RAIRO. RECHERCHE OPÉRATIONNELLE

PH. MEUNIER
E. DIDAY
J. P. RASSON

Méthode et algorithme de sélection typologique de paramètres

RAIRO. Recherche opérationnelle, tome 19, n° 4 (1985), p. 351-373

<http://www.numdam.org/item?id=RO_1985__19_4_351_0>

R.A.I.R.O. Recherche opérationnelle/Operations Research
(vol. 19, n° 4, novembre 1985, p. 351 à 373)

MÉTHODE ET ALGORITHME DE SÉLECTION TYPOLOGIQUE DE PARAMÈTRES (*)

par Ph. MEUNIER (1), E. DIDAY (2) et J. P. RASSON (1)

Résumé. — *Nous disposons d'un ensemble de N individus caractérisés par $(p+1)$ variables qualitatives nominales. p d'entre elles sont dites « explicatives »; la $(p+1)$-ième est appelée variable « à expliquer » et est notée $X^{(p+1)}$. Les méthodes de segmentation classiques essaient de résoudre le problème suivant : « expliquer » $X^{(p+1)}$ à l'aide des partitions induites par chaque variable explicative. Nous proposons une méthode qui tente de résoudre le même problème mais dans un cadre plus général en laissant plus de souplesse dans le choix des partitions. Elle sélectionne simultanément la (ou les) variable(s) explicative(s) et la partition de l'ensemble des individus en un nombre fixé de classes (non nécessairement 2) afin d'optimiser un critère qui mesure l'adéquation entre la partition, la (ou les) variable(s) explicative(s) sélectionnées(s) et la variable à expliquer. Dans cet article, nous établissons le lien existant entre notre méthode et la segmentation classique. De plus, nous montrons comment nous pouvons sélectionner plusieurs variables explicatives identiques ou non sur chaque classe de la partition.*

Mots clés : Sélection de variables; segmentation; variables explicatives; variable à expliquer; données nominales.

Abstract. — *We have a set of N objects characterized by $(p+1)$ categorized variables. p variables are said "explanatory" variables and the last one is called the "dependent" variable which is denoted by $X^{(p+1)}$. The methods of segmentation tries to solve the following problem: "to explain" $X^{(p+1)}$ with the help of partitions which are induced by each explanatory variable. We propose a method which tries to solve the same problem, but it's in a more general case because the partitions are not necessary induced by explanatory variables. Our algorithm based on this method simultaneously selects the explanatory variable(s) and the partition of the set of objects into a fixed number of classes (not necessary 2 classes) so as to optimize a criterion which expresses the fit between the selected variable(s), the partition and the dependent variable. In this paper, we establish the bond between our method and the segmentation. More, we show how can we select a few explanatory variables and how can we select explanatory variables which are different on each class.*

Keywords: Selection of variables; segmentation; explanatory variables; dependent variable; categorical data.

(*) Reçu août 1984.
(1) F.N.D.P., Département de Mathématiques, Rempart de la Vierge, 8, B-5000 Namur, Belgique.
(2) I.N.R.I.A., Domaine de Voluceau, Rocquencourt, B.P. n° 105, 78150 Le Chesnay, France.

R.A.I.R.O. Recherche opérationnelle/Operations Research, 0399-0559/85/04 351 23/$ 4.30

1. INTRODUCTION

Nous disposons d'un ensemble $\mathbb{E}$ de N individus caractérisés par $(p+1)$ variables qualitatives nominales :

$$\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^p, \mathbf{X}^{(p+1)}.$$

Lorsqu'on fait une analyse statistique d'un tel tableau de données, on est souvent amené à étudier les liaisons entre une variable $\mathbf{X}^{(p+1)}$ et un ensemble de variables $\mathbf{X}^j$ $(1 \leqq j \leqq p)$, l'objectif étant « d'expliquer » au mieux $\mathbf{X}^{(p+1)}$ à partir des p variables $\mathbf{X}^j$, considérées comme variables « explicatives ». Selon la nature des variables de chacun des 2 types, la méthode d'analyse des données utilisée sera différente. Il est donc intéressant d'étudier les techniques utilisées lorsque toutes les variables qui interviennent sont qualitatives, ces variables occupant une place de plus en plus importante dans le domaine des sciences humaines, du marketing, de la biologie, de la zoologie, de l'archéologie, de la botanique, de la médecine, de la psychiatrie, de l'agronomie, etc. Les méthodes de segmentation font partie de ces techniques.

Ainsi, dans les méthodes de segmentation classiques, on cherche d'abord les partitions induites par chaque variable explicative puis, parmi ces partitions, on retient celle qui donne la « meilleure » valeur au critère mesuré sur la variable à expliquer (*cf.* [1, 3, 5, 6, 13, 17, 19]).

Dans le but d'avoir une méthode répondant mieux au désir de l'utilisateur, notre méthode de sélection de variables (*cf.* [8]) cherche simultanément la partition et les variables les plus explicatives de $\mathbf{X}^{(p+1)}$ en termes de classification automatique où le critère à optimiser peut s'écrire sous la forme :

$$W(\mathbf{P}, \mathbf{X}^j, \mathbf{X}^{(p+1)}) = Q_1 W_1(\mathbf{P}, \mathbf{X}^j) + Q_2 W_2(\mathbf{P}, \mathbf{X}^{(p+1)}),$$

où :

$Q_1 W_1(\mathbf{P}, \mathbf{X}^j)$,	mesure l'adéquation entre la partition $\mathbf{P}$ et la variable explicative $\mathbf{X}^j$;
$Q_2 W_2(\mathbf{P}, \mathbf{X}^{(p+1)})$,	mesure l'adéquation entre la partition $\mathbf{P}$ et la variable à expliquer $\mathbf{X}^{(p+1)}$;
$\mathbf{P}$,	est une partition à k classes du segment étudié;
Q_1 et Q_2,	sont des poids qui déterminent l'importance de W_1 et de W_2 dans l'expression du critère.

Notre modélisation (*cf.* [7, 8, 9]) en termes d'optimisation de ce critère nous a permis de traiter les problèmes que résolvent les méthodes de segmentation classiques mais en laissant plus de souplesse dans le choix des partitions.

Cette modélisation nous a également permis de généraliser la méthode, aux cas suivants :

– on peut sélectionner plusieurs variables explicatives sur chaque classe de la partition obtenue lors de la « segmentation » d'un segment donné;

– la (ou les) variable(s) explicative(s) sélectionnées(s) n'est (ne sont) pas nécessairement identique(s) sur chacune des classes de cette partition.

Afin de faciliter la description de notre méthode, nous supposerons dans un premier temps que le nombre de variables explicatives sélectionnées sur chaque classe de la partition est réduit à 1, la variable explicative associée à chaque classe pouvant être identique ou non sur chacune des classes. La généralisation de la méthode au cas de plusieurs variables explicatives sélectionnées sur chaque classe de la partition se fait aisément.

Dans le but de mettre en évidence l'intérêt de cette nouvelle approche, nous présentons dans le cadre de cette introduction un petit exemple de données « artificielles » (cf. *fig.* 1.1, p.) que nous avons traitées avec un algorithme de segmentation classique et avec notre algorithme de sélection typologique de variables.

1.	1.	1.	1.	1.
1.	2.	2.	2.	1.
1.	3.	1.	3.	1.
2.	4.	3.	1.	1.
2.	3.	3.	1.	1.
1.	2.	2.	1.	2.
2.	2.	3.	2.	2.
3.	2.	1.	3.	2.
1.	2.	1.	1.	2.
3.	1.	3.	1.	3.
1.	2.	3.	2.	3.
2.	3.	3.	2.	3.

Figure 1.1. – Tableau de données artificielles.

On peut trouver un arbre de segmentation fourni par la méthode de segmentation classique E.L.I.S.E.E. à la figure 1.2, p. .

On peut également voir à la figure 1.3, p. , l'arbre de segmentation fourni par notre programme.

Lors de la description de notre méthode, nous utiliserons le formalisme des nuées dynamiques préconisées par E. Diday (*cf.* [7, 8]). Ce formalisme nécessite la définition de différents concepts tels que mesure de ressemblance,

critère à optimiser, fonction d'affectation et fonction de représentation. De plus, nous sommes amenés à introduire quelques notations que nous décrivons dans le paragraphe suivant.

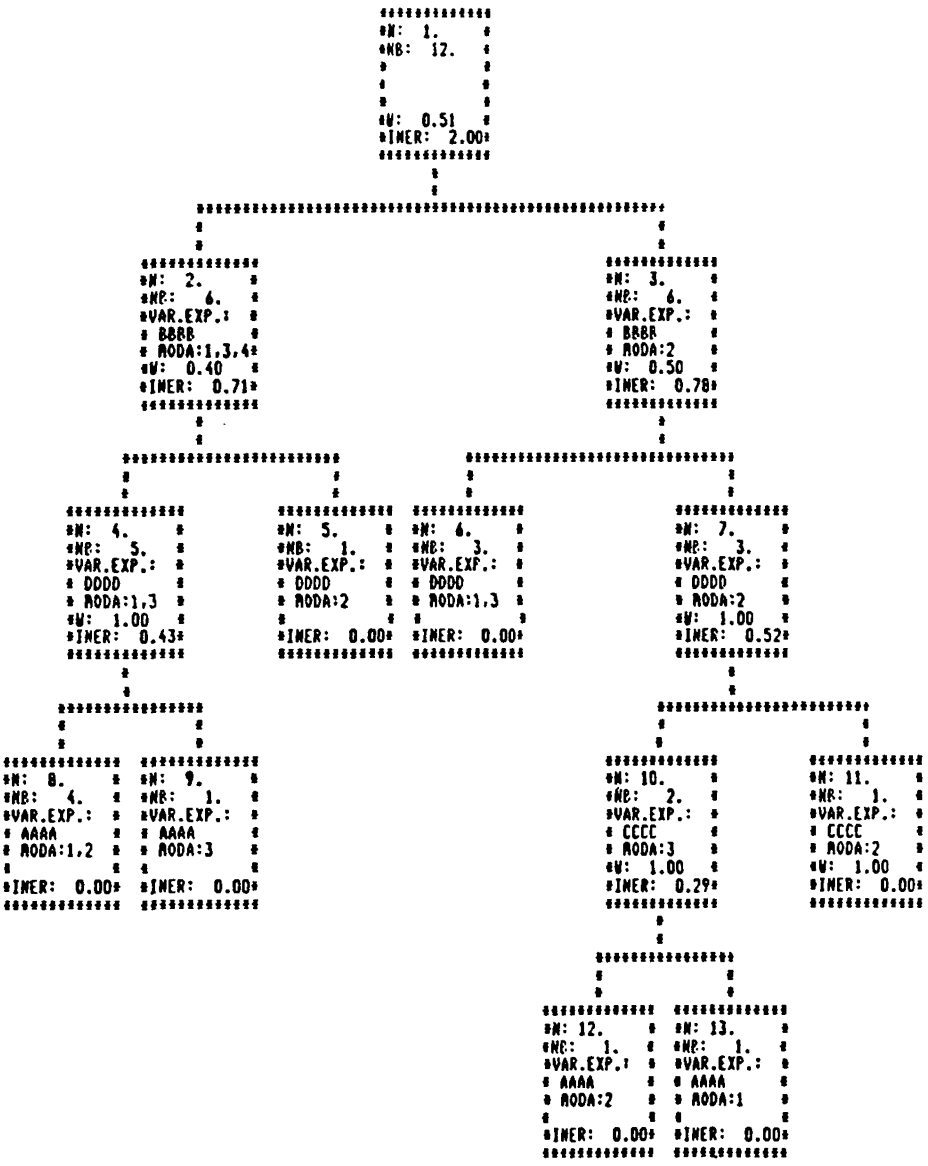

Figure 1.2. — Arbre de segmentation fourni par E.L.I.S.E.E.

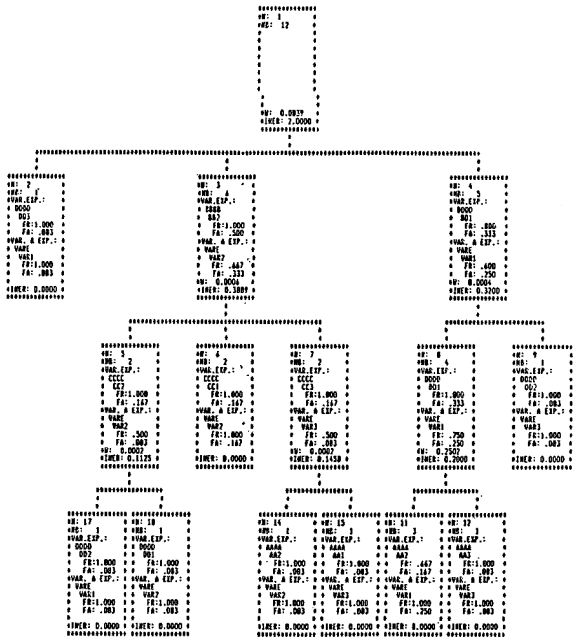

Figure 1.3. – Arbre de segmentation fourni par notre algorithme.

2. NOTATIONS

Pour rappel, nous disposons d'un tableau de données :

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

Nous notons :

$$\mathbf{X}_i=(x_i^1,\ \ldots,\ x_i^{(p+1)})^t \quad \text{le } i\text{-ième individu,}$$
$$\mathbf{X}^j=(x_1^j,\ \ldots,\ x_N^j)^t \quad \text{la } j\text{-ième variable.}$$

Nous notons également :

$\mathbb{E}=\{\mathbf{X}_i : 1\leqq i\leqq N\}$, ensemble des individus;

$J_1=\{\mathbf{X}^j : 1\leqq j\leqq p\}$, ensemble des variables explicatives;

$J_2=\{\mathbf{X}^{(p+1)}\}$, ensemble comprenant la variable à expliquer;

$\mathbb{P}$, ensemble des parties de $\mathbb{E}$;

$[Q]$, ensemble des indices des éléments de $\mathbb{E}$ qui appartiennent à Q ($Q\in\mathbb{P}$);

$\mathbb{P}_k$, ensemble des partitions à k classes de $\mathbb{E}$;

A, ensemble dépendant de J_2; cet ensemble correspond à l'espace des centres de gravité de sous-ensembles de $\mathbb{E}$, relativement à la variable à expliquer;

B, ensemble dépendant de J_1; cet ensemble correspond à l'espace des centres de gravité de sous-ensembles de $\mathbb{E}$, relativement à l'ensemble des variables explicatives;

$\mathbb{L} = J_1 \times A \times B$, espace de représentation d'une classe;

$\mathbb{L}_k = [\mathbb{L}]^k$, espace de représentation d'une partition à k classes de $\mathbb{E}$;

$]k] = \{1, \ldots, k\}$, ensembles des k premiers entiers.

De plus, à chaque individu, nous pouvons associer un poids; nous pouvons ainsi considérer l'espace probabilisé suivant : $(\mathbb{E}, \mathbb{P}, p)$ où p est une mesure de probabilité sur $(\mathbb{E}, \mathbb{P})$. Nous notons $p(\{\mathbf{X}_i\}) = p_i$. En fait, nous avons choisi $p_i = 1/N$.

La description des notations étant terminée, nous nous proposons de décrire notre méthode ainsi que l'algorithme.

3. MÉTHODE ET ALGORITHME

3.1. Définitions

Comme nous l'avons déjà dit précédemment, le formalisme des méthodes des nuées dynamiques nécessite la définition de différents concepts tels que :

- mesure de ressemblance;
- critère à optimiser;
- fonction d'affectation;
- fonction de représentation.

3.1.1. *Mesure de ressemblance*

Il nous faut définir une « quantité numérique » qui va nous permettre de mesurer l'adéquation entre une classe d'une partition et sa représentation. Comme nous le verrons lors de la définition de la fonction de représentation, la représentation d'une classe sera composée d'un ensemble de variables explicatives sélectionnées sur celle-ci et de ses centres de gravité respectivement par rapport aux variables explicatives sélectionnées et par rapport à la variable à expliquer.

Remarque : Dans la suite de l'article, nous entendrons par centre de gravité d'un groupe d'individus par rapport à un ensemble de variables qualitatives

qualitatives nominales, le vecteur « moyenne » du groupe par rapport à cet ensemble de variables qualitatives lesquelles ont été préalablement codées de façon disjonctive complète.

Cette « quantité numérique » est appelée mesure de ressemblance et est définie comme suit :

$$D: \quad \mathbb{P} \times \mathbb{L} \to \mathbb{R}^+,$$
$$(P, L) \mapsto D(P, L) = \sum_{i \in [P]} D(\{\mathbf{X}_i\}, L),$$

où :

$$L = (\mathbf{X}^j, a, b),$$
$$D(\{\mathbf{X}_i\}, L) = Q_1 D_1(\{\mathbf{X}_i\}, (\mathbf{X}^j, b)) + Q_2 D_2(\{\mathbf{X}_i\}, a),$$
$$D_1: \quad \mathbb{P} \times (J_1 \times B) \to \mathbb{R}^+,$$
$$(P, (\mathbf{X}^j, b)) \mapsto D_1(P, (\mathbf{X}^j, b))$$
$$= \sum_{i \in [P]} D_1(\{\mathbf{X}_i\}, (\mathbf{X}^j, b))$$
$$= \sum_{i \in [P]} p_i \{ d_j^2(\mathbf{X}_i, b) - ((m_j - 1) - M_1) \}$$
$$= p(P)(M_1 - (m_j - 1)) + \sum_{i \in [P]} p_i d_j^2(\mathbf{X}_i, b).$$

La mesure de ressemblance D_1 dépend uniquement de la variable explicative sélectionnée :

$$D_2: \quad \mathbb{P} \times A \to \mathbb{R}^+,$$
$$(P, a) \mapsto D_2(P, a) = \sum_{i \in [P]} D_2(\{\mathbf{X}_i\}, a)$$
$$= \sum_{i \in [P]} p_i \{ d_{(p+1)}^2(\mathbf{X}_i, a) - ((m_{(p+1)} - 1) - m_{(p+1)}) \}$$
$$= p(P) + \sum_{i \in [P]} p_i d_{(p+1)}^2(\mathbf{X}_i, a).$$

La mesure de ressemblance D_2 dépend uniquement de la variable à expliquer.

$M_1 = \sum_{j=1}^{p} m_j$, somme du nombre de modalités de chaque variable explicative;

m_j, nombre de modalités de la variable $\mathbf{X}^j$ $(1 \leqq j \leqq p+1)$.

3.1.2. *Critère à optimiser*

Dans notre méthode, le critère mesure d'une part l'adéquation entre la partition et l'ensemble des variables explicatives sélectionnées, d'autre part l'adéquation entre la partition et la variable à expliquer.

Nous définissons le critère comme suit :

$$W: \quad \mathbb{P}_k \times \mathbb{L}_k \to \mathbb{R}^+,$$

$$(\mathbf{P}, \mathbf{L}) \mapsto W(\mathbf{P}, \mathbf{L}) = \sum_{i=1}^{k} D(P_i, L_i),$$

$$\text{avec :} \quad \mathbf{P} = (P_1, \ldots, P_k),$$

$$\mathbf{L} = (L_1, \ldots, L_k).$$

3.1.3. *Fonction d'affectation*

Disposant d'une partition de l'espace des individus, comment peut-on la modifier en optimisant le critère que nous venons de définir ?

L'objet de la fonction d'affectation est de répondre à cette question.

La fonction d'affectation est définie comme suit :

$$F: \quad \mathbb{P}_k \times \mathbb{L}_k \to \mathbb{P}_k,$$

$$(\mathbf{P}', \mathbf{L}) \mapsto F(\mathbf{P}', \mathbf{L}) = \mathbf{P},$$

$$\text{avec :} \quad \mathbf{L} = (L_1, \ldots, L_k),$$

$$\mathbf{P}' = (P'_1, \ldots, P'_k),$$

$$\mathbf{P} = (P_1, \ldots, P_k),$$

où la classe P_i est définie comme suit :

$$P_i = \{\mathbf{X}_t \mid D(\{\mathbf{X}_t\}, L_i) < D(\{\mathbf{X}_t\}, L_j), j \neq i\}$$
$$\cup \{\mathbf{X}_t \mid I'_i \neq \varnothing, \mathbf{X}_t \in P'_i\}$$
$$\cup \{\mathbf{X}_t \mid I'_i \neq \varnothing, i < \text{Min}\{j \mid j \in I'_i\}, \mathbf{X}_t \notin P'_i, j \in I'_i\},$$

avec $I'_i = \{j \in]k] \setminus \{i\} \mid D(\{\mathbf{X}_t\}, L_i) = D(\{\mathbf{X}_t\}, L_j)\}$.

Le deuxième terme de l'union correspond à la règle d'affectation suivante : un individu ne change de classe que si cela « améliore » strictement le critère. Le dernier terme quant à lui correspond à une règle d'affectation arbitraire qui laisse statu-quo le critère.

3.1.4. *Fonction de représentation*

Disposant d'une partition de l'espace des individus, nous définissons un « objet » qui « représente bien » chacune des classes de la partition. Dans notre cas, le représentant d'une classe sera composé d'une variable explicative et de ses centres de gravité par rapport à la variable explicative sélectionnée et par rapport à la variable à expliquer.

L'objet de la fonction de représentation étant décrit, nous pouvons la définir plus formellement :

$$G : \mathbb{P}_k \times \mathbb{L}_k \to \mathbb{L}_k,$$
$$(\mathbf{P}, \mathbf{L}') \mapsto G(\mathbf{P}, \mathbf{L}') = \mathbf{L},$$

avec : $\mathbf{P} = (P_1, \ldots, P_k)$,
$\mathbf{L}' = (L'_1, \ldots, L'_k)$,
$\mathbf{L} = (L_1, \ldots, L_k)$,

où : $L'_i = (\mathbf{X}^{j'}, a'_i, b'_i)$,
$L_i = (\mathbf{X}^{j^*}, a_i, b_i)$,

où a_i, centre de gravité de P_i par rapport à J_2; b_i, centre de gravité de P_i par rapport à la variable explicative sélectionnée $\mathbf{X}^{j^*}$; $\mathbf{X}^{j^*}$, solution du problème suivant :

$$\text{Max}\{.(2\mathbf{X}^2(\mathbf{P}, \mathbf{X}^j)^{1/2} - (2(m_j - 1)(k-1) - 1)^{1/2}\},$$

sous les contraintes :

$$\sum_{i=1}^{k} D_1(P_i, (\mathbf{X}^j, b_i)) \leqq \sum_{i=1}^{k} D_1(P_i, (\mathbf{X}^{j'}, b'_i)), \qquad \mathbf{X}^j \in J_1,$$

où $\mathbf{X}^2(\mathbf{P}, \mathbf{X}^j)$ représente la valeur du chi-deux de contingence associé à la table de contingence définie par le croisement de la partition $\mathbf{P}$ et les modalités de la variable $\mathbf{X}^j$.

Si nous développons un tant soit peu l'expression :

$$\sum_{i=1}^{k} D_1(P_i, (\mathbf{X}^j, b_i)),$$

nous remarquons que l'on a l'égalité suivante :

$$\mathbf{X}^2(\mathbf{P}, \mathbf{X}^j) = N\left(M_1 - \sum_{i=1}^{k} D_1(P_i, (\mathbf{X}^j, b_i))\right).$$

Les concepts ainsi définis nous permettent de décrire très simplement notre méthode de sélection de variables.

3.2. Description de l'algorithme

Nous allons décrire à la figure 3.1 notre algorithme sous la forme d'un organigramme structuré (*cf.* [11]).

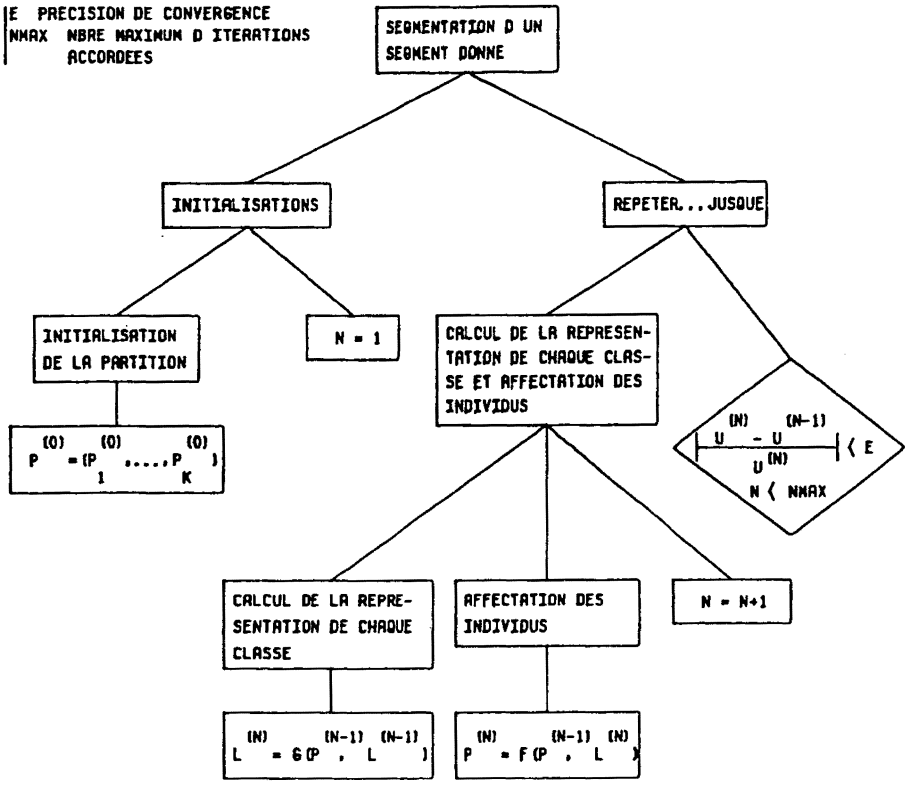

Figure 3.1. — Organigramme structuré.

Remarque : Lors du calcul de la fonction de représentation, la première contrainte assure la convergence de l'algorithme. Nous pouvons constater aisément qu'elle n'est pas utile voire même non définie lors de la première itération de l'algorithme.

3.3. Convergence de l'algorithme

Afin de démontrer la convergence de notre algorithme, nous sommes amenés à définir les suites $(u^{(n)})$, $(v^{(n)})$ comme suit :

$$-v^{(n)} = (\mathbf{P}^{(n)}, \mathbf{L}^{(n)}),$$

$$\text{avec : } \mathbf{L}^{(n)} = G(\mathbf{P}^{(n-1)}, \mathbf{L}^{(n-1)}),$$

$$\mathbf{P}^{(n)} = F(\mathbf{P}^{(n-1)}, \mathbf{L}^{(n)});$$

$$-u^{(n)} = W(v^{(n)}).$$

Il nous suffit de montrer que les suites $(u^{(n)})$, $(v^{(n)})$ sont convergentes.

PROPOSITION 1 : *La suite $(u^{(n)})$ converge en décroissant.*

Démonstration : Notons $Y^{(n)} = W(\mathbf{P}^{(n-1)}, \mathbf{L}^{(n)})$.

Nous allons montrer que :

$$u^{(n-1)} \geqq Y^{(n)} \geqq u^{(n)}.$$

Nous aurons ainsi montré que la suite $(u^{(n)})$ est décroissante. Or, elle est minorée par 0; elle est donc convergente.

On a :

$$\mathbf{L}^{(n)} = G(\mathbf{P}^{(n-1)}, \mathbf{L}^{(n-1)}).$$

D'où :

$$\sum_{i=1}^{k} D_1(P_i^{(n-1)}, (\mathbf{X}^{j(n)}, b_i^{(n)})) \leqq \sum_{i=1}^{k} D_1(P_i^{(n-1)}, (\mathbf{X}^{j(n-1)}, b_i^{(n-1)}))$$

et

$$\sum_{i=1}^{k} D_2(P_i^{(n-1)}, a_i^{(n)}) \leqq \sum_{i=1}^{k} D_2(P_i^{(n-1)}, a_i^{(n-1)}).$$

Autrement dit, nous avons :

$$\sum_{i=1}^{k} D(P_i^{(n-1)}, L_i^{(n)}) \leqq \sum_{i=1}^{k} D(P_i^{(n-1)}, L_i^{(n-1)})$$

et donc :

$$Y^{(n)} \leqq u^{(n-1)}. \tag{1}$$

Développons l'expression de $Y^{(n)}$:

$$Y^{(n)} = W(\mathbf{P}^{(n-1)}, \mathbf{L}^{(n)}) = \sum_{i=1}^{k} \sum_{t \in [P_i^{(n-1)}]} D(\{\mathbf{X}_t\}, L_i^{(n)}).$$

Soient i, t tels que $1 \leqq i \leqq k$ et $t \in [P_i^{(n)}]$.

Deux cas peuvent se présenter :

1. $t \in [P_i^{(n-1)}]$.

Dans ce cas, les termes dépendant de l'individu $\mathbf{X}_t$ dans les expressions de $Y^{(n)}$ et de $u^{(n)}$ sont identiques à $D(\{\mathbf{X}_t\}, L_i^{(n)})$.

2. $t \in [P_j^{(n-1)}]$.

Le terme dépendant de l'individu $\mathbf{X}_t$ dans l'expression de $Y^{(n)}$ (respectivement $u^{(n)}$) est $D(\{\mathbf{X}_t\}, L_j^{(n)})$ (respectivement $D(\{\mathbf{X}_t\}, L_i^{(n)})$).

Or $\mathbf{P}^{(n)} = F(\mathbf{P}^{(n-1)}, \mathbf{L}^{(n)})$.

D'où $D(\{\mathbf{X}_t\}, L_i^{(n)}) \leqq D(\{\mathbf{X}_t\}, L_j^{(n)})\}$.

Quelque soit le cas dans lequel on se trouve, nous avons donc :

$$D(\{\mathbf{X}_t\}, L_i^{(n)}) \leqq D(\{\mathbf{X}_t\}, L_j^{(n)}).$$

Nous en désuisons que :

$$u^{(n)} \leqq Y^{(n)}. \tag{2}$$

(1) et (2) assurent la décroissance de la suite $(u^{(n)})$.

PROPOSITION 2 : *La suite* $(v^{(n)})$ *est convergente.*

Démonstration : Nous avons $v^{(n)} = (\mathbf{P}^{(n)}, \mathbf{L}^{(n)})$.

Il est évident que si la suite des partitions $(\mathbf{P}^{(n)})$ est finalement stationnaire, la suite des « représentations » l'est également (*cf.* définition de la fonction de représentation G). Il nous suffit donc de montrer que la suite $(\mathbf{P}^{(n)})$ est finalement stationnaire.

La suite $(u^{(n)})$ étant définie sur un ensemble fini, atteint sa limite.

D'où $\exists No \in \mathbb{N} : u^{(No)} = u^{(No+1)}$.

Il suit que $Y^{(No+1)} = u^{(No+1)}$.

Supposons par l'absurde que $\mathbf{P}^{(No)} \neq \mathbf{P}^{(No+1)}$.

D'où $\exists i \in]k]$, $\exists \mathbf{X}_t \in P_i^{(No)} : \mathbf{X}_t \notin P_i^{(No+1)}$.

Autrement dit :

$$\exists i \in]k], \quad \exists j \in]k] \setminus \{i\}, \quad \exists \mathbf{X}_t \in P_i^{(No)} : \mathbf{X}_t \in P_j^{(No+1)}.$$

Or $\mathbf{P}^{(No+1)} = F(\mathbf{P}^{(No)}, \mathbf{L}^{(No+1)})$.

D'où :

$$D(\{\mathbf{X}_t\}, L_j^{(No+1)}) < D(\{\mathbf{X}_t\}, L_i^{(No+1)}),$$

avec :

$D(\{\mathbf{X}_t\}, L_j^{(No+1)})$ est le terme dépendant de l'individu $\mathbf{X}_t$ dans l'expression de $u^{(No+1)}$;

$D(\{\mathbf{X}_t\}, L_i^{(No+1)})$ est le terme dépendant de l'individu $\mathbf{X}_t$ dans l'expression de $Y^{(No+1)}$.

Nous en concluons que :

$$u^{(No+1)} < Y^{(No+1)} \quad \text{(absurde)}.$$

D'où la suite $(v^{(n)})$ étant finalement stationnaire est convergente.

Nous venons de traiter le cas d'une seule variable explicative. De plus, c'est la même variable qui a été sélectionnée sur les différentes classes de la partition considérée.

Le cas de plusieurs variables explicatives sélectionnées identiques et le cas de variables explicatives sélectionnées pouvant être différentes sur chacune des classes de la partition, nécessitent une modification de l'espace de représentation ainsi que de la fonction de représentation. Le traitement de ces différents cas fait l'objet du paragraphe suivant.

3.4. Autres cas

3.4.1. *Identité des variables explicatives sélectionnées*

Soit r le nombre de variables explicatives que l'on désire sélectionner. L'espace de représentation d'une classe sera défini de la façon suivante :

$$\mathbb{L} = J_1^r \times A \times B^r.$$

Quant à la fonction de représentation, nous la définissons comme suit :

$$G : \mathbb{P}_k \times \mathbb{L}_k \to \mathbb{L}_k,$$
$$(\mathbf{P}, \mathbf{L}') \mapsto G(\mathbf{P}, \mathbf{L}') = \mathbf{L},$$

avec :

$$\mathbf{P} = (P_1, \ldots, P_k),$$
$$\mathbf{L}' = (L_1', \ldots, L_k'), \quad \text{où} \quad L_i' = (\mathbf{X}^{j_1'}, \ldots, \mathbf{X}^{j_r'}, a_i', b_i^{1'}, \ldots, b_i^{r'}),$$
$$\mathbf{L} = (L_1, \ldots, L_k), \quad \text{où} \quad L_i = (\mathbf{X}^{j_1}, \ldots, \mathbf{X}^{j_r}, a_i, b_i^1, \ldots, b_i^r),$$

avec a_i, centre de gravité de P_i par rapport à J_2; b_i^m, centre de gravité de P_i par rapport à la variable explicative sélectionnée $\mathbf{X}^{j_m}$ $(1 \leqq m \leqq r)$; $\mathbf{X}^{j_m}$, solution du problème suivant :

$$\text{Max}\{(2\mathbf{X}^2(\mathbf{P}, \mathbf{X}^j)^{1/2} - (2(m_j - 1)(k-1) - 1)^{1/2}\},$$

sous les contraintes :

$$\sum_{i=1}^{k}\left\{\sum_{t=1}^{m-1} D_1(P_i, (\mathbf{X}^{j_t}, b_i^t)) + D_1(P_i, (\mathbf{X}^j, b_i))\right\} \leqq \sum_{i=1}^{k}\sum_{t=1}^{m} D_1(P_i, (\mathbf{X}^{j_{t'}}, b_i^{t'})),$$

$$\mathbf{X}^j \in J_1 \setminus \{\mathbf{X}^{j_1}, \ldots, \mathbf{X}^{j_{m-1}}\} \qquad (1 \leqq 1 \leqq r).$$

3.4.2. *Non identité de la variable explicative sélectionnée*

Dans ce cas, seule la fonction de représentation est modifiée :

$$G : \mathbb{P}_k \times \mathbb{L}_k \to \mathbb{L}_k,$$
$$(\mathbf{P}, \mathbf{L}') \mapsto G(\mathbf{P}, \mathbf{L}') = \mathbf{L},$$

avec :

$$\mathbf{P} = (P_1, \ldots, P_k),$$
$$\mathbf{L}' = (L_1', \ldots, L_k'), \qquad \text{où} \quad L_i' = (\mathbf{X}^{j_i'}, a_i', b_i'),$$
$$\mathbf{L} = (L_1, \ldots, L_k), \qquad \text{où} \quad L_i = (\mathbf{X}^{j_i}, a_i, b_i),$$

avec a_i, centre de gravité de P_i par rapport à J_2; b_i, centre de gravité de P_i par rapport à la variable explicative sélectionnée sur cette classe, X^{j_i}; $\mathbf{X}^{j_i}$, solution du problème suivant :

$$\text{Max}\{(2\mathbf{X}^2(P_i, P_i^c, \mathbf{X}^j))^{1/2} - (2(m_j - 3))^{1/2}\},$$

sous les contraintes :

$$\sum_{i=1}^{k} D_1(P_i, (\mathbf{X}^j, b_i)) \leqq \sum_{i=1}^{k} D_1(P_i, (\mathbf{X}^{j_i'}, b_i')), \qquad \mathbf{X}^j \in J_1,$$

où $\mathbf{X}^2(P_i, P_i^c, \mathbf{X}^j)$, représente la valeur du chi-deux de contingence associé à la table de contingence définie par le croisement de la partition (P_i, P_i^c) et les modalités de la variable $\mathbf{X}^j$; $(1 \leqq i \leqq k)$.

Le cas de plusieurs variables explicatives sélectionnées pouvant être différentes sur chacune des classes de la partition, il se généralise de façon analogue au cas de plusieurs variables explicatives sélectionnées identiques.

3.5. Choix du nombre (k) de classes de la partition

Nous avons remplacé la contrainte des dichotomies par une contrainte de partitions à k classes ($k \geqq 2$, fixé *a priori*). Cette idée est de loin très légitime. En effet, en paraphrasant Michaud (*cf.* [15]), « dans le cas où la variable à expliquer est à 2 modalités, l'hypothèse de dichotomisation peut sembler raisonnable mais certainement pas dans le cas contraire. Si la variable à expliquer est par exemple à 4 modalités on ne voit vraiment pas pourquoi on imposerait à la variable explicative sélectionnée d'être après structuration sous la forme d'une partition en 2 classes. Si on devait fixer un nombre de classes, celui de la variable à expliquer, ici 4, semblerait beaucoup plus logique. »

Comme nous le suggère Michaud (1983), nous prendrons k égal au nombre de modalités de la variable à expliquer.

Dans le paragraphe suivant, nous décrivons le lien existant entre notre méthode et la segmentation classique.

4. LIEN AVEC LA SEGMENTATION CLASSIQUE

Pour rappel, les méthodes de segmentation classiques cherchent d'abord les partitions induites par chaque variable explicative puis retiennent parmi ces partitions celle qui donne la meilleure valeur au critère mesuré sur la variable à expliquer (*cf.* [1, 3, 5, 6, 13, 17, 19]).

Soit $\mathbb{Q}_2$ l'ensemble des partitions à 2 classes issues de dichotomies de l'ensemble des modalités de chaque variable explicative.

La segmentation classique résout donc le problème suivant :

$$\text{Max}\{\Phi^2(\mathbf{P},\mathbf{X}^{(p+1)}) : \mathbf{P} \in \mathbb{Q}_2\}, \tag{P1}$$

où : $\Phi^2(\mathbf{P},\mathbf{X}^{(p+1)}) = \mathbf{X}^2(\mathbf{P},\mathbf{X}^{(p+1)})/N$,

N, nombre d'individus de la partition $\mathbf{P}$.

Le problème que résout notre méthode, dans le cas où $k=2$, peut s'écrire de la façon suivante :

$$\text{Min}\{W(\mathbf{P}, \mathbf{L}) : (\mathbf{P}, \mathbf{L}) \in \mathbb{P}_2 \times \mathbb{L}_2\}. \tag{P2}$$

Néanmoins le problème (P2) peut s'écrire de façon équivalente au problème (P3) :

$$\text{Max}\{Q_1 \Phi^2(\mathbf{P}, \mathbf{X}^j) + Q_2 \Phi^2(\mathbf{P}, \mathbf{X}^{(p+1)}) : (\mathbf{P}, \mathbf{X}^j) \in \mathbb{P}_2 \times J_1\}. \qquad \text{(P3)}$$

En effet :

$$W(\mathbf{P}, \mathbf{L}) = \sum_{i=1}^{2} D(P_i, L_i)$$

$$= Q_1 \sum_{i=1}^{2} \{ \sum_{t \in [P_i]} p_t d_j^2(\mathbf{X}_t, b_i) + p(P_i)(M_1 - (m_j - 1)) \}$$

$$+ Q_2 \sum_{i=1}^{2} \{ \sum_{t \in [P_i]} p_t d_{(p+1)}^2(\mathbf{X}_t, a_i) + p(P_i) \}$$

$$= Q_1 \left\{ \sum_{i=1}^{2} \sum_{t \in [P_i]} p_t d_j^2(\mathbf{X}_t, b_i) + M_1 - (m_j - 1) \right\}$$

$$+ Q_2 \left\{ \sum_{i=1}^{2} \sum_{t \in [P_i]} p_t d_{(p+1)}^2(\mathbf{X}_t, a_i) + 1 \right\}.$$

En utilisant le théorème de Huygens, nous pouvons écrire le critère sous la forme suivante :

$$W(\mathbf{P}, \mathbf{L}) = Q_1 M_1 + Q_2 m_{(p+1)} - \{Q_1 \Phi^2(\mathbf{P}, \mathbf{X}^j) + Q_2 \Phi^2(\mathbf{P}, \mathbf{X}^{(p+1)})\}.$$

Les problèmes (P2), (P3) sont équivalents.

Nous allons montrer comment notre modélisation en termes d'optimisation du critère W nous a permis de traiter les problèmes que résout la segmentation classique. Pour ce faire, nous allons montrer que les problèmes (P1), (P3) sont équivalents moyennant une condition sur les poids Q_1, Q_2.

La démonstration de cette équivalence se fera en trois étapes, chacune de ces étapes donnant naissance à une proposition.

PROPOSITION 3 : *Il existe des poids Q_1, Q_2 tels que :*

$$\text{Max}\{Q_1 \Phi^2(\mathbf{P}, \mathbf{X}^j) + Q_2 \Phi^2(\mathbf{P}, \mathbf{X}^{(p+1)}) : (\mathbf{P}, \mathbf{X}^j) \in \mathbb{P}_2 \times J_1\}$$
$$= \text{Max}\{Q_1 \Phi^2(\mathbf{P}, \mathbf{X}^j) + Q_2 \Phi^2(\mathbf{P}, \mathbf{X}^{(p+1)}) : (\mathbf{P}, \mathbf{X}^j) \in \mathbb{Q}_2 \times J_1\}.$$

Cette proposition signifie que la partition optimale, solution de notre problème (P3) est issue d'une dichotomie de l'ensemble des modalités d'une des variables explicatives.

PROPOSITION 4 : *Il existe des poids* Q_1, Q_2 *tels que : soit* $(\mathbf{P}^*, \mathbf{X}^{j^*})$ *solution du problème* (P3). *On a :*

$$\Phi^2(\mathbf{P}^*, \mathbf{X}^{(p+1)}) = \text{Max}\{\Phi^2(\mathbf{P}, \mathbf{X}^{(p+1)}) : \mathbf{P} \in \mathbb{Q}_2\}.$$

Cette proposition signifie que toute solution du problème (P3) est solution du problème (P1).

Quant à la proposition suivante, elle affirme la réciproque de la proposition 4. Toute solution du problème (P1) est solution du problème (P3).

PROPOSITION 5 : *Soit* $\mathbf{P}^*$ *solution du problème* (P1).

$\mathbf{P}^*$ *est donc issue d'une dichotomie de l'ensemble des modalités d'une des variables explicatives. Notons cette variable explicative* $\mathbf{X}^{j^*}$.

Il existe des poids Q_1, Q_2 *tels que :*

$$Q_1 \Phi^2(\mathbf{P}^*, \mathbf{X}^{j^*}) + Q_2 \Phi^2(\mathbf{P}^*, \mathbf{X}^{(p+1)})$$
$$= \text{Max}\{Q_1 \Phi^2(\mathbf{P}, \mathbf{X}^j) + Q_2 \Phi^2(\mathbf{P}, \mathbf{X}^{(p+1)}) : (\mathbf{P}, \mathbf{X}^j) \in \mathbb{P}_2 \times J_1\}.$$

Nous allons démontrer successivement les trois propositions que l'on vient d'énoncer.

Démonstration de la proposition 3 : Nous avons évidemment $\mathbb{Q}_2 \subset \mathbb{P}_2$.

Il s'en suit que :

$$\text{Max}\{Q_1 \Phi^2(\mathbf{P}, \mathbf{X}^j) + Q_2 \Phi^2(\mathbf{P}, \mathbf{X}^{(p+1)}) : (\mathbf{P}, \mathbf{X}^j) \in \mathbb{P}_2 \times J_1\}$$
$$\geqq \text{Max}\{Q_1 \Phi^2(\mathbf{P}, \mathbf{X}^j) + Q_2 \Phi^2(\mathbf{P}, \mathbf{X}^{(p+1)}) : (\mathbf{P}, \mathbf{X}^j) \in \mathbb{Q}_2 \times J_1\}.$$
$$\forall Q_1, Q_2 \in]0, 1[.$$

Nous allons imposer une condition sur les poids Q_1, Q_2 de telle sorte qu'il soit impossible d'obtenir une inégalité stricte.

Pour ce faire, nous considérons le problème suivant :

$$\text{Max}\{\Phi^2(\mathbf{P}, \mathbf{X}^{(p+1)})/(1 - \Phi^2(\mathbf{P}, \mathbf{X}^j) : (\mathbf{P}, \mathbf{X}^j) \in (\mathbb{P}_2 \setminus \mathbb{Q}_2) \times J_1\}.$$

Soit $(\mathbf{P}^{**}, \mathbf{X}^{j^{**}})$ une solution de ce problème.

Puisque $\mathbf{P}^{**} \in \mathbb{P}_2 \setminus \mathbb{Q}_2$, nous avons évidemment que $\Phi^2(\mathbf{P}^{**}, \mathbf{X}^{j^{**}}) < 1$.

De plus, en supposant par l'absurde que l'on a l'inégalité stricte dans la pénultième inégalité, nous pouvons montrer que $\Phi^2(\mathbf{P}^{**}, \mathbf{X}^{(p+1)}) > 0$.

Posons, en effet :

$$v = \Phi^2(\mathbf{P}^{**}, \mathbf{X}^{(p+1)})/(1 - \Phi^2(\mathbf{P}^{**}, \mathbf{X}^{j^{**}})).$$

Or $v > 0$ d'où $\exists Q_1, Q_2 \in]0, 1[: v = Q_1/Q_2$ et $Q_1 + Q_2 = 1$.

Sous l'hypothèse d'absurde, nous allons montrer :

$$\exists \mathbf{P}^* \in \mathbb{P}_2 \setminus \mathbb{Q}_2, \quad \exists \mathbf{X}^{j^*} \in J_1 : Q_1/Q_2 < \mathbb{O}^2(\mathbf{P}^*, \mathbf{X}^{(p+1)})/(1-\mathbb{O}^2(\mathbf{P}^*, \mathbf{X}^{j^*})).$$

Ce qui est absurde étant donnée la définition de **v**.

L'hypothèse d'absurde implique :

$$\exists \mathbf{X}^{j^*} \in J_1, \exists \mathbf{P}^* \in \mathbb{P}_2 \setminus \mathbb{Q}_2, \forall \mathbf{X}^j \in J_1, \forall \mathbf{P} \in \mathbb{Q}_2 :$$
$$Q_1 \mathbb{O}^2(\mathbf{P}^*, \mathbf{X}^{j^*}) + Q_2 \mathbb{O}^2(\mathbf{P}^*, \mathbf{X}^{(p+1)}) > Q_1 \mathbb{O}^2(\mathbf{P}, \mathbf{X}^j + Q_2 \mathbb{O}^2(\mathbf{P}, \mathbf{X}^{(p+1)}).$$

Or $\forall \mathbf{P} \in \mathbb{Q}_2, \exists \mathbf{X}^j \in J_1 : \Phi^2(\mathbf{P}, \mathbf{X}^j) = 1.$

D'où :

$$\forall \mathbf{P} \in \mathbb{Q}_2, \exists \mathbf{X}^{j^*} \in J_1 :$$
$$Q_1 \Phi^2(\mathbf{P}^*, \mathbf{X}^{j^*}) + Q_2 \Phi^2(\mathbf{P}^*, \mathbf{X}^{(p+1)}) > Q_1 + Q_2 \Phi^2(\mathbf{P}, \mathbf{X}^{(p+1)}).$$

Donc, $\exists \mathbf{X}^{j^*} \in J_1$:

$$Q_1(1-\Phi^2(\mathbf{P}^*, \mathbf{X}^{j^*})) < Q_2(\Phi^2(\mathbf{P}^*, \mathbf{X}^{(p+1)}) - \Phi^2(\mathbf{P}, \mathbf{X}^{(p+1)})).$$

De plus, $\mathbf{P}^* \notin \mathbb{Q}_2$, implique $\mathbb{Q}^2(\mathbf{P}^*, \mathbf{X}^{j^*}) < 1$.

D'où $\exists \mathbf{X}^{j^*} \in J_1$:

$$Q_1/Q_2 < (\Phi^2(\mathbf{P}^*, \mathbf{X}^{(p+1)}) - \Phi^2(\mathbf{P}, \mathbf{X}^{(p+1)}))/(1-\Phi^2(\mathbf{P}^*, \mathbf{X}^{j^*})).$$

Il s'en suit que :

$$\exists \mathbf{P}^* \in \mathbb{P}_2 \setminus \mathbb{Q}_2, \exists \mathbf{X}^{j^*} \in J_1 : Q_1/Q_2 < \Phi^2(\mathbf{P}^*, \mathbf{X}^{(p+1)})/(1-\Phi^2(\mathbf{P}^*, \mathbf{X}^{j^*})).$$

Ceci achève la démonstration de la proposition 3.

Démonstration de la proposition 4 : Soit $(\mathbf{P}^*, \mathbf{X}^{j^*}) \in \mathbb{Q}_2 \times J_1$ tel que :

$$Q_1 \Phi^2(\mathbf{P}^*, \mathbf{X}^{j^*}) + Q_2 \Phi^2(\mathbf{P}^*, \mathbf{X}^{(p+1)})$$
$$= \mathrm{Max}\{ Q_1 \Phi^2(\mathbf{P}, \mathbf{X}^j) + Q_2 \Phi^2(\mathbf{P}, \mathbf{X}^{(p+1)}) : (\mathbf{P}, \mathbf{X}^j) \in \mathbb{Q}_2 \times J_1 \}.$$

Or $\mathbf{P}^* \in \mathbb{Q}_2$ implique $\Phi^2(\mathbf{P}^*, \mathbf{X}^{j^*}) = 1$.

D'où :

$$Q_1 + Q_2 \Phi^2(\mathbf{P}^*, \mathbf{X}^{(p+1)})$$
$$= \mathrm{Max}\{ Q_1 \Phi^2(\mathbf{P}, \mathbf{X}^j) + Q_2 \Phi^2(\mathbf{P}, \mathbf{X}^{(p+1)}) : (\mathbf{P}, \mathbf{X}^j) \in \mathbb{Q}_2 \times J_1 \}.$$

Nous avons donc :

$$\Phi^2(\mathbf{P}^*, \mathbf{X}^{(p+1)}) = \text{Max}\{\Phi^2(\mathbf{P}, \mathbf{X}^{(p+1)}) : \mathbf{P} \in \mathbb{Q}_2\}.$$

Démonstration de la proposition 5 : Soit $\mathbf{P}^* \in \mathbb{Q}_2$ tel que :

$$\Phi^2(\mathbf{P}^*, \mathbf{X}^{(p+1)}) = \text{Max}\{\Phi^2(\mathbf{P}, \mathbf{X}^{(p+1)}) : \mathbf{P} \in \mathbb{Q}_2\}.$$

D'où $\exists \mathbf{X}^{j^*} \in J_1$, $\forall \mathbf{P} \in \mathbb{Q}_2$, $\forall \mathbf{X}^j \in J_1$:

$$Q_1 + Q_2 \Phi^2(\mathbf{P}^*, \mathbf{X}^{(p+1)}) \geqq Q_1 \Phi^2(\mathbf{P}, \mathbf{X}^j) + Q_2 \Phi^2(\mathbf{P}, \mathbf{X}^{(p+1)}).$$

Or $\mathbf{P}^* \in \mathbb{Q}_2$ implique $\Phi^2(\mathbf{P}^*, \mathbf{X}^{j^*}) = 1$.
D'où $\exists \mathbf{X}^{j^*} \in J_1$, $\forall \mathbf{P} \in \mathbb{Q}_2$, $\forall \mathbf{X}^j \in J_1$:

$$Q_1 \Phi^2(\mathbf{P}^*, \mathbf{X}^{j^*}) + Q_2 \Phi^2(\mathbf{P}^*, \mathbf{X}^{(p+1)}) \geqq Q_1 \Phi^2(\mathbf{P}, \mathbf{X}^j) + Q_2 \Phi^2(\mathbf{P}, \mathbf{X}^{(p+1)}).$$

Il s'en suit que :

$$Q_1 \Phi^2(\mathbf{P}^*, \mathbf{X}^{j^*}) + Q_2 \Phi^2(\mathbf{P}^*, \mathbf{X}^{(p+1)})$$
$$= \text{Max}\{Q_1 \Phi^2(\mathbf{P}, \mathbf{X}^j) + Q_2 \Phi^2(\mathbf{P}, \mathbf{X}^{(p+1)}) : (\mathbf{P}, \mathbf{X}^j) \in \mathbb{Q}_2 \times J_1\}.$$

La comparaison entre notre méthode de sélection typologique de variables et la segmentation classique vient d'être réalisée.

Nous allons consacrer le paragraphe suivant à la description du type de résultats fournis par notre programme.

5. PROGRAMME

Nous avons implémenté notre méthode de sélection typologique de variables dans le langage fortran sur l'ordinateur Dec2060 (Digital) des Facultés Notre-Dame-de-la-Paix à Namur (Belgique).

Néanmoins, nous avons implémenté le programme sous des contraintes de normalisation assez strictes (*cf.* [16]) de façon à accroître sa portabilité sur d'autres sites. Par exemple, il fonctionne sur l'ordinateur Multics de l'I.N.-R.I.A. (Institut National de Recherche en Informatique et en Automatisme) à Rocquencourt (France). Sur ce même site, nous l'avons également intégré dans un logiciel d'analyse des données appelé SICLA (*cf.* [18]).

Dans un avenir très proche, nous comptons également l'intégrer dans la bibliothèque fortran MODULAD (*cf.* [16]).

D'un point de vue temps-calcul, le programme a nécessité 2 mn 35 s sur le Dec2060 pour le traitement d'un tableau de données de 2088 individus

caractérisés par 38 variables qualitatives nominales comprenant la variable à expliquer.

Les résultats fournis par notre programme comprennent d'une part, une description de chaque segment, d'autre part, un arbre de segmentation qui offre à l'utilisateur une vue générale de la segmentation obtenue.

La description d'un segment se présente sous la forme suivante (cf. *fig.* 5.1, p. 370).

SEGMENT No: 10

EFFECTIF DU SEGMENT: 353 , 16.9%

VARIABLE(S) EXPLICATIVE(S)

VARIABLE	LIBELLE	MODALITE	LIBELLE	KHI2	EFFE	MC -- CL	MT -- N	MC -- MT
OSPO	OCCUPATIONS ET SPORTS	OV1	OUI	851.48	353	100.00	33.29	50.79

VARIABLE(S) EXPLICATIVE(S) SIGNIFICATIVE(S)

VARIABLE	LIBELLE	MODALITE	LIBELLE	KHI2	EFFE	MC -- CL	MT -- N	MC -- MT
SPOR	ACTIVITES SPORTIVES	SP3	REGULIER	332.23	281	79.60	36.93	36.45
MEAC	MEMBRE ACTIF	ME2	OUI	178.34	229	64.87	34.15	32.12

VARIABLE(S) A EXPLIQUER

VARIABLE	LIBELLE	MODALITE	LIBELLE	KHI2	EFFE	MC -- CL	MT -- N	MC -- MT
LOIS	LOISIRS	LOI9	LOISIRS 9	1506.00	294	83.29	15.37	91.59

VALEUR DE "L'HOMOGENEITE NORMALISEE": 0.0312

SEGMENT-PERE No: 1

VALEUR DU CRITERE: 0.0001

Figure 5.1. — Description d'un segment.

Nous précisons ce que signifient les paramètres MC, CL, MT, N à l'aide desquels nous avons défini les quantités présentes dans les 4 dernières colonnes des différents tableaux :

MC, effectif de la modalité considérée comme significative dans le segment;

CL, effectif su segment;

MT, effectif de cette modalité dans la population totale;

N, effectif de la population totale.

Le premier tableau correspond aux variables explicatives sélectionnées sur ce segment. Quant au deuxième tableau, il correspond aux variables explicatives significatives au sens du chi-deux sur ce segment.

Ce chi-deux est calculé sur le tableau de contingence obtenu en croisant le segment et son complémentaire par rapport à la population totale, avec les modalités de la variable explicative considérée.

Le troisième tableau fournit le même type d'informations à propos de la variable à expliquer.

Lors de l'élaboration de ces 3 tableaux, les modalités sélectionnées sont significatives au sens du chi-deux sur ce segment. Encore une fois, ce chi-deux est calculé sur le tableau de contingence obtenu en croisant le segment et son complémentaire, avec la modalité considérée et son complémentaire.

En outre, il est à remarquer que :

– l'« homogénéité normalisée » d'un segment n'est autre que le quotient de l'inertie de la variable à expliquer sur ce segment considéré comme sous-ensemble de la population totale par l'inertie de la variable à expliquer sur la population totale;

– si le segment est lui-même segmenté, nous fournissons également la valeur du critère à la convergence ainsi qu'une liste des numéros des différents segments-fils.

Quant à l'arbre de segmentation, il peut se présenter sous la forme suivante (cf. *fig.* 5.2, p. 371).

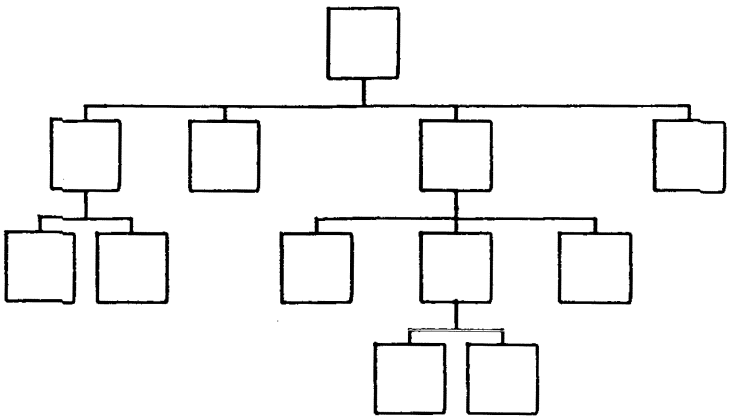

Figure 5.2. – Arbre de segmentation.

Une feuille de l'arbre est décrite comme suit :

N : 10	numéro du segment
NB : 353	effectif du segment
VAR. EXP. :	
OSPO	nom de la variable explicative sélectionnée
OV1	nom de sa modalité la plus fréquente
FR : 1.000	fréquence relative de cette modalité
FA : .169	fréquence absolue de cette modalité
VAR. A EXP. :	
LOIS	nom de la variable à expliquer
LOI9	nom de sa modalité la plus fréquente
FR : .833	fréquence relative de cette modalité
Fa : .141	fréquence absolue de cette modalité
W : 0.0001	valeur du critère à la convergence
INER : 0.0312	« homogénéité normalisée » du segment

Des sorties de ce type permettent à l'utilisateur, une interprétation plus aisée des résultats.

6. CONCLUSION

Notre modélisation en termes d'optimisation du critère W nous a permis d'étendre les méthodes de segmentation classiques au cas d'une ou plusieurs variables explicatives sélectionnées sur chaque classe d'une partition non nécessairement imposée par l'une des variables explicatives.

D'autre part, nous pouvons obtenir les mêmes variables explicatives sur chaque classe de la partition ou des variables différentes suivant l'option choisie par l'utilisateur dans le programme.

Sous la forme d'une « sortie-listing », un arbre de segmentation est fourni à l'utilisateur; ce qui lui permet d'avoir une vue plus générale et plus synthétique de la segmentation obtenue.

De façon à accroître la portabilité de notre programme sur différents sites, nous l'avons implémenté en fortran sous des contraintes de normalisation décrites dans [16]. Nous l'avons également intégré dans un logiciel d'analyse des données, (*cf.* [18]) qui a été implanté sur différents sites et notamment à l'I.N.R.I.A. Ceci permet à l'utilisateur de notre programme de bénéficier de tous les avantages qui lui sont offerts par ce logiciel.

Comme perspectives d'avenir, nous voyons le traitement de plusieurs variables à expliquer qualitatives ainsi que le traitement de variables à expliquer quantitatives.

REMERCIEMENTS

Nous remercions G. Saporta, B. Burtschy (Paris, France), G. Libert (Mons, Belgique) et A. Hardy (Namur, Belgique) pour les commentaires et les remarques qu'ils nous ont adressés à propos de ce travail.

BIBLIOGRAPHIE

1. A. Baccini, *Aspect synthétique de la segmentation et traitement de variables qualitatives à modalités ordonnées*, Thèse 3e cycle, Université Paul-Sabatier, Toulouse, 1975.
2. W. A. Belson, *Matching and Prediction on the Principle of Biological Classification*, Appl. Stat., vol. 8, 1980, p. 65-75.
3. J. M. Bouroche et M. Tenenhaus, *Quelques méthodes de segmentation*, R.I.R.O., vol. 2, 1970, p. 29-42.
4. F. Caillez et J. P. Pages, *Introduction à l'analyse des données*, Smash, Paris, 1976.
5. J. C. Cellard, B. Labbe et G. Savitsky, *Le programme E.L.I.S.E.E. — Présentation et applications*, METRA, vol. 6, n° 3, 1967, p. 503-520.
6. J. Dehedin, *Discrimination sur variables qualitatives*, Thèse 3e cycle, Université de Paris-VI, 1975.
7. E. Diday, *Nouvelles méthodes et nouveaux concepts en classification automatique*, Ph. D. Th., Thèse d'État, Paris, 1972.
8. E. Diday, *Sélection typologique de paramètres*, 188, I.N.R.I.A., Paris, 1976.
9. E. Diday *et al.*, *Optimisation en classification automatique*, I.N.R.I.A., Paris, 1982, volume 1.
10. E. Diday *et al.*, *Optimisation en classification automatique*, I.N.R.I.A., Paris, 1982, volume 2.
11. E. Diday, J. Lemaire, J. Pouget et F. Testu, *Éléments d'analyse des données*, Dunod, Paris, 1982.
12. W. D. Fisher, *On Grouping for Maximum Homogeneity*, J.A.S.A., vol. 53, 1958, p. 789-798.
13. G. V. Kass, *An Exploratory Technique for Investigating Large Quantities of Categorial Data*, Appl. Stat., vol. 29, n° 2, 1980, p. 119-127.
14. S. Levi, *Méthodes et logiciel d'aide au diagnostic par segmentation automatique et application à des données financières*, Thèse 3e cycle, Université de Paris-IX, 1981.
15. P. Michaud, *Structuration optimale d'une opinion*, F-068, IBM France, Paris, 1983.
16. MODULAD, Bibliothèque fortran pour l'analyse des données. Document Normalisation MODULAD. F-66, I.N.R.I.A., Paris, novembre 1983.
17. J. N. Morgan et J. A. Sonquist, *Problems in the Analysis of Survey Data, and a Proposal*, J.A.S.A., vol. 58, 1963, p. 415-434.
18. H. Ralambondrainy, *An Interactive System of Classification: SICLA*, Compstat 1982, 1982.
19. Vo Khac Kh et Nghien, *Étude sur les aspects théoriques et pratiques de la segmentation aux moindres carrés*, R.I.R.O., vol. 8, 1968, p. 77-90.