SUR LES CRITÈRES D'AGRÉGATION UTILISÉS EN CLASSIFICATION AUTOMATIQUE [1]

par

Michel JAMBU

SOMMAIRE

(1) Cet article fait suite à deux articles précédemment parus dans « Consommation », n° 3, 1973 et n° 2, 1974.

1. INTRODUCTION

Dans un premier article, nous avons présenté l'algorithme général des classifications ascendantes hiérarchiques. Nous avons insisté sur le fait que la classification dépendait fortement de quatre options : le choix du codage, le choix de la distance, le choix d'une stratégie d'agrégation et le choix d'un critère d'appréciation de la qualité de la représentation des données par la classification ascendante hiérarchique. Dans un second article, nous avons effectué une étude des différents indices de distances utilisables en classification. Celui-ci est l'étude des différents critères d'agrégation utilisables dans un même programme.

2. LE PRINCIPE DE L'AGRÉGATION

Ayant fait choix d'une distance entre les objets ou les variables à constituer en hiérarchie totale indicée, le principe de la construction consiste à procéder par agrégations successives, binaires, de classes de variables. Procéder ainsi consiste, à chaque étape de la construction, à réunir les deux classes en un certain sens les plus proches. Pour ce faire, il est utile de faire choix d'une stratégie ou d'un critère d'agrégation, c'est-à-dire de faire choix d'une façon d'apprécier la proximité entre deux classes : les deux classes réunies sont celles dont la « distance » entre parties est la plus petite. Les choix de critères peuvent être très divers et conduire à des résultats fort différents. Aussi convient-il de les étudier sérieusement, de façon à ce que le praticien sache quelle signification se cache derrière des formulations en apparence séduisantes, ou a priori d'égal intérêt.

3. QUELQUES RAPPELS SUR LES NOTATIONS UTILISÉES

On appelle $\partial(a, b)$ la « distance » entre parties de J. On agrège à chaque étape les deux parties de J qui réalisent, pour cette étape, le minimum de ∂. L'indice de diamètre de la classe (appelé aussi indice d'agrégation) prend la valeur de ce minimum. La construction hiérarchique nécessite, pour éviter d'avoir à recalculer toutes les « distances » ∂, de recalculer les nouvelles « distances » ∂ entre la nouvelle partie créée et les autres parties de J de la partition créée à l'étape précédente :

soit $Hn - 1$ (respt. Hn), l'ensemble de toutes les parties de J constituées par agrégations successives jusqu'à l'étape de rang $n - 1$ (respt. n);

soit Som $(Hn - 1)$ (respt. Som (Hn)), l'ensemble des éléments maximaux de $Hn - 1$ (respt. Hn).

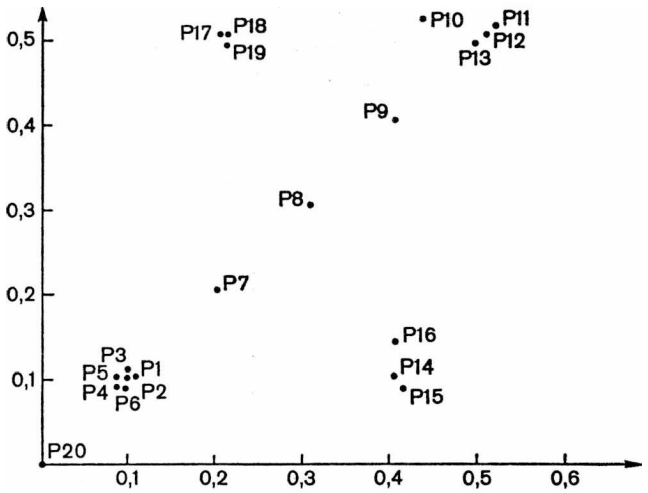
On doit recalculer $\partial(v, s \cup t)$ pour $v \in \text{Som}\,(Hn - 1) - \{ s, t \}$; les autres « distances » ∂ sont conservées pour cette étape.

Dans les paragraphes suivants, nous donnons les stratégies d'agrégation utilisables en classification hiérarchique. Pour les illustrer, nous

avons repris deux exemples. Le premier est constitué par la distribution dans un plan de vingt points auxquels on a associé la distance euclidienne usuelle (exemple 1). Le second est constitué par un tableau dit de présence-absence, image proche d'une partition de vingt variables, auquel on a associé la distance de l'analyse des correspondances (exemple 2 : cf. Consommation, n° 2, 1974, p. 80). Pour chaque stratégie d'agrégation étudiée, on donne les hiérarchies associées correspondant aux exemples.

EXEMPLE 1

Nuage de 20 points dans un plan



4. LES STRATÉGIES D'AGRÉGATION

4.1 L'ultramétrique inférieure maximale

— *Définition de la proximité entre parties de J*

$$\partial(a, b) = \inf \{ d(j, k) \quad ; \quad j \in a, k \in b \} \quad \forall\, a, b \in \mathfrak{P}(J) - \varnothing$$

$$\partial(a, b) = 0 \not\Rightarrow (a = b)$$

$$\partial(a, a) = 0$$

$$\partial(a, b) = \partial(b, a) \qquad \forall\, a, b \in \mathfrak{P}(J) - \varnothing$$

∂ ne définit pas sur $\mathfrak{P}(J) - \varnothing$ un indice de distance (l'inégalité triangulaire n'est pas forcément vérifiée).

— *Passage de l'étape de rang* $(n-1)$ *à l'étape de rang* (n)

On cherche à exprimer $\partial(v, s \cup t)$ selon une formule ne précisant que des éléments concernant les successeurs immédiats s et t de la nouvelle partie créée $s \cup t$.

$$\partial(v, s \cup t) = \inf \{ \partial(v, s), \partial(v, t) \}$$

pour $s, t, v \in \mathrm{Som}\,(Hn - 1)$ et $s \neq t \neq v$.

Le principe d'une telle stratégie est de réunir deux parties de J dont le « saut » est minimum, c'est-à-dire qu'il suffit d'un seul lien (une paire d'éléments proches) pour décider de réunir deux parties.

Remarque : Au début de la construction, on vérifie l'égalité suivante :

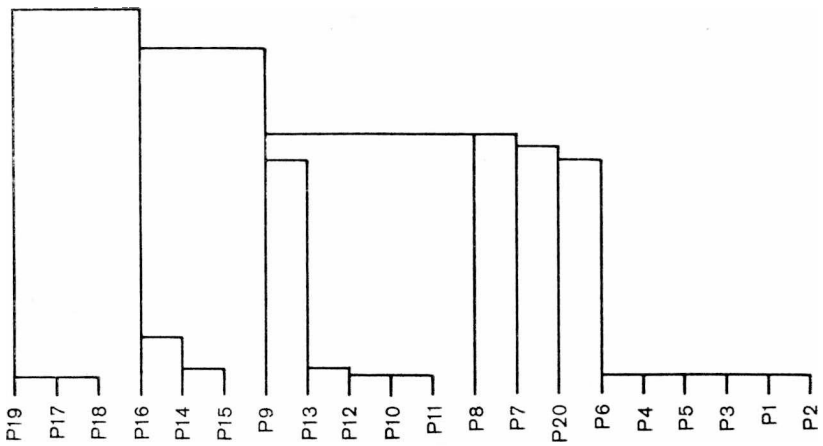
$$\partial(\{j\}, \{k\}) = d(j, k) \qquad \forall j, k \in J$$

On peut utiliser cet algorithme pour toutes les applications de $J \times J$ dans R^+ répondant à la définition d'un indice de distance.

Les graphiques suivants illustrent cet algorithme pour les deux exemples présentés au § 3.

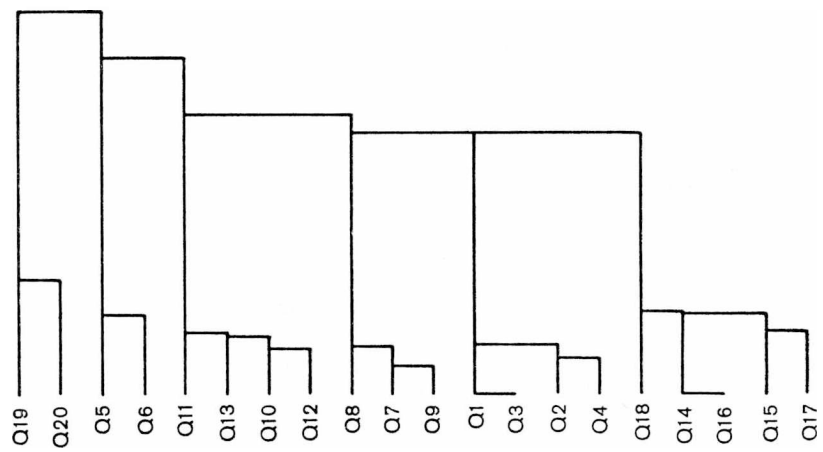
EXEMPLE 1

Nuage de 20 points dans un plan
Ultramétrique inférieure maximale



EXEMPLE 2

Questionnaire (0.1)
Ultramétrique inférieure maximale



4.2 L'ultramétrique supérieure minimale

— *Définition de la proximité entre parties de J*

$$\partial(a, b) = \sup \{ d(j, k) \quad ; \quad j \in a, k \in b \} \qquad \forall\, a, b \in \mathcal{P}(J) - \varnothing$$

$$\partial(a, b) = 0 \Rightarrow a = b$$

$$\partial(a, a) \neq 0 \qquad \text{sauf pour} \qquad a = \{ j \}$$

$$\partial(a, b) = \partial(b, a) \qquad \forall\, a, b \in \mathcal{P}(J) - \varnothing$$

∂ ne définit pas un indice de distance sur $\mathcal{P}(J) - \varnothing$ (l'inégalité triangulaire n'est pas forcément vérifiée).

— *Passage de l'étape de rang* $(n - 1)$ *à l'étape de rang* (n)

$$\partial(v, s \cup t) = \sup \{ \partial(v, s), \partial(v, t) \}$$

$$\text{pour } s, t, v \in \text{Som } (Hn - 1) \text{ et } s \neq t \neq v.$$

Le principe d'un tel critère est d'agréger deux parties de Som $(Hn - 1)$ dont le diamètre de la réunion est minimum, c'est-à-dire qu'on décide de ne réunir deux parties que si tous les liens entre elles sont uniformément assez courts.

Remarque : Au début de la construction, on vérifie l'égalité suivante :

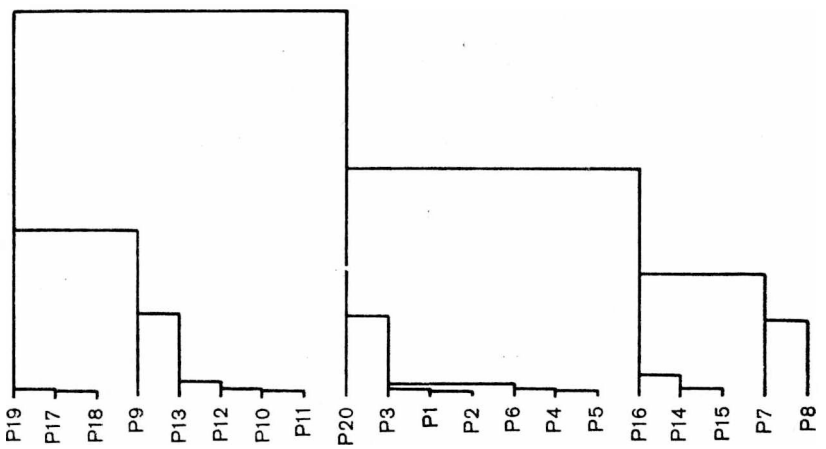
$$\partial(\{ j \}, \{ k \}) = d(j, k) \qquad \forall\, j, k \in J$$

Il n'y a pas unicité de la construction.

Les graphiques suivants illustrent cet algorithme pour les deux exemples présentés au § 3.

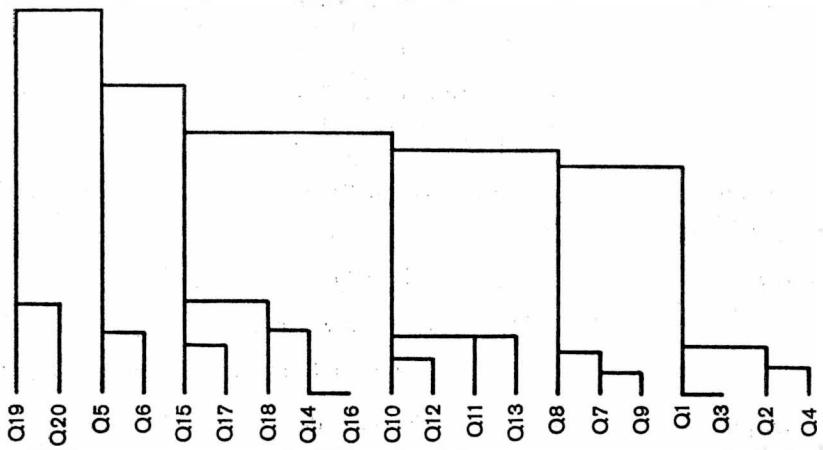
EXEMPLE 1

Nuage de 20 points dans un plan
Ultramétrique supérieure minimale



EXEMPLE 2

Questionnaire (0.1)
Ultramétrique supérieure minimale



4.3 La distance moyenne

— *Définition de la proximité entre parties de J*

$$\partial(a, b) = \left[\frac{\Sigma \{ d(j, k) \mid j \in a, k \in b \}}{\text{Card}(a) \cdot \text{Card}(b)} \right] \qquad \forall\, a, b \in \mathcal{P}(J) - \varnothing$$

Card (a) = nombre d'éléments de a

$$\partial(a, b) = 0 \not\Rightarrow a = b$$

$$\partial(a, a) \neq 0 \qquad \text{sauf si} \qquad a = \{ j \}$$

$$\partial(a, b) = \partial(b, a) \qquad \forall\, a, b \in \mathcal{P}(J) - \varnothing$$

Exemple :

soit $J = \{ 1, 2, 3 \}$

$$A(J) = \{ a = \{ 1, 2 \},\ b = \{ 2, 3 \},\ c = \{ 1, 3 \} \}$$

$$d(1, 2) = d(2, 3) = 1 \quad ; \quad d(1, 3) = 2$$

$$\partial(a, b) = \frac{3}{4} \quad ; \quad \partial(b, c) = \frac{3}{4} \quad ; \quad \delta\ (a, c) = \frac{4}{4}$$

$$\partial(a, c) = 1 \not\leq (\partial(a, b) + \partial(b, c)) = \frac{3}{2}$$

∂ ne définit pas un indice de distance sur $\mathcal{P}(J) - \varnothing$ (l'inégalité triangulaire n'est pas forcément vérifiée).

— *Passage de l'étape de rang $(n - 1)$ à l'étape de rang (n)*

$$\partial(v, s \cup t) = \Sigma \{ d(j, k) \quad ; \quad j \in a, k \in s \cup t \} \,/\, \text{Card}(a) \cdot \text{Card}(s \cup t)$$

$$(s \cap t = \varnothing) = \Sigma \{ d(j, k) \quad ; \quad j \in a, k \in s \} \,/\, \text{Card}(a) \cdot \text{Card}(s) \cdot \text{Card}(t)$$

$$\ldots + \Sigma \{ d(j, k) \quad ; \quad j \in a, k \in t \} / \text{Card}(a) \cdot \text{Card}(s) \cdot \text{Card}(t)$$

$$\partial(v, s \cup t) = \partial(a, s)/\text{Card}(t) + \partial(a, t)/\text{Card}(s)$$

$$= \frac{\text{Card}(s) \cdot \partial(a, s) + \text{Card}(t) \cdot \partial(a, t)}{\text{Card}(s) \cdot \text{Card}(t)}$$

Le principe d'une telle stratégie est de réunir deux parties de Som ($Hn - 1$) dont la distance moyenne est minimum.

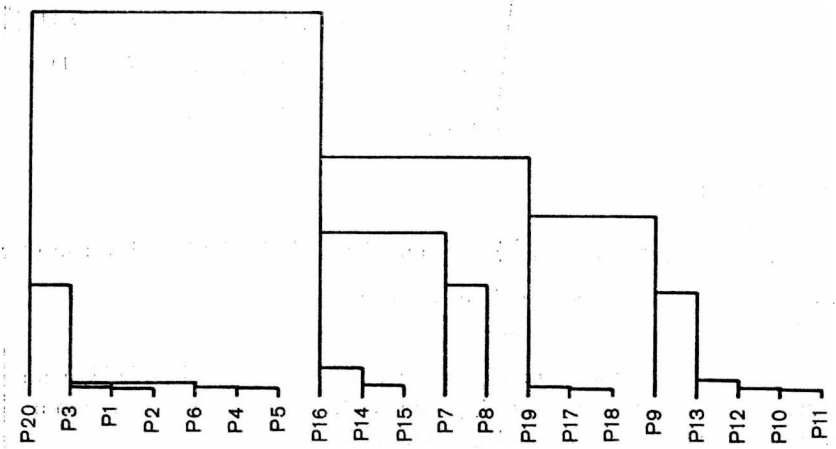
Remarque : On vérifie l'égalité suivante au début de la construction :

$$\partial(\{j\}, \{k\}) = \mathrm{d}(j, k) \qquad \forall (j, k) \in J \times J$$

Les graphiques suivants illustrent les deux exemples présentés au § 3.

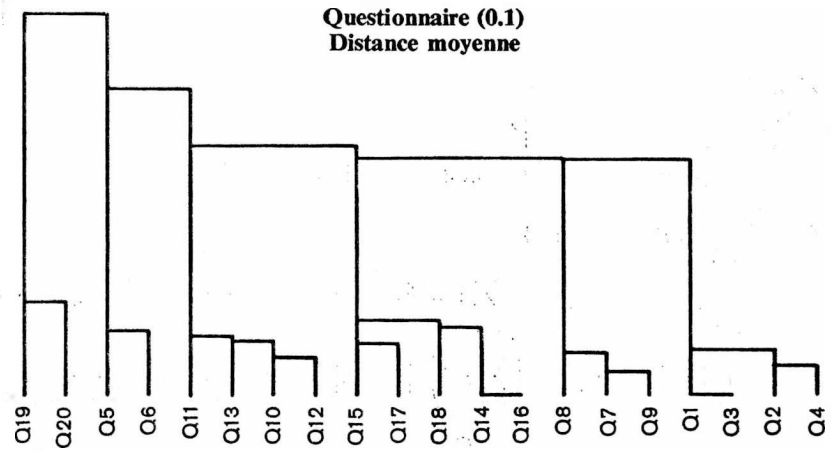
EXEMPLE 1

Nuage de 20 points dans un plan
Distance moyenne



EXEMPLE 2

Questionnaire (0.1)
Distance moyenne



4.4 Centre de gravité des classes

On suppose dans ce paragraphe J euclidien.

— *Définition de la proximité entre parties de J*

$$\partial(a, b) = \|g(a) - g(b)\| \qquad \forall\, a, b \in \mathfrak{P}(J) - \varnothing$$

où $g(a)$ et $g(b)$ sont les centres de gravité associés aux parties a et b de J.

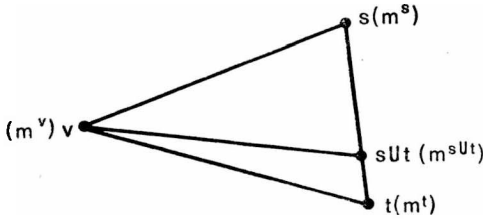
∂ ne définit pas une distance sur $\mathfrak{P}(J) - \varnothing$

$\partial(a, b) = 0 \Rightarrow g(a) = g(b) \not\Leftrightarrow a = b$

$\partial(a, a) = 0$

$\partial(a, b) = \partial(b, a)$

∂ vérifie l'inégalité triangulaire.



— *Passage de l'étape de rang* $(n-1)$ *à l'étape de rang* (n)

$$\partial(v, s \cup t) = \|g(v) - g(s \cup t)\|$$

s, t, v Som $(Hn-1)$

$s \neq t \neq v$

Définition :

$M^2(q) \overset{\text{déf}}{=} \sum_{j \in q} \{ m^j \|j - g\|^2 \}$ = moment centré d'ordre 2 de l'ensemble q

où g est le centre de gravité de l'ensemble q.

Théorème 1 : Soit J un espace euclidien de dimension finie, q un sous-ensemble d'éléments de J, g son centre de gravité, m^q la somme des masses m^j de q. Soit $v \in J, \notin q$. On a alors :

$$m^q \|v - g\|^2 + M^2(q) = \sum_{j \in q} \{ m^j \|j - v\|^2 \}$$

Dém. 1

$$\|j - v\|^2 = \|j - g + g - v\|^2$$
$$= \|j - g\|^2 + 2 \langle j - g, g - v \rangle + \|g - v\|^2$$

$$M^2(q) \overset{\text{déf}}{=} \sum_{j \in q} \{ m^j \|j - g\|^2 \}$$
$$= \sum_{j \in q} \{ m^j \|j - v\|^2 \} - \sum_{j \in q} \{ m^j \|g - v\|^2 \} - 2 \Sigma\, m^j \langle j - g, g - v \rangle$$
$$= \sum_{j \in q} m^j \|j - v\|^2 - \underbrace{\left(\sum_{j \in q} m^j \right)}_{= m^q} \|g - v\|^2 - 2 \langle \underbrace{\Sigma\, m^j (j - g)}_{= 0}, g - v \rangle$$

On a donc

$$M^2(q) + m^q \|v - g\|^2 = \sum_{j \in q} m^j \|j - v\|^2 \qquad \textit{fin 1}$$

Théorème 2 : J euclidien de dimension finie

$$2 \cdot m^J \cdot M^2(J) = \sum_{j \in J, k \in J} \{ m^j \cdot m^k \cdot \|j - k\|^2 \}$$

Dém. 2

$$\sum_{\substack{j \in J \\ k \in J}} m^j m^k \|j - k\|^2 = \sum_{j \in J} m^j \left\{ \sum_{k \in J} m^k \|j - k\|^2 \right\}$$

(d'après th. 1)
$$= \sum_{j \in J} \{ m^j M^2(J) + m^j \|j - g\|^2 \}$$
$$= \sum_{j \in J} m^j \cdot M^2(J) + \sum_{j \in J} m^j \cdot m^J \|j - g\|^2$$
$$= M^2(J) \cdot m^J + m^J \cdot \underbrace{\sum_{j \in J} m^j \|j - g\|^2}_{M^2(J)}$$
$$= 2m^J \cdot M^2(J) \qquad \textit{fin 2}$$

Dans le cas présent, nous cherchons à exprimer :

$$\partial(v, s \cup t) \overset{\text{déf}}{=} \|g(v) - g(s \cup t)\|$$

On pose $q = \{ s, t \}$, $t = g(t)$, $s = g(s)$, $v = g(v)$

D'après le théorème 1, on a :

$$m^s \|s - g(s \cup t)\|^2 + m^t \|t - g(s \cup t)\|^2 = m^s \|s - v\|^2 + m^t \|t - v\| - m^{\{s,t\}} \cdot \|g(s \cup t) - v\|^2$$

$$(m^t + m^s) \|g(s \cup t) - g(v)\|^2 = m^s \|s - v\|^2 + m^t \|t - v\|^2 - m^s \|s - g(s \cup t)\|^2 - m^t \|t - g(s \cup t)\|^2$$

(th. 2 en posant $J = \{ s, t \}$)

$$= m^s \|s - v\|^2 + m^t \|t - v\|^2 - \frac{m^s \cdot m^t}{(m^s + m^t)} \|s - t\|^2$$

(th. 1 en posant $q = \{ s, t \}$)

donc :

$$\|g(v) - g(s \cup t)\|^2 = + \frac{m^s}{(m^t + m^s)} \|s - v\|^2 + \frac{m^t}{(m^t + m^s)} \|t - v\|^2 - \frac{m^s \cdot m^t}{(m^t + m^s)^2} \|s - t\|^2$$

$$\partial(v, s \cup t) = \left[\frac{m^s}{(m^t + m^s)} \partial^2(s, v) + \frac{m^t}{(m^t + m^s)} \partial^2(v, t) - \frac{m^s \cdot m^t}{(m^t + m^s)^2} \partial^2(s, t) \right]^{1/2}$$

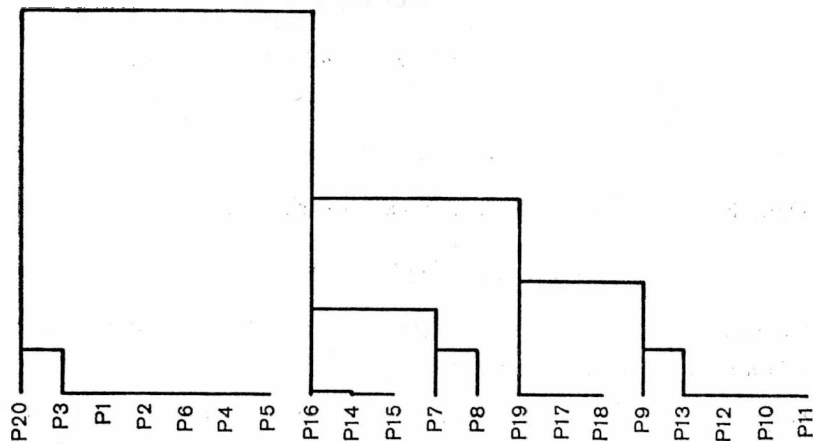
Remarque : Au début de la construction, on a :

$$\partial(\{j\}, \{k\}) = \|g(\{j\}) - g(\{k\})\| = \|j - k\| = d(j, k).$$

Les graphiques suivants illustrent cet algorithme pour les deux exemples présentés au § 3.

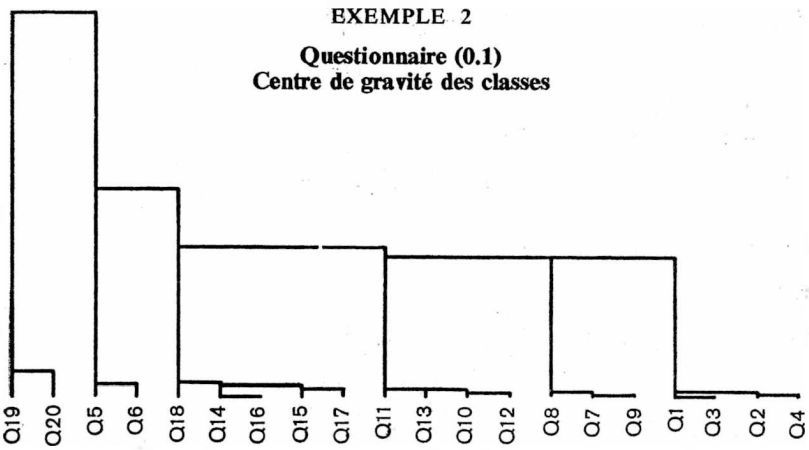
EXEMPLE 1

Nuage de 20 points dans un plan
Centre de gravité des classes



EXEMPLE 2

Questionnaire (0.1)
Centre de gravité des classes



Dans le premier exemple, on a : $\{ m^j = 1 \quad \forall j \in J \}$

Dans le deuxième exemple, on a : $m^j = p^j = \dfrac{\Sigma \{ k_{ij} \mid i \in I \}}{\Sigma \{ k_{ij} \mid i \in I, j \in J \}}$

4.5 Maximisation du moment centré d'ordre 2 d'une partition

Avant de présenter les stratégies suivantes, nous rappelons quelques définitions et propriétés utiles à l'étude des algorithmes.

4.5.1 Définitions et propriétés

Soit J un espace euclidien de dimension finie.

On pose :

$$g = g(J) = \frac{\sum_{j \in J} \{ m^j \cdot j \}}{\sum_{j \in J} m^j}$$

avec m^j masse associée au point $j \in J$

et $m^J = \sum_{j \in J} m^j$, masse de l'ensemble J

Moment centré d'ordre 2 de J :

$$M^2(J) \overset{\text{déf}}{=} \sum_{j \in J} \{ m^j \cdot \| j - g \|^2 \}$$

Variance de J :

$$V(J) \overset{\text{déf}}{=} \frac{M^2(J)}{m^J}$$

Théorème 3 : Soit Q une partition de J et q les classes qui la composent.

On a :

$$M^2(J) = M^2(Q) + \sum_{q \in Q} M^2(q)$$

avec

$$M^2(Q) = \sum_{q \in Q} m^q \| q - g \|^2 \; ; \qquad g \text{ centre de gravité de } J$$

et q centre de gravité de la classe q.

Lemme préliminaire : Soient a et b deux parties de J euclidien.

On a la relation suivante :

$$\sum_{\substack{j \in a \\ k \in b}} m^j m^k \| j - k \|^2 = m^a M^2(b) + m^b M^2(a) + m^a m^b \| g(a) - g(b) \|^2$$

avec $g(a)$ centre de gravité de $a \subset J$

$g(b)$ centre de gravité de $b \subset J$

Dém. (lemme)

$$\sum_{\substack{j\in a\\ k\in b}} m^j \cdot m^k \, \|j-k\|^2 = \sum_{j\in a} m^j \cdot \left(\sum_{k\in b} m^k \, \|j-k\|^2 \right)$$

(théorème 1) $$= \sum_{j\in a} m^j (M^2(b) + m^b \, \|j-g(b)\|^2)$$

$$= M^2(b)\left(\sum_{j\in a} m^j\right) + m^b \sum_{j\in a} m^j \, \|j-g(b)\|^2$$

(théorème 1) $$= m^a M^2(b) + m^b (M^2(a) + m^2 \, \|g(a)-g(b)\|^2)$$

$$= m^a \cdot M^2(b) + m^b \cdot M^2(a) + m^a \cdot m^b \cdot \|g(a)-g(b)\|^2$$

fin (lemme)

Dém. 3

$$2m^J \cdot M^2(J) = \left(\sum_{j\in J} \left(\sum_{k\in J} m^j \cdot m^k \cdot \|j-k\|^2 \right) \right) \qquad a, b \in Q \text{ partition de } J$$

$$= \sum_{\substack{a\in Q\\ b\in Q}} \left(\sum_{\substack{j\in a\\ k\in b}} m^j \cdot m^k \cdot \|j-k\|^2 \right)$$

(lemme) $$= \sum_{b,a\in Q} (m^a \cdot M^2(b) + m^b \cdot M^2(a) + m^a \cdot m^b \, \|g(a)-g(b)\|^2)$$

$$= \sum_{a,b\in Q} m^a \cdot M^2(b) + \sum_{a,b\in Q} m^b \cdot M^2(a)$$

$$\ldots + \sum_{a,b\in Q} m^a \cdot m^b \cdot \|g(a)-g(b)\|^2$$

$$= \sum_{a\in Q} m^a \cdot \left(\sum_{b\in Q} M^2(b) \right) + \sum_{b\in Q} m^b \cdot \left(\sum_{a\in Q} M^2(a) \right)$$

$$\ldots + \sum_{\substack{a\in Q\\ b\in Q}} M^a m^b \, \|g(a)-g(b)\|^2$$

$$= 2m^J \sum_{b\in Q} M^2(b) + 2mM^2(Q)$$

$$M^2(J) = M^2(Q) + \sum_{b\in Q} M^2(b)$$ *fin* 3

4.5.2 Définition de la stratégie d'agrégation

Soit Q_1 une partition de J, en n éléments

$$Q_1 = \{ a_i \mid a_i \subset J \mid i = 1, \ldots, n \}$$

On se propose de réunir deux parties a_j et a_k de J qui maximise le moment centré d'ordre 2 de la partition Q_2 de J construite ainsi :

$$Q_2 = \{ a_i \mid a_i \subset J \mid i = 1, \ldots, n \} \cup (a_j \cup a_k) - a_j - a_k$$

$$Q_2 = Q_1 \cup (a_j \cup a_k) - a_j - a_k$$

D'après le théorème 3, on a

$$M^2(J) = M^2(Q_1) + \sum_{a_i \in Q_1} M^2(a_i)$$

$$M^2(J) = M^2(Q_2) + \sum_{a_i \in Q_2} M^2(a_i)$$

$$\Rightarrow M^2(Q_2) + \sum_{a_i \in Q_2} M^2(a_i) = M^2(Q_1) + \sum_{a_i \in Q_1} M^2(a_i)$$

$$M^2(Q_2) = M^2(Q_1) + M^2(a_j) + M^2(a_k) - M^2(a_j \cup a_k)$$

$M^2(Q_1)$ a une valeur fixée. On choisit (a_j, a_k) tel que

$M^2(Q_2)$ soit maximum, c'est-à-dire tel que

$\{ M^2(a_j) + M^2(a_k) - M^2(a_j \cup a_k) \}$ soit maximum.

On définit alors la proximité entre 2 parties de J à partir d'une partition Q_1 de J de la façon suivante en minimisant ∂ défini ainsi :

$$\partial(a_j, a_k) = M^2(a_j \cup a_k) - M^2(a_j) - M^2(a_k)$$

$$\text{(th. 3)} = M^2 \{ a_j, a_k \}$$

or

$$M^2 \{ a_j, a_k \} = m^{a_j} \| a_j - g \|^2 + m^{a_k} \| a_k - g \|^2$$

avec $g = g(a_j \cup a_k)$ = centre de gravité de $a_j \cup a_k$

a_j = centre de gravité de la classe « a_j »

$$\Rightarrow M^2 \{ a_j, a_k \} = \frac{m^{a_j} \cdot m^{a_k}}{m^{a_j} + m^{a_k}} \| a_j - a_k \|^2$$

4.5.3 Définition de la proximité entre parties a et b de J

$$\partial(a, b) \overset{\text{déf}}{=} M^2 \{ a, b \} \overset{\text{déf}}{=} M^2(a \cup b) - M^2(a) - M^2(b)$$

$$\overset{\text{déf}}{=} \frac{m^a \cdot m^b}{m^a + m^b} \| g(a) - g(b) \|^2 \qquad \forall\, a, b \in \mathfrak{P}(J) - \varnothing$$

Remarque : $M^2 \{ a, b \} \geqslant 0$. On pourrait définir la proximité entre a et b de J d'une autre façon, en posant :

$$\partial_1(a, b) = [M^2 \{ a, b \}]^{1/2} = \sqrt{\frac{m^a \cdot m^b}{m^a + m^b}} \| g(a) - g(b) \|$$

On est alors ramené au cas de la stratégie d'agrégation liée à la distance entre centres de gravité des classes avec un coefficient de pondération $p(a, b)$ qui indique qu'à distance égale entre les centres de gravité des classes, l'agrégation porte sur les masses les plus faibles. Cette distance a l'avantage d'être homogène à la distance entre centres de gravité des classes.

— *Passage de l'étape de rang* $(n-1)$ *à l'étape de rang* (n). *Cas* ∂

$$\partial(v, s \cup t) = M^2\{v, s \cup t\} \quad \text{pour} \quad s, v, t \in \text{Som}(Hn-1) \quad \text{et} \quad v \neq s \neq t.$$

D'après les formules précédentes en posant $a = v$, et $b = s \cup t$, on a :

$$\partial(v, s \cup t) = \frac{m^v \cdot m^{s \cup t}}{m^v + m^s + m^t} \|g(v) - g(s \cup t)\|^2$$

$\|g(v) - g(s \cup t)\|^2$ a été calculé au § 4.4 « Centre de Gravité des classes »

$$\|g(v) - g(s \cup t)\|^2 = \frac{m^s}{m^t + m^s} \|s - v\|^2 + \frac{m^t}{m^t + m^s} \|t - v\|^2 - \frac{m^s \cdot m^t}{(m^t + m^s)^2} \|s - t\|^2$$

On pose $m = m^t + m^s + m^v$

$$\partial(v, s \cup t) = \frac{1}{m^t + m^s + m^v} [(m^s + m^v) \cdot \partial(s, v) + (m^t + m^v) \cdot \partial(t, v) - m^v \cdot \partial(s, t)]$$

Remarque : Au début de la construction de la hiérarchie, d'après les formules précédentes, on a :

$$\partial(\{j\}, \{k\}) = \frac{m^j \cdot m^k}{m^j + m^k} \cdot d^2(j, k)$$

On doit donc effectuer une transformation du tableau des distances initiales en posant :

$$d'(j, k) = \frac{m^j \cdot m^k}{m^j + m^k} \cdot d^2(j, k)$$

pour avoir $\partial(\{j\}, \{k\}) = d'(j, k)$.

Cette transformation introduit une déformation initiale sur la distance calculée sur J, et rend difficile la comparaison de la distance initiale d à la distance ultramétrique d^* construite à partir de la hiérarchie.

En utilisant la formule ∂_1 pour la construction de la hiérarchie, on obtient les formules suivantes :

— *Passage de l'étape de rang* $(n-1)$ *à l'étape de rang* (n). *Cas* ∂_1

$$\partial_1(v, s \cup t) = \sqrt{M^2\{v, s \cup t\}} \quad \text{pour} \quad v, s, t \in \text{Som}(Hn-1) \quad v \neq s \neq t$$

$$\partial_1^2(v, s \cup t) = \frac{m^t(m^t + m^s)}{m^t + m^s + m^v} \|g(v) - g(s \cup t)\|^2$$

$$\partial_1^2(v, s \cup t) = \frac{1}{m^t + m^s + m^v}$$

$$[(m^t + m^v) \cdot \partial_1^2(v, t) + (m^s + m^v) \cdot \partial_1^2(s, v) - m^v \cdot \partial_1^2(s, t)]$$

$$\partial_1(v, s \cup t) = \left[\frac{1}{m^t + m^s + m^v} \right.$$

$$\left. [(m^t + m^v) \cdot \partial_1^2(v, t) + (m^s + m^v) \cdot \partial_1^2(s, v) - m^v \cdot \partial_1^2(s, t)] \right]^{1/2}$$

Remarque : Au début de la construction, on effectue la transformation suivante :

$$d(j, k) \to d'(j, k) = \sqrt{\frac{m^j \cdot m^k}{m^j + m^k}} \cdot d(j, k)$$

pour avoir $\partial(\{j\}, \{k\}) = d'(j, k)$.

Cette formule permet une meilleure comparaison « numérique » de $d^*(j, k)$ à $d(j, k)$.

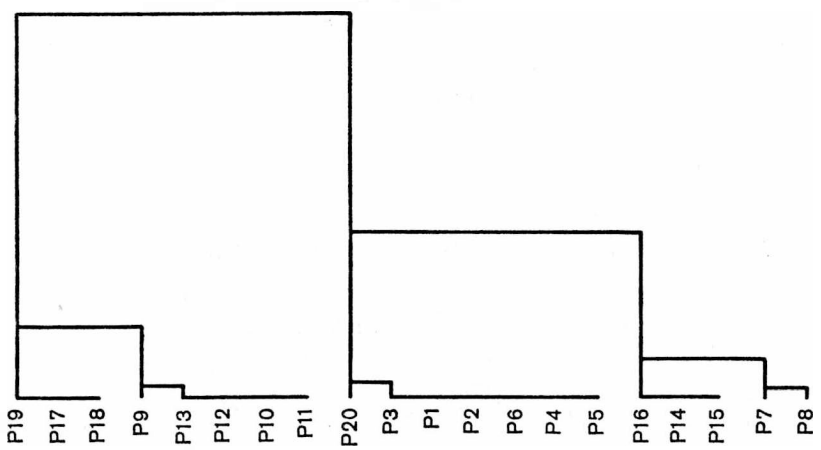
Les graphiques suivants illustrent cet algorithme (cas ∂) pour les deux exemples cités au § 3.

Dans le premier exemple, on a : $\{ m^j = 1 \quad \forall\, i \in J \}$

Dans le deuxième exemple, on a : $\left\{ m^j = p^j = \sum_{i \in I} k_{ij} \Big/ \sum_{i \in I, j \in J} k_{ij} \right\}$

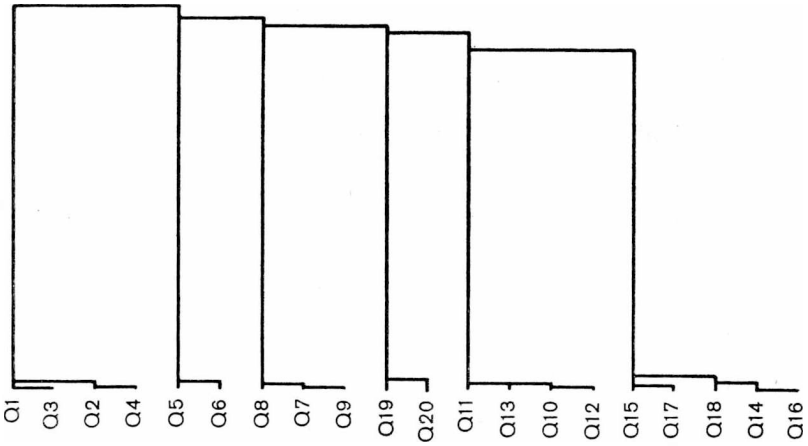
EXEMPLE 1

Nuage de 20 points dans un plan
Maximisation du moment d'ordre 2 d'une partition



EXEMPLE 2

Questionnaire (0.1)
Maximisation du moment d'ordre 2 d'une partition



4.6 Maximisation de la variance d'une partition

La construction de la hiérarchie indicée qui tient compte de la variance d'une partition (au lieu du moment centré d'ordre 2 de celle-ci) est analogue à celle exposée au § 4.5.

Nous rappelons succintement les différentes étapes et un théorème déduit du théorème 3.

Théorème 4 : Soit Q une partition de J et q les classes qui la composent.

On a :

$$V^2(J) = V^2(Q) + \sum_{q \in Q} \frac{m^q}{m^J} \cdot V^2(q)$$

avec

$$V^2(Q) = \frac{M^2(Q)}{m^J}$$

dém. 4

$$V^2(q) \overset{\text{déf}}{=} \frac{M^2(q)}{m^q} \Rightarrow m^q V^2(q) = M^2(q)$$

$$V^2(Q) \overset{\text{déf}}{=} \frac{M^2(Q)}{m^J} \Rightarrow m^J V^2(Q) = M^2(Q)$$

$$V^2(J) \overset{\text{déf}}{=} \frac{M^2(J)}{m^J} \Rightarrow m^J V^2(J) = M^2(J)$$

(th. 3)
$$\Rightarrow M^2(J) = M^2(Q) + \sum_{q \in Q} M^2(q)$$

$$m^J V^2(J) = m^J V^2(Q) + \sum_{q \in Q} m^q V^2(q)$$

$$\Rightarrow V^2(J) = V^2(Q) + \sum_{q \in Q} \frac{m^q}{m^J} \cdot V^2(q)$$

fin 4

— *Définition de la stratégie d'agrégation*

Nous cherchons comme dans le cas précédent à maximiser la variance d'une partition, étape par étape :

soit Q_1 une partition de J

$$Q_2 = Q_1 \cup (a_j \cup a_k) - a_j - a_k$$

On a :

$$V^2(Q_2) = V^2(Q_1) + \frac{m^{a_j}}{m^J} \cdot V^2(a_j) + \frac{m^{a_k}}{m^J} V^2(a_k) - \frac{m^{(a_j \cup a_k)}}{m^J} V^2(a_j \cup a_k)$$

On choisit alors : $\partial(a_j, a_k)$ de la façon suivante :

$$\partial(a_j, a_k) = \frac{1}{m^J} (\underbrace{- m^{a_j} \cdot V^2(a_j)}_{M^2(a_j)} \underbrace{- m^{a_k} \cdot V^2(a_k)}_{M^2(a_k)} + \underbrace{(m^{a_j} + m^{a_k}) V^2(a_j \cup a_k)}_{M^2(a_j \cup a_k)}$$

$$\partial(a_j, a_k) = \frac{1}{m^J} \underbrace{(M^2(a_j \cup a_k) - M^2(a_j) - M^2(a_k))}_{\partial(a_j, a_k) \text{ pour l'algorithme précédent}}$$

— *Définition de la proximité entre parties de J*

$$\partial(a, b) = \frac{1}{m^J} \left(\frac{m^a \cdot m^b}{m^a + m^b} \right) \| g(a) - g(b) \|^2 \qquad \forall\, a, b \in \mathfrak{P}(J) - \varnothing$$

La distance entre parties de J se déduit de celle calculée au § précédent à un coefficient multiplicateur près $\left(\frac{1}{m^J} \right)$

— *Passage de l'étape de rang* $(n-1)$ *à l'étape de rang* (n). *Cas* ∂

soit $\partial_M 2$ la « distance » entre parties associée à la maximisation du moment centré d'ordre 2 d'une partition;

$\partial_V 2$ la « distance » entre parties associée à la maximisation de la variance d'une partition.

$$\partial_V 2 = \frac{1}{m^J} \cdot \partial_M 2$$

L'indice de la hiérarchie associée à $\partial_V 2$ est divisé par la masse totale de l'ensemble J.

Les remarques faites sur l'algorithme précédent peuvent donc être faites sur celui-ci.

En particulier, on pourrait définir une nouvelle stratégie d'agrégation en posant

$$\partial_1(a, b) = \sqrt{\frac{1}{m^J} \cdot \frac{m^a \cdot m^b}{m^a + m^b}} \cdot \|g(a) - g(b)\|$$

Remarque :

Les hiérarchies construites par cet algorithme se déduisent de celles construites avec l'algorithme précédent, par homothétie de coefficient $(1/m^J)$. On peut considérer que les hiérarchies sont équivalentes aux précédentes (même composition des nœuds).

4.7 Moment centré d'ordre 2 des classes

— *Définition de la proximité entre parties de J*

On définit la proximité entre deux parties a et b de J de la façon suivante :

$$\partial(a, b) \overset{\text{déf}}{=} M^2(a \cup b) \qquad \forall\, a, b \in \mathfrak{P}(J) - \varnothing$$

∂ ne définit pas une distance, ni même un indice de distance sur J.

Le critère d'agrégation est donc de réunir, étape par étape, deux parties de J dont le moment centré d'ordre 2 de la réunion de ces deux parties est minimum.

$$f(a \cup b) = \inf \{ \partial(a, b), \mid a, b \in \text{Som}\,(Hn - 1) \}$$

— *Passage de l'étape de rang* $(n-1)$ *à l'étape de rang* (n)

Calcul de $\partial(v, s \cup t) = M^2(t \cup s \cup v) \qquad s,\ t, v \in \text{Som}\,(Hn-1)$

$$s \neq t \neq v$$

$a = t \cup s \cup v = (t \cup s) \cup (t \cup v) \cup (s \cup v) - s - t - v$ (ensembles deux à deux disjoints).

$$\underbrace{M^2(t \cup s \cup v)}_{= a} = \sum_{j \in a} m^j \|j - g(a)\|^2$$

D'après le théorème 2, on a :

$$2m^a M^2(a) = \sum_{j,k \in s \cup t} A + \sum_{j,k \in v \cup t} A + \sum_{j,k \in s \cup v} A - \sum_{j,k \in s} A - \sum_{j,k \in t} A - \sum_{j,k \in v} A$$

$$[A = m^j m^k \|j - k\|^2]$$

$$= 2m^{s \cup t} M^2(s \cup t) + m^{v \cup t} M^2(v \cup t) + 2m^{s \cup v} M^2(s \cup v)$$

$$\ldots - 2m^s M^2(s) - 2m^t M^2(t) - 2m^v M^2(v)$$

$$\Rightarrow (m^t + m^s + m^v)M^2(s \cup t \cup v)$$
$$= (m^s + m^t)M^2(s \cup t) + (m^t + m^v)M^2(t \cup v)$$
$$\ldots + (m^s + m^v)M^2(s \cup v) - m^sM^2(s) - m^tM^2(t) - m^vM^2(v)$$

$$\Rightarrow \partial(v, s \cup t) = \frac{1}{m^t + m^s + m^v} \ldots$$
$$\ldots [(m^s + m^t)\partial(s, t) + (m^t + m^v)\partial(t, v) + (m^s + m^v)\partial(s, v)$$
$$\ldots - m^sf(s) - m^tf(t) - m^vf(v)]$$

où $f(s), f(t), f(v)$ sont les indices des nœuds s, t, v de la hiérarchie.

Remarque 1. Au début de la construction, on transforme le tableau des distances initiales sur J en un tableau contenant les expressions suivantes :

$$d(j, k) \Rightarrow d'(j, k) = \frac{m^j \cdot m^k}{m^j + m^k} d^2(j, k)$$

On a ainsi : $\partial(\{j\}, \{k\}) = d'(j, k)$.

Remarque 2. On aurait pu définir d'une autre façon la proximité entre deux parties a et b de J, ainsi en prenant :

$$\partial_1(a, b) = \sqrt{M^2(a \cup b)} \qquad \forall\, a, b \in \mathfrak{P}(J) - \varnothing$$

Le critère d'agrégation devient alors le suivant : réunir deux parties a et b de J dont la racine carrée du moment centré d'ordre 2 de la réunion est minimum.

Cette construction modifie, par rapport au cas ∂, la valeur de l'indice hiérarchique, la valeur de d^* et la préordonnance associée à d^*.

On transforme les formules précédentes de la façon suivante :

$$\partial_1^2(v, s \cup t) = M^2(v \cup s \cup t)$$

$$\partial_1^2(v, s \cup t) = \frac{1}{m^t + m^s + m^v}$$
$$[(m^s + m^t)\partial_1^2(s, t) + (m^t + m^v)\partial_1^2(t, v) + (m^s + m^v)\partial_1^2(s, v)$$
$$- m^sf^2(s) - m^tf^2(t) - m^vf^2(v)]$$

avec $f(s), f(t), f(v)$ indices hiérarchiques des nœuds s, t, v de Som $(Hn - 1)$.

Le tableau $\{d(j, k), j, k \in J\}$ des distances initiales est transformé ainsi :

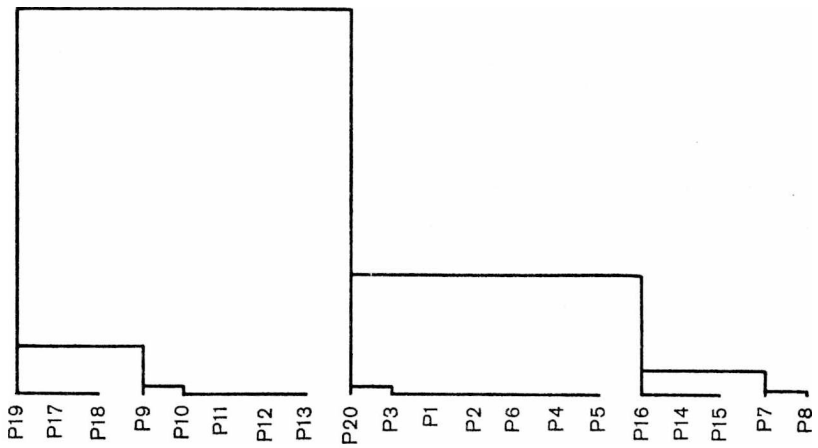
$$d(j, k) \Rightarrow d'(j, k) = \sqrt{\frac{m^j \cdot m^k}{m^j + m^k}} \cdot d(j, k)$$

pour avoir $\partial_1(\{j\}, \{k\}) = d'(j, k)$

Les graphiques suivants illustrent les résultats obtenus par cette stratégie d'agrégation (définition ∂) sur les deux exemples présentés au § 3.

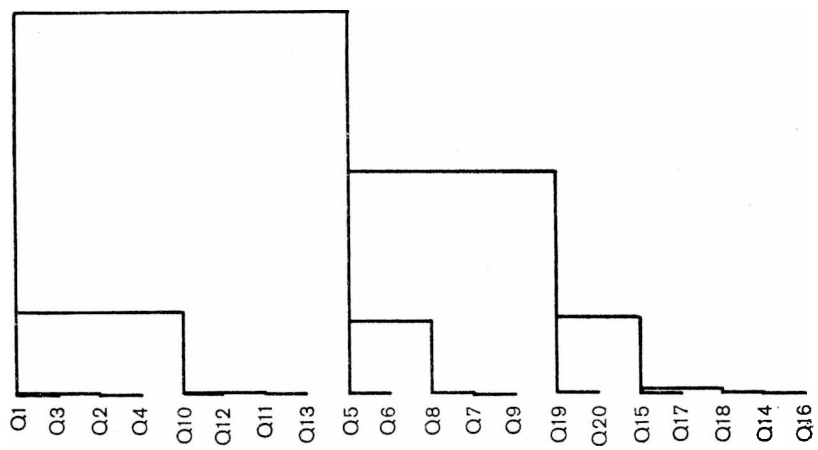
EXEMPLE 1

Nuage de 20 points dans un plan
Moment d'ordre 2 des classes



EXEMPLE 2

Questionnaire (0.1)
Moment d'ordre 2 des classes



4.8 Variance des classes

— *Définition de la proximité entre deux parties de J*

On définit la proximité entre deux parties a et b de J de la façon suivante :

$$\partial(a, b) = V^2(a \cup b) \qquad a, b \in \mathcal{P}(J) - \varnothing$$

∂ ne définit ni une distance, ni un indice de distance sur $\mathcal{P}(J) - \varnothing$.

Le critère d'agrégation est de réunir, étape par étape, deux parties de J dont la variance de l'union de ces deux parties est minimum.

$$\underset{\text{indice hiérarchique de } a \cup b}{f(a \cup b)} = \inf \{ \partial(a, b) \quad a, b \in \text{Som } (Hn - 1) \}$$

— *Passage de l'étape de rang* $(n - 1)$ *à l'étape de rang* (n)

Calcul de $\partial(v, s \cup t) = V^2(v \cup s \cup t)$ pour $s, t, v \in \text{Som } (Hn - 1)$
$s \neq t \neq v$

soit $a = s \cup t \cup v$

$$V^2(a) = \sum_{j \in a} \frac{m^j}{m^a} \| j - g(a) \|^2$$

D'après les calculs effectués au § 4.5

$$\text{avec} \quad V^2(a) = \frac{M^2(a)}{m^a}$$

on a :

$$(m^t + m^s + m^v)^2 \cdot V^2(s \cup t \cup v)$$
$$= (m^s + m^t)^2 \cdot V^2(s \cup t) + (m^t + m^v)^2 \cdot V^2(t \cup v) + (m^s + m^v)^2 \cdot V^2(s \cup v)$$
$$\ldots - (m^s)^2 \cdot V^2(s) - (m^t)^2 \cdot V^2(t) - (m^v)^2 \cdot V^2(v)$$

$$\Rightarrow \partial(v, s \cup t) = \frac{1}{(m^t + m^s + m^v)^2} \ldots$$

$$\ldots [(m^s + m^t)^2 \cdot \partial(s, t) + (m^t + m^v)^2 \cdot \partial(t, v) + (m^s + m^v)^2 \cdot \partial(s, v)$$
$$\ldots - (m^s)^2 \cdot f(s) - (m^t + m^v)^2 \cdot f(t) - (m^v)^2 \cdot f(v)]$$

où $f(s)$, $f(t)$, $f(v)$ sont les indices hiérarchiques des classes
$s,\ t,\ v \in \text{Som } (Hn - 1)$.

Remarque 1 : Calcul de $V^2(a \cup b)$

$$V^2(a \cup b) = V^2\{ a, b \} + \frac{m^a}{(m^a + m^b)} \cdot V^2(a) + \frac{m^b}{m^a + m^b} \cdot V^2(b)$$

avec $J = a \cup b$

$Q = \{ a, b \}$

(th. 4)

$$= \frac{m^a \cdot m^b}{(m^a + m^b)^2} \cdot \| a - b \|^2 + \frac{m^a}{(m^a + m^b)} V^2(a) + \frac{m^b}{(m^a + m^b)} V^2(b)$$

Ce calcul implique que pour $a = \{ j \}$, $b = \{ k \}$

$$\partial(\{ j \}, \{ k \}) = V^2(\{ j \} \cup \{ k \}) = \frac{m^j m^k}{(m^j + m^k)^2} \| j - k \|^2$$

Ceci implique une transformation du tableau des distances initiales

$$d(j,k) \xrightarrow{d'} d'(j,k) = \frac{m^j m^k}{(m^j + m^k)^2} \|j - k\|^2$$

pour avoir $$\partial(\{j\}, \{k\}) = d'(j,k).$$

Remarque 2. On aurait pu définir la proximité entre parties d'une autre façon, en posant :

$$\partial_1(a,b) = \sqrt{V^2(a \cup b)} \qquad \forall\, a, b \in \mathcal{P}(J) - \varnothing$$

Le critère d'agrégation devient alors le suivant : Réunir deux parties a et b dont la racine carrée de la variance de l'union est minimum.

Les formules précédentes sont transformées de la façon suivante :

$$\partial_1^2(v, s \cup t) = V^2(v \cup s \cup t)$$

$$\partial_1^2(s, s \cup t) = \frac{1}{(m^t + m^s + m^v)^2} \cdots$$

$$\cdots [(m^s + m^t) \cdot \partial_1^2(s,t) + (m^s + m^v) \cdot \partial_1^2(s,v) + (m^t + m^v) \cdot \partial_1^2(v,t)$$

$$\cdots - (m^s)^2 \cdot f^2(s) - (m^t)^2 \cdot f^2(t) - (m^v)^2 \cdot f^2(v)]$$

où $f(t)$, $f(s)$, $f(v)$ sont les indices hiérarchiques de s, t, $v \in \text{Som}\,(Hn - 1)$.

Au début de la construction, la transformation suivante doit être faite :

$$d(j,k) \to d'(j,k) = \sqrt{\frac{m^j \cdot m^k}{(m^j + m^k)^2}}\, \|j - k\|$$

$$d'(j,k) = \frac{d(j,k)}{(m^j + m^k)} \cdot \sqrt{m^j \cdot m^k}$$

alors

$$\partial_1(\{j\}, \{k\}) = d'(j,k)$$

Les graphiques suivants illustrent cette stratégie d'agrégation (définition ∂) sur les deux exemples présentés au § 3.

4.9 Distances angulaires

On suppose J euclidien.

— *Définition de la proximité entre deux parties a et b de J*

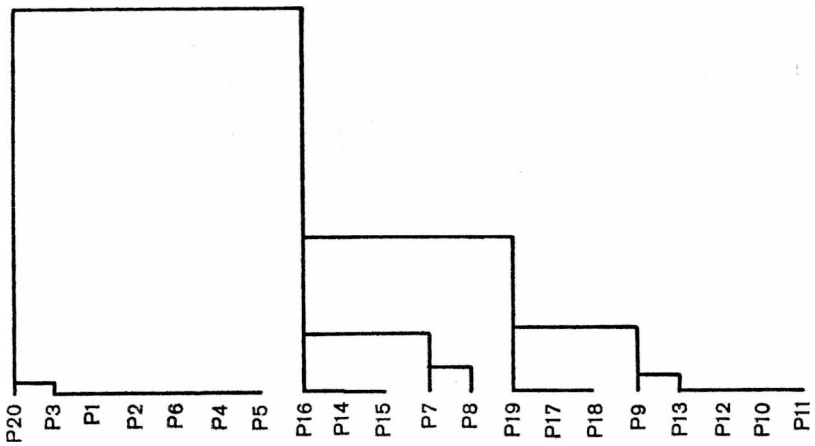
A chaque partie a de J on associe le vecteur de description de a : $V(a)$

$$V(a) = \sum_{j \in a} V(j)$$

avec $V(j)$ vecteur de description de l'élément j.

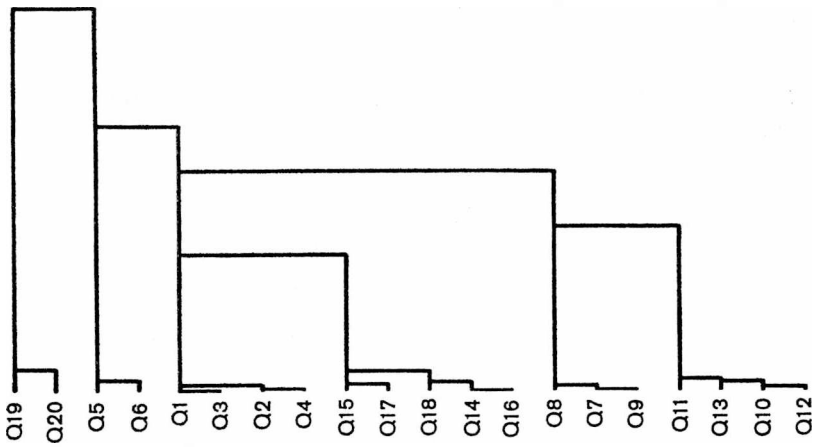
EXEMPLE 1

Nuage de 20 points dans un plan
Variance des classes



EXEMPLE 2

Questionnaire (0.1)
Variance des classes



On définit une similitude entre deux parties a, b de J de la façon suivante :

$$\text{sim} \cdot (a, b) = \frac{\langle V(a),\ V(b) \rangle}{\|V(a)\|\ \|V(b)\|} = \text{Cos}\,(V(a), V(b))$$

la « distance » prend alors la forme suivante :

$$\partial(a, b) = 1 - \text{sim}\,(a, b) \qquad \forall\ a, b \in \mathfrak{P}(J) - \varnothing$$

On a $\quad \partial(a, a) = 0$

$\partial(a, b) = 0 \quad V(a)$ et $V(b)$ sont colinéaires

$\partial(a, b) = \partial(b, a)$

∂ ne vérifie pas l'inégalité triangulaire

∂ ne définit pas une distance sur $\mathcal{P}(J) - \varnothing$.

— *Passage de l'étape de rang* $(n-1)$ *à l'étape de rang* (n)

Calcul de $\partial(v, s \cup t) = 1 - \text{sim}\,(v, s \cup t)$ pour $s, t, v \in \text{Som}\,(Hn-1)$

$$s \neq t \neq v$$

$$\text{sim}\,(v, s \cup t) = \frac{\langle \cdot\, V(v),\, V(s \cup t) \rangle}{\|V(v)\| \cdot \|V(s \cup t)\|}$$

$$V(s \cup t) = V(s) + V(t)$$

$$\Rightarrow \langle V(v), V(s \cup t) \rangle = \langle V(v), V(s) \rangle + \langle V(v), V(t) \rangle$$

$$= \|V(s)\| \cdot \|V(v)\| \cdot \text{sim}\,(s, v) + \|V(v)\| \cdot \|V(t)\| \cdot \text{sim}\,(t, v)$$

$$\|V(s \cup t)\|^2 = \|V(s) + V(t)\|^2 = \|V(s)\|^2 + 2 \langle V(s), V(t) \rangle + \|V(t)\|^2$$

$$= \|V(s)\|^2 + 2 \|V(s)\| \cdot \|V(t)\| \cdot \text{sim}\,(t, s)$$

$$+ \|V(t)\|^2$$

$$\Rightarrow \text{sim}\,(v, s \cup t) = \frac{\|V(s)\| \cdot \text{sim}\,(v, s) + \|V(t)\| \cdot \text{sim}\,(v, t)}{\|V(s) + V(t)\|}$$

Remarque : On peut noter une similitude de formule entre la stratégie d'agrégation par la moyenne et celle-ci :

$$\text{sim.}\,(a, b) = \frac{\langle V(a), V(b) \rangle}{\|V(a)\| \cdot \|V(b)\|}$$

$$\text{sim.}\,(a, b) = \frac{\langle \sum_{j \in a} V(j), \sum_{j \in b} V(k) \rangle}{\| \sum_{j \in a} V(j) \| \cdot \| \sum_{k \in b} V(k) \|}$$

$$= \frac{\sum_{j \in a . k \in b} \langle V(j), V(k) \rangle}{\| \sum_{j \in a} V(j) \| \cdot \| \sum_{k \in b} V(k) \|}$$

La similitude angulaire joue un rôle analogue à celui donné par la distance moyenne. $\|V(a)\|$ joue le rôle du poids Card (a) de la classe a, et sim (a, b) celui de la distance moyenne.

De ces résultats, nous déduisons la formule suivante :

$$\partial(v, s \cup t) = 1 - \left[\frac{\|V(s)\| \cdot (1 - \partial(s, v)) + \|V(t)\| \cdot (1 - \partial(t, v))}{\|V(s) + V(t)\|} \right]$$

avec

$$\| V(t) + V(s) \| = (\| V(s) \|^2 + 2 \| V(s) \| \cdot \| V(t) \| \ (1 - \partial(t, s)) + \| V(t) \|^2)^{1/2}$$

— *Les variantes de l'algorithme*

Des variantes de cette stratégie d'agrégation peuvent être proposées.

1) En proposant le poids de la classe comme norme d'une classe, on obtient ainsi la formulation suivante :

$$\partial(v, s \cup t) = 1 - \left[\frac{\text{Card } (s) \cdot (1 - \partial(s, v)) + \text{Card } (t) \cdot (1 - \partial(t, v))}{\| V(s) + V(t) \|} \right]$$

avec

$$\| V(t) + V(s) \| = ((\text{Card } (s))^2 + 2 \text{ Card } (s) \cdot \text{Card } (t)(1 - \partial(t, s)) + (\text{Card } (t))^2)^{1/2}$$

2) Une autre variante de cette stratégie peut être proposée en normalisant les vecteurs associés à une classe.

$$\forall \ a \in J \qquad \| V(a) \| = 1$$

On a alors les formules suivantes :

$$\partial(v, s \cup t) = 1 - \left[\frac{2 - \partial(s, v) - \partial(t, v)}{\| V(t) + V(s) \|} \right]$$

avec

$$\| V(t) + V(s) \| = [4 - 2 \cdot \partial(s, t)]^{1/2}$$

On a associé à la réunion de deux classes un vecteur unité orienté suivant la somme géométrique des vecteurs qui la composent.

Remarque : Les différents algorithmes exposés ci-dessus nécessitent, au début de la construction, d'avoir $\partial(\{j\}, \{k\}) = d(j, k)$

donc

$$\text{sim}(\{j\}, \{k\}) = \frac{\langle V(j), V(k) \rangle}{\| V(j) \| \cdot \| V(k) \|}$$

Nous ne pouvons légitimer ces trois variantes de la stratégie d'agrégation selon la moyenne angulaire que pour une distance initiale, calculée à partir d'un coefficient de similitude compris entre [— 1 et + 1].

5. CONCLUSIONS

La multiplicité des critères d'agrégation utilisables en classification automatique a de quoi rendre sceptique le praticien. On doit cependant se garder de tout jugement hâtif sur ces critères, et par-delà la multiplicité de ceux-ci, sur l'intérêt des méthodes de classification. A partir du moment

où, en classification automatique, on a admis (comment faire autrement !) qu'il n'y a pas d'optimum absolu, il convient de faire choix de critères locaux, partiels, mais cependant utiles. Il conviendra naturellement d'assortir ces constructions hiérarchiques de contrôles (nous préférons ne pas employer le terme test qui a une signification précise en statistique), contrôles a priori, et contrôles a posteriori. Les contrôles a priori dépendent essentiellement du choix effectué par le praticien, d'un indice de distance et des poids éventuels à associer à chaque couple de variables. Les contrôles a posteriori nous sont fournis par l'examen minutieux des classes constituées et par l'étude de la qualité de la représentation des données dont nous donnerons un aperçu dans un prochain exposé. Avant cela, il convient d'examiner attentivement la signification du critère d'agrégation. Celui-ci contient les germes de certaines incohérences; le choix d'une stratégie apparaît donc aussi difficile que celui d'un indice de distance. Chacune d'entre elles possède un « *biais algorithmique* », c'est-à-dire, dans certains cas, une trop grande importance du poids des variables, au fait qu'il faille un seul lien pour réunir deux ilôts, au fait qu'il faille réunir deux classes trop peu concentrées... Pour pallier ces effets, on ne saurait que trop conseiller au statisticien de pratiquer plusieurs stratégies à partir d'une même distance initiale, voire à changer de critère d'agrégation au cours de la construction hiérarchique.

Dans un prochain article, nous traiterons des critères d'appréciation de la qualité de la représentation des données que fournissent les algorithmes de classification hiérarchique.

CONSOMMATION

XXIe ANNÉE, N° 4, OCTOBRE-DÉCEMBRE 1974

RÉSUMÉS - ABSTRACTS

des articles contenus dans ce Numéro

ASPECTS GÉOGRAPHIQUES DU SYSTÈME DES SOINS MÉDICAUX; analyse des données départementales, par L. LEBART, S. SANDIER et F. TONNELLIER. *Consommation*, 4-1974, octobre-décembre 1974, pages 5 à 50.

Les nombreuses statistiques départementales contribuent à donner chaque année près de cent images du système des soins médicaux en France. Cet article présente une recherche des similitudes et une analyse des différences en vue de préciser :

— les interrelations entre la santé et les autres aspects de la vie sociale;

— les liens entre la morbidité et les soins médicaux;

— le processus d'ajustement de l'offre et de la demande de soins;

— les possibilités de substitutions entre les différents types de soins.

Après avoir fait, en recourant à l'analyse factorielle des données, un bilan global des distributions de 173 variables décrivant l'offre et la consommation de soins médicaux, le niveau et les causes de la mortalité, l'environnement socio-économique au niveau départemental, quelques conclusions se sont dégagées.

Il semble que l'industrialisation qui implique un certain mode de vie et le climat aient un effet néfaste sur l'état de santé global des départements.

Alors que la fréquentation des hôpitaux publics est très liée à la surmorbidité analysée à partir des causes de décès, les soins médicaux aux malades ambulatoires ou les hospitalisations dans le secteur privé sont plus en relation avec le niveau de développement des départements qu'avec les niveaux de morbidité.

MEDICAL CARE, GEOGRAPHICAL ASPECTS; an analysis of regional data, by L. LEBART, S. SANDIER and F. TONNELLIER. *Consommation*, 4-1974, October-December 1974, pages 5 to 50.

Every year, statistical data per region give nearly a hundred pictures of medical care in France. The authors have tried to find similarities and analyse differences to detect :

— the interrelations between health and other aspects of social life;

— the relations between morbidity and medical care;

— the process of adjustment of supply and demand of medical care;

— the eventual substitutions between the different types of care.

Non parametric factor analysis leads to a mapping of 173 variables describing supply and demand of medical care, the level and the causes of death, the economic and social environment on a regional level. Results show : industrialisation, that implies a certain way of life, and climatic factors seem to have a had effect on health.

The attendance of public hospitals is related to « over-mortality » analysed through causes of death, medical care to ambulatory cases or hospitalisation in private hospitals are more related to the level of regional development than to the level of morbidity.

On a regional level, death rates cannot be considered as indicators of the results of medical care, because the regional variations of incidental morbidity are considerable and overlap.

Supply of medical care seems to have an important effect on consumption. The availability of doctors — the same sort

Au niveau départemental, les taux de mortalité ne peuvent pas être considérés comme des indicateurs de résultats des soins médicaux, car les variations régionales de morbidité incidente sont très grandes et interfèrent de façon considérable.

L'offre de soins médicaux semble jouer un rôle très important au niveau des consommations. La disponibilité des médecins — variable analogue au taux d'occupation des lits dans les hôpitaux — semble influer de façon sensible sur les taux d'accroissement des consommations.

Il semblerait que deux types de médecines coexistent en France. L'une plus traditionnelle fait encore une large part aux visites au domicile du malade et à la prescription pharmaceutique, l'autre plus technique utilise conjointement toutes les techniques de diagnostic et de soins. Entre ces deux extrêmes, toutes les situations mixtes existent. Deux types de substitutions semblent possibles : l'hospitalisation dans le secteur privé de soins prendrait le relais des visites au domicile du malade; le niveau technique de la médecine pratiquée par le secteur privé de soins serait en relation inverse avec la fréquentation des hôpitaux publics.

Deux directions d'étude sont proposées: un recueil de monographies sur des départements représentatifs de la variété des situations; des études privilégiant les phénomènes d'évolution; par contre, il serait illusoire de vouloir chiffrer dans un modèle l'influence propre de telle ou telle variable.

of variable as the rate of occupation of hospital beds — seems to have a notable effect on the rate of increase of consumption.

Two types of care seem to co-exist in France : one, traditionnal, where visits to the patients'home and drug prescriptions hold a large part, the other, more technological, uses in the same time all the technologies for diagnosis and care. Between these two extreme cases, there is a variety of mixed situations. Two sorts of substitution seem possible : hospitalisation in the private sector might replace visits to the patients'home; the technological level of care in the private sector would then be inversely proportionnal to the attendance of public hospitals.

The authors suggest two lines for further research : monographic studies on selected regions, and dynamic studies. It would not be realistic to evaluate in a model the effect of any particular variable.

VIEILLESSE ET CLASSE SOCIALE. L'exemple des paysans bénéficiaires de l'indemnité viagère de départ et celui des petits commerçants, par P. REYNAUD et B. ZARCA. *Consommation*, 4-1974, octobre-décembre 1974, pages 51 à 80.

Nous nous proposons, dans cet article, de comparer deux ensembles d'individus repérés empiriquement par leur appartenance à une catégorie socio-professionnelle, au moment où ils entrent dans la vieillesse.

Les uns sont des paysans, les autres des petits commerçants. Nous montrons que dans le contexte historique actuel et malgré quelques ressemblances, ces deux ensembles d'individus abordent la vieillesse dans des conditions radicalement différentes :

— Les premiers, les paysans, qui ont

OLD AGE AND SOCIAL CLASSES. An example: Farmers and shop-keepers, by P. REYNAUD and B. ZARCA. *Consommation*, 4-1974, October-December 1974, pages 51 to 80.

The paper presents a comparison of two groups of inviduals who are on the eve of old age.

The first are farmers who gave up the management of their farms and benefit of a life-time aid. The second are small shop-keepers. In the present historical circumstances, though likeness do exist between the two groups, individuals belonging to them are in totally different situations :

— Farmers are « semi-retired » men, rooted in their setting. Though they are getting on in age, they still can make plans for the future, their own individual

pu prendre une « pseudo-retraite » sont des hommes enracinés dans un milieu. Bien que devenant vieux, ils peuvent encore se projeter dans l'avenir, leur projet individuel se fondant dans celui du groupe social dont ils sont des membres reconnus. L'entrée dans la vieillesse est, pour eux, un moment positif de leur vie, celui où le travail-obligation se transforme en travail-librement-choisi.

— Les seconds, les petits commerçants, sont dans l'impossibilité de se retirer et, leur travail se dévalorisant chaque jour, courent à une quasi-faillite. Ils vivent seuls leurs contradictions, chacun n'ayant avec les autres petits commerçants que des liens sériels. L'entrée dans la vieillesse est, pour eux, le moment négatif de la réalisation de l'échec de leur vie, de la perte de leur identité sociale.

Ce contraste met en lumière le fait que la vieillesse est une condition sociale qui ne peut se comprendre qu'en l'éclairant par les rapports sociaux existant dans une formation sociale.

Après avoir décrit brièvement les deux populations ayant fait l'objet de notre recherche, laquelle s'est appuyée sur des données d'enquêtes, nous situons notre problématique : nous cherchons à décrire le mode d'existence sociale de ces personnes âgées dans le cadre d'une analyse de classes. Puis, nous présentons les résultats de nos analyses : les conditions objectives d'entrée dans la vieillesse du paysan et son expérience vécue. Par contraste, celles du petit commerçant.

plan is part of plan of the social group of which they are recognized members. The beginning of old age is a positive moment of their lives : work is no longer a necessity, but it is their own free choice.

— Small shop-keepers are unable to retire, their work is daily devaluated, and consequently, they are headed to ruin. They live their own contradictions in loneliness, since they only have serial links with other small shop-keepers. The beginning of old age is a negative moment of their lives as they realize their failure and the loss of their social identity.

The contrast shows that old age is a social situation that cannot be apprehended without taking into account the social links within a group.

After a short description of the two populations based on the results of a survey, the authors attempt to describe the social insertion of these old people in the frame work of class analysis. Results show the objective setting of the farmers and the subjective evaluation of their situations, in sharp contrast with those of small shop-keepers.

SUR LES CRITÈRES D'AGRÉGATION UTILISÉS EN CLASSIFICATION AUTOMATIQUE, par M. JAMBU. *Consommation*, 4-1974, octobre-décembre 1974, pages 81 à 106.

Dans cet article, l'auteur propose un ensemble assez vaste de critères d'agrégation utilisables dans un même algorithme. De cette étude, il ressort essentiellement que le choix d'un critère d'agrégation est soumis à des arbitraires nombreux auxquels s'ajoutent les arbitraires dus aux choix d'une mesure de similarité. Malgré cela, les méthodes de classification ascendante ne sont pas à rejeter si le praticien prend le soin de choisir un critère adapté aux données qu'il veut traiter.

AGGREGATION CRITERIA USED IN AUTOMATIC CLASSIFICATION, by M. JAMBU. *Consommation*, 4-1974, October-December 1974, pages 81 to 106.

The author suggests a fairly large set of aggregation criteria to be used in the same algorithm. Results show that the choice of the aggregation criteria is very arbitrary, all the more so that the measures of similarity are also arbitrary. However, these methods of increasing classification are not be rejected if the pratician chooses criteria adapted to the data he is to analyse.

Le directeur de la publication P. BORDAS.

Dépôt légal /ED. 4e trimestre 1974. N° 029. N° de commission paritaire 29837.

Imprimé en France. — 2/1975. IMPRIMERIE NOUVELLE, ORLÉANS. N° 7128.

CONSOMMATION (ANNALES DU C. R. E. D. O. C.)

1970

N° 1. — La fréquentation des équipements collectifs. — La supériorité de la gestion collective de l'épargne mobilière : analyse méthodologique et application aux SICAV. — Le comportement des exploitants agricoles en Eure-et-Loir et en Ille-et-Vilaine.

N° 2-3. — L'Évolution de la consommation des ménages de 1959 à 1968.

N° 4. — Les services médicaux en Suède et en France. — Proposition pour une méthodologie de l'étude de la redistribution. — La consommation des boissons dans quelques pays d'Europe.

1971

N° 1. — Les familles devant l'éducation des enfants. — Nouvelle évaluation de la fortune des ménages (1959-1967). — Budget-temps et choix d'activité.

N° 2. — Enquête sur les loisirs et mode de vie du personnel de la Régie Nationale des Usines Renault. — Étude des effets différentiels des impôts sur la consommation. — La morphologie sociale des communes urbaines.

N° 3. — La consommation élargie. — Étude économique de l'activité des médecins. — Possibilités et difficultés de la régulation des problèmes d'environnement et de nuisance par entente spontanée entre les intéressés.

N° 4. — Nature et prix des soins médicaux en ville. — Quelques résultats de l'étude des bilans de petites et moyennes entreprises.

1972

N° 1. — Enquête sur les loisirs et mode de vie du personnel de la Régie Nationale des Usines Renault. — Les choix de consommation et les budgets des ménages. — Placement et Investissement. — Les budgets familiaux dans les régions de la C.E.E.

N° 2. — Les sciences humaines devant la ville et le logement. — Qualité de la vie et choix collectifs. Consommation et statut social. — Tests d'hypothèses linéaires sur un modèle de régression.

N° 3. — Le système d'indicateurs du VI• Plan. — Recherche de projections cohérentes pour des variables interdépendantes. — L'arbitrage entre salaire et temps libre.

N° 4. — L'évolution de la consommation des ménages de 1959 à 1970.

1973

N° 1. — Rôle des valeurs et politique sociale. — A qui profite l'impôt ? Mythes et réalités. — Les entreprises financières en mutation face au commerce de l'épargne. — Les leçons d'une enquête sur les petits commerçants âgés. — Cheminements aléatoires et modèles systématiques d'intervention. Bourse des valeurs de Paris. — Les dépenses de soins médicaux au Canada de 1957 à 1969.

N° 2. — Consommation des ménages et consommation publique «divisible». — Inflation et processus de décision. — Vers une description du mode de vie au moyen d'indicateurs.

N° 3. — Un indicateur de morbidité. — Rémunère-t-on les études ? — Les immigrés : réflexions sur leur insertion sociale et leur intégration juridique. — Introduction à l'analyse des données ; les méthodes de classification automatique.

N° 4. — Un premier bilan de la redistribution des revenus en France : les impôts et cotisations sociales à la charge des ménages en 1965.

1974

N° 1. — Recherche et politique sociale. — Les facteurs démographiques et la croissance des consommations médicales. — La justice civile, sa place dans la société française.

N° 2. — La consommation pharmaceutique en 1970. — Une définition des dépenses d'éducation des familles. — L'utilisation des études à long terme dans la planification française. — Sur les indices de distances en vue de la construction d'une classification hiérarchique.

N° 3. — L'essentiel ou le résidu : le cas de la planification urbaine. — Diffusion des consommations médicales de ville dans la population en 1970. — Les grèves dans l'économie française.

SOMMAIRE DES PROCHAINS NUMÉROS

Structure et inégalité des patrimoines. L'appréciation monétaire d'un surplus dans la consommation alimentaire de différentes catégories sociales. Quelques critères de comparaison des hiérarchies indicées produites en classification automatique. L'orientation du dépouillement des enquêtes.

sommaire

CENTRE DE RECHERCHES ET DE DOCUMENTATION SUR LA CONSOMMATION

45, boulevard de la Gare, PARIS-13e

Tél. 707-97-59

1974 n° 4

Octobre-Décembre