REVUE FRANÇAISE D'INFORMATIQUE ET DE RECHERCHE OPÉRATIONNELLE [SÉRIE VERTE]

VO-KHAC KHOAN
PHONG TUAN NGHIEM

Étude sur les aspects théorique et pratique de la segmentation aux moindres carrés

Revue française d'informatique et de recherche opérationnelle [Série verte], tome 2, n° V1 (1968), p. 77-90

<http://www.numdam.org/item?id=RO_1968__2_1_77_0>

R.I.R.O.
(2e année, N° 8, 1968, p. 77-90)

ETUDE SUR LES ASPECTS THEORIQUE ET PRATIQUE DE LA SEGMENTATION AUX MOINDRES CARRES

par Vo KHAC Khoan (1) et NGHIEM Phong Tuan (2)

Résumé. — *Pour décrire correctement une population composée d'individus dont certaines caractéristiques sont connues, on peut être amené à la diviser en un petit nombre de groupes homogènes vis-à-vis d'un ou de plusieurs critères. La division ainsi faite est appelée une segmentation.*

Le problème de segmentation se pose, en général, lorsque la population soumise à l'examen est trop nombreuse pour qu'on puisse étudier chacun des individus séparément, et que d'autre part, les caractéristiques auxquelles l'on s'intéresse présentent, dans l'ensemble de la population, une disparité trop grande pour que leur valeur moyenne puisse donner une information utile.

Avec les critères qui permettent de définir l'homogénéité des segments s'introduit nécessairement une notion de ressemblance entre les individus.

On peut considérer alors que le problème de segmentation n'est rien d'autre que celui d'une classification dans laquelle on rassemble les individus qui présentent une grande ressemblance mutuelle.

Les recherches effectuées ont tout d'abord permis de faire la synthèse d'un grand nombre de résultats épars, et de les généraliser ; l'étude du problème dans un espace pré-hilbertien permet notamment de prendre en compte les distances généralisées de Mahalanobis et de traiter ainsi correctement le cas de caractères notablement corrélés, ou encore d'appliquer la segmentation à des ensembles de courbes. En précisant le lien qui existe entre cette méthode et l'analyse de variance, elles conduisent à un critère permettant de fixer le nombre optimal de segments. Enfin, des méthodes de calculs ont été mises au point qui apportent dans les cas simples, une solution rigoureuse au problème.

Cet article résume succinctement les aspects théorique et pratique du problème et donne quelques exemples d'application.

INTRODUCTION

Pour décrire correctement une population composée d'individus dont certaines caractéristiques sont connues, on peut être amené à la diviser en un petit nombre de groupes homogènes vis-à-vis d'un ou de

(1) Professeur associé à la Faculté des Sciences d'Orléans-Tours. Conseiller Scientifique à la Direction des Études, Cegos.
(2) Direction des Études, Cegos.

plusieurs critères. Les groupes sont appelés ici des segments, conformément au langage utilisé dans les études de marché qui ont donné naissance aux recherches dont les résultats sont exposés dans cet article. La division ainsi faite est appelée une segmentation.

Le problème de segmentation se pose, en général, lorsque la population soumise à l'examen est trop nombreuse pour qu'on puisse étudier chacun des individus séparément, et que d'autre part, les caractéristiques auxquelles l'on s'intéresse présentent, dans l'ensemble de la population, une disparité trop grande pour que leur valeur moyenne puisse donner une information utile.

Avec les critères qui permettent de définir l'homogénéité des segments s'introduit nécessairement une notion de ressemblance entre les individus.

On peut considérer alors que le problème de segmentation n'est rien d'autre que celui d'une classification dans laquelle on rassemble les individus qui présentent une grande ressemblance mutuelle.

La vraie paternité de la segmentation aux moindres carrés doit être attribuée à Dalenius [4]. De nombreuses études ont été entreprises depuis (cf. en particulier, Dalenius et Hodges [5], Cox [3], Fisher [6], MacNaugthon-Smith, Williams et Dale [11], Guinet [8]).

Les recherches effectuées à la direction des Études de la CEGOS (cf. Vo Khac [16] [17] [18], Nghiem [12]) ont tout d'abord permis de faire la synthèse d'un grand nombre de résultats épars, et de les généraliser ; l'étude du problème dans un espace pré-hilbertien permet notamment de prendre en compte les distances généralisées de Mahalanobis et de traiter ainsi correctement le cas de caractères notablement corrélés, ou encore d'appliquer la segmentation à des ensembles de courbes. En précisant le lien qui existe entre cette méthode et l'analyse de variance, elles nous ont conduit au critère permettant de fixer le nombre optimal de segments. Enfin, des méthodes de calculs ont été mises au point qui apportent dans les cas simples, une solution rigoureuse au problème.

Les applications de la segmentation aux moindres carrés sont nombreuses, car ce problème se pose dans presque tous les domaines d'étude : biométrie [13], psychométrie [15], économétrie [7], organisation [9], publicité [18], classification des pétroles [10], phytosociologie [18], etc.

Cet article résume succinctement les aspects théorique et pratique du problème et donne quelques exemples d'application.

1. — ASPECT THEORIQUE

1.1. — Données du problème

1.1.1. — Espace économique. — Soit E un espace pré-hilbertien [17], c'est-à-dire un espace vectoriel muni d'un produit scalaire, noté $< \mid >$; la norme pré-hilbertienne associée est par définition : $\|x\| = < \vec{x} \mid \vec{x} > 1/2$. Cet espace pré-hilbertien sera appelé *espace économique.*

Dans la plupart des applications pratiques, E est un espace vectoriel à R dimensions. Le produit scalaire est donné par :

$$< \vec{x} \mid \vec{y} > = \sum_{r,s=1}^{R} \pi_{rs} x^r y^s \qquad (1)$$

où x^r et y^s sont les composantes des vecteurs x et y, et où $M = \pi_{rs}$ est une matrice carrée symétrique positive, appelée *matrice de métrisation.*

Pour rendre les différents composants comparables et pour éliminer l'influence de leurs corrélations mutuelles, cette matrice est souvent prise égale à l'inverse de la matrice de leurs variances et co-variances.

1.1.2. — Population. — Soit Ω un ensemble comportant N éléments 1, 2, ..., N. Cet ensemble Ω s'appelle la *population* et chaque élément ω ($\omega = 1, 2, ..., N$) s'appelle *une classe économique.*

1.1.3. — Fonction objective et loi de répartition. — A chaque classe économique ω, on fait correspondre :

(*i*) un vecteur $\vec{y}(\omega) \varepsilon E$, appelé *vecteur économique* (ou objectif) ;

(*ii*) un nombre $p(\omega) > 0$, appelé mesure de la classe ω (ou sa probabiblité, ou *sa fréquence de répartition*).

L'ensemble des vecteurs économiques constitue la fonction économique (ou fonction objective) ; l'ensemble des nombres $p(\omega)$ constitue la loi de répartition de la population.

1.2. — Définition des dispersions

1.2.1. — Dispersion interne. — Soit Ωj un sous-ensemble de la population Ω. On appelle mesure de Ωj le nombre $p(\Omega j) = \Sigma p(\omega)$. Le vecteur moyen de la fonction objective sur $\omega \varepsilon \Omega j$ est par définition

$$\vec{g}(\Omega j) = \frac{1}{p(\Omega j)} \sum_{\omega \varepsilon \Omega j} p(\omega) \vec{y}(\omega) \qquad (2)$$

La dispersion interne de Ωj est alors définie par :

$$\Delta(\Omega j) = \sum_{\omega \varepsilon \Omega j} p(\omega) \, \| \vec{y}(\omega) - \vec{g}(\Omega j) \|^2 \qquad (3)$$

1.2.2. — Dispersion de mélange. — On appelle dispersion de mélange des sous-ensembles $\Omega 1$, ..., ΩJ le nombre

$$\Delta(\Omega 1, ..., \Omega J) = \sum_{j=1}^{J} p(\Omega j) \, \| \vec{g}(\Omega j) - \vec{G} \|^2 \qquad (4)$$

où $G = g(\Omega)$ désigne le vecteur moyen sur la population.

1.3. — Position du problème

Soit J un nombre fixé, inférieur à N. On désire faire une partition de l'ensemble Ω en J sous-ensembles disjoints non vides $\Omega 1, ..., \Omega J$ satisfaisant simultanément aux deux conditions suivantes :

(*i*) la somme $\Delta(\Omega 1) + ... + \Delta(\Omega J)$ des dispersions internes est minimale ;

(*ii*) la dispersion de mélange $\Delta(\Omega 1, ..., \Omega J)$est maximale.

Une telle partition s'appelle *une J-segmentation optimale par la méthode des moindres carrés*, chaque sous-ensemble Ωj $(j = 1, ..., J)$ s'appelle *un segment.*

La condition (*i*) exprime l'*homogénéité à l'intérieur* de chaque segment ; la condition (*ii*) exprime la *différenciation* des divers segments.

Notons que la dispersion de mélange joue le même rôle que la *quantité d'information* [17] dans la théorie informatique, et le même rôle que le *pouvoir discriminant* dans la méthode de segmentation de Belson-Agostini [1].

1.4. — Equation de la dispersion

On peut montrer (cf. par exemple [17]) que :

$$\Delta(\Omega 1) + ... + \Delta(\Omega J) + \Delta(\Omega 1, ..., \Omega J) = \Delta(\Omega)$$

On en déduit que les conditions (*i*) et (*ii*) sont compatibles et équivalentes. Le problème posé possède donc au moins une solution.

1.5. — Critère d'optimalité

1.5.1. Distance symbolique entre deux segments

On appelle distance symbolique entre deux segments $\Omega 1$ et $\Omega 2$ le double de leur dispersion de mélange.

On démontre la relation (cf. [17]) :

$$d^2(\Omega 1, \Omega 2) = 2\Delta(\Omega 1, \Omega 2) = \frac{2p(\Omega 1)p(\Omega 2)}{p(\Omega 1) + p(\Omega 2)} \|\vec{g}(\Omega 1) - \vec{g}(\Omega 2)\|^2 \quad (6)$$

Pour $p(\Omega 1) = p(\Omega 2) = 1$, on retrouve la distance ordinaire. Ce n'est pas une distance au sens mathématique du terme parce que les axiomes de la distance ne sont pas vérifiés. Cependant cette notion permet d'exprimer très simplement les critères d'optimalité.

1.5.2. — Critère de transfert

Le problème est le suivant :

Soit deux segments $\Omega 1$ et $\Omega 2$ et un groupe de points S qu'il faut mettre en bloc ou bien dans $\Omega 1$, ou bien dans $\Omega 2$. Comment choisir entre $\Omega 1$

et Ω2 pour que la somme des dispersions internes des segments résultants soit la plus petite ?

En convenant de mesurer l'éloignement de deux segments par leur distance symbolique, le critère de choix est (cf. [17]) :

On mettra S dans le segment le plus proche.

1.5.3. — Critère d'appartenance

Du critère de transfert, on déduit le résultat (cf. [17]) :

Soit Ω1 et Ω2 deux segments et S un vrai sous-ensemble de Ω1. Pour que Ω1 et Ω2 appartiennent à une segmentation optimale il est nécessaire que soit vérifiée la relation :

$$d^2(S, \Omega 1 - S) \leqslant d^2(S, \Omega 2) \qquad (7)$$

1.5.4. — Propriétés de séparabilité convexe

La segmentation par la méthode des moindres carrés possède, en particulier la propriété suivante, qu'on peut démontrer à partir du critère d'appartenance (cf. [20]) :

Lorsque la segmentation est optimale, les enveloppes convexes des segments sont deux à deux disjointes.

Il existe d'autres méthodes de segmentations qui ne possèdent pas cette propriété.

1.6. — Algorithmes de résolution

Plusieurs algorithmes de résolution ont été étudiés (cf. [20] ou [17]) :

— algorithme de programmation dynamique, pour les problèmes à un seul caractère ;

— algorithme de division hiérarchisée ;

— algorithme d'agglomération hiérarchisée.

Seul ce dernier, qui est le plus simple, dans son principe, sera exposé ici.

1.6.1. — Première étape

(i) *Phase de calcul.* — On établit la matrice initiale (carrée, symétrique, à N colonnes) des distances entre les classes ; pour cela on utilise la formule (6) qui appliquée à deux classes ω et ω′ donne :

$$d^2(\omega, \omega') = \frac{2p(\omega)p(\omega')}{p(\omega) + p(\omega')} \left\| y(\omega) - y(\omega') \right\|^2 \qquad (8)$$

(ii) *Phase de sélection.* — On agglomère les deux classes économiques ayant la plus petite distance. On obtient ainsi une $(N - 1)$ segmentation optimale.

1.6.2. — Etape générale. — A partir d'une I-segmentation, on veut obtenir une $(I - 1)$ segmentation.

(i) *Phase de calcul.* — On établit la nouvelle matrice (symétrique, carrée, à I colonne) des distances en utilisant la formule de récurrence suivante (cf. [17]) :

$$d^2(X, A + B)$$

$$= \frac{[p(X) + p(A)]\, d^2(X, A) + [p(X) + p(B)]\, d^2(X, B) - p(X)\, d^2(A, B)}{p(X) + p(A) + p(B)}$$

(ii) *Phase de sélection.* — On réunit les deux segments dont la distance est la plus petite.

1.7. — Choix du nombre de segments

1.7.1. — Espace à une dimension

Lorsque l'espace économique a une seule dimension, l'équation de la dispersion (5) devient celle, bien connue, d'analyse de variance.

Pour un nombre de segments J donné, appelons ΔJ la somme des dispersions internes obtenue.

En analyse de variance, on teste la différenciation entre ce que nous appelons ici des segments, par la méthode de Fisher-Snedecor. Une démarche possible est la suivante :

Soit N le nombre total de classes économiques (c'est-à-dire le nombre d'observations, en analyse de variance).

Ayant $J - 1$ segments significativement différents, on teste la signification d'un J-ième segment en comparant la quantité :

$$\Phi = (N - J)\left(\frac{\Delta J - 1}{\Delta J} - 1\right) \tag{9}$$

à la valeur $F_{\alpha;1,N-J}$ prise par la fonction de Fisher-Snedecor à 1 et $N - J$ degrés de liberté, au seuil α (en général α est égal à 10 %, 5 % ou 1 %).

Si Φ est plus grand, c'est-à-dire si :

$$\frac{(N - J)(\Delta J - 1 / \Delta J - 1)}{F_{\alpha;1,N-J}} > 1$$

on conclut que le J-ième segment est significatif.

En analyse de variance, le J-ième segment est donné ; tandis qu'ici il est défini par l'échantillon. La loi de Φ n'est donc pas celle de Fisher-Snedecor. On peut néanmoins déduire, par analogie formelle, un critère pour le choix de J. Nous proposons de choisir J parmi les valeurs qui donnent un maximum local au rapport

$$\Phi / F_{\alpha;1,N-J}.$$

Au lieu de se rattacher au test du passage de $J-1$ à J, on peut aussi se rattacher à celui du passage de 1 à J ; ce qui conduit à choisir J parmi les maximums de :

$$\frac{\dfrac{N-1}{J-1}(\Delta 1/\Delta J - 1)}{F_{\alpha;J-1,N-J}} \tag{11}$$

L'expérience semble montrer que ce critère donne des résultats moins bons.

1.7.2. — Espace à plusieurs dimensions

Lorsque l'espace économique a une dimension $m > 1$ et que les composantes d'un même vecteur sont indépendantes et de même variance, on peut raisonnablement définir la distance entre deux classes ω et ω' à partir de la norme :

$$\|y(\omega) - y(\omega')\|^2 = \sum_{i=1}^{m} [y_i(\omega) - y_i(\omega')]^2 \tag{12}$$

En décomposant une distance en m composantes aléatoirement indépendantes, on peut encore aboutir au critère :

$$\frac{(N-J)(\Delta J - 1/\Delta J - 1)}{F_{\alpha;\,m,m(N-J)}} \tag{13}$$

Si les composantes sont corrélées, mais qu'on connaît leur matrice des variances et co-variances Λ, en utilisant la norme :

$$\|x\|^2 = x' \Lambda^{-1} x$$

on aboutit encore au critère (13) ci-dessus.

S'il faut estimer la matrice Λ, il faudrait, en toute rigueur, faire appel aux résultats de l'analyse multivariée. Mais les formules obtenues ne sont pas simples. Dans la pratique, on pourra toujours se contenter du critère (13).

2. — ASPECT PRATIQUE

Voici comment se rencontre le problème de segmentation en pratique.

2.1. — Ensembles de variables

On doit distinguer deux ensembles de variables : le premier ensemble est celui des variables descriptives ; le second, celui des variables objectives.

2.1.1. — Variables descriptives. — Les variables descriptives servent à construire la population. S'il existe S variables descriptives, chacune se composant de N_s éventualités ($s = 1, ..., S$), alors le nombre des classes économiques est $N = N_1 \times ... \times N_S$. L'ensemble des variables

descriptives peut donc être considéré comme une seule variable X présentant N éventualités $X_1, ..., X_N$. Chaque éventualité $X_n (n = 1, ..., N)$ est une classe économique.

2.1.2. — Variables objectives. — Les variables objectives permettent, par l'intermédiaire de la fonction objective, de définir l'homogénéité et la différenciation des divers segments en fonction de l'objectif fixé. S'il existe Q variables objectives, chacune se composant de R_q éventualités $(q = 1, ..., Q)$, alors le nombre de dimensions de l'espace économique est :

$$R = R_1 \times ... \times R_q \times ... \times R_Q.$$

L'ensemble des variables objectives peut donc être considéré comme une seule variable Y présentant R éventualités $Y^1, ..., Y^R$. Chaque éventualité $(r = 1, ..., R)$ est une dimension de l'espace économique.

REMARQUE : Selon le problème, une même variable peut être descriptive ou objective. Dans l'exemple 3.1, l'âge est une variable descriptive, alors que dans l'exemple 3.3, l'âge est une variable objective.

2.2. — Données du problème

2.2.1. — Données brutes. — Les données brutes sont celles recueillies expérimentalement sur l'ensemble des variables descriptives et objectives. Elles se présentent en général sous la forme suivante :

ÉVENTUALITÉ X_n	PROBABILITÉ DE L'ÉVENTUALITÉ X_n	Y^1	...	Y^r	...	Y^R
X_1	p_1					
.	.					
.	.					
.	.					
.	.					
X_n	p_n			a_n^r		
.	.					
.	.					
.	.					
.	.					
X_N	p_N					

Dans ce tableau, a_n^r est un nombre qui caractérise le couple (X_n, Y^r) (cf. les exemples au § 3).

D'autre part, le plus souvent les éventualités $Y^1, ..., Y^R$ sont corrélées ; dans les données brutes, peut donc se trouver une matrice de corrélation Λ.

2.2.2. — Données utiles. — De ces données brutes, il faut dégager les données utiles. La mesure (ou fréquence de répartition) de chaque éventualité X_n peut être prise comme égale à sa probabilité : $p(X_n) = p_n$.

Quant aux vecteurs économiques $y_n = (y_n^r)$, ils se déduisent des a_n^r en les multipliant par des facteurs convenables (facteurs de pondération ou facteurs de normalisation). Enfin, la matrice de métrisation dépend de la corrélation entre les diverses éventualités $Y^1, \ldots Y^R$. Si ces éventualités sont orthogonales (c'est-à-dire non corrélées), on peut admettre que la matrice de métrisation est égale à l'unité. On peut alors démontrer que dans le cas général la matrice de métrisation est égale à l'inverse de la matrice de corrélation : $M = \Lambda^{-1}$.

3. — EXEMPLES D'APPLICATIONS

Nous allons donner quelques classes d'application.

3.1. — Première classe

Dans cette première classe, on est en présence d'une seule variable objective à une éventualité. Nous allons prendre un exemple typique : la segmentation du marché [1].

Il s'agit de segmenter un échantillon de 1 000 Parisiennes sur l'utilisation d'un produit de beauté. Les critères de segmentation possibles sont l'âge (3 éventualités : 18-24 ; 25-34 ; 35-49) et la classe sociale (3 éventualités : pauvre, moyenne, riche). Le caractère choisi pour la segmentation est le niveau de consommation (quantité achetée par 100 Parisiennes).

3.1.1. — Variables. — Ici, les variables descriptives sont les critères de segmentation : âge, classe sociale. Comme chaque variable se compose de trois éventualités, le nombre de classes économiques est

$$N = N_1 \times N_2 = 3 \times 3 = 9.$$

Il n'y a qu'une seule variable objective : c'est le niveau de consommation ; cette variable ne présente qu'une seule éventualité ($R = 1$).

3.1.2. — Données. — Le Panel des consommateurs fournit le tableau suivant :

n	Classe économique X_n	Effectif de la classe : p_n	Quantités achetées : a_n	Niveau de consommation : Y_n
1	Riche, 18-24	30	14	47
2	Riche, 25-34	34	23	68
3	Riche, 35-49	51	25	49
4	Moyenne, 18-24	36	13	36
5	Moyenne, 25-34	120	38	32
6	Moyenne, 35-49	209	53	26
7	Pauvre, 18-24	81	20	24
8	Pauvre, 25-34	216	25	11
9	Pauvre, 35-49	223	27	12

Les premières colonnes sont des données brutes ; la dernière colonne est calculée par $Y_n = \frac{a_n}{p_n} \times 100$. La fréquence de répartition est prise égale à l'effectif de la classe. La matrice de métrisation est égale à l'unité.

3.1.3. — Résultats. — L'algorithme de la segmentation agglomérative hiérarchique conduit au résultat suivant.

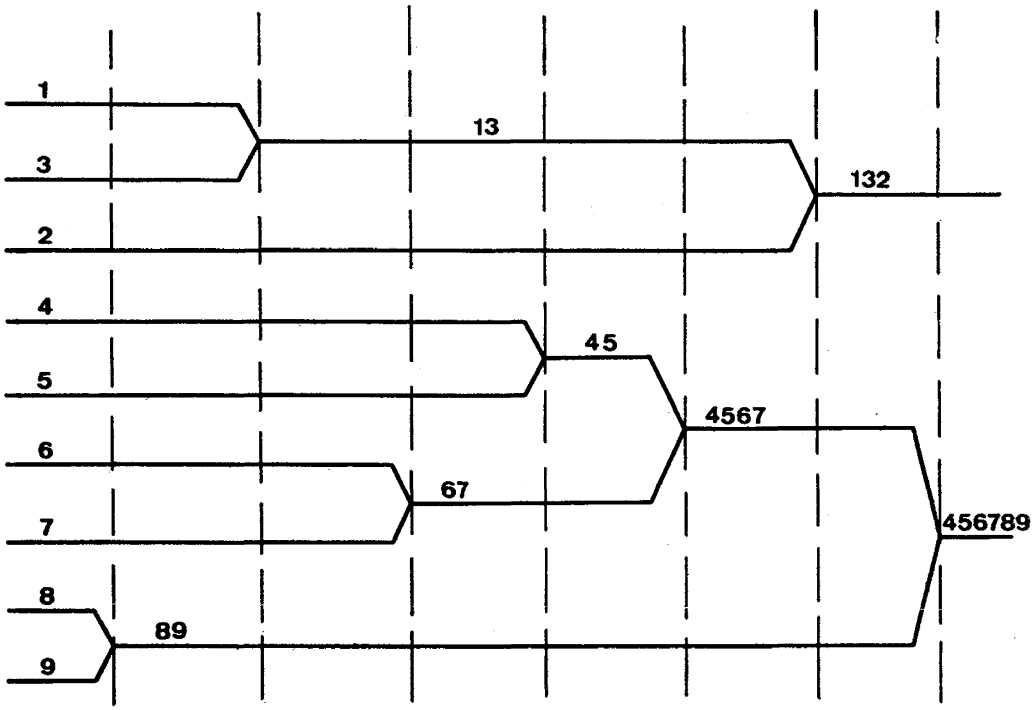

Figure 1

Le choix du nombre J des segments se fait comme dans [18]. On trouve $J = 5$. La 5-segmentation obtenue est

$$(1,3) ; (2) ; (4,5) ; (6,7) ; (8,9).$$

Les 5 segments sont donc :

I : riche, 25-34 ans
II : riche, 18-24 ans, 35-49 ans
III : moyenne, 18-24 ans
IV : moyenne, 35-49 ans ; pauvre, 18-24 ans
V : pauvre, 25-49 ans.

3.2. — Deuxième classe

Dans cette classe d'applications, il y a plusieurs variables objectives dont chacune ne présente qu'une éventualité. Comme exemples : la segmentation de l'ensemble des employés d'une usine suivant leurs qualités, de l'ensemble des produits pétroliers suivant leurs propriétés, du marché suivant plusieurs caractères. L'exemple typique est l'étude anthropométrique des tribus indiennes [12].

On désire faire une segmentation de l'ensemble de douze tribus indiennes selon les neufs caractères suivants :

Y^1 : longueur de la tête
Y^2 : largeur de la tête
Y^3 : largeur bizygomatique
Y^4 : hauteur du nez
Y^5 : largeur du nez
Y^6 : profondeur du nez
Y^7 : stature
Y^8 : hauteur assise
Y^9 : largeur du front.

3.2.1. — Variables. — Ici, il y a une seule variable descriptive : le nom des tribus indiennes, se composant d'éventualités : Basti, Brahmin... Par contre, il existe 9 variables objectives, chacune présentant une seule éventualité. La dimension de l'espace objectif est donc $R = 9 \times 1 = 9$.

3.2.2. — Données brutes. — Les données brutes sont rassemblées dans le tableau suivant :

n	Tribus	Effectifs	Y^1	Y^2	Y^3	Y^4	Y^5	Y^6	Y^7	Y^8	Y^9
1	Basti	86	191,92	139,88	133,36	51,24	36,55	25,49	164,51	86,43	104,74
2	Brahmin ...	92	191,35	139,50	132,68	50,40	36,13	24,74	165,07	86,25	104,46
3	Chattri.	139	192,58	131,72	131,70	52,72	35,64	24,73	163,33	82,25	103,98
4	Muslim.....	168	190,78	137,40	131,40	51,38	36,36	24,49	162,45	81,83	103,28
5	Bhatu.....	150	186,10	138,58	133,58	52,06	35,65	25,09	163,38	84,49	99,34
6	Habru	124	181,94	137,40	131,16	50,30	35,82	24,19	164,91	85,53	100,18
7	Bril.	186	181,87	137,62	131,18	48,60	37,49	24,05	162,92	82,60	103,36
8	Dom.	113	186,87	137,52	132,64	50,34	38,11	25,33	166,53	84,19	104,16
9	Ahir	68	187,45	138,12	131,70	48,98	35,60	24,19	161,37	84,35	12,76
10	Kurmi	94	188,86	137,86	131,82	49,22	36,21	24,03	161,35	83,41	102,62
11	Kahar	57	188,83	136,28	130,70	48,62	36,51	24,08	160,53	81,47	101,68
12	Autres tribus. ...	173	187,69	136,84	131,30	48,72	36,27	23,13	161,34	83,09	102,44

Ces neufs caractères sont corrélés; la matrice de corrélation est :

	Y^1	Y^2	Y^3	Y^4	Y^5	Y^6	Y^7	Y^8	Y^9
Y^1	1	0,20	0,28	0,18	0,19	0,15	0,27	0,27	0,23
Y^2		1	0,54	0,17	0,14	0,13	0,19	0,21	0,45
Y^3			1	0,19	0,27	0,16	0,29	0,30	0,49
Y^4				1	0,04	0,29	0,20	0,22	0,11
Y^5					1	0,11	0,14	0,12	0,18
Y^6						1	0,18	0,21	0,12
Y^7							1	0,58	0,22
Y^8								1	0,30
Y^9									1

3.2.3. — Données utiles. — La fréquence de répartition de chaque classe économique est égale à leur effectif. Pour obtenir les vecteurs économiques, nous utilisons les variables centrées réduites :

$$Y_n^r = \frac{Y_n^r - m(Y^r)}{\sigma(Y^r)}$$

où $m(Y^r)$ et $\sigma(Y^r)$ sont respectivement la moyenne et l'écart-type de Y^r. La matrice de métrisation est prise égale à l'inverse de la matrice de corrélation.

3.2.4. — Résultats. — La segmentation se fait alors aisément appliquant la méthode indiquée dans [18], on trouve que :

(*i*) le nombre de segments qu'il faut prendre est $J = 5$.

(*ii*) la 5-segmentation obtenue est Basti-Brahmin ; Chattri-Muslim ; Bhatu-Habru ; Bril-Dom ; Ahir-Kurmi — Kahar — autres tribus.

Nota : Pratiquement, on procède autrement. On remplace les caractères donnés par des caractères fictifs et orthogonaux (et réduits), et on prend la matrice de métrisation égale à l'unité. Ce procédé a été programmé sur ordinateur.

3.3. — Troisième classe d'application

Dans cette classe, il existe plusieurs variables objectives dont chacune présente diverses éventualités. L'exemple typique est la typologie de périodiques basée sur l'analyse de leur clientèle.

3.3.1. — Variables. — Il n'existe qu'une seule variable descriptive : c'est le titre (50 éventualités : Réalité, la Maison Française, Historia,... Clair Foyer). Les variables objectives sont nombreuses et chacune d'elles présente plusieurs éventualités. Pour simplifier, nous supposons qu'il n'existe que deux variables objectives.

Sexe : deux éventualités (masculin, féminin).

Age : trois éventualités (15-30 ; 31-50 ; 51-80).

Alors, la dimension de l'espace économique est $R = 2 \times 3 = 6$.

Autrement dit, chaque vecteur économique possède six composantes :

Y^1 : hommes de 15 à 30 ans ; Y^4 : femmes de 15 à 30 ans ;
Y^2 : hommes de 31 à 50 ans ; Y^5 : femmes de 31 à 50 ans ;
Y^3 : hommes de 51 à 80 ans ; Y^6 : femmes de 51 à 80 ans.

3.2.2. — Données brutes. — Les données brutes sont contenues dans le tableau suivant :

n	Titre	Total des lecteurs	Y^1	Y^2	Y^3	Y^4	Y^5	Y^6
1	Réalité	1 500	100	200	500	200	100	400
2	Maison française ...	2 000	200	100	300	500	500	400
3	Historia	2 500	300	400	900	100	300	500
.								
.								
.								
49	Télé 7 jours.....	2 000	100	200	200	400	500	600
50	Télémagazine	1 000	50	100	100	200	250	300

3.2.3. — Données utiles. — Chaque titre est une classe économique dont la fréquence de répartition p_n est égale au nombre de lecteurs de ce titre. Pour obtenir les vecteurs économiques, on divise chaque colonne par la première colonne :

$$Y_n^r = \frac{Y_n^r}{p_n}$$

En effet, on fait la segmentation non pas sur le nombre de lecteurs, mais sur la proportion des lecteurs (dans l'exemple donné, *Télé 7 jours* et *Télémagazine* doivent être considérés, comme ayant une dispersion de mélange nulle). Pour plus de détail, cf. [18].

Nota : Bergonier et Boucharenc [2] ont donné une autre méthode basée sur la théorie de l'information, les deux méthodes donnent sensiblement les mêmes résultats.

SOURCES REFERENTIELLES

[1] Agostini (J. M.), *Une méthode de segmentation du marché,* Esomar-Wapor Conférence (1966), Dublin,

[2] Bergonier (H.) et Boucharenc (L.), *Une méthode de segmentation basée sur la théorie de l'information,* Prix Marcel Dassault (1966).

[3] Cox (D. R.), *Note on grouping* J. Amer. Stat. Ass. 52 (1957), 543-547.

[4] Dalenius (T.), *The problem of optimum stratification,* Skandinavisk Aktnarietidskrift (1950), 203-213.

[5] Dalenius (T.) et Hogdes (J. L. Fr.), *Minimum variance stratification* J. Amer. Strat. Assoc. 54 (1959), 88-101.

[6] Fischer (W.), *On grouping for maximum homogeneity,* J. Amer. Stat. Assoc. 53 (1958), 789-798.

[7] Green (P. E.), Frank (R. E.) et Robinson (P. J.), Cluster Analysis in Test Market Selection *Management Science.* B, 13 (1967), 387-400.

[8] Guinet (J.), *Recherche d'une segmentation optimale,* AUROC (1966).

[9] Howard (R. N.), *Classifying a population into homogen groups* ESOMAR-WAPOR Conference (1965), Cambridge.

[10] Jizba, *A Contribution to statistical theory of classification,* Standford Univ. Publ. Geol. Sci. 9 (1964), 729-756.

[11] MacNaughton-Smith (P.), Williams (W. T.) et Dale (M. B.), On objective method of weighting in similarity analysis, *Nature,* 201 (1964), 426.

[12] Nghiem (Ph. T.), *La segmentation aux moindres carrés et l'analyse de la variance D.E.,* CEGOS (1968).

[13] Rao (C.R.), *Advanced statistical methods in biometric research,* John Wiley and Sons, New York (1952).

[14] Sneath (Ph. A) et Sokal (R.R.), *Principles of numerical taxonomy,* Freeman, San Francisco et Londres (1963).

[15] Thorndike (R.L.), Who belongs to the family ? *Psychometrika,* 18 (1953), 267-276.

[16] Vo-Khac (K), *Segmentation aux moindres carrés dans un espace pré-hilbertien,* C. R. Acad. Sci. Paris, novembre 1967.

[17] Vo-Khac (K), *Étude théorique du problème de segmentation,* D. E. CEGOS, 1966.

[18] Vo-Khac (K), *Applications pratiques de la segmentation par la méthode des moindres carrés,* D. E. CEGOS, 1966-1967.

[19] Yaglom (A. M.) et Yaglom (I. M.), *Probabilité et information,* Dunod, Paris (1959).

[20] Document de synthèse sur la méthode de segmentation aux moindres carrés D. E. CEGOS (1967).