quel que soit le type du tableau de données, nous sommes parvenus à l'établissement d'un indice de proximité entre variables descriptives ou bien entre objets décrits, qui se réfère à une échelle $[0,1]$ de probabilité, où on introduit la notion de vraisemblance dans celle de "ressemblance". Désignons par L l'ensemble à classifier ; il peut s'agir de l'ensemble V des variables descriptives ou bien, de celui E des objets décrits et soit

$$\{Q(e,f)/\{e,f\} \in P_2(L)\} \quad , \quad (1)$$

les valeurs de l'indice "centré réduit" sur l'ensemble des paires d'éléments distincts de L. C'est théoriquement la formule

$$P(e,f) = \Phi\left[Q(e,f)\right] \quad , \quad (2)$$

où Φ est la fonction de répartition de la loi normale centrée réduite, qui permet de passer à l'indice se référant à une échelle de probabilité. Toutefois la variance des données dans les

Sciences Humaines est telle, que pratiquement nous nous rapprochons des conditions de l'hypothèse d'absence de lien, au niveau global des proximités en remplaçant les indices Q(e,f) par ceux Q'(e,f) qui s'en déduisent au moyen des formules

$$(\forall\{e,f\} \in P_2(L)), Q'(e,f) = \left[Q(e,f)-\overline{Q}\right] / \sigma_Q \quad , (3)$$

où $\overline{Q}$ et σ_Q^2 sont la moyenne et la variance de la distribution de Q sur $P_2(L)$.

Cette transformation a permis d'organiser de façon plus nuancée et plus cohérente les liens "faibles" entre classes formées à un niveau élevé de l'arbre des classifications (cf.[4])

Cet arbre des classifications est construit de façon ascendante en démarrant de la partition la plus fine et en réunissant à chaque pas la paire ou les paires de classes les plus proches. Pour obtenir un tel arbre, il y a donc lieu d'étendre la notion de proximité entre deux éléments de L à celle entre deux parties disjointes C et D de L.

On part de l'indice de base

$$p(C,D) = \max \{P(c,d)/(c,d) \in C \times D\} \quad , (4)$$

où, rappelons-le,

$$P(c,d) = \phi\left[Q'(c,d)\right] \text{ pour tout } (c,d) \in C \times D.$$

L'indice final qu'on retiendra résulte de la distribution de la v.a. p(C',D') associée à (4), où C' et D' sont respectivement associés à C et D dans le cadre d'une hypothèse d'absence de lien tenant compte des cardinaux de C et de D. L'indice auquel on aboutit prend la forme suivante

$$P(C,D) = \left[p(C,D)\right]^{lm} \quad , (5)$$

où l = card(C) et m = card(D).

L'algorithme qui construit la représentation polonaise de l'arbre des classifications à partir de l'indice (5) a été appelé "Algorithme de la Vraisemblance du Lien" (A.V.L.). La première analyse de cet algorithme des points de vue informatique et comportement statistique, a été réalisée par Mme NICOLAÜ (cf.[29]).

Le passage de l'indice (4) à celui (5) a constitué un progrès décisif pour l'apparition des classes de faible cohésion qui n'en ont pas moins une certaine cohérence interne. Les éléments de telles classes se trouvaient avec l'indice (4), éparpillés en raison de l'attraction de quelques noyaux de forte cohésion. En effet, dans le cas de l'indice (5), une valeur donnée de la proximité p(C,D) (cf. formule (4)) "comptera" bien davantage dans (5) si elle concerne deux classes C et D de faible densité que si elle concerne deux classes C et D de forte densité.

VII-CONDENSATION DE L'ARBRE A SES NOEUDS SIGNIFICATIFS.

Une étape décisive de la méthode consiste à condenser l'arbre aux niveaux où se produit un noeud "significatif" détecté à partir du comportement d'une statistique de proximité entre une certaine forme de l'information quant aux ressemblances de l'ensemble L à classifier et l'association entre deux classes correspondante au noeud.

On se ramène à la comparaison de deux structures de même type en ne retenant que l'indice de proximité P sur L que le préordre total associé sur l'ensemble $F = P_2(L)$ des paires d'éléments distincts de L, appelée préordonnance sur L et notée $\omega(L)$:

$$(\forall (p,q) \in F \times F) \; ; \; p < q \; (\text{pour } \omega(L)) \Leftrightarrow P(p) < P(q) \; , \; (1).$$

$\omega(L)$ sera représenté par la partie $gr\left(\omega(L)\right)$ de F x F définie par

$$gr\left(\omega(L)\right) = \{(p,q) \in F \times F / p < q \text{ et non } q < p, \text{ pour } \omega(L)\} \; , \; (2)$$

Une même partition π, éventuellement produite à un niveau de l'arbre des classifications, est regardée comme définissant un préordre total sur F à deux classes $S(\pi)$ et $R(\pi)$ où $S(\pi)$ est l'ensemble des paires séparées et où $R(\pi)$ est celui des paires réunies par la partition π. $S(\pi) < R(\pi)$ pour l'ordre quotient : les composantes d'une paire séparée (resp. réunie) sont considérées éloignées (resp. proches) du point de vue de la partition.

L'indice "brut" entre la préordonnance $\omega(L)$ et la partition π sera dans ces conditions

$$\sigma(\pi,\omega) = card\left[gr(\omega) \cap (S(\pi) \times R(\pi))\right] \quad , \; (3)$$

J.P. BENZECRI a introduit ce cardinal sous la forme encore trop métrique du "nombre d'inégalités entre les distances spécifiées par la partition".

Soit ω un ordre total sur F. ω est choisi compatible avec la préordonnance $\omega(L)$ qui d'ailleurs se réduit souvent, quasiment à une ordonnance (ordre total sur F) compte tenu de la "finesse" de l'indice P choisi pour l'établir. Nous démontrons que la distribution de $\sigma(\pi',\omega)$, où π' est un élément aléatoire dans l'ensemble muni d'une probabilité uniformément répartie des partitions d'un même type, est dans des conditions assez générales, asymptotiquement normale, de moyenne r.s/2 et de variance, sensiblement r.s(f+1)/12, où $r = card\left[R(\pi')\right]$ et $s = card\left[S(\pi')\right]$ sont des invariants liés au type de la partition et où f = card(F) = r+s.

La statistique Σ obtenue en centrant et en réduisant (3) définit la "mesure" d'adéquation globale de la partition. La distribution des valeurs 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 noeud "significatif". En attachant à chaque niveau i, le taux d'accroissement $\theta_i = (\Sigma_i - \Sigma_{(i-1)})$; un tel noeud, qui correspond à l'achèvement à un certain degré d'une classe, apparaît comme un maximum local de la distribution deθ sur la suite des niveaux.

Ainsi Σ joue le rôle d'un critère de jugement permettant l'interprétation dynamique de l'arbre des classifications lequel est formé à partir du critère local d'agrégation P(C,D) (cf. (5) § VI). Ne peut-on pas inverser les rôles ?

A notre avis non ! En effet, le critère de formation des classes doit avoir un caractère local qui tienne intimement compte de la nature des données ; alors que le critère d'appréciation doit avoir un caractère global faisant intervenir toute l'information concernant les ressemblances. Bien que Σ ait d'intéressantes propriétés d'invariance résultant du degré de généralité où se situe sa définition, on aurait pu envisager un autre critère global que Σ ; par exemple l'inertie expliquée. D'ailleurs, dans la méthode des "pôles d'attraction" (cf. [23]) que nous esquisserons bientôt, on a mis en oeuvre deux critères de jugement ; le premier s'apparente à Σ et le second est celui de l'inertie expliquée. Mais ce dernier n'est pas toujours facile à concevoir et à mettre en oeuvre ; il suffit par exemple de songer un instant au cas de la classification d'une famille d'échelles discrètes.

Pour ce dernier cas, nous avons une méthode combinatoire et métrique (cf. [21] Chap.8) qui permet de dégager derrière une même classe d'échelles ayant une cohésion suffisante pour justifier l'existence d'une même variable unidimensionnelle sous jacente, l'"échelle hiérarchique optimale" ordonnant totalement l'ensemble des modalités n'occupant pas la position initiale des différents items de la classe (cf.[16] et Chap.8 de [21]).

VIII-DEGRE DE NEUTRALITE D'UN ELEMENT, SERIATION, CLASSIFICATION PAR LA METHODE DES POLES D'ATTRACTION.

Le point départ de cette partie de la recherche a été l'élaboration d'un indice statistique permettant avant toute classification de nettoyer les données des éléments trop "neutres" ou "mal typés" ; sinon, de "comprendre" une fois la classification obtenue, pourquoi tel élément peut mal cadrer dans la classe où il se retrouve. L'idée de base a été d'associer à chaque élément e de l'ensemble à classifier L la distribution des proximités ou des distances à e ; soit $\{Q(e,x)/x \in L - \{e\}\}$ où $\{D(e,x)/x \in L - \{e\}\}$, (1) ; où la distance D correspond à la métrique définie par Q.

L'analyse du phénomène à détecter nous conduit à proposer comme indice de neutralité d'un même élément e la variance des proximités ou bien des distances des différents objets de L à e ; soit

$$\mathcal{V}(e) = \frac{1}{(p-1)} \sum_{x \in L-\{e\}} \left[Q(e,x)-Q(e)\right]^2 \qquad (2)$$

où p = card (L) et où $Q(e) = \frac{1}{(p-1)} \sum_{x \in L-\{e\}} Q(e,x)$ est la moyenne des proximités à e.

Une formule analogue peut être proposée en remplaçant les proximités Q par les distances D.

L'expérience a permis d'établir que plus $\mathcal{V}(e)$ est grand plus e intervient intimement dans la formation de la classe où il apparaît (cf. [21] Chap.3).

D'où l'idée de chercher par une analyse de la variance des proximités ou des distances (de chacun des points de L à ses autres points), une suite de sommets de L ayant chacun un fort "pouvoir attractif" et relativement indépendants (resp. éloignés) les uns des autres. Cette suite de sommets est la suite $(p_1,p_2,\ldots,p_k,..)$ des "pôles d'attraction" qui est déterminée de proche en proche de façon récurrente (cf. [21] Chap. 7 et [23]).

La première stratégie qui nous est apparue pour déterminer cette suite de pôles d'attraction est la suivante :

p_1 maximise la quantité critère (2).

p_2 doit être choisi relativement indépendant de p_1 ($Q(p_1,p_2)$ faible) et ayant un fort degré dans la discrimination de ses proximités aux différents autres points de L ($\mathcal{V}(p_2)$ fort). p_2 est déterminé de façon à rendre maximum la quantité sous le signe [ ], de même dimension que (2)

$$\max_{e \notin \{p_1\}} \left[\mathcal{V}(e)/Q(e,p_1)\right]^2 \qquad , (3)$$

Le (k+1) ème pôle est obtenu à partir des k premiers par la règle max min :

$$\max\left[\min\left\{\left(\frac{\mathcal{V}(e)}{Q(e,p_1)}\right)^2, \left(\frac{\mathcal{V}(e)}{Q(e,p_2)}\right)^2, \ldots, \left(\frac{\mathcal{V}(e)}{Q(e,p_k)}\right)^2\right\}\right] \quad , \quad (4).$$

La détermination d'un couple de pôles d'attraction permet une représentation euclidienne dont les axes se trouvent respectivement entraînés par les deux pôles. Cette représentation que nous considérons le plus souvent seulement autour de (p_1,p_2), a surtout un intérêt local. L étant un ensemble A d'attributs de description ; dans le cas de l'existence d'une échelle d'attitude (problème de Psychologie) ou d'une "sériation" (problème d'Archéologie) ; cette dernière représentation (autour de (p_1,p_2) permet d'extraire la structure ordinale sous jacente. Dans la mesure où une telle structure est "statistiquement nette" nous démontrons qu'elle peut être dégagée par la seule détermination du premier pôle d'attraction (cf. [21] Chap.7).

La détermination d'une suite de pôles d'attraction au niveau de l'ensemble à classifier conduit à une riche famille d'algorithmes de classification (thèse de 3ème cycle de H.Leredde en cours de rédaction) qui est obtenue

(i) en variant le critère tel que (4) et ceci, en travaillant avec les distances au lieu de travailler avec les proximités et avec les moments absolus d'ordre 2 au lieu des variances. L'expression du critère dépend de la nature de l'ensemble à classifier.

(ii) en variant le critère d'affectation d'un même élément à l'une des classes en cours de formation autour des différents pôles d'attraction.

(iii) en attachant à une même classification en k classes (autour de k pôles), la valeur d'une statistique de signification qui peut être de même nature que celle du paragraphe VII ou bien, basée sur l'inertie expliquée. L'évolution de chacun de ces critères pour k variant permet de définir un test d'arrêt sinon, de retenir les quelques meilleures classifications.

Pour terminer, signalons que par rapport à une méthode de re-allocation telle que celle des "nuées dynamiques" ; en partant d'un système de "noyaux" qui soit formé de pôles d'attraction, l'algorithme converge en une seule étape (résultat expérimental toujours constaté) ; de sorte que les notions de forme "forte" et de forme "faible" attachées à cette méthode semblent davantage liées au caractère aléatoire du système initial de noyaux.

IX-AUTRES ASPECTS

Nous nous sommes également intéressés à des aspects statistiques fondamentaux tel que celui de la stabilité des résultats de la classification des variables (il s'agissait en l'occurrence d'attributs descriptifs) par rapport à l'accroissement de l'échantillon aléatoire des individus. Cette étude a été effectuée d'un point de vue expérimental dans le cadre d'un stage de D.E.A. à la faveur de l'analyse d'un important fichier de bilans de santé de la Sécurité Sociale. Il nous est apparu que la rapidité de convergence vers une stabilité parfaite des résultats de la classification dépend de deux facteurs : le premier est la fréquence des différents attributs et le second est l'aptitude à être classifié conformément aux proximités de l'ensemble des attributs.

Cette aptitude à être classifié d'un ensemble conformément aux ressemblances entre ses éléments est dans notre méthode caractérisée par une distribution et "mesurée" par un indice qui évalue la distorsion de la structure de la préordonnance associée aux ressemblances par rapport à un état ultramétrique de cette dernière (cf. [21] Chap. 3 et [20]). Il faut souligner que la notion de "Classificabilité" est relativement indépendante de celle de cohésion des classes formées par un bon algorithme de classification.

X-PROGRAMMATION ET DONNEES REELLES.

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. Relativement à la classification on distingue actuellement

a) une chaîne de programmes correspondante à la méthode de classification hiérarchique

b) un ensemble de trois programmes correspondants aux différentes stratégies de la méthode des "pôles d'attraction".

La chaîne de programmes (a) comprend sous forme modulaire quatre étapes qui sont enchaînées :

(1) PROX dont le rôle est d'établir le tableau des proximités entre caractères descriptifs où l'ensemble des modalités d'un même caractère est totalement ordonné.

(2) ORDON dont le rôle est

(α) d'établir le tableau des proximités lorsque V est formé de variables de la première catégorie (cf. § II ci-dessus) (i.e. attributs descriptifs ou bien variables numériques)

(β) d'établir et d'éditer le tableau des valeurs croissantes de l'indice de neutralité de chacun des éléments de l'ensemble à classifier (cf. formule (2) § VII).

(γ) d'établir l'ordonnance associée à l'indice de proximité (c'est sa fonction la plus importante).

(3) POLON dont le rôle est

(α) d'établir la représentation polonaise de l'arbre pour différents critères de formation ascendante des classes dont celui de "A.V.L." (cf. § VI) et celui plus classique de l'inertie expliquée en comprenant le cas du traitement des lignes (resp. colonnes) d'un tableau de contingence.

(β) de calculer pour chacun des niveaux les valeurs des statistiques globale et locale des niveaux (cf. § VII) et d'établir la représentation polonaise de l'arbre condensé des classifications.

(4) ARBRE dont le rôle est d'éditer directement l'arbre condensé. Mais sur cette représentation on retrouve toute l'information puisque chacun des noeuds de l'arbre détaillé se trouve représenté à la position adéquate de l'arbre réduit, accompagné du numéro du niveau de l'arbre total où il s'est formé ; d'autre part et c'est important, le noeud se trouve affecté du signe (*) s'il est significatif.

Les chercheurs qui ont travaillé à des degrés d'ailleurs très divers à cette chaîne de programmes sont : P.Achard (1968), I.C.Lerman (1970), N.Nicolaü(1971), Mme M.H.Nicolaü (1971), I.Cohen (1974), M.Morel (1976), C.Chauré (1977) et T.Chantrel (1978).

Pour une situation non encore intégrée à la chaîne, l'établissement du tableau des indices réduits doit se faire à partir d'un programme séparé et on entrera dans la chaîne à travers une option de l'étape ORDON. Nous prévoyons un important programme "TAUX" où on cherchera à prévoir la quasi totalité des situations possibles quant à la structure du tableau des données pour établir la table des valeurs de l'indice de proximité entre variables descriptives ou bien, entre objets. Un tel programme modifierait la structure de la chaîne.

H.Leredde est l'auteur exclusif de l'ensemble des trois programmes relatifs à la méthode des "pôles d'attraction" à laquelle il a contribué et dont il a analysé le comportement au niveau de nombreux et gros fichiers de données. L'ensemble de ces programmes s'adresse à des tableaux de données où les variables descriptives sont de la première catégorie (attributs ou bien variables numériques).

Nous avons exprimé au paragraphe VIII le type de critère utilisé pour extraire les pôles d'attraction dans le premier programme qui est davantage orienté vers la classification des paramètres descriptifs et vers la construction de représentations euclidiennes, généralement autour des deux premiers pôles pour, notamment, dégager des "sériations". L'affectation d'un élément à l'une des classes en cours de formation se fait ici selon le critère de la plus grande proximité entre un point extérieur à la réunion des classes déjà constituées et l'une des classes en cours de formation.

Le deuxième algorithme permet de déterminer la classification classe après classe ; une même étape de l'algorithme est définie par la constitution d'une classe qu'on entraîne autour d'un pôle d'attraction par l'affectation à ce dernier de tous les éléments dont la distance est inférieure à un certain seuil S qu'on détermine par un algorithme statistique simple. Ici nous travaillons avec la distance associée à la métrique que suppose notre indice de proximité et nous adoptons comme quantité critère le moment absolu d'ordre 2.

L'algorithme sous jacent au troisième programme est le plus consistant. Qu'il s'agisse de la classification des variables ou bien des objets ; on a ici un traitement unique, moyennant le remplacement des mesures brutes par celles "centrées réduites" par rapport à la distribution des différentes variables descriptives. La quantité critère utilisée pour la détermination des pôles d'attraction ainsi d'ailleurs que pour la répartition autour de ces derniers des différents éléments de l'ensemble L à classifier, est basée sur le moment d'inertie. Ici on tente de choisir les différents pôles d'attraction aussi "éloignés" que possible les uns des autres.

Comme nous le signalions ci-dessus (§ VII et VIII) on associe à la suite des classifications (premier et troisième programmes) la suite des valeurs de deux statistiques de signification, ponctuée par l'édition de deux histogrammes. L'examen de ces deux distributions permet de retenir les "meilleures" classifications.

Pour terminer signalons l'ensemble des programmes sur la classificabilité que nous avons nous-mêmes mis au point. Cet ensemble de programmes où on travaille au niveau de la préordonnance est séparé des autres programmes sur la classification (cf. § IX).

D'autre part, on dispose d'un ensemble cohérent de programmes de recherche de "l'échelle hiérarchique" sous jacente à une classe d'items dont chacun définit un caractère à l'ensemble totalement ordonné des modalités (programmation réalisée par C.Riso-Lévy) (cf. [21] Chap. 8).

Du point de vue des données réelles nous avons travaillé par rapport à de nombreuses disciplines des sciences humaines, économiques, biologiques et médicales : Sociologie, Psycho-Sociologie, Psycho-Pédagogie, Psychologie génétique, Pédagogie mathématique, Socio-Economie, Economie rurale, Sociologie médicale, Médecine, Préhistoire, Fiabilité des composants électroniques,... (cf. [3], [4], [5], [7], [14], [20], [24], [25], [26], [27], [29], [30], [33] et [34]).

BIBLIOGRAPHIE

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

[2] J.P. BENZECRI,
"L'Analyse des Données", "Dunod", Paris, 1974.

[3] A. BERGE, G. DENJEAN,
"Comportement Digestif et Fonctionnement Intellectuel",
"Revue de Neuropsychiatrie Infantile", 1974, 22(6), pp. 355-370.

[4] 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, Sept. 1977.

[5] I. COHEN,
"Classification d'une Famille d'Echelles 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 3ème cycle, Université de Paris VI (I.S.U.P.), Fév. (1977).

[6] H.E. DANIELS,
"The Relation between Measures of Correlation in the Universe of Sample Permutations", Biometrika, vol. 33, 1944.

[7] 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.

[8] L.A. GOODMAN, W.H. KRUSKAL,
"Measures of Association for Cross Classifications", J.A.S.A., 49, déc. 1954, pp. 732-764.

[9] L.A. GOODMAN, W.H. KRUSKAL,
"Measures of Association for Cross Classifications",
Approximate Sampling Theory, J.A.S.A., 58, June 1963, pp. 310-364.

[10] D.A.S. FRASER,
"Non Parametric Methods in Statistics", John Wiley, New York, 1967 (third edition).

[11] J.A. HARTIGAN,
"Clustering Algorithms", John Wiley, New York, 1975.

[12] N. JARDINE, R. SIBSON,
"Mathematical Taxonomy", John Wiley, New York, 1971.

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

[14] 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 3ème cycle soutenue le 21/9/78, Université de Rennes I.

[15] G. LECALVE,
"Problèmes d'analyse des données", 2ème partie d'une thèse d'état , Université de Rennes I, Nov. 1976.

[16] I.C. LERMAN,
"Analyse Hiérarchique", Revue Mathématiques et Sciences Humaines n°17, Paris 1967 ; repris et complété dans le chapitre 8 de [21]

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

[18] I.C. LERMAN,
"Analyse du phénomène de la "sériation" ", Revue Mathématiques et Sciences Humaines, n° 38, Paris 1972 ; repris et complété dans le chapitre 7 de [21].

[19] I.C. LERMAN,
"Etude 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.

[20] 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.

[21] I.C. LERMAN,
"Reconnaissance et Classification des structures finies en analyse des données", rapport I.R.I.S.A. n° 70, Université de Rennes I, 1976-77 (500 p.).

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

[23] I.C. LERMAN et H. LEREDDE,
"La méthode des pôles d'attraction" colloque international I.R.I.A., "Analyse des données en Informatique", Versailles, Sept. 1977.

[24] 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.

[25] 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. Janv. 1977.

[26] 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.

[27] J.R. MASSE,
"Adaptation de méthodes d'analyse des données à l'étude des mesures statiques sur circuits intégrés logiques", Université de Rennes I, Lab. de Statistique, Juin 1977. Ce travail a donné par ailleurs lieu à un rapport de recherche interne au C.N.E.T. (Dépt. "Fiabilité").

[28] M. MOTOO,
"On the Hoeffding's Combinatorial Central Limit Theorem", Ann. Inst. Stat. Math. 8, (1957), pp. 145-154.

[29] F. and M.H. NICOLAU,
"Analyse d'un Algorithme de Classification" et "Contributions au Traitement Automatique de Données", 2 Thèses de 3ème cycle, Université Paris VI, I.S.U.P., Nov. 1972.

[30] G. PIERAUT, LE BONNIEC and K. VAN METER,
"Etude Génétique de la Construction d'une Propriété Relationnelle : la Relation de Passage", Monographies Françaises de Psychologie, n° 35, C.N.R.S., Paris, 1976.

[31] S. REGNIER,
"Sur Quelques Aspects Mathématiques des Problèmes de la Classification Automatique", I.C.C. Bull, vol. 4, 1965.

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

[33] 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, Etudes et Travaux, 1976.

[34] B. TALLUR,
"Etude régionale de l'Agriculture Française", rapport I.R.I.S.A. n°103 Université de Rennes I, 1978.

[35] W.F. DE LA VEGA,
"Techniques de Classification Automatique Utilisant un Indice de Ressemblance", Revue Française de Sociologie, Paris, 1967.

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