REVUE FRANÇAISE D'INFORMATIQUE ET DE RECHERCHE OPÉRATIONNELLE. SÉRIE VERTE

JEAN-MARIE BOUROCHE
MICHEL TENENHAUS

Quelques méthodes de segmentation

Revue française d'informatique et de recherche opérationnelle. Série verte, tome 4, n° V2 (1970), p. 29-42

<http://www.numdam.org/item?id=RO_1970__4_2_29_0>

R.I.R.O.
(4e année, V-2, 1970, p. 29-42)

QUELQUES METHODES DE SEGMENTATION (1)

Jean-Marie BOUROCHE (2), Michel TENENHAUS (3)

Résumé. — *Les éléments d'une population sont décrits au moyen de plusieurs caractéristiques. On cherche à expliquer l'une des caractéristiques, notée Y, à l'aide des autres. Pour cela, on se place dans un espace métrique et l'on cherche à constituer des groupes les plus homogènes possible d'une part et les plus différents possible deux à deux d'autre part. On montre, à l'aide du théorème de Kœnig-Huygens, que ces deux propriétés sont équivalentes.*

Dans le cas ou Y prend des valeurs discrètes non ordonnées, l'espace métrique est muni de la distance du χ^2 proposée par J.-P. Benzecri, et on aboutit ainsi à la méthode Elisee.

Lorsque Y prend des valeurs continues, l'espace métrique est identifié à R et on aboutit à la méthode A.I.D.

D'autres méthodes sont proposées ou citées permettant de résoudre le cas ou Y est discret et ordonné et le cas ou Y prend ses valeurs dans R^n.

INTRODUCTION

Les problèmes qu'on rencontre dans la pratique peuvent fréquemment se formaliser de la manière suivante :

On a une population E d'individus ou d'objets.

A chaque élément e de E on associe un ensemble de caractéristiques $(x_1, ..., x_n)$. On peut noter $\mathfrak{X}_i$ l'ensemble des x_i (i-ième caractéristique) possibles et $X_i : E \to \mathfrak{X}_i$ l'application ou variable qui à un individu e associe $X_i(e)$, sa i-ième caractéristique.

Par exemple :

— E : ensemble des employés d'une entreprise.

— $\mathfrak{X}_1$: sexe, $\mathfrak{X}_2$: statut, $\mathfrak{X}_3$: formation, $\mathfrak{X}_4$: salaire, $\mathfrak{X}_5$: ancienneté, $\mathfrak{X}_6$: fonction, etc...

(1) Les auteurs remercient Mlle Ulmo, professeur à l'Institut de Statistique de l'Université de Paris, et M. Bertier, directeur à la SEMA (Metra International), pour les nombreuses discussions qu'ils ont eues avec eux et qui leur ont permis de mieux comprendre les problèmes de segmentation.

(2) Chargé de Recherches à la Direction Scientifique de SEMA (Metra International).

(3) Assistant à la Faculté des Lettres et Sciences Humaines de Nanterre.

Lorsqu'une des variables X_i joue un rôle privilégié on s'intéresse aux deux problèmes suivants :

1) on cherche à « expliquer » la variable privilégiée à l'aide des autres variables. Par exemple on peut chercher à « expliquer » le salaire (X_4) par le sexe (X_1), le statut (X_2), ...

On a l'habitude de noter la variable à « expliquer » (ou critère) Y.

On appelle alors les autres variables X_i des variables « explicatives » (ou facteurs).

2) on cherche une partition (ou segmentation) de E à l'aide des variables X_i de telle sorte que :

a) deux parties (ou segments) soient les plus « différentes » possibles vis-à-vis de Y,

b) chaque partie soit aussi « homogène » que possible vis-à-vis de Y.

On distingue variables qualitatives et variables quantitatives.

Une variable qualitative prend ses valeurs dans un ensemble discret ordonné ou non (exemple : sexe, formation).

Une variable quantitative prend ses valeurs dans l'ensemble des nombres réels (exemple : salaire, ancienneté).

Les méthodes de segmentation, qui permettent de résoudre, dans une certaine mesure, les problèmes 1) et 2) dépendent de la structure (qualitative ou quantitative) de la variable à expliquer Y, les variables explicatives étant toujours qualitatives [1] (ou, quand elles sont quantitatives, réduites en classes).

Après quelques considérations théoriques sur la segmentation nous présenterons plusieurs méthodes.

I. APPROCHE THEORIQUE DU PROBLEME DE LA SEGMENTATION

I.1. Recherche d'une segmentation optimale

On cherche une partition de E vérifiant les deux conditions suivantes (Bertier P. et Boutin A. M. [2]).

I.1.1. Deux parties disjointes sont les plus « différentes » possibles vis-à-vis de Y.

I.1.2. Chaque partie est aussi « homogène » que possible vis-à-vis de Y.

Le problème consiste à trouver une bonne mesure de la « différence de deux parties (distance inter-groupe) et de l'homogénéité d'une partie (distance intra-groupe).

---

(1) Lorsque les variables explicatives sont toutes quantitatives il existe d'autres méthodes : analyse discriminante (critère qualitatif), régression multiple (critère quantitatif), etc.

I.2. Distance inter-groupe et distance intra-groupe

Pour définir une distance inter-groupe et une distance intra-groupe on peut procéder de la manière suivante (Vo Khac Kh. et Nghiem Ph. T. [6]).

A l'espace E on associe un espace pré-hilbertien H par l'intermédiaire d'une application Y^* (nous donnerons dans les parties II et III des exemples d'application Y^* et d'espace H).

La distance entre éléments de E, ou entre parties de E, sera définie à l'aide d'une forme quadratique sur H.

Notons $\langle x, y \rangle$ le produit scalaire défini sur H, $\| x \|^2 = \langle x, x \rangle$ et $d^2(x, y) = \| x - y \|^2$. On définit la « distance » entre deux éléments de E par :

$$d^2(e, e') = d^2(Y^*(e),\ Y^*(é)).$$

On définit une application $\bar{Y}^* : \mathfrak{P}(E) \to H$, qui à $G \in \mathfrak{P}(E)$, associe $\bar{Y}^*(G)$ telle que :

$$\bar{Y}^*(G) = \frac{1}{n(G)} \sum_{g \in G} Y^*(g) \text{ avec } n(G) = \text{Card } G$$

$\bar{Y}^*(G)$ représente le barycentre des points $Y^*(g)$ munis de la masse $\dfrac{1}{n(G)}$

On définit la distance entre deux parties G_1 et G_2 de E par

$$d^2(G_1, G_2) = d^2(\bar{Y}^*(G_1), \bar{Y}^*(G_2))$$

I.2.1. Distance intra-groupe

Soit $G \in \mathfrak{P}(E)$. On définit la distance intra-groupe $D(G)$ par

$$D(G) = \frac{1}{n(G)} \sum_{g \in G} d^2(\bar{Y}^*(G), Y^*(g)).$$

I.2.2. Distance inter-groupe

Soient G_1 et $G_2 \in \mathfrak{P}(E)$. On définit la distance inter-groupe par

$$D(G_1, G_2) = P(G_1)d^2(\bar{Y}^*(G_1), \bar{Y}^*(G_1 \cup G_2)) + P(G_2)d^2(\bar{Y}^*(G_2), \bar{Y}^*(G_1 \cup G_2))$$

avec $P(G_i) = \dfrac{n(G_i)}{n(E)}$

La distance inter-groupe s'étend de manière naturelle au cas de

$$k \text{ groupes : } D(G_1, \ldots, G_k) = \sum_{i=1}^{k} P(G_i)d^2\left(\bar{Y}^*(G_i),\ \bar{Y}^*\left(\bigcup_{i=1}^{k} G_i \right) \right)$$

REMARQUE : On peut montrer que si $G_1 \cap G_2 = \varnothing$

$$D(G_1, G_2) = \frac{P(G_1)P(G_2)}{P(G_1) + P(G_2)} d^2(G_1, G_2).$$

I.3. Relation entre la distance intra-groupe et la distance inter-groupe : Théorème de Kœnig-Huygens

On déduit du théorème Koenig-Huygens la relation suivante entre distance inter-groupe et distance intra-groupe de deux parties disjointes G_1 et G_2 de E :

$$\boxed{D(G_1 \cup G_2) = P(G_1)\, D\,(G_1) + P(G_2)\, D\,(G_2) + D(G_1, G_2)} \quad \text{(KH1)}$$

Dans le cas particulier où on cherche à dichotomiser E en deux parties disjointes E^0 et E^1 on a :

$$D(E) = P(E^0)D(E^0) + P(E^1)D(E^1) + D(E^0, E^1)$$

Maximiser la distance inter-groupe entre E^0 et E^1 est équivalent à minimiser la moyenne des distances intra-groupes. On peut ainsi trouver une dichotomie de E qui réponde à la condition I.1.1 et, en moyenne, à la condition I.1.2.

La relation (KH1) se généralise au cas de k parties disjointes de E.

$$D\left(\bigcup_{i=1}^{k} G_i\right) = \sum_{i=1}^{k} p(G_i)D(G_i) + D(G_1, \ldots, G_k) \qquad \text{(KH2)}$$

I.4. Procédure de segmentation par dichotomisations successives

I.4.1. Caractéristiques des partitions recherchées

On s'intéresse aux partitions de E qu'il est possible de décrire à l'aide des variables X_i, c'est-à-dire aux partitions $E_1, \ldots, E_k$ de E induites par les partitions $\mathfrak{X}^1, \ldots \mathfrak{X}^k$ de $\mathfrak{X} = \mathfrak{X}_1 \times \ldots \times \mathfrak{X}_n$.

Les méthodes de segmentation utilisent une procédure séquentielle. La première étape consiste à dichotomiser la population E en deux sous populations E^0 et E^1 à l'aide d'une dichotomie sur $\mathfrak{X}_i$.

Une dichotomie sur $\mathfrak{X}_i$ est équivalente à la donnée d'une application de $\mathfrak{X}_i$ sur $\{0,1\}$.

Notons $\Delta_i = \{\delta_i\}$ l'ensemble de ces applications.

On a, $\forall \delta_i \in \Delta_i$: $\qquad \mathfrak{X}_i = \mathfrak{X}_i^0 + \mathfrak{X}_i^1$

où $\qquad \delta_i^{-1}(0) = \mathfrak{X}_i^0 \qquad$ et $\qquad \delta_i^{-1}(1) = \mathfrak{X}_i^1$

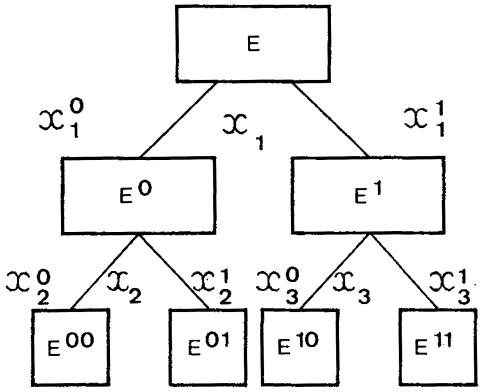
La dichotomie induite sur E par δ_i est donnée par :

$$E^0 = X_i^{-1}(\delta_i^{-1}(0)) \qquad \text{et} \qquad E^1 = X_i^{-1}(\delta_i^{-1}(1)).$$

A l'étape suivante on dichotomise E^0 et E^1, l'ensemble de caractéristiques associé à E^0(Resp. E^1) devenant

$$\mathfrak{X}_1 \times \ldots \times \mathfrak{X}_i^0 \times \ldots \times \mathfrak{X}_n (\text{Resp. } \mathfrak{X}_1 \times \ldots \times \mathfrak{X}_i^1 \times \ldots \times \mathfrak{X}_n).$$

La procédure se poursuit sur les sous-populations obtenues. On obtient ainsi un arbre.



EXEMPLE

$\mathfrak{X}_1^0$: sexe masculin

$\mathfrak{X}_1^1$: sexe féminin

$\mathfrak{X}_2^0$: ouvrier, $\mathfrak{X}_2^1$ non ouvrier

$\mathfrak{X}_3^0$: études supérieures

$\mathfrak{X}_3^1$: pas d'étude supérieure

et E^{00} représente les employés de sexe masculin et ouvrier.

E^{10} représente les employées ayant fait des études supérieures. etc...

I.4.2. Choix de la dichotomie

On cherche à dichotomiser E en deux parties disjointes E^0 et E^1 à l'aide d'une dichotomie d'un espace $\mathfrak{X}_i$.

On désire maximiser la distance inter-groupe entre E^0 et E^1. Autrement dit on cherche $\delta_i^* \in \Delta_i$ vérifiant

$$D(X_i^{-1}(\delta_i^{*-1}(0)), X_i^{-1}(\delta_i^{*-1}(1))) = \max_{\delta_i \in \Delta_i} D(X_i^{-1}(\delta_i^{-1}(0)), X_i^{-1}(\delta_i^{-1}(1))$$

Pour chaque variable X_i on obtient la dichotomie optimum. On cherche ensuite la variable X_{i^*} vérifiant

$$D(X_{i^*}^{-1}(\delta_{i^*}^{*-1}(0)), X_{i^*}^{-1}(\delta_{i^*}^{*-1}(1))) = \max_i D(X_i^{-1}(\delta_i^{*-1}(0)), X_i^{-1}(\delta_i^{*-1}(1)))$$

Ainsi la dichotomie $\delta_{i^*}^*$ sur la variable X_{i^*} induit sur E deux sous-populations

$$E^0 = X_{i^*}^{-1}(\delta_{i^*}^{*-1}(0)) \qquad \text{et} \qquad E^1 = X_{i^*}^{-1}(\delta_{i^*}^{*-1}(1))$$

dont la distance inter-groupe est maximum. De plus la moyenne des

distances intra-groupes de E^0 et E^1 est minimum (voir I.3). Il est clair qu'on ne cherche à maximiser la distance inter-groupe qu'entre les parties E^0 et E^1 de E générées à l'aide des dichotomies permises.

La procédure se poursuit sur les sous-populations obtenues. Celles-ci sont alors considérées comme deux populations de départ et on redéfinit sur elles l'espace H et le produit scalaire sur H.

Les règles d'arrêt sont en général liées :

— à la taille des segments,

— au rapport de la distance intra-groupe des sous-populations à la distance intra groupe de la sous-population totale,

— à différents tests statistiques.

I.4.3. Liaison entre les variables

Les variables X_i qui ont été choisies pour effectuer les dichotomies sur E s'interprétent généralement comme les plus « liées » à la variable à expliquer Y.

L'ordre dans lequel elles ont été choisies peut s'interpréter comme une hiérarchie sur la force de liaison.

Cette interprétation est liée à la métrique choisie ; elle n'a donc pas de valeur générale.

Nous la justifierons dans une certaine mesure pour la méthode Elisée et la méthode AID.

II. APPLICATION AU CAS D'UN CRITERE QUALITATIF LA METHODE ELISEE

II.1. Critère nominal

Supposons que Y prenne J valeurs notées 1, 2, ... j, ... J.

Prenons comme espace H, associé à E, R^J muni du produit scalaire (Benzecri J. P. [1]) :

$$\langle x, y \rangle = \sum_{j=1}^{J} \frac{1}{p(j)} x_j y_j$$

où $p(j) = P(Y = j)$

La distance entre deux éléments de R^J est :

$$d^2(x, y) = \sum_{j=1}^{J} \frac{1}{p(j)} (x_j - y_j)^2$$

I.1.1. Définition des applications Y^* et $\bar{Y}^*$

On pose :

$$Y^*(e) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$Y^*(e)$ est un vecteur de R^J ayant toutes ses coordonnées nulles sauf la $Y(e)$-ième qui est égale à 1.

On peut encore écrire :

$$Y^*(e) = \begin{pmatrix} p(1 \mid e) \\ p(2 \mid e) \\ \vdots \\ p(j \mid e) \\ \vdots \\ p(J \mid e) \end{pmatrix}$$

Soit $G \in \mathfrak{T}(E)$, la définition du § I.2 donne :

$$\bar{Y}^*(G) = \frac{1}{n(G)} \sum_{g \in G} Y^*(g)$$

$$= \frac{1}{n(G)} \begin{pmatrix} n_1(G) \\ n_2(G) \\ \vdots \\ n_j(G) \\ \vdots \\ n_J(G) \end{pmatrix}$$

où $n_j(G)$ est le nombre d'éléments g de G tels que $Y(g) = j$, finalement :

$$\bar{Y}^*(G) = \begin{pmatrix} p(1 \mid G) \\ p(2 \mid G) \\ \vdots \\ p(J \mid G) \end{pmatrix}$$

C'est la loi de probabilité de Y restreint à G.

II.1.2. Définition des distances

— *distance entre parties de E*

$$d^2(G_1, G_2) = d^2(\bar{Y}^*(G_1), \bar{Y}^*(G_2))$$
$$= \sum_{j=1}^{J} \frac{1}{p(j)} (p(j \mid G_1) - p(j \mid G_2))^2$$

— *distance intra-groupe*

$$D(G) = \frac{1}{n(G)} \sum_{g \in G} d^2(\bar{Y}^*(G), Y^*(g))$$
$$= \frac{1}{n(G)} \sum_{g \in G} \sum_{j=1}^{J} \frac{1}{p(j)} (p(j \mid G) - p(j \mid g))^2$$

— *distance inter-groupe*

$$D(G_1, G_2) = P(G_1) d^2(\bar{Y}^*(G_1), \bar{Y}^*(G_1 \cup G_2))$$
$$+ P(G_2) d^2(\bar{Y}^*(G_2), \bar{Y}^*(G_1 \cup G_2))$$
$$= P(G_1) \sum_{j=1}^{J} \frac{1}{p(j)} (p(j \mid G_1) - p(j \mid G_1 \cup G_2))^2$$
$$+ P(G_2) \sum_{j=1}^{J} \frac{1}{p(j)} (p(j \mid G_2) - p(j \mid G_1 \cup G_2))^2$$

et si $G_1 \cap G_2 = \varnothing$ on peut écrire :

$$D(G_1, G_2) = \frac{P(G_1)P(G_2)}{P(G_1) + P(G_2)} \sum_{j=1}^{J} \frac{1}{p(j)} (p(j \mid G_1) - p(j \mid G_2))^2$$

II.1.3. Application du théorème de Kœnig-Huygens

Le théorème de Koenig-Huygens s'écrit pour une dichotomie E^0, E^1 de E.

$$D(E) = P(E^0)D(E^0) + P(E^1)D(E^1) + D(E^0, E^1)$$

Explicitons les différents termes de cette équation :

1) Montrons que $D(E) = J - 1$

$$D(E) = \frac{1}{n(E)} \sum_{e \in E} \sum_{j=1}^{J} \frac{1}{p(j)} (p(j) - p(j \mid e))^2$$

On a $p(j|e) = 0$ pour $j \neq Y(e)$

et $p(j|e) = 1$ pour $j = Y(e)$

Par conséquent

$$\sum_{j=1}^{J} \frac{1}{p(j)} (p(j) - p(j \mid e))^2$$

$$= \sum_{j \neq Y(e)} \frac{1}{p(j)} (p(j))^2 + \frac{1}{p(Y(e))} (p(Y(e)) - 1)^2$$

$$= 1 - p(Y(e)) + \frac{1}{p(Y(e))} (p(Y(e)) - 1)^2$$

$$= \frac{1 - p(Y(e))}{p(Y(e))}$$

Ensuite $\frac{1}{n(E)} \sum_{e \in E} \sum_{j=1}^{J} \frac{1}{p(j)} (p(j) - p(j \mid e))^2$

$$= \frac{1}{n(E)} \left(\sum_{e \in E} \frac{1}{p(Y(e))} \right) - 1$$

$$= \frac{1}{n(E)} \sum_{j=1}^{J} \left(\sum_{e : Y(e) = j} \frac{n(E)}{n_j(E)} \right) - 1$$

$$= J - 1$$

2) Montrons que $D(E^0, E^1)$ est égal à un φ^2 tel qu'il est défini par Cramer (voir pour les propriétés du φ^2, Cramer p. 282 et p. 441-444, [4]).

Y	E^0	E^1	
1	$p(1, E^0)$	$p(1, E^1)$	$p(1)$
j	$p(j, E^0)$	$p(j, E^1)$	$p(j)$
J	$p(J, E^0)$	$p(J, E^1)$	$p(J)$
	$p(E^0)$	$p(E^1)$	

On a :

$$\varphi^2 = \sum_j \left[\frac{(p(j, E^0) - p(E^0)p(j))^2}{p(E^0)p(j)} + \frac{(p(j, E^1) - p(E^1)p(j))^2}{p(E^1)p(j)} \right]$$

et d'autre part

$$D(E^0, E^1) = P(E^0) \sum_j \frac{1}{p(j)} (p(j \mid E^0) - p(j))^2$$

$$+ P(E^1) \sum_j \frac{1}{p(j)} (p(j \mid E^1) - p(j))^2$$

On montre alors que :

$$D(E^0, E^1) = \varphi^2$$

Lorsque les $p(j, E^i)$ sont remplacés par $\frac{n_j(E^i)}{n(E)}$, on a :

$$\varphi^2 = \frac{\chi^2}{n(E)}$$

II.1.4. La méthode Élisée :

On cherche à dichotomiser E en deux parties E^0 et E^1 selon la procédure décrite au § I.4. Pour une variable X_i et une dichotomie $\delta_i \in \Delta_i$, on a :

$$E^0 = X_i^{-1}(\delta_i^{-1}(0))$$

$$E^1 = X_i^{-1}(\delta_i^{-1}(1))$$

et

$D(E^0, E^1) = \varphi^2[Y, \delta_i \circ X_i]$ où $\delta_i \circ X_i$ est la variable dichotomisée qui applique E sur $\{0,1\}$.

On cherche la variable X_{i^*} et la dichotomie $\delta_{i^*}^*$ telle que :

$$\varphi^2[Y, \delta_{i^*}^* \circ X_{i^*}] = \max_i \max_{\delta_i \in \Delta_i} \varphi^2[Y, \delta_i \circ X_i]$$

La procédure se poursuit sur les populations E^0 et E^1 ainsi obtenues.

On peut trouver des exemples d'application dans Cellard, Labbé et Savitsky [3].

II.1.5. Quelques remarques sur la méthode

Le fait de maximiser le φ^2 comporte les avantages suivants :

a) Maximiser le φ^2 revient à maximiser la distance entre les histogrammes de Y sur les parties de E obtenues à partir de dichotomies sur les $\mathfrak{X}_i$.

b) On minimise, en moyenne, les dispersions intra-groupe de E^0 et E^1.

c) Le maximum, qui est égal à 1, de $\varphi^2[Y, \delta_i \circ X_i]$ est atteint lorsqu'il y a liaison fonctionnelle entre Y et $\delta_i \circ X_i$. Les modalités du critère prises par les éléments de E^0 sont alors disjointes des modalités prises par les éléments de E^1.

d) La distance intra-groupe des sous-populations ne peut que diminuer. Lorsqu'elle est nulle, cela signifie qu'une seule modalité du critère est représentée dans la sous-population.

e) Les règles d'arrêt sont liées au test du χ^2. Lorsque χ^2 est petit, cela signifie qu'il n'y a pas une grande distance entre les groupes E^0 et E^1 et que les variables Y et $\delta_i \circ X_i$ peuvent être considérées comme indépendantes.

II.1.6. Conclusion

On s'aperçoit que la métrique du χ^2 utilisée dans la méthode Elisée résoud en grande partie les différents problèmes exposés dans l'introduction. On construit, en effet, des segments homogènes à l'aide des variables les plus liées au critère Y.

II.2. Critère ordinal

Nous supposons que Y peut prendre J valeurs ordonnées : $1 < 2 < \ldots < J$. Nous prenons comme espace H, associé à E, R^J muni du produit scalaire

$$\langle x, y \rangle = \sum_{j=1}^{J} \frac{1}{K(j)} x_j y_j \qquad \text{où} \qquad K(j) = P(Y \geqslant j)$$

Pour tenir compte de la structure ordinale de $\mathcal{Y}$ on peut choisir comme application $Y^*(e)$

$$Y^*(e) = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$Y^*(e)$ est un vecteur de $\mathbf{R}^J$ ayant toutes ses coordonnées de rang inférieur ou égal à $Y(e)$ égales à 1 et les suivantes nulles.

On peut aussi écrire :

$$Y^*(e) = \begin{bmatrix} K(1|e) \\ K(2|e) \\ \vdots \\ K(j|e) \\ \vdots \\ K(J|e) \end{bmatrix}$$

où $K(j|e) = P(Y \geqslant j|e)$

Définissons $\bar{Y}^*$:

$$\bar{Y}^*(G) = \frac{1}{n(G)} \sum_{g \in G} Y^*(g)$$

$$= \frac{1}{n(G)} \begin{bmatrix} N_1(G) \\ N_2(G) \\ \vdots \\ N_j(G) \\ \vdots \\ N_J(G) \end{bmatrix}$$

où $N_j(G)$ est le nombre d'éléments g de G tels que $Y(g) \geqslant j$

Donc

$$\bar{Y}^*(G) = \begin{bmatrix} K(1|G) \\ K(2|G) \\ \vdots \\ K(j|G) \\ \vdots \\ K(J|G) \end{bmatrix}$$

où $K(j|G) = P(Y \geqslant j|G) = 1 - F(j|G)$

où $F(j|G)$ est la fonction de répartition de Y restreint à G.

La distance $d(G_1, G_2)$ représente une distance entre les fonctions de répartition de Y restreint à G_1 et Y restreint à G_2 :

$$d^2(G_1, G_2) = d^2(\bar{Y}^*(G_1), \bar{Y}^*(G_2))$$

$$= \sum_{j=1}^{J} \frac{1}{K(j)} (K(j \mid G_1) - K(j \mid G_2))^2$$

$$= \sum_{=1}^{J} \frac{1}{K(j)} (F(j \mid G_1) - F(j \mid G_2))^2$$

Les résultats du chapitre I s'appliquent à ce cas particulier.

On cherche à maximiser :

$$D(E^0, E^1) = P(E^0)P(E^1) \sum_{j=1}^{J} \frac{1}{K(j)} (F(j \mid E^0) - F(j \mid E^1))^2$$

où E^0, E^1 est induit par une dichotomie sur un $\mathfrak{X}_i$.

III. APPLICATION AU CAS D'UN CRITERE QUANTITATIF

III.1. Critère quantitatif à une dimension : la méthode A.I.D.

Le critère Y est à la valeur dans R.

On prend $H = R$ et $Y^* = Y$.

La distance sur R est la distance habituelle : $d(e, e') = |Y(e) - Y(e')|$.

L'application $\bar{Y}^*$ est définie par la moyenne $\bar{Y}$:

$$\bar{Y}(G) = \frac{1}{n(G)} \sum_{g \in G} Y(g)$$

(moyenne de Y restreint à G).

La distance entre deux parties est la distance entre les moyennes :

$$d(G_1, G_2) = |\bar{Y}(G_1) - \bar{Y}(G_2)|$$

La distance intra-groupe est égale à la variance de Y restreint à G :

$$D(G) = \frac{1}{n(G)} \sum_{g \in G} (Y(g) - \bar{Y}(G))^2$$

et la distance inter-groupe

$$\begin{aligned} D(G_1, G_2) &= P(G_1)(\bar{Y}(G_1) - \bar{Y}(G_1 \cup G_2))^2 \\ &\quad + P(G_2(\bar{Y})(G_2) - \bar{Y}(G_1 \cup G_2))^2 \end{aligned}$$

ce qui s'écrit dans le cas où $G_1 \cap G_2 = \varnothing$

$$D(G_1, G_2) = \frac{P(G_1)P(G_2)}{P(G_1) + P(G_2)} (\bar{Y}(G_1) - \bar{Y}(G_2))^2$$

Le théorème de Koenig-Huygens donne dans le cas d'une dichotomie E^0, E^1 de E l'équation de l'analyse de la variance :

$$\begin{aligned} &\sum_{e \in E} (Y_{\backslash e}) - \bar{Y}(E))^2 \\ &= \sum_{e \in E^0} (Y(e) - \bar{Y}(E^0))^2 + \sum_{e \in E^1} (Y(e) - \bar{Y}(E^1))^2 \\ &+ \frac{n(E^0)n(E^1)}{n(E)} (\bar{Y}(E^0) - \bar{Y}(E^1))^2 \end{aligned}$$

La méthode A.I.D. (Morgan J.N. and Sonquist J.A. [5]) consiste à appliquer la procédure de segmentation décrite au § I.4 en prenant :

$$D(E^0, E^1) = \frac{n(E^0)n(E^1)}{n(E)^2} (\bar{Y}(E^0) - \bar{Y}(E^1))^2$$

Cette quantité représente ce que l'on appelle habituellement la variance expliquée. La dichotomie $\delta_i \circ X_i$ que l'on choisit est celle qui maximise la variance expliquée et on retrouve une liaison entre Y et $\delta_i \circ X_i$ au sens de l'analyse de la variance.

III.2. Critère quantitatif multidimensionnel

Le critère Y est à valeur dans $\mathbf{R}^k$: on pose $Y = (Y_1, \ldots Y_i, \ldots Y_k)$

On prend $H = \mathbf{R}^k$. On peut munir $\mathbf{R}^k$ du produit scalaire $\langle x, y \rangle = x'\Lambda^{-1}y$ où Λ est la matrice de covariance des Y_i (cf. [6]).

La distance entre deux éléments de $\mathbf{R}^k$ est alors :

$$d^2(x, y) = (x - y)' \Lambda^{-1} (x, - y).$$

Les définitions du paragraphe I peuvent alors se généraliser à ce cas particulier.

BIBLIOGRAPHIE

[1] Benzecri (J. P.), Analyse factorielle des correspondances 5e leçon, Publication I.S.U.P.

[2] Bertier (P.) et Boutin (A. M.), La structuration des données, *METRA*, vol. 8, n° 3, septembre 1969.

[3] Cellard (J. C.), Labbé (B.) et Savitsky (G.), Le programme Élisée, présentation et application, *METRA*, vol. 6, n° 3, septembre 1967.

[4] Cramer (H.), Mathematical methods of statistics, Princeton University Press, 1946.

[5] Morgan (J. N.) and Sonquist (J. A.), Problems in the analysis of survey data, and a proposal, *J.A.S.A.*, vol. 58, n° 302, June 1963.

[6] Vo Khac Kh et Nghiem (Ph. T.), Étude sur les aspects théoriques et pratiques de la segmentation aux moindres carrés, *R.I.R.O.*, 2e année, n° 8, 1968.