QUELQUES CRITÈRES DE COMPARAISON DES HIÉRARCHIES INDICÉES PRODUITES EN CLASSIFICATION AUTOMATIQUE (1)

par

Michel JAMBU

SOMMAIRE

---

(1) Cet article fait suite à trois articles précédemment parus dans la revue *Consommation*, n° 3-1973, n° 2-1974, n° 4-1974.

1. INTRODUCTION

Dans trois précédents articles sur la classification automatique on a présenté l'algorithme général des constructions ascendantes hiérarchiques, assorti d'une gamme assez large de formules de calculs de similarités et des principaux critères d'agrégation actuellement utilisés. Comme nous avions fait remarquer qu'à partir d'un même tableau de données on pouvait obtenir des représentations arborescentes assez différentes, la question qui vient tout naturellement à l'esprit du lecteur concerne le crédit qu'il doit accorder à de telles méthodes. En d'autres termes, existe-t-il des critères formels simples qui permettent de choisir, même *a posteriori*, telle méthode plutôt que telle autre. Et pour cela, il convient de définir, en premier lieu, ce que l'on entend par comparaisons des critères d'agrégation.

2. COMMENT COMPARER LES CRITÈRES D'AGRÉGATION ?

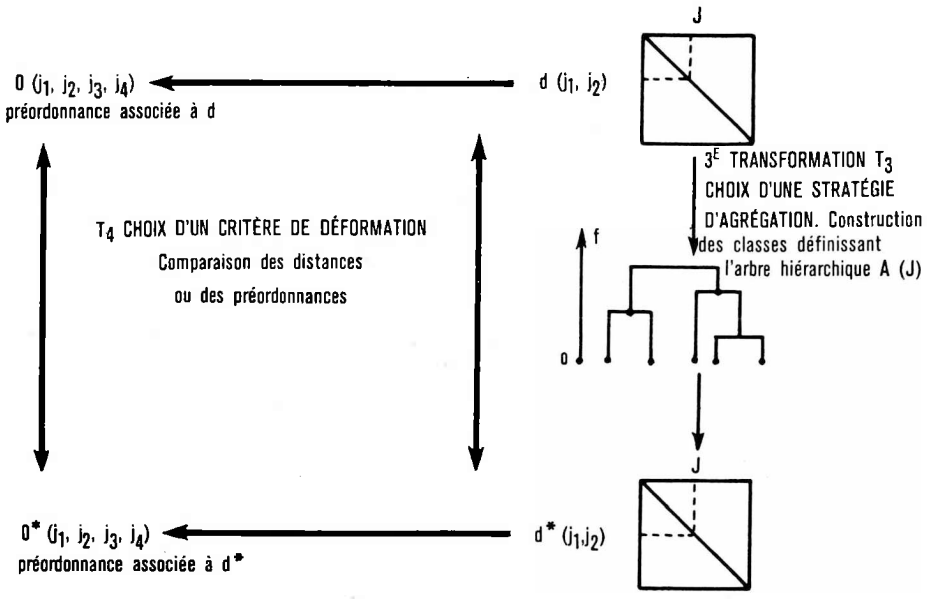
On se souvient de la démarche du taxinomiste : observer une population selon certains critères, structurer ces observations en constituant un tableau de données puis, faisant choix d'une mesure de similarité, formaliser les rapports entre critères ou entre les éléments de la population. En dernier, il doit représenter cette distance; ce qui est fait en faisant choix d'un critère (ou stratégie) d'agrégation. En résumé, le but du taxinomiste est de faire passer les observations dans des filtres successifs qui transforment la réalité première, inconnue ou difficile à lire, en une autre réalité, plus facile à déchiffrer, mais obtenue au prix de certaines déformations. Son but est de réduire, à chaque étape, ces déformations. Pour que les comparaisons aient un sens, elles doivent être replacées par rapport à chacune des étapes de la structuration des données. Ainsi, dans le présent article, on s'attachera à l'étude des comparaisons des hiérarchies indicées associées à une même distance, à la comparaison des mesures de similitude associées à un même tableau de données. On pourrait s'attacher aux divers codages possibles d'un tableau de données. Ce point, fort important puisqu'il est le point de départ de l'analyse, ne sera pas traité dans le présent article.

3. COMPARAISON DES HIÉRARCHIES INDICÉES TOTALES, ASSOCIÉES A UNE MÊME MESURE DE SIMILARITÉ s OU A UN MÊME INDICE DE DISTANCE d

3.1. Quelques rappels sur les notations

On supposera, dans ce paragraphe, qu'on connaît une mesure de similarité s sur un ensemble J ou un indice de distance d sur ce même ensemble. On uniformise les notations en considérant pour un s donné, l'indice $d = 1—s$.

FIGURE 1



A l'entrée du « filtre » caractérisé par la transformation T_3, on trouve l'indice de distance d, et à la sortie du filtre, une « représentation arborescente » à partir de laquelle on calcule sur J une distance d^* de la façon suivante :

$$d^*(j, k) = f(a(j, k)),$$

où f est l'indice de la hiérarchie et $a(j, k)$ la plus petite partie de la hiérarchie qui contient à la fois j et k.

La distance d^* est ultramétrique.

Comme on sait, d'autre part, qu'il est équivalent de se donner sur un ensemble J une distance ultramétrique d^* et une hiérarchie de parties de J, totale et indicée (*cf. Consommation* n° 3-1973, p. 113, théorème 2), toutes les comparaisons entre hiérarchies de parties indicées totales sur J relatives à un indice de distance sur J, reviennent à comparer deux objets de même nature : l'indice de distance d et la distance ultramétrique d^* (ce qui est illustré par la figure 1). On distinguera, par la suite, deux types de comparaisons : en premier lieu, celles fondées sur la comparaison des valeurs respectives de d et de d^*; en second lieu, celles fondées sur la comparaison des inégalités entre d et d^*. Chaque comparaison est illustrée par les calculs correspondant à des exemples présentés précédemment (*cf. Consommation* n° 2-1974).

3.2. Les comparaisons de type numérique. Formules

On considère

$$J^* = \{ p = (j, k) \in J \times J \text{ tel que } j > k \}$$
$$\text{Card}(J^*) = \text{card}(J) * (\text{card}(J) - 1)/2$$

comme d et d^* sont telles que

$$d(j, k) = d(k, j), \qquad \forall (j, k) \in J \times J$$

et

$$d^*(j, k) = d^*(k, j), \qquad \forall (j, k) \in J \times J$$

et que

$$d(j, j) = d^*(j, j) = 0, \qquad \forall j \in J,$$

on étudie les applications d et d^* sur J^*.

En fait, on cherche un indice de déformation de la distance (ou de l'indice d) par l'application de la stratégie d'agrégation qui a conduit à la définition de d^*. On notera DEF (d, d^*) un tel nombre auquel on exige seulement qu'il prenne la valeur nulle si les distances coïncident. Il prendra la valeur 1 pour certaines formules (celles déduites des similarités).

Voici la liste d'un certain nombre d'indices de déformation utilisables en classification automatique et par toute méthode qui a pour base de départ une mesure de similarité ou un indice de distance.

3.2.1. *Formules*

• *moyenne quadratique des écarts entre d et d* :*

$$\text{(D 1)} \qquad \begin{cases} \text{DEF}(d, d^*) = \left[\dfrac{\sum\limits_{(j,k) \in J^*} (d(j, k) - d^*(j, k))^2}{\text{Card}(J^*)} \right]^{1/2}, \\ (d = d^*) \Rightarrow \text{DEF}(d, d^*) = 0. \end{cases}$$

• *moyenne quadratique pondérée des écarts entre d et d* :*

$$\text{(D 2)} \qquad \text{DEF}(d, d^*) = \left[\frac{\sum\limits_{(j,k) \in J^*} p(j, k).(d(j, k) - d^*(j, k))^2}{\text{Card}(J^*)} \right]^{1/2},$$

où $p(j, k)$ est un coefficient de pondération associé au couple (j, k) d'éléments de J^*;

$$(d = d^*) \Rightarrow \text{DEF}(d, d^*) = 0.$$

• *moyenne de la différence des carrés des distances d et d* :*

$$\text{(D 3)} \qquad \begin{cases} \mathrm{DEF}(d, d^*) = \left[\dfrac{\sum\limits_{(j,k) \in J^*} \left| d^2(j,k) - d^{*2}(j,k) \right|}{\mathrm{Card}(J^*)} \right], \\ \text{GOWER (1970)}, \\ (d = d^*) \Rightarrow \mathrm{DEF}(d, d^*) = 0. \end{cases}$$

• *moyenne pondérée de la différence des carrés des distances d et d* :*

$$\text{(D 4)} \qquad \begin{cases} \mathrm{DEF}(d, d^*) = \left[\dfrac{\sum\limits_{(j,k) \in J^*} p(j,k) \left| d^2(j,k) - d^{*2}(j,k) \right|}{\mathrm{Card}(J^*)} \right], \\ (d = d^*) \Rightarrow \mathrm{DEF}(d, d^*) = 0. \end{cases}$$

• *distance angulaire :*

$$\text{(D 5)} \qquad \begin{cases} \mathrm{DEF}(d, d^*) = \left[\dfrac{\sum\limits_{(j,k) \in J^*} d(j,k).d^*(j,k)}{\left[\left(\sum\limits_{(j,k) \in J^*} d^2(j,k) \right) . \left(\sum\limits_{(j,k) \in J^*} d^{*2}(j,k) \right) \right]^{1/2}} \right], \\ \text{GUTTMAN (1968)}, \\ (d = d^*) \Rightarrow \mathrm{DEF}(d, d^*) = 1. \end{cases}$$

• *moyenne (ou moyenne pondérée) de la valeur absolue des différences des distances :*

$$\text{(D 6)} \qquad \begin{cases} \mathrm{DEF}(d, d^*) = \left[\dfrac{\sum\limits_{(j,k) \in J^*} p(j,k) \left| d(j,k) - d^*(j,k) \right|}{\mathrm{Card}(J^*)} \right], \\ (d = d^*) \Rightarrow \mathrm{DEF}(d, d^*) = 0; \end{cases}$$

$$\text{(D 7)} \qquad \begin{cases} \mathrm{DEF}(d, d^*) = \dfrac{1}{\mathrm{Card}(J^*)} \left[\sum\limits_{(j,k) \in J^*} \dfrac{\left| d(j,k) - d^*(j,k) \right|}{d(j,k) + d^*(j,k)} \right], \\ (d = d^*) \Rightarrow \mathrm{DEF}(d, d^*) = 0; \end{cases}$$

$$\text{(D 8)} \qquad \begin{cases} \mathrm{DEF}(d, d^*) = \sup\limits_{(j,k) \in J^*} \left| d(j,k) - d^*(j,k) \right|, \\ (d = d^*) \Rightarrow \mathrm{DEF}(d, d^*) = 0; \end{cases}$$

$$\text{(D 9)} \qquad \begin{cases} \mathrm{DEF}(d, d^*) = \left[\dfrac{\sum\limits_{(j,k) \in J^*} \dfrac{(d(j,k) - d^*(j,k))^2}{\sup \left| d(j,k) - d^*(j,k) \right|}}{\mathrm{Card}(J^*)} \right]^{1/2}, \\ (d = d^*) \Rightarrow \mathrm{DEF}(d, d^*) = 0; \end{cases}$$

(D 11)
$$\begin{cases} \underset{\lambda \in]0,1]}{\mathrm{DEF}}(d, d^*) = \left[\dfrac{\sum\limits_{(j,k) \in J^*} |d(j,k) - d^*(j,k)|^{1/\lambda}}{\mathrm{Card}(J^*)} \right]^{\lambda}, \\ \lambda = \dfrac{1}{2}, \quad [\mathrm{D}11] = [\mathrm{D}1], \quad \lambda = 1, \quad [\mathrm{D}11] = [\mathrm{D}12], \\ (d = d^*) \Rightarrow \mathrm{DEF}(d, d^*) = 0; \end{cases}$$

(D 12)
$$\begin{cases} \mathrm{DEF}(d, d^*) = \left[\dfrac{\sum\limits_{(j,k) \in J^*} (d(j,k) - d^*(j,k))}{\mathrm{Card}(J^*)} \right], \\ (d = d^*) \Rightarrow \mathrm{DEF}(d, d^*) = 0; \end{cases}$$

• *corrélation des distances :*

(D 13)
$$\begin{cases} \mathrm{DEF}(d, d^*) = \dfrac{\sum\limits_{(j,k) \in J^*} (d(j,k) - \bar{d})(d^*(j,k) - \bar{d}^*)}{\left\{ \left[\left(\sum\limits_{(j,k) \in J^*} (d(j,k) - \bar{d}) \right)^2 \right. \right. \\ \left. \left. \times \left(\sum\limits_{(j,k) \in J^*} (d^*(j,k) - \bar{d}^*) \right)^2 \right] \right\}^{1/2}}, \\ (d = d^*) \Rightarrow \mathrm{DEF}(d, d^*) = 1; \end{cases}$$

où $\bar{d}$ et $\bar{d}^*$ représentent les moyennes des distances d et d^*

soient $S(d) = \sum\limits_{(j,k) \in J^*} d(j,k)$, $\quad S^*(d^*) = \sum\limits_{(j,k) \in J^*} d^*(j,k)$.

• *Distance entre profils :*

(D 14)
$$\begin{cases} \mathrm{DEF}(d, d^*) \\ \quad = \left[\sum\limits_{(j,k) \in J^*} \dfrac{1}{d(j,k) + d^*(j,k)} \left[\dfrac{d(j,k)}{S(d)} - \dfrac{d^*(j,k)}{S^*(d^*)} \right]^2 \right]^{1/2}, \\ (d = d^*) \Rightarrow \mathrm{DEF}(d, d^*) = 0. \end{cases}$$

Cette définition est utile pour l'étude des hiérarchies indicées dont les indices varient très fortement. Cette formule est fondée sur la comparaison des « profils » des systèmes de distances.

(D 15)
$$\begin{cases} \mathrm{DEF}(d, d^*) = \left[\dfrac{\sum\limits_{(j,k) \in J^*} \left[\dfrac{d^2(j,k)}{d^{*2}(j,k)} \right]}{\mathrm{Card}(J^*)} \right] \\ \text{SHEPARD (1969), CAROLL (1969),} \\ (d = d^*) \Rightarrow \mathrm{DEF}(d, d^*) = 1; \\ \mathrm{DEF}(d, d^*) \neq \mathrm{DEF}(d^*, d); \end{cases}$$

(D 16)
$$\left\{\begin{array}{l} \mathrm{DEF}(d, d^*) = \left[\dfrac{\sum\limits_{(j,k)\in J^*}\left[\dfrac{d(j,k)}{d^*(j,k)}\right]}{\mathrm{Card}(J^*)}\right] \\ \text{CAROLL (1969), KRUSKALL (1969),} \\ (d = d^*) \Rightarrow \mathrm{DEF}(d, d^*) = 1, \\ \mathrm{DEF}(d, d^*) \neq \mathrm{DEF}(d^*, d). \end{array}\right.$$

Ces deux formules sont surtout utilisées en analyse des proximités pour comparer une distance et son approximation d^*.

Les calculs de déformation sont, en général, des calculs de moyenne ou de somme sur les paires de points de J et s'expriment souvent sous la forme générale :

$$\mathrm{DEF}(d, d^*) = \sum_{(j,k)\in J^*} \{p(j,k).f(d(j,k) - d^*(j,k))\}$$

$p(j,k)$ = pondération associée au couple (i,k)

très souvent, $p(j,k) = \dfrac{m_j . m_k}{m_j + m_k}$ si m_j, m_k sont les masses associées aux points j et k.

Les formules suivantes servent à préciser la position de la distance d^* vis-à-vis de la distance d.

On calcule les quantités suivantes :

a = nombre de fois où $d^*(j,k) > d(j,k)$,
b = nombre de fois où $d^*(j,k) = d(j,k)$,
c = nombre de fois où $d^*(j,k) < d(j,k)$,

(D 17)
$$\mathrm{DEF}(d, d^*) = \frac{a-c}{\mathrm{Card}(J^*)},$$
$$\mathrm{DEF}(d, d^*) = 0 \Leftrightarrow (a = c);$$

(D 18)
$$\mathrm{DEF}(d, d^*) = \frac{a}{c} \quad \text{avec} \quad c \neq 0,$$
$$\mathrm{DEF}(d, d^*) = 1 \Leftrightarrow (a = c).$$

3.2.2. *Quelques remarques sur les formules*

Avant d'illustrer ces formules par des calculs sur les différents exemples que nous proposons dans ce texte, nous devons faire quelques remarques d'ordre général sur les indices des hiérarchies. En effet, les distances ultramétriques d^* peuvent avoir des valeurs très différentes selon les stratégies d'agrégation. Celles-ci sont déduites, rappelons-le, de l'indice de la hiérarchie de la façon suivante :

$$d^*(j,k) = f(a(j,k)),$$

où f est l'indice de $a(j, k)$, plus petite partie de la hiérarchie qui contient j et k.

Or, les valeurs de l'indice $f(a(j, k))$ dépendent également de la technique de construction et non uniquement du caractère intrinsèque des données.

Certaines techniques d'agrégation exigent parfois une transformation initiale du tableau des distances $\{d(j, k)\}$.

Il en est ainsi pour les stratégies suivantes (moment d'ordre 2 d'une partition, variance d'une partition, moment d'ordre 2 d'une partie, variance d'une partie, barycentre), afin d'avoir, au début de la construction :

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

avec

$$d'(j, k) = p(j, k).d^2(j, k)$$
$$\text{ou} \qquad p(j, k).d(j, k).$$

Ainsi, les distances ultramétriques d^*, construites à partir de l'indice de la hiérarchie f, n'expriment pas nécessairement la déformation réelle d'une distance d. Aussi, l'interprétation de certains calculs de déformation qui mettent trop en relief l'aspect purement numérique de différences entre distances est difficile. Pour pallier cet inconvénient, on peut alors calculer une déformation sur des distances transformées, définies comme suit :

• 1re *transformation possible :*

On pose

$$d^{**}(j, k) = \frac{d^*(j, k)}{\sup\limits_{(j,k) \in J^*} \{d^*(j, k)\}},$$

$$d'(j, k) = \frac{d(j, k)}{\sup\limits_{(j,k) \in J^*} \{d(j, k)\}};$$

on calcule ainsi DEF (d', d^{**}) au lieu de DEF (d, d^*) sur des distances transformées. On a alors

$$\sup_{(j,k) \in J^*} \{d'(j, k)\} = \sup_{(j,k) \in J^*} \{d^{**}(j, k)\} = 1;$$

on représente souvent les hiérarchies indicées sous une même forme en faisant en sorte que l'indice de J soit égal à l'unité.

• 2e *transformation possible :*

On pose

$$d^{**}(j, k) = \frac{d^*(j, k)}{\sum\limits_{(j,k) \in J^*} \{d^*(j, k)\}},$$

$$d'(j, k) = \frac{d(j, k)}{\sum_{(j,k) \in J^*} \{d(j, k)\}};$$

on calcule alors DEF (d', d^{**}) (sauf pour les formules de calculs qui tiennent déjà compte des profils).

- 3e *transformation possible :*

On pose

$$d^{**}(j, k) = d^*(j, k) - \bar{d}^*,$$

$$d'(j, k) = d(j, k) - \bar{d}.$$

$\bar{d}^*$ et $\bar{d}$ sont les moyennes des distances sur J associées respectivement à d^* et d.

La transformation consiste à centrer les distances et calculer les déformations sur les distances centrées.

3.3. Comparaisons de type numérique. Exemples

Les résultats suivants illustrent sur deux exemples précédemment cités (*cf. Consommation* n° 2-1974), les formules ci-dessous décrites. Dans le premier exemple (nuage de 20 points dans un plan), la distance de départ est la distance euclidienne usuelle. Dans le deuxième exemple, la distance de départ est la distance associée à l'analyse des correspondances. On a joint à ces calculs une représentation graphique des distances de départ et des distances ultramétriques associées. (En abscisse, les éléments de J^* ordonné par l'ordre lexicographique usuel; en ordonnée, la valeur de la distance — la ligne brisée rejoignant l'ensemble des points donne une représentation de la distance.)

- *Exemple* 1 : figure 1 [Nuage de 20 points dans un plan]

code des stratégies d'agrégation :

$k = 1$ — UIM = ultramétrique inférieure maximale;
$k = 2$ — USM = ultramétrique supérieure minimale;
$k = 3$ — MOY = distance moyenne;
$k = 4$ — MAX = maximisation du moment d'ordre 2 d'une partition;
$k = 5$ — MOM = moment d'ordre 2 des classes;
$k = 6$ — VAR = variance des classes;
$k = 7$ — CEN = centre de gravité des classes;
D. MIN = déformation minimale pour le critère de déformation considéré.

Tableau des déformations DEF (d, d_k^*)

	UIM.	USM.	MOY.	MAX.	MOM.	VAR.	CEN.	D.MIN
[D1]	0.016	0.016	0.008	0.009	0.023	0.023	0.018	0.
[D3]	- 0.106	0.174	0.005	- 0.020	0.258	- 0.130	- 0.118	0.
[D5]	0.915	0.966	0.951	0.945	0.931	0.913	0.935	1.
[D6]	0.167	0.168	0.085	0.098	0.250	0.274	0.214	0.
[D7]	0.288	0.185	0.125	0.343	0.422	0.793	0.578	0.
[D8]	0.580	0.580	0.350	0.315	0.722	0.646	0.552	0.
[D9]	0.021	0.021	0.014	0.016	0.027	0.029	0.025	0.
[D12]	- 0.167	0.168	0.012	- 0.043	0.178	- 0.274	- 0.213	0.
[D13]	0.587	0.857	0.786	0.820	0.800	0.715	0.753	1.
[D14]	2.345	1.213	1.539	1.325	0.927	2.339	2.014	0.
[D15]	0.405	2.746	1.517	0.817	2.708	0.027	0.145	1.
[D16]	0.589	1.552	1.123	0.735	1.296	0.125	0.305	1.
[D17]	- 0.884	0.889	0.011	- 0.263	0.074	- 1.000	- 0.937	0.

Les nombres soulignés indiquent pour chaque mode de calcul de la déformation les deux stratégies « les plus proches » de la distance initiale *d*.

Pour réduire l'écart excessif entre distances, nous avons effectué certains calculs de déformation sur des distances transformées.

1er *cas* : $d'(j, k) \overset{T_1}{\to} \dfrac{d(j, k)}{S(d)}$, $\quad d^{**}(j, k) \overset{T_1}{\to} \dfrac{d^*(j, k)}{S(d^*)}$.

Tableau des déformations DEF (d', d_k^{**})

	UIM.	USM.	MOY.	MAX.	MOM.	VAR.	CEN.	D. MIN
[D5]	0.915	0.966	0.951	0.945	0.931	0.913	0.935	1.
[D6]	0.002	0.001	0.001	0.002	0.002	0.002	0.002	0.
[D7]	0.185	0.125	0.126	0.339	0.376	0.360	0.299	0.
[D8]	0.007	0.005	0.006	0.006	0.007	0.007	0.006	0.
[D13]	0.587	0.857	0.786	0.820	0.800	0.715	0.753	1.
[D15]	1.799	1.178	1.411	1.093	1.112	1.408	1.319	1.
[D16]	1.241	1.017	1.083	0.850	0.830	0.901	0.920	1.
[D17]	0.411	- 0.179	- 0.126	- 0.263	- 0.200	- 0.316	- 0.326	0.

2e *cas* : $d'(j, k) \overset{T_2}{\to} \dfrac{d(j, k)}{\sup\limits_{(j,k) \in J^*} \{d(j, k)\}}$,

$$d^{**}(j,k) \xrightarrow{T_2} \frac{d^{*}(j,k)}{\sup\limits_{(j,k)\in J^{*}} \{d(j,k)\}}.$$

Tableau des déformations DEF (d', d_k^{**})

	UIM.	USM.	MOY.	MAX.	MOM.	VAR.	CEN.	D. MIN
[D1]	0.027	0.022	0.024	0.022	0.023	0.024	0.023	0.
[D3]	0.352	0.334	0.332	0.286	0.268	0.261	0.287	0.
[D5]	0.915	0.966	0.951	0.945	0.931	0.913	0.935	1.
[D6]	0.321	0.232	0.240	0.239	0.256	0.256	0.233	0.
[D7]	0.312	0.185	0.212	0.363	0.402	0.369	0.308	0.
[D8]	0.709	0.804	0.813	0.804	0.804	0.813	0.813	0.
[D9]	0.032	0.025	0.026	0.025	0.025	0.027	0.026	0.
[D12]	0.277	0.232	0.240	0.162	0.134	0.143	0.183	0.
[D13]	0.587	0.857	0.786	0.820	0.800	0.715	0.753	1.
[D14]	0.889	0.875	0.885	0.751	0.742	0.769	0.784	0.
[D15]	4.765	2.746	3.362	2.042	1.888	2.471	2.638	1.
[D16]	2.019	1.552	1.672	1.162	1.082	1.193	1.301	1.
[D17]	0.695	0.889	0.995	0.237	0.068	0.068	0.132	0.

D'après ces calculs, trois stratégies semblent être assez souvent celles qui minimisent les déformations. Ce sont les agrégations selon l'ultramétrique supérieure minimale, la distance moyenne, la maximisation du moment centré d'ordre 2 d'une partition.

- *Exemple* 2 : figure 2. Questionnaire (0.1)

Tableau des déformations DEF (d, d_k^{*})

	UIM.	USM.	MOY.	MAX.	MOM.	VAR.	CEN.	D. MIN
[D1]	0.024	0.031	0.017	0.171	0.099	0.049	0.522	0.
[D3]	- 1.723	2.285	0.690	- 8.704	- 0.553	1.778	91.886	0.
[D5]	0.999	0.998	0.998	0.982	0.895	0.982	0.959	1.
[D6]	0.278	0.328	0.159	2.210	1.115	0.514	5.771	0.
[D7]	0.049	0.052	0.027	0.658	0.319	0.156	0.447	0.
[D8]	1.085	1.085	0.813	3.965	3.507	1.592	14.373	0.
[D9]	0.023	0.029	0.019	0.086	0.053	0.038	0.138	0.
[D12]	- 0.276	0.326	0.087	- 2.210	- 0.334	0.092	5.705	0.
[D13]	0.988	0.980	0.983	0.862	0.518	0.936	0.884	1.
[D14]	0.184	0.165	0.172	0.268	0.158	0.156	0.080	0.
[D15]	0.830	1.253	1.058	0.060	0.942	0.977	8.239	1.
[D16]	0.909	1.115	1.026	0.216	0.825	0.929	2.639	1.
[D17]	- 0.879	0.879	0.284	- 0.989	- 0.084	0.158	0.737	0.

De la lecture de ces différents résultats, il ressort que les comparaisons de type numérique sont difficiles à interpréter. Pour certaines d'entre elles, un biais est immédiatement introduit (MAX, MOM, CEN, VAR) en pondérant la distance initiale. On ne doit donc comparer les critères d'agrégation que s'ils appartiennent à un même groupe. Pour le premier groupe (UIM, USM, MOY), le critère de la moyenne semble minimiser tous les indices de déformation proposés. Dans le second groupe (MAX, MOM,

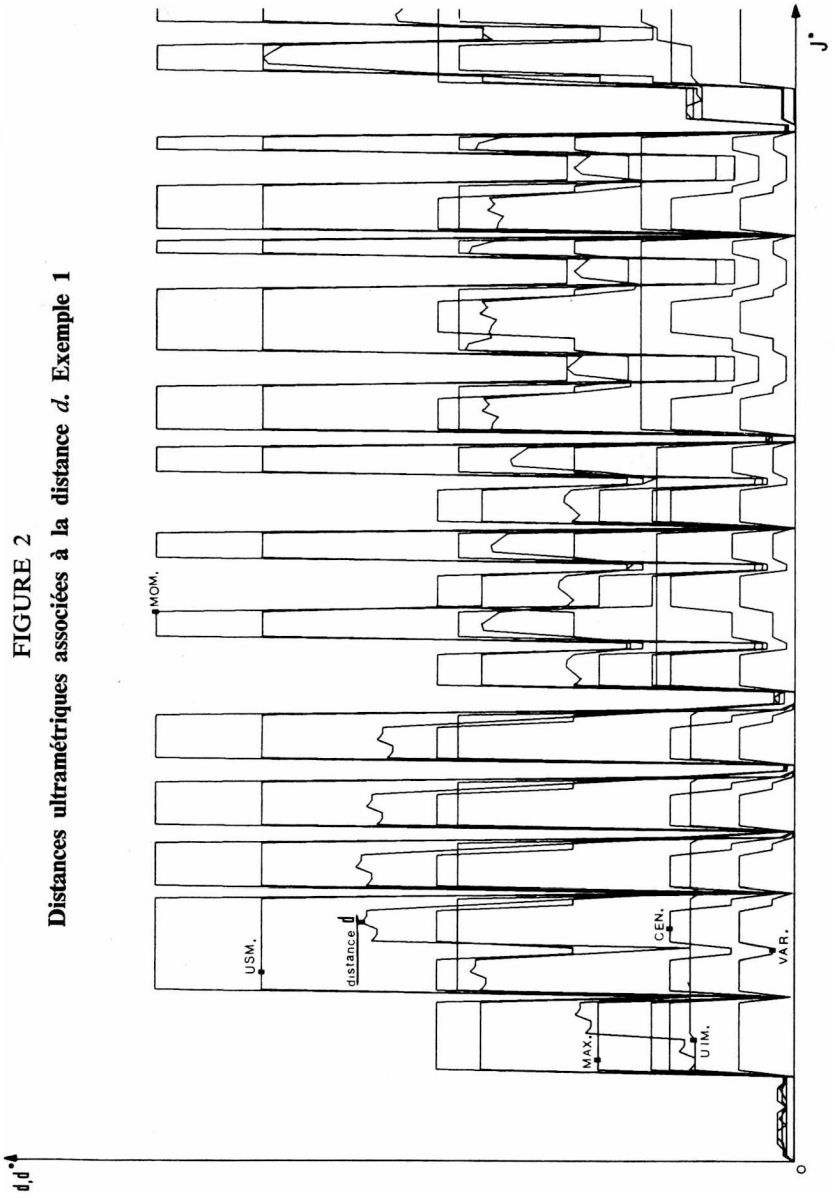


FIGURE 2
Distances ultramétriques associées à la distance *d*. Exemple 1

CEN, VAR), le critère d'agrégation par la maximisation de la variance d'une partition minimise les indices de déformation dans le cas du premier exemple, et le critère de la variance dans le second exemple.

Pour pallier les différents inconvénients évoqués ci-dessus, nous proposons un ensemble d'indices de déformation fondés sur les comparaisons des inégalités entre distances.

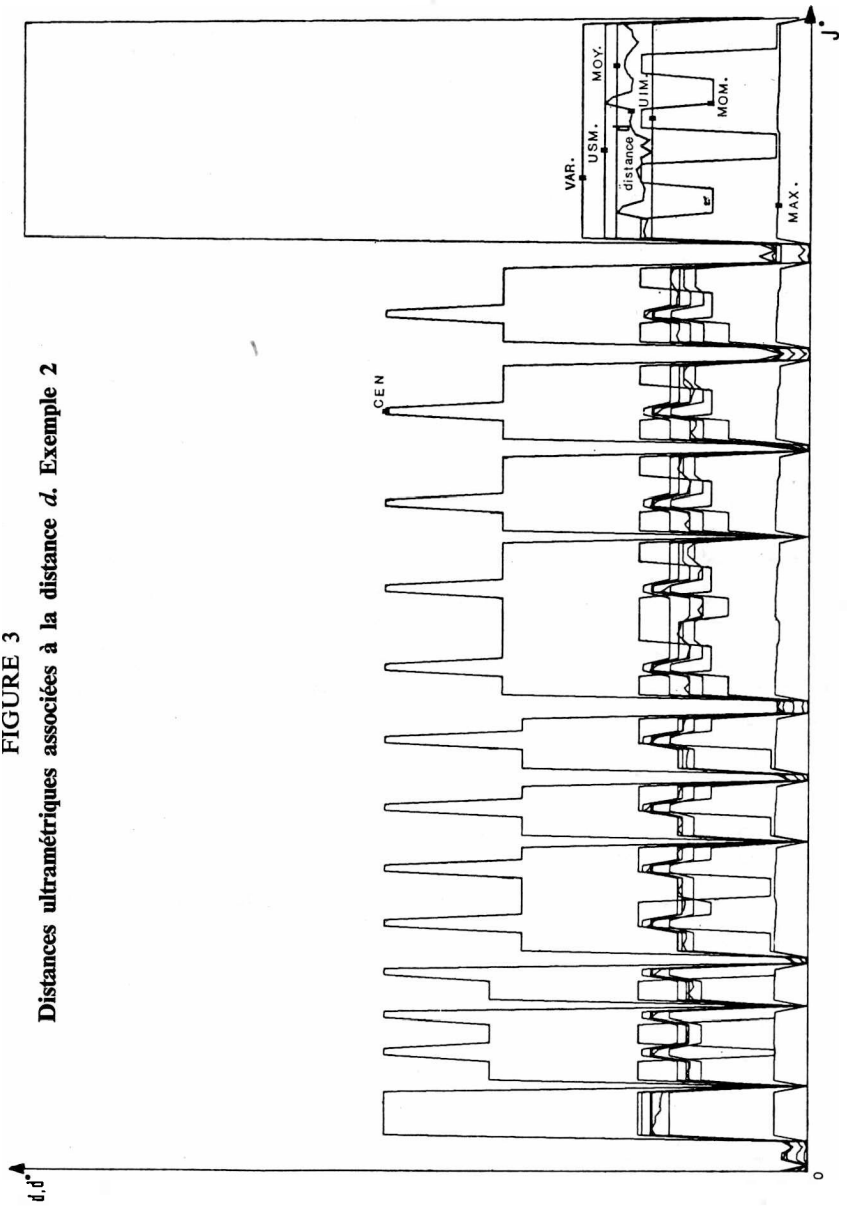


FIGURE 3

Distances ultramétriques associées à la distance *d*. Exemple 2

soient :

$$J^* = \{ p = (j, k);\ (j, k) \in J \times J;\ j > k,\ j \neq k \},$$
$$D = \{ d(j, k) \mid (j, k) \in J^* \},$$
$$D^* = \{ d^*(j, k) \mid (j, k) \in J^* \}.$$

3.4.1. *Coefficient de corrélation des rangs de Spearman*

A chaque valeur $d(j, k)$ d'un couple (j, k) de J^*, on associe son rang dans J^*, $r(j, k)$, l'élément de J^*, $p = (j, k)$ qui prend la plus petite valeur aura le rang 1, la plus grande valeur card (J^*). On fait de même pour $d^*(j, k)$.

La déformation est exprimée par le coefficient de corrélation des rangs de Spearman.

$$\text{(D 18)} \qquad \begin{cases} = sp[(r(j, k), r^*(j, k))], \\ (d = d^*) \Rightarrow (r = r^*) \Rightarrow \mathrm{DEF}(d, d^*) = 1; \end{cases}$$

$$\text{(D 18')} \qquad \begin{cases} \mathrm{DEF}(d, d^*) = \dfrac{6. \sum\limits_{(j,k) \in J^*} (r(j, k) - r^*(j, k))}{\mathrm{Card}(J^*).[\mathrm{Card}^2(J^*) - 1]}, \\ (d = d^*) \Rightarrow \mathrm{DEF}(d, d^*) = 0; \\ \text{(D 18')} = 1 - \text{(D 18)}. \end{cases}$$

Si les distances ont des rangs identiques, la déformation est nulle.

L'inconvénient d'un tel calcul provient de la notation des rangs. Les distances ultramétriques ont un nombre important d'égalités. Le calcul du coefficient de corrélation des rangs peut prendre des valeurs différentes selon la façon de noter les rangs des paires situés au même niveau hiérarchique.

3.2.4. *Coefficient de corrélation des rangs de Kendall*

Le coefficient de corrélation des rangs de Kendall tient compte dans son calcul de la préordonnance associée à une distance.

Le calcul de ce coefficient s'effectue sur l'ensemble des couples de d ou de d^*; il y a donc

$$\frac{\mathrm{Card}(J^*).(\mathrm{Card}(J^*) - 1)}{2}$$

couples d'observations $((j, k), (l, m))$.

Le calcul est effectué à partir des rangs $r(j, k)$ et $r^*(j, k)$.

On note le nombre de fois où $(r(j, k)$—$r(l, m))$ est de même signe que $(r^*(j, k)$—$r^*(1, m))$ et le nombre de fois où ces expressions sont de signe

différent; soient n^+ le premier de ces nombres, n^- le second de ces nombres. Le coefficient de corrélation de Kendall s'établit ainsi :

$$\tau = \frac{2(n^+ - n^-)}{\text{Card}(J^*).[\text{Card}(J^*)-1]};$$

si les expressions $[r(j, k)—r(1, m)]$ et $[r^*(j, k)—r^*(1, m)]$ sont de même signe, on a

$$\left.\begin{array}{l} n^+ = \dfrac{\text{Card}(J^*)[\text{Card}(J^*)-1]}{2} \\ n^- = 0 \end{array}\right\} \Rightarrow \quad \tau = 1,$$

si les expressions sont toutes de signe différent, alors $\tau = —1$.

La déformation associée à ce coefficient de corrélation est la suivante :

$$\text{(D 19)} \qquad \left\{\begin{array}{c} \text{DEF}(d, d^*) = \tau, \\ (d = d^*) \Rightarrow \text{DEF}(d, d^*) = 1; \end{array}\right.$$

$$\text{(D 19')} \qquad \left\{\begin{array}{c} \text{DEF}(d, d^*) = 1-\tau, \\ (d = d^*) \Rightarrow \text{DEF}(d, d^*) = 0. \end{array}\right.$$

Dans le cas présent, les égalités des distances entre elles posent également un problème de notation des rangs. Si des couples $((j, k), (l, m))$ sont tels que $d(j, k) = d(l, m)$, on transforme la formulation précédente de la façon suivante :

Soient n le nombre de couples pour lesquels les distances sont égales et n^* le nombre de couples pour lesquels les distances d^* sont égales.

$$\text{(D 20)} \qquad \left\{\begin{array}{c} \text{DEF}(d, d^*) = 1 - \dfrac{(n^+ - n^-)}{\sqrt{\text{Card}(J^*)-n}.\sqrt{\text{Card}(J^*)-n^*}}, \\ (d = d^*) \Rightarrow \text{DEF}(d, d^*) = 0; \end{array}\right.$$

on peut utiliser une autre formulation; soit A le nombre de $[(j, k), (l, m)]$ pour lesquels on a

$$d(j, k) = d(l, m) \qquad \text{ou} \qquad d^*(j, k) = d^*(l, m);$$

on a alors la formulation suivante :

$$\text{(D 21)} \qquad \left\{\begin{array}{l} \text{DEF}(d, d^*) = 1 - \dfrac{(n^+ - n^-)}{A}, \\ A \leqq \dfrac{\text{Card}(J^*)[\text{Card}(J^*)-1]}{2}. \end{array}\right.$$

Si les distances d et d^* ont des valeurs distinctes, alors (D 21) = (D 19').

3.4.3. *Accords et désaccords entre préordonnances*

A des distances d et d^* sur J, on peut associer leurs préordonnances totales O et O^* définies ainsi :

$$O\,[(j,k),(l,m)] \overset{\text{déf}}{\Leftrightarrow} [d(j,k) \leqq d(l,m)];$$
$$O^*[(j,k),(l,m)] \overset{\text{déf}}{\Leftrightarrow} [d^*(j,k) \leqq d(l,m)].$$

On définit $\bar{O}$ par (non O), $\bar{O}^*$ par (non O^*);

$$\bar{O}[(j,k),(l,m)] \Leftrightarrow [d(j,k) > d(l,m)];$$
$$\bar{O}^*[(j,k),(l,m)] = [d^*(j,k) > d(l,m)].$$

L'idée de ce critère est de comparer les préordonnances totales associées à chacune des distances d et d^*. Deux possibilités ont été proposées, la première par M. Benzecri, la seconde par MM. Regnier et De La Wega. On pose les définitions suivantes :

$$G = \{((j,k),(l,m)) \,|\, (j,k),(l,m) \in J \times J \,|\, j > k;\ j \neq k \,|\, l > m;\ l \neq m;$$
$$\ldots\ O[(j,k),(l,m)]\}.$$

$$G^* = \left\{((j,k),(l,m)) \,|\, (j,k),(l,m) \in J \times J \left\{\begin{matrix} j > k;\ j \neq k \\ l > m;\ l \neq m \end{matrix}\right\} \right.$$
$$\left. \text{et}\ O^*[(j,k),(l,m)]\right\}.$$

On ne conserve du tableau des distances que le triangle inférieur (sans la diagonale).

• 1re *formulation :* comparer d et d^* équivaut à comparer G et G^*. On mesure l'accord entre G et G^* ainsi :

(D 22) $$\text{DEF}(d, d^*) = \text{Card}(G \cap G^*),$$
(BENZECRI),

$$G \cap G^* = \{((j,k),(l,m)) \,|\, O \wedge O^*\},$$
$$= \{[(j,k),(l,m)] \,|\, [d(j,k) \leqq d(l,m)]$$
$$\text{et}\ [d^*(j,k) \leqq d^*(l,m)]\};$$

• 2e *formulation :*

$$G \Delta G^* = \{((j,k),(l,m)) \,|\, (O \wedge \bar{O}^*) \vee (\bar{O} \wedge O^*)\},$$

= ensemble des paires de couples de points de J pour lesquels il y a désaccord exprimé pour la forme suivante,

$$G \Delta G^* = \{((j,k),(l,m)) \,|\, [d(j,k) \leqq d(l,m)]$$
$$\text{et}\ [d^*(j,k) > d^*(l,m)]\},$$

$$\cup \{((j, k), (l, m)) \mid [d(j, k) > d(l, m)] \text{ et } [d^*(j, k) \leqq d^*(l, m)]\};$$

On pose

(D 23) $$\text{DEF}(d, d^*) = \text{Card}(G \Delta G^*)$$
REGNIER-DE LA WEGA.

La seconde formulation est plus usitée; la déformation est d'autant plus petite que les désaccords sont moins nombreux; la déformation est nulle s'il n'y a pas de « désaccord ».

La formule (D 23) définit une distance sur $\mathcal{G} = \{ G \}$:

$$D(G, G^*) = \text{Card}(G \Delta G^*),$$
$$D(G, G^*) \geqq 0,$$
$$D(G, G^*) = 0 \Leftrightarrow \text{Card}(G \Delta G^*) = 0,$$
$$\Leftrightarrow G \Delta G^* = \varnothing,$$
$$\Leftrightarrow G = G^*,$$
$$D(G, G^*) = D(G^*, G).$$

Remarque : L'introduction du coefficient de corrélation des rangs de Kendall, dans le calcul des déformations, fait apparaître, d'une certaine façon, les préordonnances totales associées aux distances d et d^*.

D'après les notations précédentes, on a

n^+ = nombre de fois où $[r(j, k) - r(l, m)]$ est de même signe que $[r^*(j, k) - r^*(l, m)]$;

n = nombre de fois où les 2 expressions précédentes sont de signe contraire.

Or les rangs $\{ r(j, k) \}$ associés à un système de distances $\{ d(j, k) \}$ sont déterminés par les inégalités entre les distances.

Ainsi

$$(r(j, k) < r(l, m)) \Leftrightarrow (d(j, k) < d(l, m)).$$

Donc n^+ représente le nombre de fois où l'on a les accords sur les expressions suivantes :

(*a*) $[d(j, k) < d(l, m)]$ et $[d^*(j, k) < d^*(l, m)]$;

(*b*) $[d(j, k) > d(l, m)]$ et $[d^*(j, k) > d^*(l, m)]$.

On supprime les couples $[(j, k), (l, m)]$ pour lesquels on a au moins une égalité des distances pour (j, k) et (l, m) distincts.

Or l'expression (*a*) fait intervenir la préordonnance associée à la distance d et d^* de la façon suivante :

Par définition

$$[d(j, k) \leqq d(l, m)] = O[(j, k), (l, m)],$$

$$[d^*(j, k) \leqq d^*(l, m)] = O^*[(j, k), (l, m)]$$

DEF (d, d^*) (D 23) $= n^-$ (un des deux termes qui intervient dans l'élaboration du coefficient de corrélation des rangs de Kendall)

3.5. Comparaisons de type ordinal. Exemples :

Nous reprenons les 2 exemples cités précédemment (*cf.* § 3.3). Nous donnons, dans ce paragraphe, les résultats obtenus à partir du coefficient de corrélation des rangs de Spearman, de Kendall et la formule de comparaison des préordonnances (D 23).

Exemple 1 – Nombre de couples [(j,k), (l,m)] de J* x J* = 17 955

	UIM.	USM.	MOY.	MAX.	MOM.	VAR.	CEN.	D. MIN
Préordonnance	6179	3749	3950	3751	3781	3948	3945	0.
Rangs Kendall	0.31	0.71	0.62	0.71	0.71	0.62	0.62	1.
Rangs Spearman	0.34	0.83	0.74	0.83	0.83	0.74	0.74	1.

Exemple 2 – Questionnaire (0.1)

	UIM.	USM.	MOY.	MAX.	MOM.	VAR.	CEN.	D. MIN
Préordonnance	1839	2326	1771	7412	6766	2268	2895	0.
Rangs Kendall	0.83	0.76	0.84	0.16	0.20	0.77	0.69	1.
Rangs Spearman	0.93	0.90	0.94	0.25	0.31	0.91	0.81	1.

3.6. Conclusion

Les résultats obtenus sur les exemples pour les indices de déformation de type ordinal confirment, dans l'ensemble, ceux obtenus par les indices de déformation de type numérique. Il reste cependant certain que la multiplicité des indices de déformation a de quoi, encore une fois, décourager le taxinomiste le plus sérieux.

Il n'est naturellement pas question de privilégier un indice de déformation par rapport à un autre. Bien au contraire, on devra tenir compte dans l'étude des résultats des points de vue exprimés par chacun de ces indices, et ce d'autant plus que ces indices s'expriment souvent sous forme de moyenne par rapport à l'ensemble de la hiérarchie. Il faudrait pouvoir étudier sur quelles paires de points les déformations les plus importantes interviennent. Car, en effet, certaines stratégies d'agrégation conduisent à des déformations moyennes importantes et, cependant, elles représentent parfois mieux la structure des données (raccourcissements relatifs des petites distances et allongements relatifs des grandes distances).

La comparaison des hiérarchies indicées totales relatives à une même distance s'avère difficile. En l'absence de critères formels sûrs, le statisticien pourra se servir de ces indices pour se faire une idée de la déformation globale apportée par la stratégie d'agrégation. Pour s'assurer de la validité de résultats obtenus en classification, il convient de procéder d'une autre façon en examinant attentivement les classes constituées.

Puisqu'il paraît illusoire, dans l'état actuel, de vouloir chercher la meilleure méthode, cherchons plutôt dans quelle mesure les résultats obtenus en classification automatique se ressemblent. Pour cela, on cherchera à comparer les hiérarchies indicées entre elles, les indices entre eux, les partitions obtenues à certains niveaux de l'arbre, les hiérarchies supérieures (celles obtenues après une coupe dans la représentation arborescente).

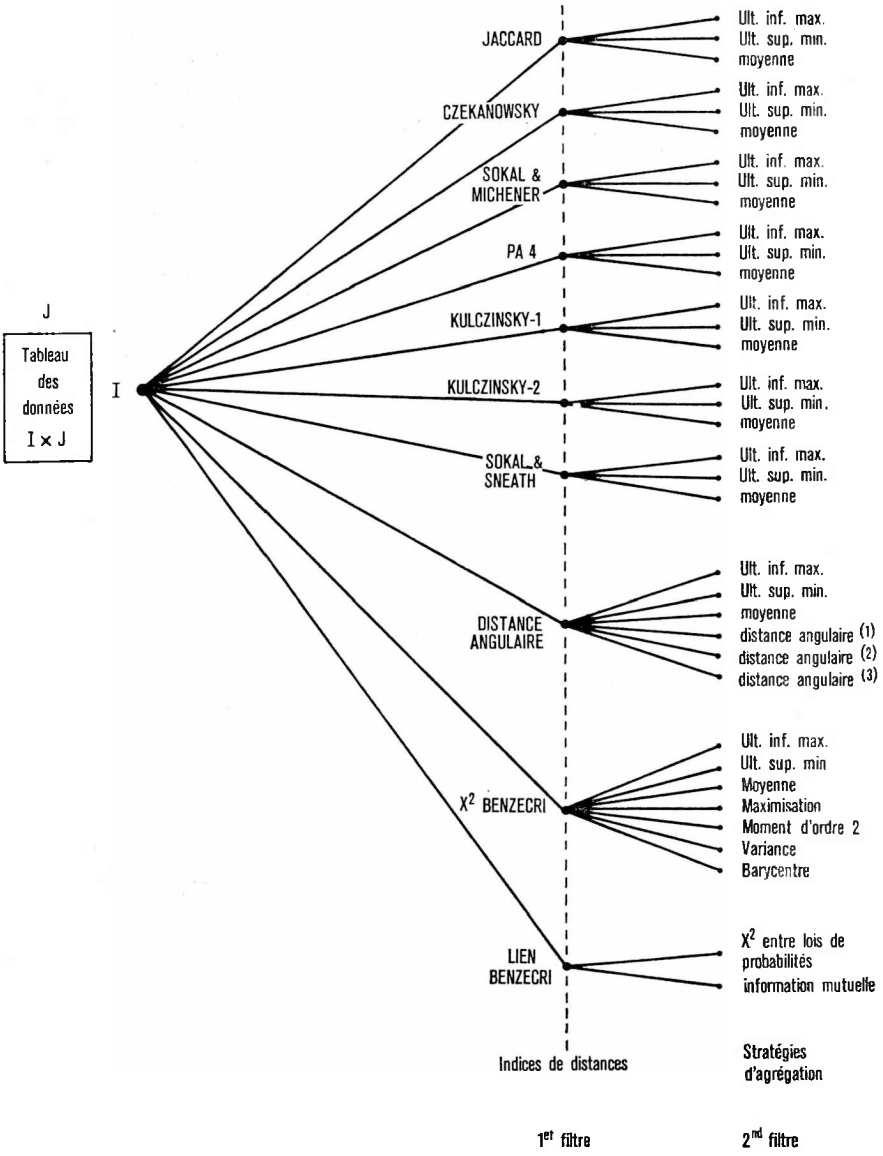
4. COMPARAISON SIMULTANÉE DES INDICES DE DISTANCES ET DES HIÉRARCHIES INDICÉES TOTALES ASSOCIÉES

Dans le cas de données exprimées sous forme d'un tableau de fréquences de mesures, ou de notes d'intensité, la comparaison des indices et des ultramétriques associées ne s'impose pas, dans la mesure où les formules de calcul des indices sont peu nombreuses (*cf. Consommation* nº 2-1974, « Sur les indices de distance en vue de la construction d'une classification hiérarchique »).

Dans le cas de tableaux de données exprimés sous la forme de tableaux logiques, les formules de calcul sont plus nombreuses. La comparaison peut alors avoir un sens. Le tableau suivant illustre cette possibilité.

Sur un exemple de questionnaire fictif (exemple 2), image grossière d'une partition en cinq classes, nous avons effectué les calculs suivants :

• Soit d_i un indice de distance et $\{ d^*_{i,k} \}$ les distances ultramétriques qui lui sont associées par différentes stratégies d'agrégation : on calcule, par les formules présentées au paragraphe 3, DEF $(d_i, d^*_{i,k})$.



Les deux tableaux suivants donnent la valeur des nombres DEF $(d_i\, d^*_{i,k})$ pour les indices de déformation fondée sur le calcul de corrélations et sur les désaccords entre préordonnances.

De ces différents calculs et formules on serait tenté de dégager un critère de choix d'un indice de distance d_i et d'une stratégie d'agrégation $d^*_{i,k}$, en proposant le couple $(d_i, d^*_{i,k})$ qui minimise DEF $(d_i, d^*_{i,k})$. Ces mêmes calculs nous montrent en même temps que ce choix est inefficace et qu'on risque, en se basant sur ces résultats, de se priver de méthodes de classi-

Tableau des corrélations entre indices et distances ultramétriques

	UIM.	USM.	MOY.	MAX.	MOM.	VAR.	CEN.
BENZECRI	0.98	0.97	0.98	0.86	0.56	0.93	0.88
JACCARD	0.98	0.99	0.98				
CZEKANOWSKY	0.99	0.99	0.98				
SOKAL, MICHENER	0.99	0.99	0.99				
PA4	0.98	0.99	0.98				
KULCZINSKY −1	0.98	0.98	0.97				
KULCZINSKY −2	0.96	0.96	0.95				
SOKAL, SNEATH	0.98	0.98	0.98				
DISTANCE ANGULAIRE	0.99	0.99	0.99				

Tableau des désaccords entre préordonnances

	UIM.	USM.	MOY.	MAX.	MOM.	VAR.	CEN.
BENZECRI	4 227	5 541	4 157	15 252	14 876	5 463	6 679
JACCARD	5 725	6 392	5 780				
CZEKANOWSKY	5 725	6 392	5 780				
SOKAL, MICHENER	2 818	2 648	2 652				
PA4	3 562	3 438	4 431				
KULCZINSKY −1	2 818	2 648	2 652				
KULCZINSKY −2	5 484	6 469	5 400				
SOKAL, SNEATH	5 725	6 392	5 780				
DISTANCE ANGULAIRE	4 911	5 371	4 905				

fication fort utiles. Les raisons qu'on peut invoquer pour éviter ce choix systématique sont les suivantes :

- Certaines ultramétriques peuvent être très déformées (pour un indice de déformation fixé) sans pour autant mal représenter les données. L'exemple nous est donné par la méthode dite de maximisation du moment centré d'ordre 2 d'une partition. L'indice de déformation est élevé et cependant la représentation arborescente est plus fidèle à l'image d'une partition en cinq classes que toutes les autres représentations (absence de niveaux intermédiaires, meilleure mise en évidence des classes). La déformation de la distance s'effectue dans le sens de la classification (*cf.* exemple 2).
- Deux ultramétriques peuvent être semblables (i. e. : avoir un indice de déformation égal) sans pour autant que la composition des classes soit semblable.
- Deux indices de distance peuvent être semblables et conduire à des ultramétriques différentes.
- En général, les calculs effectués ne donnent pas de différence significative pour pouvoir *a posteriori* décider de conserver tel indice de distance et telle stratégie d'agrégation.

Conclusion

Le choix formel semble illusoire. Dans la pratique, le statisticien effectue rarement toutes les comparaisons possibles d'algorithmes d'agrégation et d'indices de distances. Le paragraphe précédent nous a éclairés sur les difficultés que soulèvent de telles comparaisons. L'expérimentateur utilise quelques algorithmes en fonction de leurs propriétés propres et cherche plus à obtenir des partitions des variables, à reconnaître des classes, à ordonner ces classes, qu'à interpréter l'arbre dans sa totalité et sa complexité. L'étude de toutes les branches est souvent fastidieuse et de peu d'intérêt. C'est une des raisons pour laquelle l'utilisateur de telles techniques, plutôt que d'effectuer des comparaisons de hiérarchies comme celles définies au paragraphe précédent, cherche à construire des classes, des nomenclatures stables et facilement caractérisables, et à comparer ces classes ou ces partitions obtenues par quelques techniques d'agrégation différentes. C'est ce que nous tentons d'expliquer dans le prochain paragraphe.

5. COMPARAISON DES PARTITIONS EXTRAITES DES CLASSIFICATIONS ASCENDANTES HIÉRARCHIQUES

A partir d'un arbre représentatif d'une hiérarchie de parties, on peut construire une partition de l'ensemble des variables hiérarchisées en « sciant » l'arbre à une hauteur fixée. Comme les indices des hiérarchies ont des significations différentes selon les algorithmes d'agrégation, le choix de la hauteur de la « coupe » est fait de telle façon que le nombre

d'éléments qui constituent la partition soit fixé (soit Card (P) ce nombre pour la partition P).

Soient P, Q deux partitions de J et Card (P) et Card (Q) le nombre de classes de chacune des partitions.

On construit le tableau de correspondances suivant :

$$\{K(p, q),\ p = 1, \text{Card}(P),\ q = 1, \text{Card}(Q)\}.$$

$K(p, q)$ = nombre d'éléments de J,
communs aux deux classes p de P et q de Q ;

$$K(p, q) = \text{Card}\{j \in J ; f_P(j) = p \text{ et } f_Q(j) = q\},$$

où f_p et f_q sont les fonctions des partitions P et Q, qui, à un élément j de J, associent la classe à laquelle il appartient dans P (respectivement dans Q); si les partitions P et Q étaient indépendantes, la corrélation entre f_p et f_q serait nulle, la correspondance serait l'expression d'une loi produit.

$$K.K(p, q) = K(p).K(q), \qquad \forall (p, q) \in P \times Q,$$

avec

$$K(p) = \sum_{q \in Q} K(p, q), \qquad K(q) = \sum_{p \in P} K(p, q),$$

$$K = \sum_{p \in P,\, q \in Q} K(p, q).$$

On calcule la quantité suivante :

$$x^2(P, Q) = K.\left[\left[\sum_{(p, q) \in P \times Q}\left[\frac{K^2(p, q)}{K(p).K(q)}\right]\right] - 1\right],$$

x^2 (P, Q) est alors comparé à un χ^2 à [card (P)—1].[Card (Q)—1] degrés de liberté.

Le test s'effectue de la façon suivante :

L'hypothèse H_0 à tester est : « P et Q sont deux partitions indépendantes »

si	$x^2(P, Q) > \chi^2_{(\alpha)}$	$[\text{Card}(P)-1].[\text{Card}(Q)-1]$
	$\Rightarrow$ H_0 est rejetée $\overset{\text{déf}}{\Leftrightarrow}$ P et Q ne sont pas étrangères;	
si	$x^2(P, Q) < \chi^2_{(\alpha)}$	$[\text{Card}(P)-1].[\text{Card}(Q)-1]$,
	$\Rightarrow$ H_0 est acceptée $\overset{\text{déf}}{\Leftrightarrow}$ P et Q sont indépendantes.	

Nous avons appliqué ce test à la comparaison des partitions déduites d'une représentation arborescente sur les deux exemples cités au paragraphe 3.2. Ces partitions comportent le même nombre de classes (5). Dans chaque cas, l'hypothèse nulle a été rejetée (les partitions ne sont pas étrangères deux à deux).

6. COMPARAISON DES HIÉRARCHIES INDICÉES TOTALES OU TRONQUÉES

Le but de telles comparaisons est d'essayer de dégager des ressemblances entre hiérarchies indicées ou partitions produites à partir de critères d'agrégation différents. L'esprit n'est plus à la recherche de la meilleure stratégie d'agrégation, mais de pouvoir, à partir de techniques différentes, apprécier la ressemblance entre représentations arborescentes. Pour ce faire, il existe trois voies possibles dont deux sont déduites des paragraphes précédents.

6.1. Comparaison de type numérique

Pour apprécier la proximité entre représentations arborescentes, il suffit de se reporter aux formules de calcul d'un indice de déformation présentées au paragraphe 3.

Supposons avoir utilisé N critères d'agrégation.

soit $H_j(J)$ et $H_k(J)$ deux hiérarchies de parties indicées totales sur le même ensemble J

et $d(H_j, H_k)$ la « distance » entre les hiérarchies H_j et H_k.

Pour apprécier les distances relatives entre hiérarchies indicées, il suffit de dresser le tableau :

$$d(H_j, H_k) = \mathrm{DEF}(d_j^*, d_k^*) \qquad \text{pour} \quad j > k \quad \text{et} \quad j = 2, N.$$

6.2. Comparaison de type ordinal

On effectue la même démarche que précédemment en utilisant les indices de déformation cités au paragraphe 3.4.

On dresse alors le tableau :

$$d(H_j, H_k) = \mathrm{DEF}(d_j^*, d_k^*), \qquad \text{pour} \quad j > k \quad \text{et} \quad j = 2, N$$

6.3. Comparaison sur la structure hiérarchique des classes

6.3.1. *Une distance entre hiérarchies de parties ; $d^0(A, B)$:*

Soit A et B deux hiérarchies de parties, totales, indicées, sur le même ensemble J.

On pose

$$d^0(A, B) \overset{\text{déf}}{=} \sum_{i=\mathrm{Card}\,J+1}^{2.\mathrm{Card}\,J-1} \mathrm{Card}\left\{N_i(A) \Delta N_i(B)\right\};$$

(Formule 1),

où

$N_i(A)$ [respectivement $N_i(B)$], est le i-ième nœud créé de la hiérarchie. $A(J)$ [respectivement $B(J)$],

Axiome (81) :

$$A(J) = B(J) \Rightarrow N_i(A) \Delta N_i(B) = \varnothing,$$
$$\forall i \in [\text{Card}\, J + 1, 2.\text{Card}\,(J) - 1],$$
$$\Rightarrow d^0(A, B) = O;$$

$$d^0(A, B) = 0 \Rightarrow N_i(A) \Delta N_i(B) = \varnothing,$$
$$\forall i \in [\text{Card}\, J + 1, 2.\text{Card}\,(J) - 1],$$
$$\Rightarrow A = B;$$

Axiome (82) :

$$d^0(A, B) = d^0(B, A);$$

Axiome (83) :

$$d^0(A, B) \leqq d^0(A, C) + d^0(C, B);$$

propriété de la différence symétrique.

$$\text{Card}\{N_i(A) \Delta N_i(B)\} \leqq \text{Card}\{N_i(A) \Delta N_i(C)\} + \text{Card}\{N_i(C) \Delta N_i(B)\},$$
$$\Rightarrow d^0(A, B) \leqq d^0(A, C) + d^0(C, B); \qquad \forall A, B, C.$$

d^0 définit une distance sur l'ensemble des hiérarchies construites sur J.

6.3.2. *Propriétés de la distance d^0*

Cette distance ne permet pas de comparer la distance initiale d à la distance ultramétrique d^* associée à une hiérarchie mais de chercher parmi toutes les hiérarchies construites celles qui sont constituées de la même façon (mêmes composants, même niveau hiérarchique). Ainsi deux hiérarchies composées des mêmes nœuds à des niveaux distincts, $(f(N_i(A)) \neq f(N_i(B))$, sont semblables.

On élimine ainsi l'aspect numérique ou ordinal pour ne conserver que la structuration des données en partitions successives. Ce qui semble plus utile au statisticien qui souhaite plus comparer des classes ou des partitions entre elles, avec des algorithmes différents, que d'avoir des calculs savants sur les ultramétriques, mais qui le renseignent peu sur les classes.

6.3.3. *La distance d^{00}*

On pose

$$d^{00}[A(J), B(J)] = \sum_{i=m_1}^{i=m_2} \underbrace{(i - \text{Card}\, J)}_{r(i)} \text{Card}\{N_i(A) \Delta N_i(B)\}$$

(Formule 2)

avec $m_1 = \text{Card}(J)+1,$

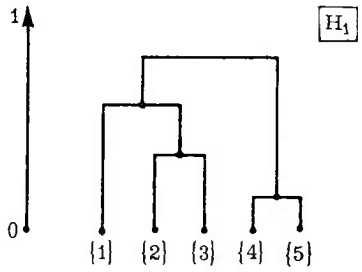
$m_2 = 2.\text{Card}(J)-1;$

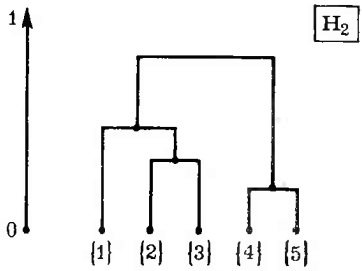
où $r\,(\text{Card}\,J+1) = 1$ (1er nœud créé),

$r\,(2.\text{Card}\,J-1) = \text{Card}\,J-1$ (dernier nœud créé).

Ainsi, deux hiérarchies composées des mêmes éléments seront d'autant plus proches que les inversions dans les constructions se seront produites au début des agrégations. Cette distance donne donc plus de poids au fait que les changements de classes se sont produits dans le bas de la hiérarchie.

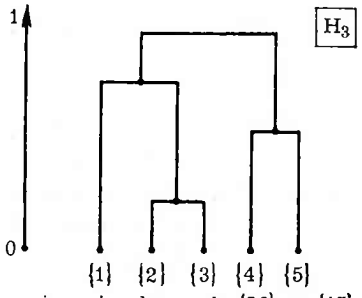
6.3.4. *Quelques exemples de calculs de d°.*





même composant à des niveaux hiérarchiques différents.

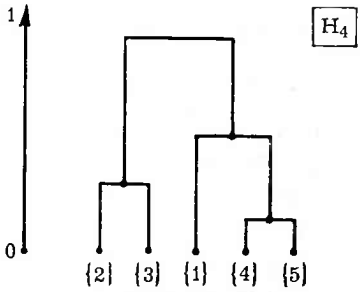
d° (H1, H2) = 0



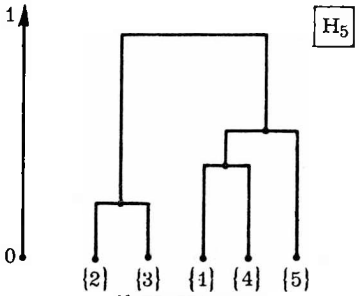
inversion de nœuds {2,3} et {4,5}

d° (H1, H3) = 4

Pas de modification fondamentale de la structure.

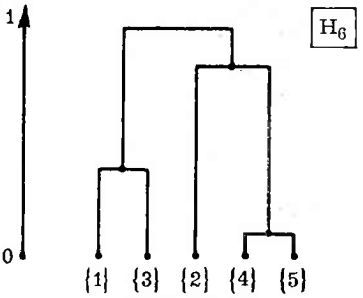


d° (H4, H1) = 4



d° (H5, H1) = 12

bouleversement de la structure initiale H1.



d° (H6, H1) = 6

6.4. Comparaisons, sur un exemple, des hiérarchies de parties, indicées (totales ou tronquées)

L'exemple retenu est celui constitué par la représentation euclidienne d'un nuage de vingt points dans un plan (*cf.* § 3.3). Nous avons retenu cinq mesures de ressemblances entre hiérarchies indicées totales :

— la distance d^0;
— le coefficient de corrélation;
— le coefficient de corrélation des rangs de Spearman;
— le coefficient de corrélation des rangs de Kendall;
— le nombre des désaccords entre préordonnances.

6.4.1. *La distance d^0*

Tableau des Card $\{ N_i (H_j) \Delta N_i (H_k) \}$ $i \in I$, $j \in K$, $k \in K$ et $j > k$

i \ k,j	2.1	3.1	3.2	4.1	4.2	4.3	5.1	5.2	5.3	5.4	6.1	6.2	6.3	6.4	6.5	7.1	7.2	7.3	7.4	7.5	7.6
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	5	5	0	5	0	0	5	0	0	0	5	0	0	0	0	5	0	0	0	0	0
23	6	6	0	6	0	0	6	0	0	0	6	0	0	0	0	6	0	0	0	0	0
24	7	7	0	7	0	0	7	0	0	0	7	0	0	0	0	7	0	0	0	0	0
25	5	5	0	5	0	0	4	5	5	5	5	0	0	0	5	5	0	0	0	5	0
26	3	3	0	3	0	0	8	5	5	5	3	0	0	0	5	3	0	0	0	5	0
27	1	1	0	1	0	0	5	6	6	6	1	0	0	0	6	1	0	0	0	6	0
28	5	6	5	6	5	0	6	5	6	6	6	5	0	0	6	6	5	0	0	6	0
29	0	5	5	5	5	0	0	0	5	5	5	5	0	0	5	9	9	8	8	9	0
30	8	8	0	8	0	0	6	10	10	10	8	0	0	0	10	0	8	8	8	6	8
31	0	0	0	0	0	0	10	10	10	10	0	0	0	0	10	0	0	0	0	10	8
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	7	0	7	7	0	7	7	0	7	0	12	9	12	9	9	12	9	12	9	9	0
34	0	0	0	12	12	12	12	12	12	0	12	12	12	0	0	9	9	9	7	7	7
35	13	8	7	1	12	9	1	12	9	0	8	7	0	9	9	13	0	7	12	12	7
36	10	10	0	10	0	0	10	0	0	0	10	0	0	0	0	10	0	0	0	0	0
37	12	12	0	12	0	0	12	0	0	0	12	0	0	0	0	12	0	0	0	0	0
38	5	10	15	5	0	15	5	0	15	0	10	15	0	15	15	10	15	0	15	15	0
$d^0(H_j H_k)$	87	86	39	93	34	43	99	65	100	47	108	53	24	33	80	108	55	44	59	80	30

I = Ensemble des nœuds des hiérarchies numérotés de Card $(J) + 1$ à $2.\text{Card}(J) - 1$;
K = Ensemble des stratégies d'agrégation;
k = 1 ultramétrique inférieure maximale;
k = 2 ultramétrique supérieure minimale;
k = 3 agrégation par la moyenne;
k = 4 agrégation par la maximisation du moment centré d'ordre 2 d'une partition;
k = 5 agrégation par le moment d'ordre 2 d'une classe;
k = 6 agrégation par la variance d'une classe;
k = 7 agrégation par le centre de gravité des classes.

Tableau des distances d° (Hj, Hk), j et k ∈ K

	UIM 1	USM 2	MOY 3	MAX 4	MOM 5	VAR 6	CEN 7
UIM 1	0	87.	86.	93	99	108	108
USM 2		0.	39	34	65	53	54
MOY 3			0.	43	100.	24	44
MAX 4				0.	47	33	59
MOM 5					0.	80	80
VAR 6						0.	30.
CEN 7							0.

6.4.2. *L'indice* (D 13), *le coefficient de corrélation*

Tableau des indices DEF (d^*_j, d^*_k)

	UIM 1	USM 2	MOY 3	MAX 4	MOM 5	VAR 6	CEN 7
UIM i	1.	0.65	0.62	0.52	0.48	0.44	0.52
USM 2		1.	0.76	0.98	0.96	0.66	0.72
MOY 3			1.	0.68	0.62	0.96	0.98
MAX 4				1.	0.99	0.61	0.66
MOM 5					1.	0.55	0.60
VAR 6						1.	0.99
CEN 7							1.

d^*_j ultramétrique associée à la hiérarchie H_j;
d^*_k ultramétrique associée à la hiérarchie H_k.

6.4.3. *Le coefficient de corrélation des rangs de Spearman*

Tableau des indices DEF (d^*_j, d^*_k)

	UIM 1	USM 2	MOY 3	MAX 4	MOM 5	VAR 6	CEN 7
UIM 1	1.	0.46	0.33	0.46	0.46	0.33	0.33
USM 2		1.	0.73	0.99	0.99	0.73	0.73
MOY 3			1.	0.73	0.73	0.99	0.99
MAX 4				1.	0.99	0.73	0.73
MOM 5					1.	0.73	0.73
VAR 6						1.	0.99
CEN 7							1.

6.4.4. *Le coefficient de corrélation des rangs de Kendall*

Tableau des indices DEF (d^*_j, d^*_k)

	UIM 1	USM 2	MOY 3	MAX 4	MOM 5	VAR 6	CEN 7
UIM 1	1.	0.41	0.26	0.41	0.40	0.26	0.26
USM 2		1.	0.61	1.	0.99	0.61	0.61
MOY 3			1.	0.61	0.61	1.	0.99
MAX 4				1.	0.99	0.61	0.61
MOM 5					1.	0.60	0.60
VAR 6						1.	1.
CEN 7							1.

6.4.5. *Les désaccords entre préordonnances*

Tableau des indices DEF (d^*_j, d^*_k)

	UIM 1	USM 2	MOY 3	MAX 4	MOM 5	VAR 6	CEN 7
UIM 1	0	10 199	12 526	10 191	10 294	12 526	12 537
USM 2		0	7 001	56	145	6 953	6 966
MOY 3			0	6 945	7 050	48	77
MAX 4				0	105	6 993	7 006
MOM 5					0	7 098	7 111
VAR 6						0	29
CEN 7							0

Ces différents calculs sont, à notre avis, nécessaires au statisticien. La lecture ou l'interprétation d'une classification hiérarchique est difficile. Le traitement « des données » par divers algorithmes d'agrégation ayant même base de départ et la comparaison des hiérarchies obtenues après traitement nous paraissent utiles, si on se souvient qu'il ne faut pas, dans une classification hiérarchique, attribuer une importance trop grande au fait qu'un objet appartienne à une classe ou à une autre, qui lui est proche dans l'arbre. L'interprétation d'une arborescence est globale et se fait en « descendant » dans l'arbre, en vue d'apprécier les oppositions réalisées dans chaque partition (de la plus grossière à la plus fine).

7. CONCLUSION

A partir du moment où on a admis que les méthodes de classification ascendante hiérarchique ne peuvent fournir un optimum absolu, et qu'une méthode n'exprime qu'un point de vue déformé de la réalité, l'art et la sagesse du statisticien seront de confronter, synthétiser plusieurs de ces réalités, en enrichissant ces confrontations de différents calculs de déformation, imparfaits certes, mais qui, en tout état de cause, permettent quand même d'approfondir la connaissance des données. Ainsi, les méthodes de classification complètent, par le point de vue qu'elles adoptent, les méthodes de statistique descriptive plus classiques.

CONSOMMATION

XXIIe ANNÉE, No 1, JANVIER-MARS 1975

RÉSUMÉS - ABSTRACTS

des articles contenus dans ce Numéro

STRUCTURE ET INÉGALITÉ DES PATRIMOINES, par D. STRAUSS-KAHN. *Consommation*, 1-1975, janvier-mars 1975, pages 5 à 31.

Les choix patrimoniaux des ménages sont mal connus. Dans cet article, on a essayé de lier la détention de tel ou tel actif à certaines caractéristiques des ménages comme l'âge, les ressources, l'instruction, etc. Il apparaît alors que l'âge opère une distinction entre les actifs détenus en fonction du service rendu (actifs de jouissance ou actifs financiers), alors que les ressources sont à l'origine d'un clivage qui dépend de la classe de risque de l'actif considéré.

L'inégalité que présente la distribution des patrimoines et dont on sait qu'elle est beaucoup plus marquée que celle des revenus, peut elle-même être analysée comme une inégalité en fonction de l'âge et une inégalité en fonction des ressources. Quant aux actifs patrimoniaux pris un à un, leur détention est parfois concentrée entre très peu de mains (c'est le cas des valeurs de portefeuille) ou au contraire largement diffusée (comme pour les comptes de chèques, par exemple) et, ici encore, on peut décomposer l'inégalité en une inégalité selon l'âge et une inégalité suivant le revenu.

STRUCTURE AND INEQUALITIES OF WEALTH, by D. STRAUSS-KAHN. *Consommation*, 1-1975, January-March 1975, pages 5 to 31.

Little is known about families' choices about their property. In the present paper, the author has attempted to find out whether the possession of different types of assets is related to certain characteristics such as age, level of income, level of education. Age seems to have an effect on the property of assets according to the service expected of them : assets like houses of personnal use to their owners or motorcars vs financial assets. The level of incomes makes a difference in the ownership of assets that depends on the risk incured.

The inequality of the distribution of fortunes that is much more important than the inequality of incomes distribution, may be analysed as inequality according to age and inequality according to incomes. If one considers each type of assets separately, the property of some assets is very concentrated in very few families (shares for instance) or on the contrary, widely spread (such as cheque accounts). There again, the inequality may be analysed as inequality according to age and inequality according to incomes.

L'APPRÉCIATION MONÉTAIRE D'UN SURPLUS DANS LA CONSOMMATION ALIMENTAIRE DE DIFFÉRENTES CATÉGORIES SOCIALES, par P. NAVEAU et P. PETIT. *Consommation*, 1-1975, janvier-mars 1975, pages 33 à 54.

L'estimation monétaire d'un surplus dans la consommation alimentaire de diverses catégories socio-professionnelles

THE MONETARY VALUATION OF A SURPLUS IN THE FOOD CONSUMPTION OF DIFFERENT SOCIAL GROUPS, by P. NAVEAU and P. PETIT. *Consommation*, 1-1975, January-March 1975, pages 33 to 54.

The monetary value of a surplus in the food consumption of different social groups has been determined by using

a été effectuée en utilisant le critère économique de minimisation d'une dépense à prix donnés avec comme contrainte, successivement, deux normes arbitraires de besoins en nutriments tirées des consommations observées. Trois résultats principaux se dégagent de ce travail :

- une caractérisation des productivités nutritionnelles des biens;
- la reconnaissance d'une liaison entre les caractéristiques nutritionnelles d'un bien et la variance de son prix ;
- une évaluation de l'ordre de grandeur par catégorie sociale de ce que l'on a appelé le surplus.

the economic criterium of reducing the consumption cost to a minimum with regard to a technical constraint about an arbitrary standard of nutritional needs, estimated from the observed levels of consumption. Three principal results have been obtained in the study:

- the appraisal of the nutritional productivity of food consumption;
- the main outlines of the relation between the nutritional characteristics of food goods and their price's deviation;
- a valuation of the surplus for different social groups.

QUELQUES CRITÈRES DE COMPARAISON DES HIÉRARCHIES INDICÉES PRODUITES EN CLASSIFICATION AUTOMATIQUE, par M. JAMBU. *Consommation*, 1-1975, janvier-mars 1975, pages 55 à 84.

Dans cet article, l'auteur a proposé de nombreux critères de qualité des représentations hiérarchiques de données. Des comparaisons sont proposées sur des exemples concrets. L'auteur aboutit à la conclusion que les comparaisons formelles de critères d'agrégation s'avèrent assez inefficaces, et que la solution au problème du choix d'une méthode repose essentiellement, faute de mieux, sur le choix axiomatique d'un indice de distance qui conditionne toute représentation géométrique ou hiérarchique des données.

SOME CRITERIA OF COMPARISON OF INDEXED HIERARCHIES PRODUCED IN AUTOMATIC CLASSIFICATION, by M. JAMBU. *Consommation*, 1-1975, January-March 1975, pages 55 to 84.

The paper presents numerous criteria of the quality of hierarchical representation of data. Examples are given. The author finds that formal comparisons of criteria of aggregation are not very efficient, and that the answer to the question of the choice of a method lies mostly, through lack of anything better, on the axiomatic choice of an index of distances which conditions all geometric representation or hierarchy of data.

Le directeur de la publication P. BORDAS
Dépôt légal/ED. 1er trimestre 1975. N° 029. N° de commission paritaire 29837.
Imprimé en France. — 5/1975. IMP. GAUTHIER-VILLARS, MONTREUIL. N° 2092.

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

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^{e} 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 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.

N° 4. — Aspects géographiques du système des soins médicaux. Analyse des données départementales. — Vieillesse et classe sociale. L'exemple des paysans bénéficiaires de l'I.V.D. et celui des petits commerçants. — Sur les critères d'agrégation utilisés en classification automatique.

SOMMAIRE DES PROCHAINS NUMÉROS

Vers une évaluation de la consommation réelle des ménages. La justice distributive de l'école. L'orientation du dépouillement de certaines enquêtes par l'analyse des correspondances multiples.

sommaire

CENTRE DE RECHERCHES ET DE DOCUMENTATION SUR LA CONSOMMATION

45, boulevard de la Gare, PARIS-13e

Tél. 707-97-59

1975 n° 1

Janvier-Mars