Statistique et analyse des données

Ph. Meunier
P. Baufays
J. P. Rasson

Nouveau critère de segmentation pour des variables à expliquer qualitative ordinale et quantitative multidimensionnelle

Statistique et analyse des données, tome 10, n° 3 (1985), p. 50-67
<http://www.numdam.org/item?id=SAD_1985__10_3_50_0>

Statistique et Analyse des Données
1985 - vol 10 n°3 pp 50-67

NOUVEAU CRITERE DE SEGMENTATION POUR DES VARIABLES A EXPLIQUER QUALITATIVE ORDINALE ET QUANTITATIVE MULTIDIMENSIONNELLE

PH. MEUNIER, P. BAUFAYS ET J.P. RASSON

F.N.F.P., Département de Mathématique
Rempart de la Vierge, 8
B-5000 Namur Belgique

Résumé : *Nous disposons d'un ensemble de* N *individus caractérisés par* (P+1) *variables.* p *variables sont dites "explicatives". La dernière, notée* $\underline{X}^{(p+1)}$ *, est appelée variable "à expliquer". Les méthodes de segmentation classiques essayent d'expliquer* $\underline{X}^{(p+1)}$ *à l'aide d'un arbre dichotomique qui est basé sur des partitions induites par certaines variables explicatives, qualitatives.*

Nous proposons un critère qui semble bien adapté pour le traitement d'une variable à expliquer qualitative ordinale. Si la variable à expliquer est quantitative, alors nous proposons un critère basé sur la mesure de Lebesgue d'enveloppes convexes de groupes d'individus qui sont uniquement décrits par la variable à expliquer. De plus, nous vérifions que ces critères satisfont à certaines propriétés d'invariance.

Abstract : *We have a set of* N *objects characterized by* (P+1) *variables.* p *variables are said "explanatory variables". The last one denoted by* $\underline{X}^{(p+1)}$ *is called the "dependent variable". The classical methods of segmentation try to explain* $\underline{X}^{(p+1)}$ *with the help of a binary tree which is based on partitions induced by some explanatory categorical variables.*

We propose a criterion which seems well adapted for the treatment of a dependent ordinal variable. If the dependent variable is quantitative, then we propose a criterion based on the Lebesgue measure of convex hulls of sets of individuals which only are described by the dependent variable. More, we verify that those criteria fulfil some invariance properties.

Manuscrit reçu le 10 juin 1985, révisé le 28 janvier 1986

Mots clés : *méthodes de segmentation, critère ordinal, critère des surfaces, mesure de Lebesgue.*

Indices de classification I.S.I. : *06-010, 06-020, 06-030.*

1 - INTRODUCTION

On dispose d'un ensemble E de N individus caractérisés par $(p+1)$ variables. p d'entre elles, qualitatives, sont dites "explicatives"; la $(p+1)$-ième est appelée variable "à expliquer" et est notée $\underline{x}^{(p+1)}$. Les méthodes de segmentation essaient de résoudre le problème suivant : "expliquer" $\underline{x}^{(p+1)}$ à l'aide des variables "explicatives". Cela consiste à rechercher, sur base des partitions induites par chaque variable explicative, une partition de l'ensemble des individus en sous-ensembles qui soient entre eux "les plus différents possible", et chacun "le plus homogène possible", ceci relativement à $\underline{x}^{(p+1)}$, et en un sens à préciser. En effet, suivant l'interprétation mathématique qui sera donnée des termes "différent" et "homogène", on obtiendra diverses méthodes de segmentation.

Pour obtenir la partition désirée de l'ensemble des individus, une méthode de segmentation réalise des dichotomies successives de cet ensemble, chacune étant induite par une dichotomie de l'ensemble des modalités de l'une des variables explicatives.

Pour chaque niveau de segmentation, et pour chaque sous-ensemble de E figurant à ce niveau, la méthode de recherche de la dichotomie optimale de cet ensemble est toujours la même : déterminer la variable explicative et la dichotomie de ses modalités optimisant un certain critère. C'est la raison pour laquelle la recherche d'une dichotomie optimale sera dans la suite toujours décrite sur l'ensemble E lui-même.

Dans le cadre de cet article, nous présentons un critère qui semble bien adapté au traitement d'une variable à expliquer qualitative ordinale, puis nous proposons un critère basé sur l'enveloppe convexe des groupes d'individus pour le traitement d'une variable à expliquer quantitative multidimensionnelle.

Par commodité mathématique, nous allons introduire quelques notations.

2 - NOTATIONS

Nous disposons d'un tableau de données

$$X = (x_i^j) \quad (1 \leqslant i \leqslant N \; ; \; 1 \leqslant j \leqslant p+1) \; .$$

$\underline{X}_i = (x_i^1, \ldots, x_i^{(p+1)})$ représente le i-ième individu;

$\underline{X}^j = (x_1^j, \ldots, x_n^j)$ représente la j-ième variable.

Nous notons également :

$\mathbb{E} = \{\underline{X}_i : 1 \leqslant i \leqslant N\}$: ensemble des individus;

$(\mathbb{E}, \mathbb{P}, p)$: espace probabilisé sur $\mathbb{E}$ où $\mathbb{P}$ est la tribu des parties $\mathbb{E}$ et p la probabilité définie sur $(\mathbb{E},\mathbb{P})$ par $p(\{\underline{X}_i\}) = p_i$, $(1 \leqslant i \leqslant N)$, les p_i étant des "poids" associés aux individus;
en général, on prend une distribution uniforme sur $\mathbb{E}$, c'est-à-dire $p_i = 1/N$ $(1 \leqslant i \leqslant N)$;

(P_1, P_2) : partition à 2 classes de $\mathbb{E}$ issue d'une dichotomie des modalités d'une des variables explicatives;

m_j : nombre de modalités de la variable $\underline{X}^j$;

$(E_1^j, \ldots, E_{m_j}^j)$: partition triviale de $\mathbb{E}$ induite par la variable (qualitative) $\underline{X}^j$, de telle sorte que $E_t^j = \{\underline{X}_i \in \mathbb{E} : x_i^j = t\}$ représente l'ensemble des individus de $\mathbb{E}$ présentant la t-ième modalité de la j-ième variable, $\underline{X}^j$;

$G^j(P)$: centre de gravité de P par rapport à la variable $\underline{X}^j$, où $P \in \mathbb{P}$; autrement dit, $G^j(P) = \Sigma\{p_i \, \underset{\sim}{X}_i^j : \underline{X}_i \in P\}$ où $\underset{\sim}{X}_i^j$ représente le vecteur codé de façon disjonctive complète (respectivement de façon additive) correspondant à l'individu $\underline{X}_i$ caractérisé par la variable $\underline{X}^j$ qualitative nominale (respectivement ordinale).

Nous allons consacrer le paragraphe suivant au traitement d'une variable à expliquer qualitative ordinale.

3 - CRITERE ORDINAL

Baccini (1975) énonce deux propriétés auxquelles devrait satisfaire une méthode de segmentation traitant le cas d'une variable à expliquer ordinale :

- invariance de la méthode si l'on remplace la relation d'ordre donnée sur l'ensemble des modalités de $\underline{X}^{(p+1)}$ par la relation d'ordre de sens inverse;
- indépendance de la méthode par rapport à l'ordre défini sur les modalités de la variable à expliquer dans le cas où elle est dichotomique; autrement dit, le critère est dans ce cas équivalent à celui décrit dans la méthode E.L.I.S.E.E.

De plus, il aborde la segmentation sous un aspect nouveau et synthétique qui lui permet d'englober dans un cadre unique les méthodes de segmentation aux moindres carrés (E.L.I.S.E.E., A.I.D., A.I.D. généralisée). Cette nouvelle approche lui permet également de proposer une méthode qui est mieux adaptée au cas où la variable à expliquer est ordinale dans la mesure où elle satisfait aux deux propriétés précitées.

Nous proposons une méthode qui tient compte de la structure ordinale de $\underline{X}^{(p+1)}$ et qui de plus satisfait aux deux propriétés de Baccini.

Notons D_1 (respectivement D_2) l'indice intergroupe associé à l'ordre naturel défini par les valeurs des modalités de $\underline{X}^{(p+1)}$ (respectivement l'ordre inverse).

Nous définissons notre indice intergroupe à maximiser comme suit :

$$
\begin{aligned}
& D(P_1,P_2) \\
& \quad = D_1(P_1,P_2) + D_2(P_1,P_2) \\
& \quad = p(P_1)\, p(P_2) \sum_{t=1}^{m(p+1)} (1/p(E_t^{(p+1)})) \{(p(V_t \mid P_2) - p(V_t \mid P_1))^2 \\
& \qquad\qquad + (p(U_t \mid P_2) - p(U_t \mid P_2))^2\}
\end{aligned}
$$

$$V_t = U \{E_k^{(p+1)} : t \leqslant k \leqslant m_{(p+1)}\} ;$$

$$U_t = U \{E_k^{(p+1)} : 1 \leqslant k \leqslant t\} ;$$

$$(1 \leqslant t \leqslant m_{(p+1)}) .$$

L'invariance de la méthode associée à cet indice intergroupe par rapport au sens de l'ordre défini sur l'ensemble des modalités de $\underline{X}^{(p+1)}$ est évidemment assurée. De plus, la seconde propriété est également vérifiée.

En effet, si $m_{(p+1)} = 2$, nous avons :

$$D(P_1,P_2) = p(P_1)\, p(P_2) \sum_{t=1}^{2} (1/p(E_t^{(p+1)}))\, (p(E_t^{(p+1)} \mid P_2) - p(E_t^{(p+1)} \mid P_1))^2 .$$

Ce critère n'est autre que celui employé dans la méthode E.L.I.S.E.E. lorsque la variable à expliquer est dichotomique et qualitative nominale.

De plus, si on code la variable $\underline{X}^{(p+1)}$ de façon additive relativement à l'ordre naturel défini par les valeurs des modalités, on peut associer à cet indice intergroupe D , une distance quadratique d de métrique M_d où

$$D(P_1,P_2) = p(P_1)\, p(P_2)\, d^2(G^{(p+1)}(P_1) , G^{(p+1)}(P_2)) ;$$

$$M_d = \text{Diag}\ (h_1,\ldots,h_{m_{(p+1)}}) ;$$

$$h_1 = 1 / p(E_1^{(p+1)}) ;$$

$$h_t = 1 / p(E_{(t-1)}^{(p+1)}) + 1 / p(E_t^{(P+1)}) ;$$

$$(2 \leqslant t \leqslant m_{(p+1)}) .$$

Cette distance est définie comme suit :

$$d : \mathbb{R}^{m(p+1)} \times \mathbb{R}^{m(p+1)} \longrightarrow \mathbb{R}^{+}$$

$$(\underline{X}_i, \underline{X}_j) \longrightarrow d_{(p+1)}(\underline{X}_i, \underline{X}_j)$$

$$= \{(\underline{X}_i - \underline{X}_j)^t M_d (\underline{X}_i - \underline{X}_j)\}^{1/2}$$

où $\underline{X}_i = ((\underline{X}_i)_1, \ldots, (\underline{X}_i)_{m_{(p+1)}})^t$ représente par abus de langage le vecteur codé de façon additive du i-ième individu relativement à la variable $\underline{X}^{(p+1)}$, c'est-à-dire $\underline{X}_i$ est la somme des r premiers vecteurs de la base canonique de $\mathbb{R}^{m(p+1)}$ si le i-ième individu présente la r-ième modalité de $\underline{X}^{(p+1)}$.

Si nous explicitons le critère $D(P_1, P_2)$ en termes d'effectifs, il s'exprime comme suit :

$$D(P_1, P_2) = \frac{1}{N\, N_{1.}\, N_{2.}} \sum_{t=2}^{m(p+1)} \frac{N.(t-1) + N.t}{N.(t-1)\, N.t} \left(\sum_{j=t}^{m(p+1)} (N_{1j}\, N_{2.} - N_{2j}\, N_{1.}) \right)^2$$

où $N_{i.} = \text{Card}\,(P_i)$

$N.j = \text{Card}\,(E_j^{(p+1)})$

$N_{ij} = \text{Card}\,(P_i \cap E_j^{(p+1)})$

En outre, la distance entre les individus $\underline{X}_{i_1}$, $\underline{X}_{i_2}$ est invariante par rapport au sens de l'ordre défini sur les modalités de $\underline{X}^{(p+1)}$. La preuve de ce résultat est évidente.

Nous allons illustrer notre approche sur un exemple tiré de Baccini (1975). Considérons une seule variable explicative $\underline{X}^j$ à trois modalités non ordonnées (trois dichotomies de $\mathbb{E}$ lui sont associées), et la variable à expliquer $\underline{X}^{(p+1)}$ à trois modalités ordonnées (ordre naturel). Supposons de plus que la population $\mathbb{E}$ (300 individus) se répartisse comme suit :

$\underline{x}^{(p+1)}$ / $\underline{x}^j$	1	2	3	
1	100	0	0	100
2	0	100	0	100
3	0	0	100	100
	100	100	100	300

Notons D_B (respectivement D_N , D_{BA}) l'indice intergroupe proposé par Bouroche et Tenenhaus (1970) (respectivement le nôtre, celui de Baccini). Il est facile de montrer que l'on a le tableau suivant :

(P_1 , P_2)	$D_B(P_1,P_2)$	$D_N(P_1,P_2)$	$D_{BA}(P_1,P_2)$
E_1^j , $E_2^j \cup E_3^j$	1/2	5/3	0
E_2^j , $E_1^j \cup E_3^j$	1/4	2/3	3/4
E_3^j , $E_1^j \cup E_2^j$	3/4	5/3	0

En outre, il est à remarquer que si l'on permute les lignes et colonnes du premier tableau, on obtient le tableau suivant :

$\underline{X}^{(p+1)}$ / $\underline{X}^j$	3	2	1	
3	100	0	0	100
2	0	100	0	100
1	0	0	100	100
	100	100	100	300

Par conséquent, le tableau des indices intergroupes s'écrit comme suit (lorsqu'on considère l'ordre de sens inverse sur les modalités de $\underline{X}^{(p+1)}$).

(P_1 , P_2)	$D_B(P_1,P_2)$	$D_N(P_1,P_2)$	$D_{BA}(P_1,P_2)$
E_1^j , $E_2^j \cup E_3^j$	3/4	5/3	0
E_2^j , $E_1^j \cup E_3^j$	1/4	2/3	3/4
E_3^j , $E_1^j \cup E_2^j$	1/2	5/3	0

Contrairement au critère (D_B) de Bouroche et Tenenhaus, notre critère (D_N) ainsi que celui de Baccini (D_{BA}) , qui est à minimiser, sont invariants par rapport au sens de l'ordre défini sur les modalités de $\underline{X}^{(p+1)}$.

4 - CRITERE BASE SUR LA MESURE DE LEBESGUE

4.1 - Introduction

Les critères utilisés dans la plupart des méthodes de classification sont souvent basés sur le calcul de dissimilarités entre couples de points (centres de gravité, individus, etc...) de l'espace dans lequel sont représentés les individus.

Dans notre unité de recherche, une méthode de classification automatique (cf. Hardy et Rasson (1982), Hardy et Rasson (Compstat 82), Hardy, Meunier et Rasson (Compstat 82)), a été élaborée non pas en termes de dissimilarités, comme il est de coutume, mais en termes de mesures de Lebesgue d'enveloppes convexes de groupes de points.

Autrement dit, le problème de classification automatique peut être décrit succintement comme suit : disposant d'un échantillon d'individus représentés dans un espace euclidien multidimensionnel, il s'agit de trouver une partition de cet échantillon en k classes dont les enveloppes convexes sont disjointes et dont la somme de leurs mesures de Lebesgue est minimale.

Pratiquement, si l'espace de base est $\mathbb{R}$, on cherche les k intervalles disjoints contenant tous les points représentant les individus et tels que la somme de leurs longueurs soit minimale. Dans $\mathbb{R}^2$, on recherche les k groupes de points tels que la somme des aires de leurs enveloppes convexes soit minimale. Dans $\mathbb{R}^3$ (respectivement $\mathbb{R}^q$), on essaie de trouver les k groupes de points tels que la somme des volumes (respectivement hypervolumes) de leurs enveloppes convexes soit minimale.

Pour plus de détails à propos des algorithmes qui tentent de résoudre ce problème de classification automatique, on peut se référer à Hardy (1983).

Un exemple de données dans le plan traité par l'un de ces algorithmes est présenté à la figure 1.

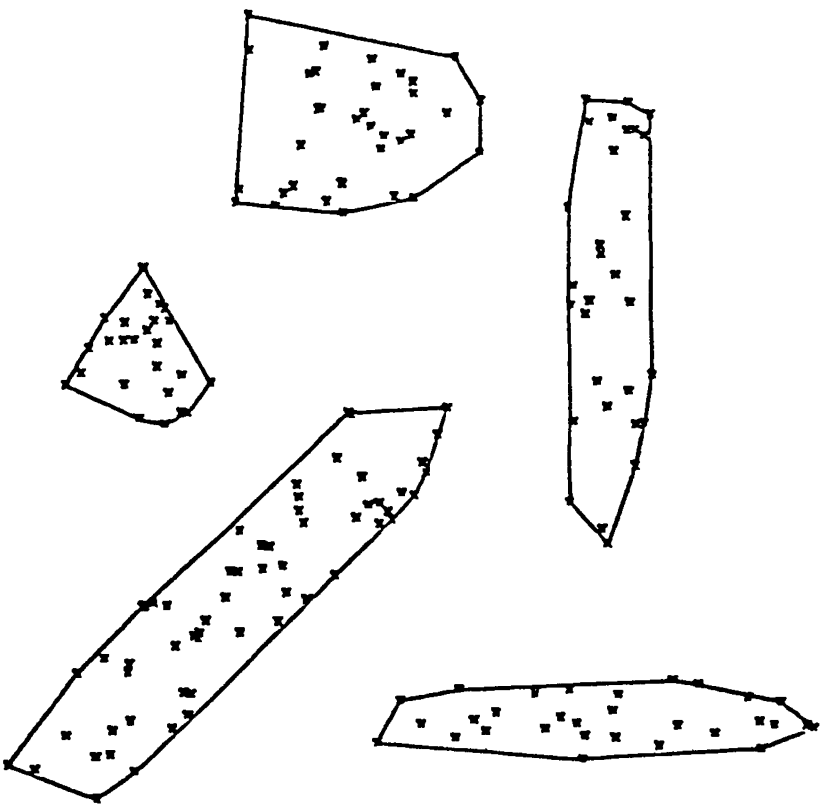

FIGURE 1 : Exemple de données dans le plan

4.2 - Critère de sélection de variables en segmentation

A la suite de ces travaux, dans le cas d'une variable à expliquer quantitative multidimensionnelle, nous allons présenter deux critères de segmentation qui s'expriment également en termes de mesures de Lebesgue d'enveloppes convexes de groupes de points.

Pour ce type de données, la littérature relative à la segmentation propose la méthode A.I.D. généralisée. Cette méthode de segmentation classique tente d'expliquer $\underline{X}^{(p+1)}$ à l'aide des variables explicatives. Pour ce faire, elle réalise des dichotomies successives de l'ensemble $\mathbf{E}$, chacune étant induite par une dichotomie de l'ensemble des modalités de l'une des variables explicatives. Pour chaque niveau de segmentation, et pour chaque segment figurant à ce niveau, la méthode de recherche de la dichotomie optimale de cet ensemble est toujours la même : déterminer la variable explicative et la dichotomie de ses modalités minimisant la somme des inerties de chacun des deux groupes obtenus, relativement à $\underline{X}^{(p+1)}$ et à la métrique de Mahalanobis. En d'autres termes, on dispose d'un ensemble $\mathcal{Q}_2$ de partitions à 2 classes des individus parmi lesquelles on sélectionne celle qui résoud le problème suivant :

$$\text{Min } I(P_1) + I(P_2)$$

sous la contrainte

$$(P_1, P_2) \in \mathcal{Q}_2$$

où $$I(P_i) = \sum_{t \in [P_i]} p_t \, (\underline{x}_t^{(p+1)} - \underline{G}_i)^t \, V^{-1} \, (\underline{x}_t^{(p+1)} - \underline{G}_i) \; ;$$

$\underline{G}_i$ représente le centre de gravité de P_i relativement à $\underline{X}^{(p+1)}$;

V représente la matrice des covariances empiriques de $\underline{X}^{(p+1)}$ sur $\mathbf{E}$ $(= P_1 \cup P_2)$.

Le choix de cette métrique résulte du fait que l'on désire rendre les composantes de la variable à expliquer non corrélées. De plus, le critère optimisé dans A.I.D. généralisée satisfait à différentes propriétés intéressantes.

Fisher et Van Ness (1971) ont proposé différentes conditions auxquelles un algorithme de classification doit satisfaire pour être décrété admissible. Néanmoins, seulement certaines de ces propriétés sont indépendantes de la méthode d'optimisation du critère. Nous avons uniquement passé en revue les propriétés "strictement applicables" au critère.

Les conditions d'admissibilité sont :

1. *Stabilité par rapport aux proportions de points*

Définition : Une procédure est dite stable par rapport aux proportions de points si une duplication d'un ou plusieurs points ne modifie pas la frontière des classes.

Proposition : La méthode A.I.D. généralisée ne satisfait pas à cette propriété.

2. *Propriétés d'invariance*

Par contre, elle est invariante par rapport aux

- translations;
- rotations;
- homothéties.

Nous allons maintenant présenter deux nouveaux critères de segmentation tous deux induits d'un modèle statistique. Désignons le vecteur des réalisations $(\underline{X}_1,\ldots,\underline{X}_N)$ par $\underset{\sim}{X}$ et l'ensemble inconnu de $\mathbf{R}^q$ auquel appartiennent les points $(\underline{X}_i)$ représentant les individus, exprimés relativement à $\underline{X}^{(p+1)}$ (q-dimensionnelle), par P . Parmi les partitions (P_1,P_2) appartenant à $\mathcal{Q}_2$, on recherche celle qui optimise un certain critère exprimé en termes de mesures de Lebesgue.

L'ensemble P de $\mathbf{R}^q$ est constitué par l'union de deux ensembles convexes comprenant tous les points. Il est à remarquer que les classes P_1 , P_2 d'une partition appartenant à $\mathcal{Q}_2$ seront considérées par abus de langage comme les enveloppes convexes des points appartenant respectivement à chacune des classes. (Cf. Rasson, 1979 et Hardy, 1983).

Le premier modèle s'interprète comme un mélange de densités uniformes. La densité au point x s'écrit

$$f(x) = p_1 f(x \mid P_1) + p_2 f(x \mid P_2)$$

où $f(x \mid P_i)$ est la densité uniforme sur P_i et p_i représente la probabilité a priori de P_i ;
autrement dit,

$$f(x \mid P_i) = (1 / m_i) 1_{P_i}(x)$$

où m_i est la mesure de Lebesgue de P_i .

De plus, nous imposons que les probabilités a priori p_1 , p_2 soient proportionnelles aux mesures de Lebesgue des ensembles P_1 , P_2 . Par conséquent, on a évidemment :

$$p_i = m_i / (m_1 + m_2) .$$

La densité au point x s'écrit donc :

$$f(x) = (1 / (m_1 + m_2)) \{1_{P_1}(x) + 1_{P_2}(x)\} .$$

Nous allons estimer P_1 , P_2 par maximum de vraisemblance. Nous considérons les individus $\underline{X}_i$ comme des réalisations indépendantes et identiquement distribuées de variable aléatoire de densité f . La vraisemblance s'écrit :

$$L(P_1 , P_2 ; X) = \prod_{i=1}^{N} f(\underline{X}_i)$$

$$= (1 / (m_1 + m_2))^N \prod_{i=1}^{N} (1_{P_1}(\underline{X}_i) + 1_{P_2}(\underline{X}_i)) .$$

Notons le nombre de points appartenant à la fois à P_1 et P_2 par r .

La vraisemblance s'écrit alors comme suit

$$L(P_1 , P_2 ; X) = (1 / (m_1 + m_2))^N \, 2^r .$$

Le problème de segmentation s'écrit

$$\text{Max } L(P_1 , P_2 ; X)$$

sous la contrainte

$$(P_1 , P_2) \in \mathcal{Q}_2 .$$

Il s'exprime également comme suit

$$\text{Min } (m_1 + m_2) / 2^{(r/N)}$$

sous la contrainte

$$(P_1 , P_2) \in \mathcal{Q}_2 .$$

Autrement dit, parmi les partitions induites des dichotomies de l'ensemble des modalités des variables explicatives, on détermine celle qui minimise la somme des mesures de Lebesgue des enveloppes convexes des points de chacune des classes, pondérées par un terme dépendant entre autres du nombre de points communs aux deux classes.

Cette nouvelle méthode de segmentation ne satisfait qu'à certaines des conditions d'admissibilité de Fisher et Van Ness.

1. Stabilité par rapport aux proportions de points

Elle ne satisfait pas à cette propriété. En effet, le terme $2^{(r/N)}$ n'est pas invariant par rapport à n'importe quelle duplication de points.

2. Propriétés d'invariance

Néanmoins, elle est invariante par rapport aux

- translations;
- rotations;
- homothéties.

Nous proposons maintenant un autre critère, associé à un autre modèle probabiliste, et pour lequel la méthode de segmentation associée satisfait aux propriétés d'invariance ainsi qu'à la propriété de stabilité par rapport aux proportions de points Dans ce cas, nous supposons que les points sont "distribués indépendamment et uniformément" dans l'ensemble P considéré comme l'union de deux ensembles convexes, comprenant tous les points $\underset{\sim}{X}_i$.

La densité au point x s'écrit

$$f(x) = (1 / m(P))\ 1_P(x)$$

où m(P) représente la mesure de Lebesgue de l'ensemble convexe P .

Quant à la vraisemblance, elle peut s'écrire comme suit :

$$L(P ; \underset{\sim}{X}) = (1 / m(P))^N \prod_{i=1}^{N} 1_P(\underset{\sim}{X}_i) .$$

Par conséquent, le problème de segmentation s'écrit

$$\text{Max } L(P ; \underset{\sim}{X})$$

sous les contraintes

$$P = P_1 \cup P_2$$

$$(P_1 , P_2) \in \mathcal{Q}_2 .$$

Autrement dit, le problème peut également s'écrire

$$\text{Min } m(P_1 \cup P_2)$$

sous la contrainte

$$(P_1 , P_2) \in \mathcal{Q}_2 .$$

Il est à remarquer que $m(P_1 \cup P_2)$ peut s'exprimer comme $m(P_1) + m(P_2) - m(P_1 \cap P_2)$.

Du fait que ce critère s'exprime uniquement en termes de mesures de Lebesgue de sous-ensembles de $\mathbf{R}^q$, la méthode de segmentation associée satisfait à la propriété de stabilité par rapport aux proportions de points et est invariante par rapport aux translations, aux rotations et aux homothéties.

Bien que le premier critère soit plus facile à calculer, nous lui préférerons le second qui contrairement au premier, vérifie la propriété de stabilité par rapport aux proportions de points.

4 - CONCLUSIONS

Dans un premier temps, dans le cadre du traitement d'une variable à expliquer qualitative ordinale, nous avons élaboré un critère qui satisfait aux propriétés énoncées par Baccini (1975). C'est dans ce sens que l'on entend que ce critère semble bien adapté à ce type de données.

Dans un second temps, dans le cadre du traitement d'une variable à expliquer quantitative multidimensionnelle, nous avons établi deux modèles statistiques desquels découlent deux critères basés tous deux sur la mesure de Lebesgue d'enveloppes convexes de groupes de points. En outre, ces critères satisfont à certaines propriétés d'invariance.

REFERENCES

[1] BACCINI, A., Aspect synthétique de la segmentation et traitement des variables qualitatives à modalités ordonnées, Thèse troisième cycle, Université Paul Sabatier, Toulouse, 1975.

[2] BACCINI, A. et POUSSE, A.," Segmentation aux moindres carrés : un aspect synthétique",Revue de statistique appliquée, 23, 1975.

[3] BELSON, W.A., "Matching and prediction on the principle of biological classification", Applied Statistics, 8, 65-75, 1980.

[4] BERTIER, P. et BOUROCHE, J.M., Analyse des données multidimensionnelles, P.U.F., 1975.

[5] BOUROCHE, J.M. et TENENHAUS, M.,"Quelques méthodes de segmentation", R.I.R.O., 2, 29-42, 1970.

[6] CAILLEZ, F. et PAGES, J.P., Introduction à l'analyse des données, Smash, Paris, 1976.

[7] CELLARD, J.C., LABBE, B. et SAVITSKY, G., "Le programme E.L.I.S.E.E. - Présentation et applications", METRA 6(3), 503)520, 1967.

[8] DIDAY, E., LEMAIRE, J., POUGET, J. et TETSU, F., Eléments d'analyse des données, Dunod, Paris, 1982.

[9] HARDY, A., et RASSON, J.P., A global algorithm and a criterion for grouping data, In Caussinus, H., Ettinger, P. et Tomassone, R. (editors), Proc. of Fifth COMPSTAT Meeting, Toulouse, Springer Verlag, 1982.

[10] HARDY, A., MEUNIER, PH. et RASSON, J.P., A new algorighm for cluster analysis, In Caussinus, H., Ettinger, P. et Tomassone, R. (Editors) Proc. of Fifth COMPSTAT Meeting, Toulouse, Springer Verlag, 1982.

[11] HARDY, A. et RASSON, J.P., "Une nouvelle approche des problèmes de classification automatique", Statistique et Analyse des Données, 7, 41-56, 1982.

[12] HARDY, A., Statistique et classification automatique : Un modèle - un nouveau critère - des algorithmes - des applications, PhD Thesis, Facultés Universitaires de Namur, Belgique, 1983.

[13] LERMAN, I.C., Les bases de la classification automatique, Gauthier-Villars Paris, 1970.

[14] MEUNIER, PH., DIDAY, E. et RASSON, J.P., "Méthode et algorithme de sélection typologique de paramètres", sous presse dans R.A.I.R.O., vol. 19, 1985.

[15] MORGAN, J.N. et SONQUIST, J.A., "Problems in the analysis of survey data, and a proposal", J.A.S.A., 58, 415-434, 1963.

[16] RASSON, J.P., "Estimation de formes convexes du plan", Statistique et analyse des données, 1, 31-46, 1979.

[17] VO KHAC, KH. et NGHIEN, "Etude sur les aspects théoriques et pratiques de la segmentation aux moindres carrés", R.I.R.O., 8, 77-90, 1968.