Statistique et analyse des données

H. Al Nachawati

Segmentation multidimensionnelle

Statistique et analyse des données, tome 9, n° 2 (1984), p. 1-30.

<http://www.numdam.org/item?id=SAD_1984__9_2_1_0>

Statistiques et Analyse de Données
1984 - Vol. 9 n° 2 - pp.1-30

SEGMENTATION MULTIDIMENSIONNELLE

AL NACHAWATI H.

Equipe de Reconnaissance des Formes
et Microscopie Quantitative
Laboratoire TIM 3
C.E.R.M.O.
BP N°68
38402 St-Martin-d'Hères Cédex

Résumé : *La segmentation est définie depuis 1960 comme une méthode de partitionnement d'un ensemble E, par dichotomies successives, au moyen de variables explicatives et par référence à une variable "à expliquer". Après avoir résumé la démarche de plusieurs méthodes de segmentation un élargissement de la définition initiale est proposé. Cette définition permet d'associer une étape de fusion à chaque étape de partitionnement et d'utiliser plus d'une variable " à expliquer ".*

L'utilisation d'un processus d'interrogation séquentiel non arborescent permet d'améliorer la typologie des individus en regroupant les sous-ensembles qui ne sont pas significativement différents. Par ailleurs, plusieurs critères d'optimisation de l'étape de fusion sont présentés et font appel à la technique de l'analyse de la variance multidimensionnelle (ou analyse de la dispersion), pour comparer des partitions n'ayant pas le même nombre de classes.

La méthode est interactive et permet, par l'utilisation d'un terminal graphique, de contrôler l'évolution des résultats à travers une analyse en composantes principales.

Abstract : *Since 1960, the segmentation has been defined as a method of partitioning a set E, which is sucessively divided by dichotomy using explicative variables and referring to a variable "to be explained".*

The main features of many of these are reviemed, an extension of these approaches is proposed. The new definition allows to associate any partitioning step to a fusion steps by referring to more than one variable "to be explained".
The use of an interrogation sequential non arborescent process leads to improve the typology of individuals by regrouping the eventual non significant subsets. On the other hand, several optimisation criterion of the fusion steps are presented and allow to compare the partitions which have not the same number of classes, on the basis of the multivariate analysis theory.
When using a graphic terminal, the proposed approach is interactive and , the user can control evolution of the results by applying the principal components analysis.

Mots clés : *Segmentation, processus nonarborescent, analyse de la variance multidimensionnelle.*

I - INTRODUCTION

Soit E un ensemble d'individus ou d'objets. On dispose des variables $Q^1, \ldots, Q^m$ (caractères), divisées en deux groupes :

- $J_1 = \{ Q^1, \ldots, Q^s \}$ ensemble dit "à expliquer", les variables $Q^1, \ldots, Q^s$ prennent leurs valeurs dans l'ensemble des réels (variables quantitatives), ou dans un ensemble discret fini (variables qualitatives).
- $J_2 = \{ Q^{s+1}, \ldots, Q^m \}$ ensemble dit "explicatif", les variables (questions) Q^j ($j = s+1, \ldots, m$) sont qualitatives (1)
  Chaque variable de J_2 définit une partition $p^j = (q_1^j, \ldots, q_{\ell_j}^j)$ de E.

A toute partition $P = \{ E_1, \ldots, E_i, \ldots, E_r \}$ de l'ensemble E des individus en r classes, on associe les variables X_{ijv} égales à la valeur du caractère Q^j, $Q^j \in J_1$ pour l'individu v de l'ensemble E_i .

où :
$i = 1, \ldots, r$
$j = 1, \ldots, s$
$v = 1, \ldots, \text{card}(E_i) = n_i$

Remarque (1) : Si les variables explicatives de l'ensemble J_2 sont quantitatives, il suffit de les rendre qualitatives en utilisant l'un des algorithmes classiques de discrétisations comme par exemple l'algorithme de FISHER. W.D.[19].

Dans la pratique :

1) on veut construire une partition de l'ensemble E, en sous-ensembles $E_1, \ldots, E_i, \ldots, E_r$ à l'aide des questions $O^j \in J_2$ de telle façon que :

a) chaque ensemble E_i soit aussi "homogène" que possible vis-à-vis de l'ensemble des variables de J_1.

b) les ensembles E_i soient aussi "hétérogènes" que possible entre eux, vis-à-vis de l'ensemble des variables de J_1.

2) On souhaite de plus utiliser un nombre minimum de questions de J_2 en choisissant les plus discriminantes.

3) On veut pouvoir affecter les individus supplémentaires aux classes de cette partition, par exemple en utilisant une méthode de reconnaissance des variables.

On rencontre ce type de problèmes notamment dans les sciences humaines où le besoin, est grand de pouvoir mettre en relation des variables de comportement et des variables d'attitude, et trouver une partition de E expliquée par les variables privilégiées, ainsi que dans les domaines biologique et médical et économique, etc ..., MOUSTAFA M. et BRUGAL G. [29], GAUVAIN C. [21],OPFERMANN M. [30].

Exemple :

Les incidents banquiers :

Les deux groupes des caractères considérés sont :

- le groupe J_1 des incidents banquiers relatifs à n clients et concernant :
  . le nombre de chèques sans provision ;
  . les retards de paiement ;
  . le nombre d'erreurs ;
  . etc ...

- le groupe J_2 des caractéristiques des clients, concernant :

  . la catégorie socio-professionnelle ;
  . l'ancienneté ;
  . la région d'origine ;
  . etc ...

Le but sera ici :

1) de sélectionner si possible, les variables de J_2 (caractéristiques des clients) qui permettent de décrire le mieux possible le groupe J_1 (les incidents banquiers).

2) De faire une partition de E, qui fournit une typologie de la clientèle.

3) de pouvoir affecter un individu anonyme, par exemple un nouveau client, à l'une de ces parties, en utilisant les variables (questions) les plus discriminantes que l'on a choisies de l'ensemble J_2.

2 - QUELQUES METHODES DE SEGMENTATION

Le développement des méthodes de segmentation est relativement récent. On peut le dater, en gros, autour des années 1960.

La segmentation est, au départ, définie comme une méthode de partitionnement d'une population par dichotomies successives au moyen de variables explicatives et par référence à une variable "à expliquer". En 1976, cette définition a été élargie, pour obtenir des partitions à plus de deux classes et pouvoir utiliser plusieurs variables explicatives à chaque étape du processus de partitionnement. Plus récemment en 1978, de nouvelles méthodes ont été développées, elles utilisent non seulement des étapes de partitionnement mais aussi des étapes de fusion entre les ensembles des partitions précédemment obtenues. Rappelons brièvement ces méthodes.

2.1 - Méthodes utilisant un processus arborescent

Sélection typologique des variables DIDAY E. [18, 17, 16]
Il s'agit de détecter simultanément le (ou les) questions de l'ensemble J_2 les plus explicatives de l'ensemble des variables de J_1 et la partition de E en r classes (r ici est un nombre de classes donné à l'avance) qui optimisent un certain critère. Cette méthode contient plusieurs méthodes classiques comme par exemple :

- La méthode de ELISEE
(Exploitation des Liaisons et Interactions par Segmentation d'un Ensemble Expérimental) BOURROCHE J.M. et TENENHAUS M. [9,8]

Cette méthode s'applique dans le cas où la variable "à expliquer" est qualitative ; elle contient à son tour la méthode de BELSON W.A.[6]où la variable "à expliquer" n'a que deux modalités.

- La méthode de AID
(Automatic Interaction Detector) MORGAN J.N et SONQUIST J.A. [28]

Cette méthode s'applique dans le cas où la variable "à expliquer" est quantitative.

Pour ces deux dernières méthodes, nous renvoyons le lecteur à la thèse de BACCINI A. [5]

La méthode de WILLIAM W.J et LAMBERT J.M [37]
Les variables ou les questions de l'ensemble J_2 ont chacune deux modalités. On veut sélectionner la question la plus discriminante, c'est-à-dire, celle qui apporte le plus d'information sur les autres variables de J_2.
Cette méthode est en fait une segmentation sans variable "à expliquer" ou plutôt une segmentation où toutes les variables sont "à expliquer", (c'est-ce qu'on appelle une méthode monothétique sans variable "à expliquer").
Le principe consiste à sélectionner la variable qui fournit la meilleure dichotomie des objets à classer ; les critères à optimiser sont fondés sur des calculs de X^2 (WILLIAM et LAMBERT [37] ou sur l'information mutuelle (MOLLER F. [27] et CAPECCHI V. [11]).

2.2 - Méthodes utilisant un processus non arborescent

Il s'agit d'expliquer une variable à l'aide des questions de l'ensemble J_2 et de trouver une partition optimale en minimisant un certain critère.

La méthode de reconnaissance d'une variable continue TERRENOIRE M. et TOUNISSOUX D. [33]
On a vu que la méthode de AID cherche la dichotomie la plus "liée" à la variable "à expliquer". Le contrôle de la qualité de l'estimation des valeurs moyennes de la variable "à expliquer" sur les parties de E, conduit à renoncer à la structure arborescente et à construire un processus de partitionnement de E non arborescent qui améliore les estimations.
TERRENOIRE et TOUNISSOUX ont introduit une étape de fusion entre les étapes de partitionnement, arrivant ainsi à une partition optimale par rapport à un certain critère.
Ils ont proposé d'utiliser le test du FISCHER-SNEDECOR pour fusionner les parties qui ne sont pas significativement différentes.

La méthode de reconnaissance d'une variable discrète ROUTIER J.L. [32] *TOUNISSOUX D.* [33]
Pour la même raison que précédemment, ROUTIER J.L. a modifié la méthode d'ELISEE, qui cherche, comme on l'a vu les classes d'individus les plus "explicatives" des modalités d'une variable qualitative particulière "à expliquer" et a utilisé un processus non arborescent pour arriver à une décision optimale.
Ils ont proposé d'utiliser le test du X^2 comme critère de fusion.

3 - LA METHODE DE SEGMENTATION MULTIDIMENSIONNELLE

Cette méthode a déjà été présentée dans ALNACHAWATI H. [1, 2, 3, 4].
Elle va permettre de traiter les deux problèmes suivants :

3.1 - Reconnaissance de plusieurs variables "à expliquer"

On veut déterminer les variables de J_2 qui expliquent le mieux, pour un certain critère, les variables de l'ensemble J_1 (card $(J_1) \geq 1$) à l'aide d'une partition de l'ensemble E dont le nombre de classes n'est pas fixé. Le cas d'une variable "à expliquer" qu'on a rappelé précédement, est un cas particulier de cette méthode.

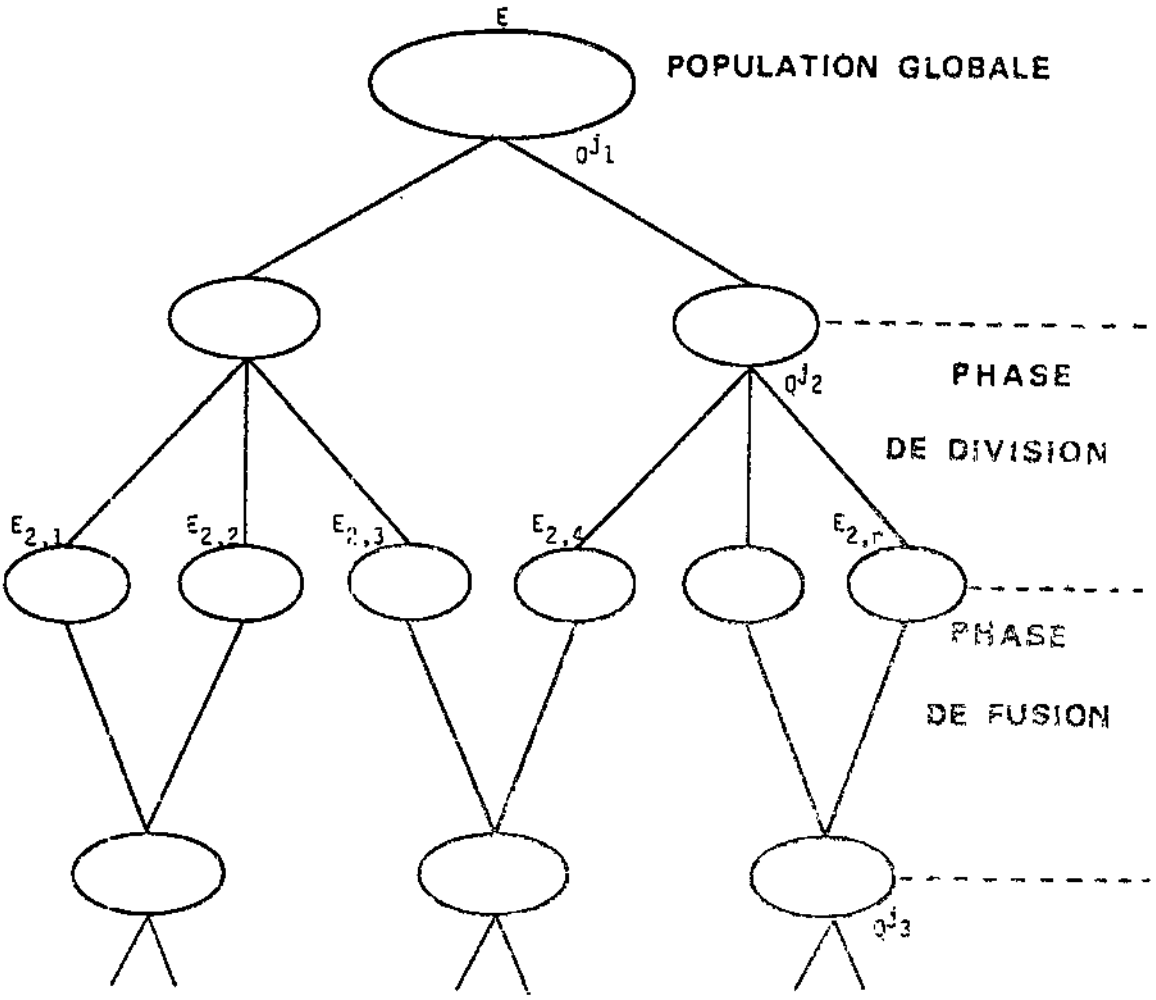
3.2 - Segmentation typologique d'un ensemble d'individus

L'ensemble des variables $Q^1, \ldots, Q^m$ $(m \geq 1)$ n'est pas ici partagé en deux sous-ensembles J_1 et J_2. On veut plus simplement construire une partition de E, en utilisant un nombre minimum de questions de J (où $J = J_1 \cup J_2$) et en optimisant un certain critère.

On va utiliser pour cela un processus d'interrogation séquentielle non arborescent, (voir graphique (1)), qui consiste à engendrer une suite de partitions $P_{k,\alpha_k} = \{E_{k,1}, \ldots, E_{k,\alpha_k}\}$ de E, chaque partition étant déduite de la partition précédente en deux phases :

- l'une "descendante" pour sélectionner les questions discriminantes de l'ensemble J_2 qui donnent pour chaque partie ou partition, une partition optimale pour un certain critère sur l'ensemble J_1. Cela se fait pour une étape k par le croisement entre $P_{k-1,\alpha_{k-1}}$ et la partition $P^{j_k} = \{q^{j_k}_1, \ldots, q^{j_k}_{\varsigma_{j_k}}\}$ déduite de la question Q^{j_k}, où $(\varsigma_{j_k} \in N)$.

- l'autre "ascendante" pour fusionner les parties qui ne sont pas significativement différentes ou qui ne caractérisent pas une partie (par exemple si card $(E_{k,i}) < n_0$) au sens statistique du terme.



Graphique 1 : Schéma graphique de la méthode

- on utilise la question $Q^{j}1$ pour engendrer la partition $E_{1,1}$, $E_{1,2}$ de E
- .... $Q^{j}2$ ...
- ....
- on utilise $Q^{j}K$ pour engendrer la partition $E_{k,1}$,..., E_{k,α_k} de E .

Ainsi, division et fusion sont réalisées alternativement, pour sélectionner les questions discriminantes de l'ensemble de J_2 et arriver à une partition optimale de l'ensemble E, vis-à-vis de l'ensemble des variables de J_1, dont le nombre de parties est ni trop petit ni trop grand, et bien significatif au sens statistique de ce terme.
On va proposer deux ensembles de critères, le premier pour le cas où les variables de J_1 sont quantitatives (table des mesures), le deuxième pour le cas où elles sont qualitatives (table de questionnaires)

4 - A PROPOS DES CRITERES "A OPTIMISER" EN SEGMENTATION

Les critères dans les méthodes de segmentation doivent prendre en compte les idées suivantes :

- comment mesurer les proximités entre un groupe de variables de J_1 (quantitatives ou qualitatives) et une variable qualitative, c'est-à-dire une partition de E ?

- comment comparer ces mesures si toutes les partitions ont le même nombre de classes ou si ce nombre varie d'une partition à l'autre ?

4.1 - La proximité entre des variables quantitatives et une variable qualitative

Le cas où les partitions ont le même nombre de classes
Les critères qu'on va proposer seront utilisés dans les phases descendantes, où on compare des partitions ayant le même nombre de classes.
Supposons que, à l'étape k, on dispose des partitions $(P^j_{k,r}\ ;\ j \in L_k)$, L_k est un ensemble fini d'entiers, r est le nombre de classes et on veut comparer ces partitions au sens d'un certain critère calculé sur l'ensemble de variables de J_1 pour choisir la meilleure de ces partitions.
Les critères à optimiser découlent de la relation de HUYGENS. La dispersion totale T de la population E est égale à la somme de la dispersion intra-

classe W et de la dispersion inter-classe B

$$T = W + B$$

avec $W = \sum_{i=1}^{r} W_i$ où W_i est la dispersion de la classe $E_{k,i}$

On a donc :

$$T_{lh} = \sum_{i=1}^{r} \sum_{\nu=1}^{n_i} (X_{il\nu} - X_{.l.})(X_{ih\nu} - X_{.h.})$$

$$W_{lh} = \sum_{i=1}^{r} \sum_{\nu=1}^{n_i} (X_{il\nu} - X_{il.})(X_{ih\nu} - X_{ih.})$$

$$B_{lh} = \sum_{i=r}^{r} n_i (X_{il.} - X_{.l.})(X_{ih.} - X_{.h.})$$

FRIEDMAN H.P., RUBIN J. [20] ont proposé et discuté pour leur méthode les cinq critères suivant, qu'on peut utiliser :

Minimisation de la trace de la matrice W :

$$\Delta P_{j_k} = \min_{j \in L_K} \{\text{trace } (W) = \sum_{il\nu} (X_{il\nu} - X_{il.})^2\}$$

où L_K est un ensemble fini d'entiers

Minimisation du coefficient de WILKS :

$$\Delta P_{j_k} = \min_{j \in L_k} \left(\frac{|W|}{|T|}\right) \text{où } \frac{|W|}{|T|} = \prod_{l=1}^{p} \nu_l \quad \text{et} \quad \left|I + W^{-1}B\right| = \left|\frac{T}{W}\right|$$

ν_l est valeur propre de $T^{-1}W$

Maximisation de la trace $\sum_{l=1}^{p} \nu_l$ de $W^{-1}B$.

Maximisation de la plus grande valeur propre ν_1 de $W^{-1}B$

Minimisation de $|W| = \prod_{i=1}^{R} |W_i|$ (Scott et Symons, 1971)

Le choix de l'un de ces critères dépend du temps de calcul, de la standardisation des variables et de la connaissance que l'on a a priori de la structure des groupes $E_{k,i}$

L'utilisation de la trace de (W) implique que l'on cherche des groupes sphériques avec standardisation préalable des variables.
Les critères $|\frac{W}{T}|$, trace $(W^{-1}B)$ sont invariants par les transformations linéaires des données (CHANDON et PINSON [12] donnent plus de détails sur ces critères).

Remarque :
Si les variables de l'ensemble de J_1 sont corrélées et si on connaît la matrice T^{-1} où T est la matrice de covariance empirique de ces variables, la norme $||X|| = X' T^{-1}X$ se ramène au critère trace (W) ci-dessus.

Cas où les partitions n'ont pas toutes le même nombre de classes.
Les critères qu'on va proposer peuvent s'utiliser soit dans un processus arborescent, soit dans les phases descendantes d'un processus non-arborescent dans les cas où les partitions produites ont des nombres de classes différents.
Ces critères mesurent aussi l'homogénéité à l'intérieur des classes et l'hétérogénéité entre les classes.
Supposons qu'on ait, à étape K, des partitions possibles P_{k,α_1}, P_{k,α_2}, ..., $P_{K,r}$, ..., P_{k,α_K} et qu'on veut comparer ces partitions au sers d'un certain critère calculé sur l'ensemble de variables de J_1 pour chcisir la meilleure de ces partitions.
A toute partition $P_{k,r}$, on associe les variables $X_{ij\nu}$ comme on l'a fait précédement et on pose :

$$Y_{i\nu} = \begin{bmatrix} X_{i1\nu} \\ \vdots \\ X_{is\nu} \end{bmatrix} = \mu_i + \varepsilon_{i\nu} \qquad \begin{matrix} i \in \{1,\ldots,r\} \\ \nu \in \{1,\ldots,n_i\} \end{matrix}$$

où les vecteurs aléatoires $\varepsilon_{i\nu}$ sont supposés indépendants, gaussiens centrés et, pour tout ν, de matrice de variance $\mathrm{Var}(\varepsilon_{i\nu}) = \Sigma_i$

On va utiliser diverses statistiques de test de l'hypothèse
$H_o : \mu_1 = \mu_2 = \ldots = \mu_r$ contre sa négation.

(avec l'hypothèse supplémentaire $\Sigma_1 = \Sigma_2 = \ldots \Sigma_r = \Sigma$)

A chacune de ces statistiques de test $\tilde{R}(r)$ de loi probabilité $Q_{(r)}$, on associe le nombre $\tilde{\pi}_r = Q_r$ ([0, R(r)]) où R (r) est la valeur observée de $\tilde{R}(r)$ sur l'échantillon (à r classes).

Nous suggérons les statistiques de test suivantes :

Test de WILKS (DAGNELIE [14])
Le test de WILKS utilise la statistique $|\frac{W}{T}|$ où W et B sont indépendantes et suivent des lois de WISHART. On peut utiliser l'approximation de RAO : la loi de $|\frac{W}{T}|$ peut être assimilée à la loi :

$$\beta_2 \left(s(r-1),\ m\xi + 1 - s(r-1)/2\right) \text{ avec}$$

$$m = N-1-(s+r)/2 \qquad \xi^2 = \frac{s^2(r-1)^2-4}{s^2+(r-1)^2-5}$$

on pose $\tilde{\pi}_r = \beta_2 \left(s(r-1),\ m\xi+1-s(r-1)/2\right) [0, R(r)]$

où

$$\hat{R}(r) = \frac{\left|\frac{W}{T}\right|^{1/\xi}}{1-\left|\frac{W}{T}\right|^{1/\xi}} \times \frac{s(r-1)}{m\xi + 1-s(r-1)/2}$$

et on prend la partition $\tilde{P}_{k,t} = \{E_{k,1}, \ldots, E_{k,t}\}$ qui vérifie

$$\tilde{\pi}_t = \text{Max}\{ \tilde{\pi}_r,\ r = a_1, \ldots, a_\varsigma \}$$

Test de ROY : (J. POGET [16] , DAGNELIE [14])
Ce test utilise la plus grande valeur propre ν_1 du produit BW^{-1} ; alors :

$$\hat{R}(r) = \frac{\nu_1}{1+\nu_1}$$

On pose $\tilde{\pi}_r$ = ROY (min (r-1, s), $\frac{|r-1-s|-1}{2}$, $\frac{N-r-s-1}{2}$) [0, R(r)]

et on retient la partition $\tilde{P}_{r,t}$ telle que :

$$\tilde{\pi}_t = \text{Max } (\tilde{\pi}_r \; ; \; r = \alpha_1, \ldots, \alpha_k).$$

Test de LAWLEY et HOTELLING (DAGNELIE, [14])
Ce test utilise la somme des valeurs propres de BW^{-1} c'est-à-dire la statistique :

$$\hat{R}(r) = \text{trace } (BW^{-1}) = \sum_{i=1}^{m} \nu_i$$

où m = min (r-1, s).

La loi de $\hat{R}(r)$ est une loi de Hotelling généralisée (DAGNELIE. [14] p.72.)
On procède comme précédemment pour le choix de $\tilde{\pi}_t$

Choix de $\tilde{\pi}_r$ utilisant un modèle d'analyse de la variance à deux facteurs (AL NACHAWATI [1])
A toute partition $P_{k,r}$ on associe les variables $X_{ij\nu}$ définies précédemment et on suppose ici que :

$$X_{ij\nu} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij\nu}$$

avec $\sum_i \alpha_i = 0$; $\sum_j \beta_j = 0$ et $\sum_{i,j} \gamma_{ij} = 0$ et où les $\varepsilon_{ij\nu}$ sont des variables aléatoires supposées gaussiennes indépendantes de même variance σ^2. Soit $\hat{R}(r)$ la statistique de test usuelle pour le test de l'hypothèse $H_o : \alpha_1 = \ldots = \alpha_r = 0$ contre sa négation définie par :

$$\hat{R}(r) = \frac{i\Sigma_{j\nu}(X_{ij\nu} - X_{ij.})^2}{s \sum_{i=1}^{r} n_i \; (X_{i..} - X...)^2}$$

On sait que $\hat{R}(r)$ est une statistique de loi de probabilité.

$$\beta_2 \ (r-1,\ s \sum_{i=1}^{r} n_i - rs)$$

On pose $\tilde{\Lambda}_r = \beta_2(r-1,\ s \sum_{i=1}^{r} n_i - rs)\ [0,\ R(r)]$

et on retient la position $\gamma_{k,t}$ telle que :

$$\tilde{\Lambda}_t = \text{Max}\ (\tilde{\Lambda}_r,\ r = \alpha_1, \ldots, \alpha_k)$$

4.2. Proximité entre des variables qualitatives et une variable qualitative

Cas où les partitions ont le même nombre de classes

Les critères qu'on va proposer utilisent soit le calcul du χ^2 soit la théorie de l'information pour choisir la meilleure partition de $(P^j_{k,r}, j \in L_k)$ où L_k est ensemble fini d'entiers, r est le nombre de classes.

Critère de maximisation du χ^2

On pose :

$$\Delta P_{j_k} = \underset{j \in L_k}{\text{Max}} \{\Sigma \chi^2_{jl} / l=1, \ldots, s\}$$

$$\Delta P_{j_k} = \underset{j \in L_k}{\text{Max}} \left(\frac{1}{n..} \sum_{l=1}^{s} \sum_{t=1}^{r} \sum_{i=1}^{m_l} \frac{n^l_{..}}{n^l_{i.}\ n^l_{.t}} \left(n^l_{it} - \frac{n^l_{i.}\ n^l_{.t}}{n^l_{..}}\right)^2\right)$$

$n_{..} \Delta P_{j_k}$ est un χ^2 d'ordre $(r-1)\ (\sum_{l=1}^{s} m_l - s)$

où L_k est un ensemble fini d'entiers.

n^l_{it} est la fréquence d'association entre deux modalités

$$n^l_{i.} = \sum_t n^l_{it} \quad ; \quad n^l_{..} = \sum_{i,t} n^l_{it}\ , \ n.. = n^l_{..}$$

m_l est le nombre de modalités du paramètre l

Critères utilisant la théorie de l'information

On pose :

$$\Delta P j_k = \underset{j \in L_k}{\text{Max}} \left\{ \sum_{l=1}^{s} h_{jl} = \sum_{l=1}^{s} \sum_{t=1}^{r} \sum_{l=1}^{m_l} f^l_{it} \lg_2 \frac{f^l_{it}}{f^l_{i.} f^l_{.t}} \right\}$$

où

$$f^l_{it} = \frac{n^l_{it}}{n} \quad ; \quad n = \sum_{i,t} n^l_{it}$$

h_{jl} est l'information mutuelle entre j et l

Référence : LANCE et WILLIAM [23]

Critère de MOLLER [25]

On pose :

$$\Delta P_{j_k} = \underset{j \in L_k}{\text{Max}} \left\{ I(j) = \frac{1}{S} \left[\left(\sum_{l=1}^{s} \frac{h_{jl}}{H_l} \right) \left(\sum_{l=1}^{s} \frac{h_{jl}}{Hj} \right) \right]^{1/2} \right\}$$

où

h_{jl} est l'information mutuelle entre j et l

H_j est l'entropie de j

L'entropie définie par DAROCZY [15]

On pose :

$$P_{j_k} = \underset{j \in L_k}{\text{Min}} \left\{ \sum_{l=1}^{s} \left[\sum_{t=1}^{r} \left(\frac{n^l_{.t}}{n..} \right)^{\beta} \left[\left(\sum_{i=1}^{m_l} \frac{n^l_{it}}{n^l_{.t}} \right)^{\beta} - 1 \right] \right] \right\}$$

où $\beta \in]0,1[$

TERRENOIREet TOUNISSOUX [34] donnent un exemple d'utilisation de ce critère pour une variable "à expliquer" qualitative.

Cas où les partitions n'ont pas toutes le même nombre de classes :
Les critères qu'on va proposer peuvent s'utiliser soit dans un processus arborescent, soit dans les phases descendantes d'un processus non arborescent dans le cas où les partitions utilisées ont des nombres de classes différents.
On utilise pour celà l'un des critères proposés par la statistique inférentielle, par exemple celui de la probabilité du χ^2 à $(r-1)(\sum_{l=1}^{s} m_l - s)$ degrés de liberté, d'obtenir un χ^2 supérieur au χ^2 calculé sous l'hypothèse " d'indépendance" des variables.

Transformation des variables qualitatives de J_l en des variables quantitatives
Pour faire cette transformation, on applique l'analyse des correspondances multiples qui définit un codage quantitatif et on applique les critères et les tests paramétriques proposés précédemment aux variables quantitatives obtenues.
Remarque :
On peut faire la transformation en sens contraire et transformer les variables quantitatives de J_l en variables qualitatives par les méthodes de discrétisation classiques et appliquer les critères proposés ci-dessus.

5. LES ALGORITHMES

Nous allons décrire successivement deux algorithmes de segmentation. Le premier est arborescent et n'utilise que des étapes de partitionnement de E avec des partitions de plus en plus fines. Le second est non arborescent et intercale une étape de fusion de sous-ensembles de E entre deux étapes de partitionnement.

5.1 - Algorithme "descendant"

On fait des partitions successives. On répartit l'ensemble E en classes de plus en plus fines. A chaque étape, on sélectionne la question Q^j_k qui

explique le mieux l'ensemble de variables de J_1.

Ainsi, à l'étape k, on dispose d'une partition $P_{k-1,\alpha_{k-1}} = \{E_{k-1,1}, \ldots, E_{k-1,\alpha_{k-1}}\}$ en α_{k-1} sous-ensembles non vides, construite aux étapes précédentes à l'aide des questions Q^{j_1}, ..., $Q^{j_{k-1}}$ extraites de l'ensemble J_2. On construit une partition $P_{k,\alpha_k} = \{E_{k,1}, \ldots, E_{k,\alpha_k}\}$ de E à l'aide d'une nouvelle question Q^j

Par croisement de $P_{k-1,\alpha_{k-1}}$ et de $P^{j_k} = \{q_1^{j_k}, \ldots, q_{\varsigma_{j_k}}^{j_k}\}$ la partition déduite de la question Q^{j_k} où $j_k \in \{s+1, \ldots, m\} - \{j_1, \ldots, j_{k-1}\}$

choisie de façon à optimiser un certain critère, tel que celui qu'on a proposé précédemment.

5.2 - Algorithme non arborescent

A chaque étape de l'algorithme descendant, on fusionne les sous-ensembles de la partition P_{k,α_k} que l'on considère comme homogènes ou qui ne sont pas caractéristiques au sens statistique du terme, par exemple si card $(E_{k,i}) \leq n_o$. On peut utiliser les critères de fusion suivants :

Critère utilisant un test statistique

Pour tout entier r, $2 \leq r \leq \alpha_k - 1$, on construit toutes les partitions en r sous-ensembles possibles déduites de P_{k,α_k} par fusion de parties et on choisit, pour chaque r, une partition $\hat{P}_{k,r}$ en r classes qui optimise l'un des critères proposés (§ 4.1 ou 4.2): à toute partition $\hat{P}_{k,r}$ on associe un nombre $\tilde{\pi}_r$, qui est un indice d'homogénéité des parties de $\hat{P}_{k,r}$ et d'hétérogénéité entre les parties et on choisit la meilleure partition pour ce critère.

Critère utilisant les distances entre les vecteurs moyens.

On fusionne les deux parties $E_{k,h}$, $E_{k,u}$, si la distance dans $\mathbb{R}^s$ entre les

vecteurs $u_h = (X_{h1}, \ldots, X_{hs})$ et $u_u = (X_{u1}, \ldots, X_{us})$ est inférieure à une seuil fixé a priori pour le critère utilisé.

Fusion des parties trop petites
Pour ne pas avoir de parties trop petites avec un effectif inférieur à un nombre n_0 fixé à l'avance, on fusionne une telle partie avec la partie la plus proche pour le critère utilisé.

6 - CRITERE D'ARRET

6.1 - Dans le cas où on utilise un test statistique :

Dans un processus non arborescent, l'algorithme s'arrête de lui-même quand à l'issue de l'étape K+1, la partition obtenue après la fusion est identique à la partition de l'étape précédente.

Dans un processus arborescent, on donne un seuil d'arrêt sur la valeur du critère à optimiser ou sur le nombre des étapes.

6.2 - Dans les autres cas, on peut se donner un seuil "a priori" sur la valeur du critère à optimiser ou sur le nombre des étapes, avec toutes les difficultés que posent toujours le choix d'un tel seuil selon la nature des données étudiées.

7 - INTRODUCTION D'INDIVIDUS SUPPLEMENTAIRES

L'application de la méthode précédente sur E fournit un ensemble J_2' des questions discriminantes. Le problème de reconnaissance consiste pour tout individu i (on connait les valeurs de i dans l'ensemble J_2') à affecter cet individu à une classe et à donner les valeurs en i des variables appartenant à J_1 selon le critère utilisé.
Pour répondre à ce problème, il suffit de suivre la procédure précédente,

en affectant l'individu à la classe d'identification de la partition de E, et on peut estimer les valeurs en i de la variable $Q^j \in J_1$, par X_{ij}. (si l'on utilise la trace de W comme critère).

Remarques :

Pour optimiser le temps de calcul quand le nombre de variables à expliquer J_1 est important, on peut déterminer au préalable les P(en choisissant P au mieux) premières composantes principales de ces variables et appliquer les critères à la base associée aux P axes principaux ainsi construits.

- au lieu de poser la même question pour toutes les partitions de E pour une étape donnée, on pourrait poser une question différente pour chaque partie de E.

- on peut faire la même étude pour r facteurs.

- on peut appliquer cette méthode sur les données d'un tableau de contingence .

- on pourrait étudier l'influence du facteur question et fusionner les questions qui ne sont pas significativement différentes par la même technique.

-Si l'ensemble des variables de J_2 est quantitatif, au lieu de transformer J_{2_j} en des variables qualitative, c'est-à-dire en des partitions $p^j_1, q^j_1, \ldots, q^{J_j}_{\zeta_j}$, on pourrait fabriquer, toujours à l'aide de $Q^j \in J_2$ une partition P_{k,α_k} en partitionnant les sous-ensembles de la partition $P_{k-1,\alpha_{k-1}}$ par l'utilisation d'une des méthodes de discrimination.

8 - LE PROGRAMME

Le système interactif de la méthode de segmentation multidimensionnelle, permet l'utilisation d'un terminal graphique grâce auquel l'utilisateur peut se rendre compte de l'évolution des résultats (en appliquant ACP) au cours des phases.

9 - EXEMPLE

Nous allons donner une idée sur les avantages du processus non arborescent en segmentation en traitant un exemple élémentaire.

Supposons qu'on ait deux groupes de variables $J_1 = \{Q^1, Q^2\}$ et $J_2 = \{Q^3, Q^4\}$ observés sur 25 individus.

Q^1	Q^2	Q^3	Q^4
4	9	1	1
6	10	1	2
5	9	1	1
10	4	2	1
13	4	2	1
16	5	2	1
19	6	2	1
13	5	2	1
14	7	2	1
17	5	2	1
19	6	2	1
11	7	2	1
13	14	2	2
11	13	2	2
17	15	2	2
13	14	2	2
16	14	2	2
16	12	2	2
6	10	1	2
4	10	1	2
4	7	1	1
5	8	1	1
4	9	1	1
3	8	1	1
6	7	1	1

Dans un premier essai, on a utilisé un processus arborescent (des partitions successives) en optimisant le critère trace (W) sur J_1 , pour sélectionner les questions discriminantes de l'ensemble de J_2. Les résultats sont schématisés dans les figures 1 et 2 ; la première figure (1) présente la première étape, c'est-à-dire, la première partition et pour laquelle Q^3 est la

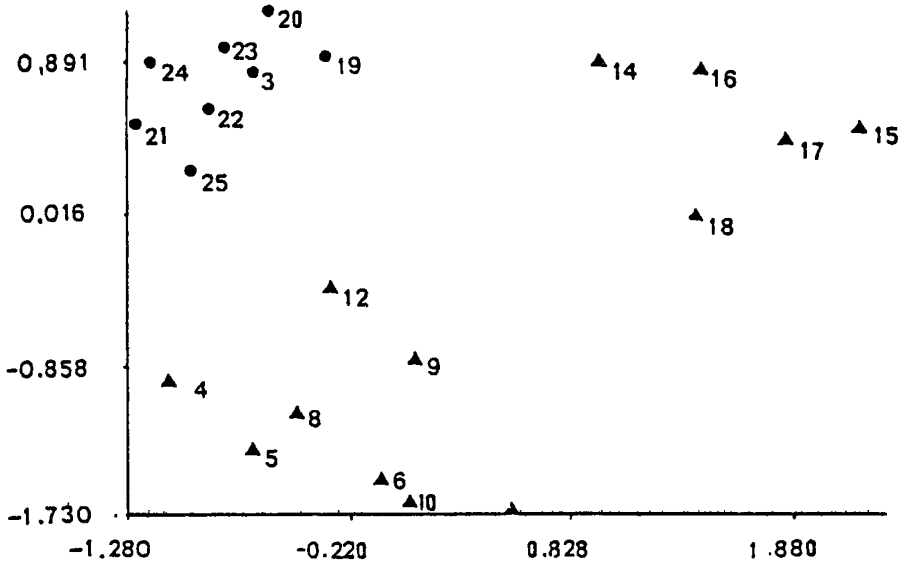


Figure 1 : Le plan des axes 1-2 pour la première étape.

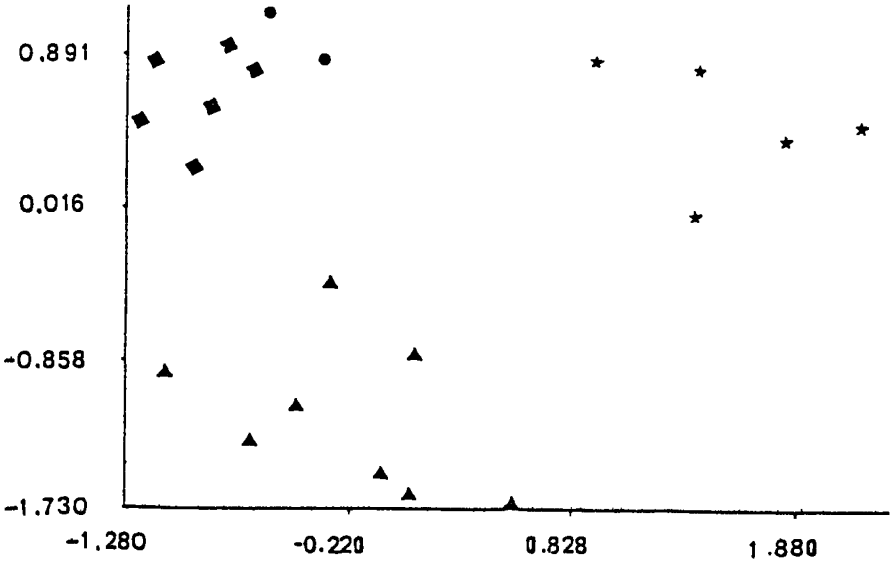


Figure 2 : Le plan des axes 1-2 pour la deuxième étape.

meilleure question discriminante, de l'ensemble J_2 ; la figure (2) présente une partition en 4 classes obtenue à la deuxième étape à l'aide de la question Q^4.

On peut résumer les résultats dans le graphique 2 suivant :

N° Étape	trace W	$\tilde{\pi}$	Résultat du processus
0	9.750		E, card(E)=25, Q^3
1	3.948	4.98E-17	$E_{1,1}$: 10, Q^4 ; $E_{1,2}$: 15, Q^4
2	1.421	2.04E-25	$E_{2,1}$: 3 ; $E_{2,2}$: 7 ; $E_{2,3}$: 6 ; $E_{2,4}$: 9

Graphique 2 : graphique représentant le déroulement du processus arborescent en utilisant les 2 questions Q^3, Q^4.

Dans un deuxième essai, on a utilisé un processus non arborescent qui fusionne les classes qui ne sont pas significativement différentes, en optimisant le critère du test de WILKS (approximation de RAO). Les résultats sont schématisés dans les figures 1 et 3.

A la première étape, on a obtenu les mêmes classes $E_{1,1}$, $E_{1,2}$ que précédemment (figure 1). A la deuxième étape, la phase descendante redonne les mêmes ensembles $E_{2,1}$, $E_{2,2}$, $E_{2,3}$, $E_{3,4}$ que précédemment, mais la fusion regroupe

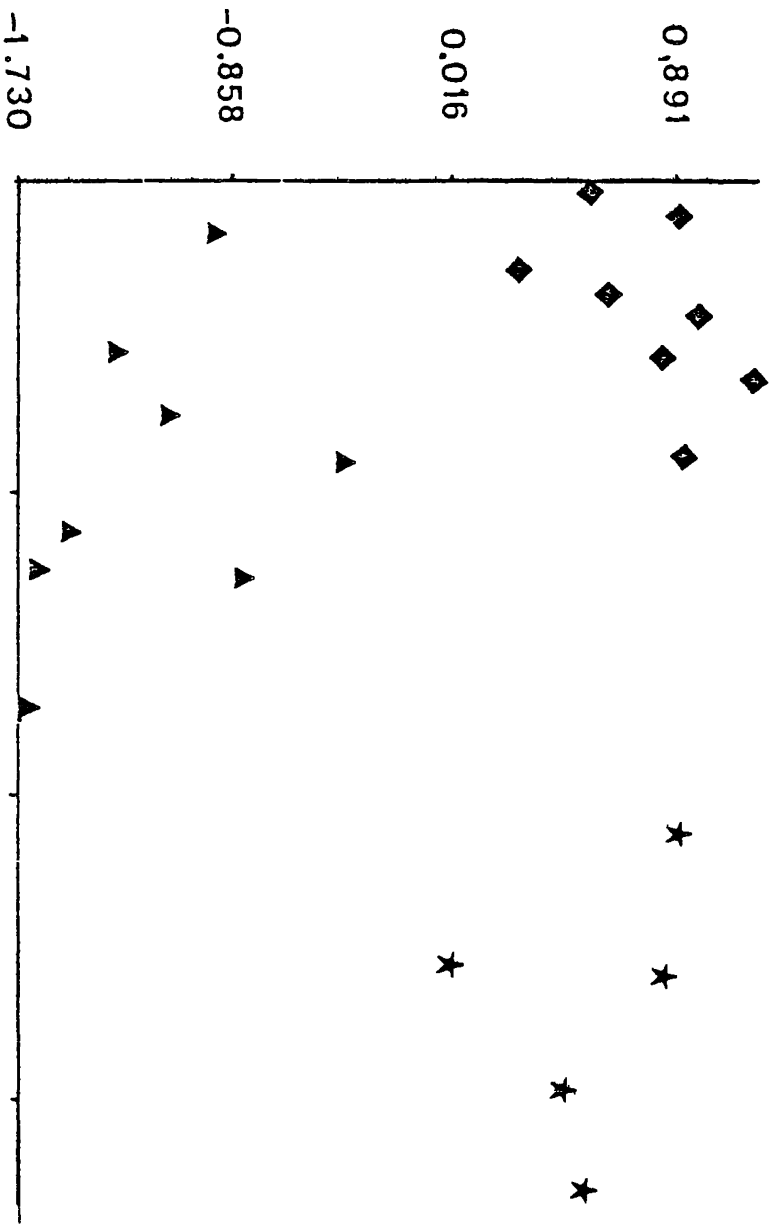
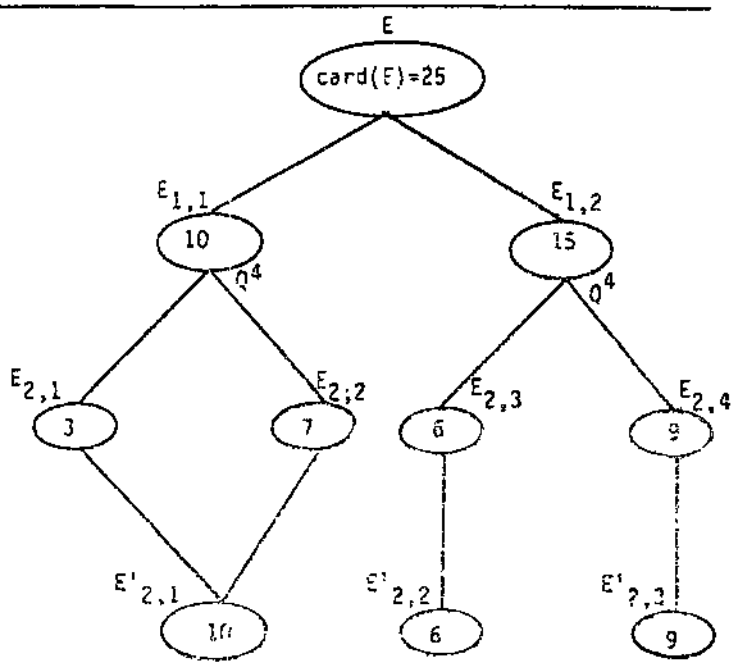


Figure 3 : Le plan des axes 1-2 pour la deuxième étape.

$E_{2,1}$ et $E_{2,2}$. Le graphique ci-dessous résume ces résultats.

N° Et.	Trace W	$\tilde{\tau}$	Résultat du processus
0	9.750		
1	3.948	4.98E-17	
2a	1.421	2.04E-25	
2b	1.510	1.58E-25	

Graphique 3 : représentation du déroulement du processus non arborescent en utilisant les questions Q^3, Q^4.

Le test de WILKS* qui mesure la signification d'hétérogénéité entre les classes et l'homogénéité à l'intérieur des classes conduit au regroupement des classes $E_{2,1}$ et $E_{2,2}$. Ce regroupement étant celui qui minimise le critère parmi tous les regroupements possibles.

* On teste si les moyennes d'au moins deux groupes parmi les groupes sont différentes de façon significative.

10 - CONCLUSION

Après avoir rappelé l'essentiel de la démarche de plusieurs méthodes de segmentation et avoir donné un élargissement de la définition de la segmentation pour traiter plus d'une variable "à expliquer", nous avons montré que l'utilisation d'une fusion de sous-ensembles obtenus après une étape de partitionnement dans l'algorithme, permet d'améliorer la typologie des individus en regroupant les sous-ensembles non significativement différents. Par ailleurs, plusieurs critères d'optimisation de la fusion ont été présentés et illustrés. L'utilisation des méthodes et des critères issus de l'analyse de la variance multidimensionnelle permet d'appliquer cet algorithme non arborescent de partitionnement pour comparer des partitions n'ayant pas le même nombre de classes.

11 - REMERCIEMENTS

Je tiens à exprimer mes plus vifs remerciements à M. Bernard VAN CUTSEM pour le soin avec lequel il a examiné ce travail et son aide précieuse ; ainsi qu'à l'Equipe de Microscopie Quantitative où ce travail a été réalisé.

12 - BIBLIOGRAPHIE

[1] AL NACHAWATI (h) (1983)
"Classification et sélection des paramètres à l'aide de l'analyse de la variance à deux facteurs" - I.M.A.G. R.R. N°355

[2] AL NACHAWATI (h) (1983)
"Classification et sélection des questions à l'aide de l'analyse de la variance"
Congrès européens des sociétés de classification PARIS

[3] AL NACHAWATI (h) (1983)
"Segmentation multidimensionnelle"
IMAG-TIM3, R.R. n°413

[4] AL NACHAWATI (h) (1984)
"Segmentation multidimensionnelle" Journées statistiques MONTPELLIER

[5] BACCINI (A) (1975)
"Aspect synthétique de la segmentation et traitement de variables qualitatives à modelites ordonnées"
Université Paul Sabatier de Toulouse - Thèse de 3ème cycle

[6] BELSON (W.A.) (1959)
"Matching and prediction on the principle of biological classification"
Applied Statistics, vol. III

[7] BENZECRI (J.P) & Collaborateurs (1980)
"L'analyse des données" Dunod

[8] BERTIER (P) et BOUROCHE (J.M.) (1975)
"Analyse des données multidimensionnelles"
Presses universitaires de France

[9] BOUROCHE (J.M.) et TENENHAUS (M) (1970)
"Quelques méthodes de segmentation"
Revue française d'Informatique et de Recherche Opérationnelle - Vol. V-2.

[10] CAILLIEZ (F) et PAGÈS (J.P.) (1976)
"Introduction à l'analyse des données"
SMASH, Paris

[11] CAPECCHI (V) (1964)
"Une méthode de classification fondée sur l'entropie".
Revue Française de Sociologie - Vol. V. 290-306

[12] CHANDON (J.L.) et PINSON (S) (1981)
"Analyse typologique"
MASSON

[13] CHAPOUILLE (P) (1973)
"Planification et analyse des expériences"
MASSON

[14] DAGNIELIE (1970-1975)
"Théorie et méthodes de statistiques"
3Vol. GEMBLOUX PRESSE AGRONOMIQUE

[15] DAROCZY (1970)
"Generalized information functions"
Information and control, 16, page 16

[16] DIDAY (E), LEMAIRE (J), POUGET (J) et TESTU (F) (1982)
"Eléments d'analyse de données"
DUNOD

[17] DIDAY (E) (1972)
"Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance de formes"
Thèse d'état, Paris VI.

[18] DIDAY (E) (1976)
"Selection Typologique des parametres"
Rapport Laboria 188

[19] FISHER (W.D.) (1958)
"On grouping for maximum homogeneity"
JASA Vol. 53 PP 789-798

[20] FRIEDMAN (H.D.) et RUBIN (J) (1967)
"On some invariant criteria for grouping data"
JASA -Vol.62 ,1159-1178.

[21] GAUVAIN (C) (1984)
"Application de l'instrumentation et des méthodes d'analyse d'images à l'étude de la maturation et de la prolifération dans la lignée érythroblastique humaine normale".
Thèse de 3ème cycle - Grenoble 1

[22] HUGUES (M) (1970)
"Segmentation et typologie"
Bordas

[23] JANCE (G.N.) et WILLIAMS (W.T.) (1968)
"Note on a new information statistic classification program"
The computer Journal, 11, 195-197

[24] LAUMON (B) (1979)
"Méthode de reconnaissance de formes pour l'estimation d'une variable continue : application à la docimologie"
Thèse de Docteur-Ingénieur LYON

[25] LECHEVALLIER (Y) (1976)
"Classification automatique optimale sous contrainte d'ordre total"
Rapport Laboria 200, INRIA

[26] MARTEL (1979)
Cours DEA LYON 1 , (non publié).

[27] MOLLER (F) (1972)
"An experimental design for evaluation in cluster analysis"
"First National Meeting of the operation Research Society"
NEW ORLEANS

[28] MORGAN (J.N.) et SONQUIST (J.A.) (1963)
"Problems in the analysis of survey data and a proposal"
JASA Vol. 58 N°302 - PP. 414-433

[29] MOUSTAFA (Y) et BRUGAL (G) (1984)
"Image analysis of cell proliferation and differentiation in the thymus of the new pleurodeles watlii, Michah. by SAMBA 200 cell image processor "
W. ROUX Arch. Dep. Biol (sous presse)

[30] OPFERMANN (M) (1984)
"Application des méthodes de la Cytologie Quantitative aux tumeurs de seins"
Rapport DEA Biologie Cellulaire et Moléculaire TIM3 GRENOBLE

[31] ROUTHIER (J.L.) (1978)
"Un processus d'interrogation latticiel - application à l'aide du diagnostic des modules thyroïdiens"
Thèse de 3ème cycle - LYON 1

[32] SCHEFFE (h) (1959)
"The analysis of variance"- J. WILEY

[325] SCOTT (A.J) SYMONS (M.J) (1971) Biometrics, 27, 217-219

[33] TERRENOIRE (M) et TOUNISSOUX (D) (1979)
"Processus non arborescent pour la reconnaissance d'une variable continue" - 2ème congrès AFCET-IRIA "Reconnaissance des formes et intelligence artificielle"
TOULOUSE PP. 410-417

[34] TERRENOIRE (M) et TOUNISSOUX (D) (1981)
"Sample size sensitive entropy" in "pattern recognition and application"
Edited by Dr. J. KITTLER et al., Reidel publishing company

[35] TOUNISSOUX (D) (1980)
"Processus sequentiel adaplatif de reconnaissance de formes pour l'aide au diagnostic"
Thèse d'état LYON 1

[36] VO KHAC (K) et NGHIEM (PH.T.) (1968)
"Etude sur les aspects théorique et pratique de la segmentation aux moindres carrés"
Revue française d'Informatique et de Recherche opérationnelle,
Vol. V-1.

[37] WILLIAMS (W.T.) and LAMBERT (J.M.) (1959)
"Multivariate methods in plant ecology"
Journal of Ecology, vol; 47, P. 83-101