REVUE DE STATISTIQUE APPLIQUÉE

J.-J. DENIMAL

Analyse des interactions entre k partitions prises 2 à 2. Théorie et application en biologie

Revue de statistique appliquée, tome 42, n° 1 (1994), p. 19-40

<http://www.numdam.org/item?id=RSA_1994__42_1_19_0>

Rev. Statistique Appliquée, 1994, **XLII** (1), 19-40

ANALYSE DES INTERACTIONS ENTRE k PARTITIONS PRISES 2 À 2 THÉORIE ET APPLICATION EN BIOLOGIE.

J.-J. Denimal

Laboratoire de Statistique et Probabilités
Université des Sciences et Technologies de Lille
59655 – Villeneuve d'Ascq Cedex, France

RÉSUMÉ

A partir d'un tableau de contingence k_{IJ}, l'ensemble I étant muni de n partitions $\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_n$, affectées des poides $a_1, a_2, ..., a_n$, on se propose, dans le cadre de l'analyse des correspondances, de définir une méthode permettant l'analyse des «interactions» entre ces différentes partitions prises 2 à 2. Cette méthode dessinera d'autre part, un cadre théorique où se retrouveront comme cas particuliers : les analyses factorielles intra-classe, inter-classe, ainsi que l'analyse des correspondances multiples. De nouveaux critères d'agrégation seront également établis, permettant d'édifier sur I une classification ascendante hiérarchique sous contraintes. Enfin, une application à un jeu de données viendra illustrer les différentes méthodes proposées.

AMS Classification : *– 62XX07 –*

Mots-clés : *Interactions multiples, analyses des correspondances, analyse factorielle intra-classe, inter-classe, interne, classifications sous contrainte*

SUMMARY

This paper introduces a new notion of «interactions» between n partitions $\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_n$ (with weights $a_1, a_2, ..., a_n$) on a set I, with respect to a contingency table k_{IJ}, where J is a set of categorical variables. We propose a method, based on correspondence analysis, to study these interactions. This method covers a large area, including factor analysis of structured tables (between classes analysis and within classes analysis) and multiple correspondance analysis. Some new hierarchical clustering algorithms (under constraints) are also given. Finally, this new methods are applied to a set of biological data.

AMS Classification : *– 62XX07 –*

Key-words : *Interactions, correspondence analysis factor analysis of structured tables, classifications under constraints.*

1. Introduction

L'objet de la méthode développée dans cet article, est donc de définir, puis d'analyser les «interactions» entre k partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k$ définies sur un nuage $\mathcal{N}(I)$, construit à partir de k_{IJ} (ce dernier étant un tableau de contingence, ou tout tableau pouvant être soumis à l'analyse des correspondances).

A partir d'une formule générale de décomposition du produit scalaire, vérifiée dans tout espace euclidien, et appliquée ici dans $\mathbf{R}^{\text{card } I}$ (muni de la métrique classique du CHI2), une quantité notée $\mathcal{A}(\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k)$ sera définie et mesurera «l'interaction» entre les k partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k$, (ces dernières étant supposées munies de poids $a_1, a_2, \ldots, a_k$). L'analyse de ces interactions sera réalisée, en soumettant à l'analyse des correspondances, un tableau kr_{IJ} construit à partir du tableau initial k_{IJ} et d'inertie $\mathcal{A}(\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k)$. L'interprétation des représentations obtenues sera précisée, en particulier, par l'obtention de formules de transition.

Cette analyse, appelée dans la suite analyse factorielle des interactions entre $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k$, généralise dans le cas de 2 partitions $\mathcal{P}_1$ et $\mathcal{P}_2$ munis de poids égaux à 1, d'une part l'analyse factorielle intra-classe d'Escofier que l'on obtient lorsque l'une des partitions $\mathcal{P}_1$ et $\mathcal{P}_2$ est celle réduite à une classe, et d'autre part, l'analyse factorielle inter-classe que l'on retrouve lorsque l'une des partitions $\mathcal{P}_1$ et $\mathcal{P}_2$ est celle composée des singletons. Un cinquième paragraphe sera consacré aux applications à la classification automatique. On définira, en particulier, un nouveau critère d'agrégation. Ce dernier se présentera sous la forme d'un critère de Ward modifié, vérifiera également l'axiome de la médiane, et permettra d'édifier sur l'ensemble I, à partir du tableau k_{IJ}, une classification ascendante hiérarchique, qui, à chaque niveau, construira une partition où l'influence de n partitions (c'est-à-dire de variables qualitatives) définies sur I et données a priori sera minimisée.

Enfin, un sixième paragraphe sera consacré aux applications des méthodes proposées, dans le cadre d'une étude, menée conjointement avec l'Institut Pasteur de Lille et le C.R.E.S.G.E.* et dont l'objet était l'estimation de la consommation d'alcool à partir de paramètres biologiques.

2. Les données et les hypothèses

2.1 Présentation des données

La méthode présentée dans cet article est développée à partir d'un tableau k_{IJ}, appelé dans la suite le tableau de contingence initial, k_I et k_J désignant les marges de ce tableau. Celui-ci placé dans le cadre de l'analyse des correspondances, donne classiquement naissance aux deux nuages :

$$\mathcal{N}(I) = \{(f_J^i, f_i)/i \in I\} \text{ et } \mathcal{N}(J) = \{(f_I^j, f_j)/j \in J\}.$$

* Centre de Recherches Economiques, Sociologiques et de Gestion 1, Rue Norbert Segard – B.P. 109 – 59016 Lille Cedex

Chaque élément i de I (resp. j de J) est donc représenté par son profil f_J^i (resp. f_I^j) et admet le poids $f_i = \frac{t(i)}{k}$ (resp. $f_j = \frac{k(j)}{k}$); les distances utilisées étant celles de CHI2.

On suppose, de plus, que l'ensemble I est muni de k partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k$ munies de poids $a_1, a_2, \ldots, a_k$ supposés positifs et de somme notée r.

Enfin, pour la suite de l'article, indiquons que tout produit scalaire entre éléments de $\mathbf{R}^{\text{card}I}$ (resp. $\mathbf{R}^{\text{card}J}$) sera, sauf mention express du contraire, celui définissant la métrique du CHI2, c'est-à-dire celui défini par la métrique diagonale dont les éléments diagonaux valent $(1/f_i)_{1\leqslant i\leqslant \text{card } I}$ (resp. $(1/f_j)_{1\leqslant j\leqslant \text{card } J}$). Cette remarque s'applique également aux différents projecteurs qui seront employés.

2.2 Les sous-espaces $F_1, F_2, F_3, \ldots, F_k$ de $\mathbf{R}^{\text{card}I}$, associés aux k partitions $\mathcal{P}_1, \mathcal{P}_2 \ldots, \mathcal{P}_k$.

2.2.1. Définition $\forall \ell \in [1,k]$, à chaque partition $\mathcal{P}_\ell$ de I, est associé le sous-espace F_ℓ de $\mathbf{R}^{\text{card } I}$ engendré par le système $\{e_p / p \in \mathcal{P}_\ell\}$ défini comme suit :

$$\begin{cases} \forall i \in p & e_p(i) = \dfrac{f_i}{\sqrt{f_p}} \\ \forall i \notin p & e_p(i) = 0 \end{cases} \quad (\text{avec } f_p = \sum_{r\in p} f_r).$$

2.2.2. Propriétés. La métrique étant celle du CHI2, on vérifie facilement que :

a) Les vecteurs $\{e_p / p \in \mathcal{P}_\ell\}$ sont de norme 1 et orthogonaux 2 à 2.

b) f_I^j représentant le profil de l'élément j de J, la projection orthogonale de f_I^j sur F_ℓ, notée $P_{F_\ell}(f_I^j)$, s'écrit :

$$P_{F_\ell}(f_I^j) = \left(\frac{k(p^i, j)k(i)}{k(p^i)k(j)} \right)_{i\in I}$$

p^i étant la classe de la partition $\mathcal{P}_\ell$ contenant i, et $k(p^i, j) = \sum_{r\in p_i} k(r, j)$ (et de même, $k(p^i) = \sum_{r\in p^i} k(r)$).

c) $f_I = (f_i)_{i\in I}$ appartient à chacun des sous-espaces F_ℓ lorsque ℓ varie dans $[1,k]$. En effet, $\forall \ell \in [1,k]$, $f_I = \sum_{p_\ell \in \mathcal{P}_\ell} \sqrt{f}_{p_\ell} e_{p_\ell}$.

3. Les données et les hypothèses.

3.1. Lemme de décomposition

Soit k sous espaces $F_1, F_2, \ldots, F_k$ d'un espace euclidien E quelconque, affectés respectivement de k réels $a_1, a_2, \ldots, a_k$ de somme r supposée non nulle.

$P_{F_i}(1 \leqslant i \leqslant k)$ désignant la projection orthogonale sur F_i, on obtient alors la décomposition suivante du produit scalaire $< x, y >$: $\forall x \in E,\ \forall y \in E$,

$$< x, y >=< A(x), A(y) > + \frac{2}{r^2} \sum_{i=1}^{k} \sum_{j=1}^{k} a_i a_j < P_{F_i}(x) - P_{F_j}(x), P_{F_i}(y) - P_{F_j}(y) >$$

où A est l'opérateur : $A = \frac{2}{r} \sum_{i=1}^{k} a_i P_{F_i} - I$ (I étant l'application identité de E).

Démonstration : (voir annexe).

3.2 Formule de décomposition de l'inertie de k_{IJ}.

3.2.1 Propriété : L'inertie In du tableau initial k_{IJ} se décompose comme suit :

$$In = \sum_{i \in I} f_i \cdot \| f_J^i - f_J \|^2 = \sum_{i \in I} f_i \| \frac{2}{r} \sum_{\ell=1}^{k} a_\ell \cdot f_J^{p_\ell^i} - f_J^i - f_J \|^2$$
$$+ \frac{2}{r^2} \sum_{\ell=1}^{k} \sum_{\ell'=1}^{k} a_\ell a_{\ell'} (\sum_{i \in I} f_i \| f_J^{p_\ell^i} - f_J^{p_{\ell'}^i} \|^2)$$

($f_i, f_J^i, f_J^{p_\ell^i}, f_J^{p_{\ell'}^i}$ représentant respectivement le poids de i et les profils de i et des centres de gravité des classes p_ℓ^i et $p_{\ell'}^i$ des partitions $\mathcal{P}_\ell$ et $\mathcal{P}_{\ell'}$ (contenant l'individu $i \in I$), calculés à partir du tableau initial k_{IJ}).

Démonstration : (voir annexe).

3.2.2 Définition. A partir de la décomposition de l'inertie totale du tableau k_{IJ} (présentée ci-dessus), en 2 termes dont l'un mesure un «écart moyen» entre les partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k$ prises 2 à 2, il est naturel de prendre l'autre terme, noté $\mathcal{A}(\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k)$ comme mesure de l'interaction multiple entre les k partitions considérées; ces dernières étant munies des poids positifs $a_1, a_2, \ldots, a_k$ de somme égale à r. Autrement dit, nous poserons :

$$\mathcal{A}(\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k) = \sum_{i \in I} f_i \| \frac{2}{r} \sum_{\ell=1}^{k} a_\ell \cdot f_J^{p_\ell^i} - f_J^i - f_J \|^2$$

3.3 Propriétés de l'interaction

3.3.1. Propriété : L'interaction multiple $\mathcal{A}(\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k)$ entre les k partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k$ définies sur I, s'écrit encore sous la forme :

$$\mathcal{A}(\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k) = \sum_{j \in J} f_j \cdot \| \frac{2}{r} \cdot \sum_{\ell=1}^{k} a_\ell \cdot (P_{F_\ell}(f_I^j - f_I)) - (f_I^j - f_I) \|^2$$

(P_{F_ℓ} représentant la projection orthogonale sur le sous-espace F_ℓ associé à la partition $\mathcal{P}_\ell$).

Démonstration : (elle apparaît, au cours de celle de la propriété §3.2.1. donnée en annexe).

3.3.2 Remarque. La propriété précédente montre que les modalités j ($j \in J$) jouant un rôle prépondérant dans l'interaction $\mathcal{A}(\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k)$ entre les k partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k$ sont celles pour lesquelles le quotient

$$\frac{f_j \cdot \| \frac{2}{r} \sum_{\ell=1}^{k} a_\ell \cdot P_{F_\ell}(f_I^j - f_I) - (f_I^j - f_I) \|^2}{\mathcal{A}(\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k)}$$ prend les valeurs les plus élevées.

Nous dirons qu'il s'agit des modalités $j (j \in J)$ marquant le mieux l'interaction entre les k partitions considérées.

Par contre, lorsque le quotient précédent est proche de 0, autrement dit lorsque l'on a : $f_I^j - f_I \simeq \frac{2}{r} \sum_{\ell=1}^{k} a_\ell \cdot P_{F_\ell}(f_I^j - f_I)$, nous dirons qu'il y a additivité entre la modalité j et les k partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k$.

Enfin, un quotient analogue (provenant de la définition §3.2.2.) pour les modalités i ($i \in I$) pourraît être également établi.

4. L'analyse factorielle des interactions multiples

4.1 Le tableau kr_{IJ} associé au tableau initial k_{IJ}.

4.1.1. Définition Le tableau kr_{IJ} croisant les ensembles I et J, associé au tableau initial k_{IJ}, l'ensemble I étant muni de k partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k$ affectées des poids positifs $a_1, a_2, \ldots, a_k$ de somme égale à r, se définit comme suit :

$$\forall i \in I,\ \forall j \in J : \quad kr(i,j) = \frac{2}{r} \sum_{\ell=1}^{k} a_\ell \cdot \frac{k(p_\ell^i, j) k(i)}{k(p_\ell^i)} - k(i,j)$$

p_ℓ^i étant la classe de la partition P_ℓ contenant i ($i \in I$).

4.1.2. Propriétés immédiates :

α) Les deux tableaux kr_{IJ} et k_{IJ} ont les mêmes marges sur I et sur J.

β) Les profils fr_I^j et f_I^j d'un élément j de J associés respectivement aux tableaux kr_{IJ} et k_{IJ} sont liés par la relation :

$$\forall j \in J : \quad fr_I^j = A(f_I^j) \ \text{ avec } \ A = \frac{2}{r} \sum_{\ell=1}^{k} a_i \cdot P_{F_\ell} - I$$

(P_{F_ℓ} étant la projection orthogonale sur le sous-espace de F_ℓ de $\mathbf{R}^{\text{card } I}$ associé à la partition $\mathcal{P}_\ell$, et I l'application identité).

$\gamma)$ *L'inertie du tableau* kr_{IJ} *est égale à* $\mathcal{A}(\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k)$.

4.2 Notations 1.

A partir du tableau initial k_{IJ} sont définis le tableau kr_{IJ} ainsi que les tableaux $k_{\mathcal{P}_\ell J}$ et k_{IJ_ℓ}, $\forall \ell \in [1,k]$. Le tableau ci-dessous précisera leurs définitions et les notations employées.

Tableaux	Définition	Notations des profils-lignes	Notations des profils-colonnes
k_{IJ}	tableau initial	$f_J^i = \left(\frac{k(i,j)}{k(i)}\right)_{j\in J}$	$f_I^j = \left(\frac{k(i,j)}{k(j)}\right)_{i\in I}$
kr_{IJ}	voir § 4.1.1.	$f_{rJ}^i = \left(\frac{kr(i,j)}{k(i)}\right)_{j\in J}$	$f_{rI}^j = \left(\frac{kr(i,j)}{k(j)}\right)_{i\in I}$
$\forall \ell \in [1,k]$ $k_{\mathcal{P}_\ell J}$	$\forall p \in \mathcal{P}_\ell \forall j \in J$ $k_{P_\ell J}(p,j) = \sum_{i\in p} k(i,j)$	$f_J^{p_\ell} = \sum_{i\in p_\ell} \frac{k(i)}{k(p_\ell)} \cdot f_J^i$	
$\forall \ell \in [1,k]$ k_{IJ_ℓ}	$\forall i \in I, \forall j \in J$ $k_{IJ_\ell} = \frac{k(p_\ell^i, j) k(i)}{k(p_\ell^i)}$		$f_I^{j\ell} = P_{F_\ell}(f_I^j)$

(P_{F_ℓ} étant la projection orthogonale sur le sous-espace F_ℓ associé à la partition $\mathcal{P}_\ell$).

4.3 Définition.

Nous appellerons analyse factorielle des interactions multiples, l'analyse des correspondances du tableau kr_{IJ}, les tableaux k_{IJ} et $k_{\mathcal{P}_\ell J}$ $\forall \ell \in [1,k]$ étant placés en lignes supplémentaires, et les tableaux k_{IJ} et k_{IJ_ℓ} $\forall \ell \in [1,k]$ en colonnes supplémentaires.

4.4. Notations 2.

A partir de l'axe d'ordre α issu de l'analyse des correspondances du tableau kr_{IJ}, nous précisons ci-dessous les notations des coordonnées sur cet axe, des différentes lignes et colonnes intervenant dans l'analyse factorielle des interactions multiples entre les k partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k$.

Tableaux	Coordonnées sur l'axes α	
	d'une ligne	d'une colonne
K_{IJ}	$F_\alpha(i)$	$G_\alpha(j)$
Kr_{IJ}	$Fr_\alpha(i)$	$Gr_\alpha(j)$
$\forall \ell \in [1,k]$ $k_{\mathcal{P}_\ell,J}$	$F_\alpha(p_\ell)$	
$\forall \ell \in [1,k]$ k_{IJ_ℓ}		$G_\alpha(j_\ell)$

4.5. Propriétés.

En utilisant les notations 1 et 2 précédentes (§ 4.2. et 4.4.), λ_α désignant la valeur propre d'ordre α issue de l'analyse des correspondances de kr_{IJ}, on obtient :

a) $$\lambda_\alpha = \sum_{i\in I} f_i F_\alpha^2(i) - \frac{2}{r^2}\sum_{\ell=1}^{k}\sum_{\ell'=1}^{k} a_\ell . a_{\ell'} (\sum_{i\in I} f_i [F_\alpha(p_\ell^i) - F_\alpha(p_{\ell'}^i)]^2)$$

b) $$\lambda_\alpha = \sum_{j\in J} f_j Gr_\alpha^2(j) = \sum_{j\in J} f_j [\frac{2}{r}\sum_{\ell=1}^{k} a_\ell G_\alpha(j_\ell) - G_\alpha(j)]^2$$

(où j_ℓ désigne $P_{F_\ell}(f_I^j)$).

c) $$\forall p_\ell \in \mathcal{P}_\ell \ , \ F_\alpha(p_\ell) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j\in J} \frac{k(p_\ell, j)}{k(p_\ell)} Gr_\alpha(j)$$

d) $$\forall i \in I, \ Fr_\alpha(i) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j\in J} \frac{kr(i,j)}{k(i)} Gr_\alpha(j)$$

$$= \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j\in J} [\frac{2}{r}\sum_{\ell=1}^{k} a_\ell \cdot \frac{k(p_\ell^i, j)}{k(p_\ell^i)} - \frac{k(i,j)}{k(i)}] \cdot Gr_\alpha(j)$$

$$e) \qquad \forall j \in J,\ Gr_\alpha(j) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i \in I} \frac{kr(i,j)}{k(j)} Fr_\alpha(j)$$

$$= \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i \in I} [\frac{2}{r} \sum_{\ell=1}^{k} a_\ell (P_{F_\ell}(f_I^j))_i - f_i^j] \cdot Fr_\alpha(j)$$

Démonstration : (voir annexe).

4.6 Interprétation des représentations obtenues.

Notons $\mathcal{N}r(I)$ et $\mathcal{N}r(J)$ (resp. $\mathcal{N}(I)$ et $\mathcal{N}(J)$) les nuages associés au tableau kr_{IJ} (resp. k_{IJ}).

La propriété b) du § 4.5. montre que les axes factoriels dans $\mathbf{R}^{\text{card}I}$ (issus de l'analyse de kr_{IJ}) sont construits de façon à maximiser la quantité $\sum_{j \in J} f_j [\frac{2}{r} \sum_{\ell=1}^{k} a_\ell \cdot G_\alpha(j_\ell) - G_\alpha(j)]^2$, autrement dit sont construits à partir des modalités j de J, marquant le mieux l'interaction entre les partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k$. (cf. § 3.3.2.). Une remarque analogue pourrait être faite pour les modalités i de I, intervenant dans la construction de ces différents axes factoriels. Enfin, les formules d) et e) du § 4.5. ne sont autres que les formules de transition classiques liant les 2 nuages $\mathcal{N}r(I)$ et $\mathcal{N}r(J)$.

Quant à la propriété a) du § 4.5., le terme $\sum_{i \in I} f_i F_\alpha^2(i)$ représentant l'inertie projetée du nuage $\mathcal{N}(I)$ sur l'axe α, les axes factoriels dans $\mathbf{R}^{\text{card}J}$, issus de l'analyse de kr_{IJ}, sont donc construits de façon à simultanément maximiser l'inertie projetée du nuage $\mathcal{N}(I)$ et minimiser l'«écart projeté moyen" entre les classes des différentes partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k$ prises 2 à 2. Cette dernière condition ayant pour objet de mettre en évidence les variations communes de ces différentes partitions suivant les modalités de J.

4.7. Remarques :

L'analyse factorielle des interactions multiples (§4.3.) généralise les analyses classiques :

4.7.1. Lorsque I est muni de 2 partitions $\mathcal{P}_1$ et $\mathcal{P}_2$ ($k = 2$) affectées de poids égaux à 1, le tableau kr_{IJ} (§ 4.1.) s'écrit alors :

$$kr(i,j) = \frac{k(p_1^i, j)k(i)}{k(p_1^i)} + \frac{k(p_2^i, j)k(i)}{k(p_2^i)} - k(i,j)$$

p_1^i et p_2^i étant les classes contenant i des 2 partitions $\mathcal{P}_1$ et $\mathcal{P}_2$ définies sur I.

a) Si la partition $\mathcal{P}_2$ est la partition de I réduite à la seule classe I, le tableau kr_{IJ} devient :

$$kr(i,j) = \frac{k(p_1^i,j)k(i)}{k(p_1^i)} - k(i,j) + \frac{k(j)k(i)}{k}$$

l'analyse des correspondances de kr_{IJ} représente donc l'analyse factorielle intra-classe de k_{IJ}, I étant muni de la partition $\mathcal{P}_1$.

b) Si la partition $\mathcal{P}_2$ est la partition de I composée de singletons, le tableau kr se réduit à :

$$kr(i,j) = \frac{k(p_1^i,j)k(i)}{k(p_1^i)}$$

son analyse est équivalente (par application du principe d'équivalence distributionnelle) à celle du tableau $k_{\mathcal{P}_1 J}$, croisant $\mathcal{P}_1$ et J ($k_{\mathcal{P}_1 J}(p,j) = \sum_{i\in p} k(i,j) = k(p,j), \forall p \in \mathcal{P}_1, \forall j \in J$) et par conséquent représente l'analyse factorielle inter-classe de k_{IJ}, I étant muni de la partition $\mathcal{P}_1$.

4.7.2. On démontre (voir bibliographie, Denimal 92) que, lorsque le tableau initial k_{IJ} est le tableau identité k_{II}, I étant muni de partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k$ affectées des poids $a_1, a_2, \ldots, a_k$ bien choisis, l'analyse du tableau kr_{IJ} associé est équivalente à l'analyse des correspondances multiples associée à ces k partitions définies sur I.

5. Applications à la classification automatique

5.0

Nous allons utiliser la décomposition (§3.2.1.) de l'inertie totale de k_{IJ}, pour élaborer une classification ascendante hiérarchique sur l'ensemble des éléments de I. Nous verrons que l'on peut définir un nouveau critère d'agrégation, pouvant se présenter sous la forme d'un critère de Ward modifié et vérifiant toujours l'axiome de la médiane; ce qui permettra de lui appliquer les techniques d'agrégations accélérées (graphes réductibles, voisins réciproques).

Différentes propriétés seront tout d'abord données dans le cadre de $k = 2$ partitions munies de poids égaux à 1 (§5.1.), puis étendues au cas général (§5.2.). La définition et les propriétés d'un nouveau critère d'agrégation sous contraintes seront ensuite présentées au §5.3.

5.1

On considère, donc, un tableau de contingence initial k_{IJ}, l'ensemble I, étant muni de 2 partitions $\mathcal{P}_1$ et $\mathcal{P}_2$ affectées de poids égaux à 1.

Dans ces conditions, la formule de décomposition (§3.2.1.) s'écrit :

$$\text{Inertie de } k_{IJ} = \mathcal{A}(\mathcal{P}_1, \mathcal{P}_2) + \sum_{i\in I} f_i \| f_J^{p_1^i} - f_J^{p_2^i} \|^2$$

où p_1^i et p_2^i sont les classes de $\mathcal{P}_1$ et $\mathcal{P}_2$ contenant i.

5.1.1. Théorème. Soit $\mathcal{P}_1$ et $\mathcal{P}_1'$ deux partitions de I, $\mathcal{P}_1'$ se déduisant de $\mathcal{P}_1$ par l'agrégation de 2 classes r et s de $\mathcal{P}_1$. On a alors :

a)
$$\mathcal{A}(\mathcal{P}_1', \mathcal{P}_2) - \mathcal{A}(\mathcal{P}_1, \mathcal{P}_2) = \sum_{i\in r} f_i \cdot \|f_J^r - f_J^{p_2^i}\|^2 + \sum_{i\in s} f_i \|f_J^s - f_J^{p_2^i}\|^2 - \sum_{i\in r\cup s} f_i \cdot \|f_J^{r\cup s} - f_J^{p_2^i}\|^2$$

b)
$$\mathcal{A}(\mathcal{P}_1', \mathcal{P}_2) - \mathcal{A}(\mathcal{P}_1, \mathcal{P}_2) = \frac{f_r \cdot f_s}{f_r + f_s} \cdot \left[\|(\overline{f_J^r} - f_J^r) - (\overline{f_J^s} - f_J^s)\|^2 - \|\overline{f_J^r} - \overline{f_J^s}\|^2\right]$$

où $\overline{f_J^r}$ désigne le reconstitué de r à partir de la partition $\mathcal{P}_2$, autrement dit $\overline{f_J^r} = \sum_{i\in r} \frac{f_i}{f_r} f_J^{p_2^i}$. Les notations $\overline{f_J^s}$ et $\overline{f_J^{r\cup s}}$ auront des significations analogues.

La démonstration est donnée dans Denimal (1992).

5.1.2. Définition et propriété.

a) I étant muni de la partition $\mathcal{P}_2$, on pose pour r et s deux sous-ensembles disjoints de I (avec les notations du §5.1.1.)

$$\delta(r,s) = \frac{f_r \cdot f_s}{f_r + f_s}\left[\|(\overline{f_J^r} - f_J^r) - (\overline{f_J^s} - f_J^s)\|^2 - \|(\overline{f_J^r} - \overline{f_J^s})\|^2\right]$$

b) soit t, r, s 3 sous-ensembles de I, disjoints 2 à 2, on montre alors que :

$$\delta(t, r\cup s) = \frac{1}{f_t + f_r + f_s}\left[(f_t + f_s)\delta(t,s) + (f_t + f_r)\delta(t,r) - f_t\delta(r,s)\right]$$

Démonstration : (voir annexe).

5.1.3. Conséqences. δ va donc vérifier l'axiome de la médiane. Autrement dit, r, s, t étant 3 sous-ensembles de I, disjoints 2 à 2, on obtient :

$$[\delta(r,s) \leqslant \inf[\delta(r,t), \delta(s,t)]] \Rightarrow [\mathrm{Inf}[\delta(r,t), \delta(s,t)] \leqslant \delta(r\cup s, t)]$$

5.2. Généralisation au cas des interactions multiples.

On considère, dans ce §, un tableau initial k_{IJ}, I étant cette fois muni de k partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k$ affectées des poids respectifs $a_1, a_2, \ldots, a_k$ de somme r. La décomposition de l'inertie de k_{IJ} s'écrit alors (§3.2.1.).

$$\text{Inertie de } k_{IJ} = \mathcal{A}(\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k) + \frac{2}{r^2}\sum_{\ell=1}^{k}\sum_{\ell'=1}^{k} a_\ell \cdot a_{\ell'}\left(\sum_{i\in I} f_i \cdot \|f_J^{p_\ell^i} - f_J^{p_{\ell'}^i}\|^2\right)$$

5.2.1. Théorème. Sous les hypothèses précédentes, $\mathcal{P}'_k$ étant une partition de I, se déduisant de $\mathcal{P}_k$ par l'agrégation de 2 classes r et s de $\mathcal{P}_k$, on a alors :

a)

$$\mathcal{A}(\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_{k-1}, \mathcal{P}'_k) - \mathcal{A}(\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_{k-1}, \mathcal{P}_k) = \frac{4}{r^2} \sum_{\ell=1}^{k-1} a_\ell . a_k . \delta^\ell(r, s)$$

avec

- $r = \sum_{\ell=1}^{k} a_\ell$
- $\delta^\ell(r, s) = \dfrac{f_r . f_s}{f_r + f_s} . \left[\| (\overline{f_J^r}^\ell - f_J^r) - (\overline{f_J^s}^\ell - f_J^s) \|^2 - \| \overline{f_J^r}^\ell - \overline{f_J^s}^\ell \|^2 \right]$

où $\overline{f_J^r}^\ell$ et $\overline{f_J^s}^\ell$ représentent les reconstitués de r et s à partir de la partition $\mathcal{P}_\ell$.

b) t, r, s étant 3 sous-ensembles de I, 2 à 2 disjoints, en posant :

$$\delta(r, s) = \frac{4}{r^2} \sum_{\ell=1}^{k-1} a_\ell . a_k . \delta^\ell(r, s)$$

on obtient alors :

$$\delta(t, r \cup s) = \frac{1}{f_t + f_r + f_s} . [(f_t + f_s) . \delta(t, s) + (f_t + f_r)\delta(t, r) - f_t \delta(r, s)]$$

c) δ vérifie l'axiome de la médiane.

Démonstration : (*cf.* Denimal (1992)).

5.3. Un nouveau critère d'agrégation sous contraintes.

Le critère d'agrégation sous contraintes proposé dans cet article, (qui se présentera sous la forme d'un critère de Ward modifié) est un cas particulier de l'indice δ du §5.2. .

Nous supposerons, ici, que l'ensemble I du tableau k_{IJ} est muni de $(k-1)$ partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_{k-1}$ données a priori et affectées de poids connus $a_1, a_2, \ldots, a_{k-1}$, telles que :

- $k \geqslant 3$
- $\mathcal{P}_{k-1} =$ partition réduite à 1 classe.

Sous ces conditions, on se propose de construire une hiérarchie sur I, qui, à chaque niveau n élabore une partition $\mathcal{P}_k^{(n)}$ de poids donné a_k et d'indice $\nu(n)$ égal à :

$$\nu(n) = \mathcal{A}(\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_{k-1}, \mathcal{P}_k^{(n)}) - \mathcal{A}(\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_{k-1}, \mathcal{P}_k^{(n-1)})$$

D'après la propriété (§5.2.1.), $\mathcal{P}_k^{(n)}$ se déduisant de $\mathcal{P}_k^{(n-1)}$ par l'agrégation des classes r et s on peut écrire que :

$$\nu(n) = \frac{4a_k}{r^2} \sum_{\ell=1}^{k-1} a_\ell \cdot \delta^\ell(r,s) \quad \text{avec} \quad r = \sum_{\ell=1}^{k} a_\ell$$

Or, $\mathcal{P}_{k-1}$ étant la partition réduite à une classe, nous en déduisons, pour $\ell = k-1$, que :

$$\delta^\ell(r,s) = \frac{f_r \cdot f_s}{f_r + f_s} \| f_J^s - f_J^r \|^2$$

puisque, dans ce cas : $\overline{f_J^r}^\ell = \overline{f_J^s}^\ell = f_J$. Ce qui est le critère de Ward classique, que nous noterons Ward(r,s).

En conséquence, le critère d'agrégation proposé s'écrit alors sous la forme :

$$\delta(r,s) = \frac{4a_k}{r^2} [\sum_{\ell=1}^{k-2} a_\ell \delta^\ell(r,s) + a_{k-1} . \text{Ward}(r,s)]$$

Quant à l'interprétation de ce critère, nous la présentons ci-dessous dans le cadre des hypothèses : $k = 3$ et $a_1 = a_2 = a_3 = 1$, (l'extension au cas général s'en déduit alors facilement).

Sous ces conditions,

$$\delta(r,s) = \frac{4}{9} [\delta^1(r,s) + \text{Ward}(r,s)]$$

où

$$\begin{aligned} \delta^1(r,s) &= \mathcal{A}(\mathcal{P}_1, \mathcal{P}_3^{(n)}) - \mathcal{A}(\mathcal{P}_1, \mathcal{P}_3^{(n-1)}) \\ &= \sum_{i \in I} f_i \| f_J^{p_1^i} - f_J^{p_3^i} \|^2 - \sum_{i \in I} f_i \| f_J^{p_1^i} - f_J^{p_3'^i} \|^2 \end{aligned}$$

($p_1^i, p_3^i, p_3'^i$ étant les classes des partitions $\mathcal{P}_1, \mathcal{P}_3^{(n-1)}, \mathcal{P}_3^{(n)}$ contenant i) (voir §5.1).

En conséquence, les classes r et s, qui sont agrégées, doivent minimiser $\delta(r,s)$ et donc vérifient les 2 conditions :

- miminiser Ward(r,s)
- minimiser $\delta^1(r,s)$, c'est-à-dire encore maximiser $\sum_{i \in I} f_i \| f_J^{p_1^i} - f_J^{p_3'^i} \|^2$. Autrement dit, la nouvelle partition construite $\mathcal{P}^{(n)}$ devra être la plus «éloignée» possible de la partition $\mathcal{P}_1$ donnée *a priori* (cette partition $\mathcal{P}_1$ représentant une variable qualitative dont on veut minimiser l'influence).

Enfin, notre indice d'agrégation peut prendre des valeurs négatives, mais comme il vérifie l'axiome de la médiane, la suite des indices de niveau sera forcément croissante.

6. Un exemple d'application

6.1. Introduction :

Les données utilisées sont extraites d'une étude que j'ai menée conjointement avec l'Institut Pasteur de Lille et le CRESGE[(1)] (*cf.* Denimal (1991)), se présentent sous la forme d'un tableau de contingence, calculé à partir de 11000 individus, croisant un ensemble I de modalités (définies à partir des variables sexe, âge, CSP, niveau de consommation d'alcool) et un ensemble J composé des modalités de 6 marqueurs biologiques.

Les paragraphes 6.2., 6.3., 6.4., ci-dessous seront consacrés respectivement aux définitions des ensembles I et J, ainsi qu'à celle des 2 partitions $\mathcal{P}_1$ et $\mathcal{P}_2$ définies sur I.

6.2. Définition et notations des modalités de I.

TABLEAU 1

Elements de I	SIGNIFICATIONS		EFFECTIFS	NOMBRE MOYEN DE GRAMMES D'ALCOOL ABSORBE PAR SEMAINE
NOTATIONS	SEXE*AGE*CSP	CONSOMMATIONS		
F302	femme d'age	non nulle	239	122
F301	inferieur à 30	nulle	1723	0
F403	femme	forte	104	321
F402	d' age	faible	410	122
F401	[30;40[	nulle	843	0
F603	femme	forte	216	303
F602	d'age	faible	617	131
F601	[40;60[	nulle	964	0
F>62	femme d'age	non nulle	225	155
F>61	superieur à 60	nulle	270	0
HIN2	homme d'age	non nulle	103	205
HIN1	[0;30[inactif	nulle	486	0
HEM3	homme d'age	forte	118	444
HEM2	[0;30[ouvrier employe	moyenne	378	214
HEM1	cadre moyen	nulle	414	0
HCS2	homme d'age [0;30[cadre	non nulle	54	253
HCS1	superieur	nulle	37	0
HOV5	homme	tres forte	150	594
HOV4	d'age [30;40[	forte	168	378
HOV3	ouvrier	moyenne	143	268
HOV2	employe	faible	139	154
HOV1	inactif	nulle	211	0
HCM4	homme d'age	tres forte	94	505
HCM3	[30;40[	forte	191	297
HCM2	cadre moyen	moyenne	179	150
HCM1	et superieur	nulle	164	0
HEC4	homme d'age	tres forte	283	543
HEC3	[40;60[ouvrier	forte	408	371
HEC2	employe inactif	moyenne	387	182
HEC1	cadre moyen	nulle	268	0
HSP4	homme d'age	tres forte	104	542
HSP3	[40;60[	forte	223	342
HSP2	cadre	moyenne	138	162
HSP1	superieur	nulle	94	0
H>64	homme	tres forte	68	523
H>63	d'age	forte	209	328
H>62	superieur	moyenne	114	150
H>61	à 60	faible	64	0

6.3. Définition et notations des modalités de J.

TABLEAU 2

ELEMENTS DE J	SIGNIFICATION DES VARIABLES	DEFINITIONS DE LEURS MODALITES
CHL1	CHOLESTEROL en mmol / l	[2 ; 5 [
CHL2		[5 ; 5,7 [
CHL3		[5,7 ; 6,5 [
CHL4		[6,5 ; 13 [
ACD1	ACIDE URIQUE en μ mol / l	[100 ; 200 [
ACD2		[200 ; 240 [
ACD3		[240 ; 280 [
ACD4		[280 ; 800 [
GLU1	GLUCOSE en mmol / l	[3 ; 4,9 [
GLU2		[4,9 ; 5,3 [
GLU3		[5,3 ; 5,8 [
GLU4		[5,8 ; 20 [
GGT1	GGT en mu /ml	[0 ; 10 [
GGT2		[10 ; 15 [
GGT3		[15 ; 20 [
GGT4		[20 ; 33 [
GGT5		sup. à 33
VGM1	VOLUME GLOBULAIRE MOYEN (en fl)	[0 ; 86 [
VGM2		[86 ; 93 [
VGM3		sup. à 93
LGO1	TOTAL LEGO	0
LGO2		]0 ,2]
LGO3		3
LGO4		sup. à 3

6.4. Les partitions P_1 et P_2 de I.

$P_1 = \{F < 30, F < 40, F < 60, F > 60, H < 30, H < 40, H < 60, H > 60\}$

$P_2 = \{ALC1, ALC2, ALC3, ALC4, ALC5\}$

TABLEAU 3

PARTITIONS	CLASSES	CONTENUS DES CLASSES
P1	F<30	F301 F302
	F<40	F401 F402 F403
	F<60	F601 F602 F603
	F>60	F>61 F>62
	H<30	HIN1 HIN2 HEM1 HEM2 HEM3 HCS1 HCS2
	H<40	HOV1 HOV2 HOV3 HOV4 HOV5 HCM1 HCM2 HCM3 HCM4
	H<60	HEC1 HEC2 HEC3 HEC4 HSP1 HSP2 HSP3 HSP4
	H>60	H>61 H>62 H>63 H>64
P2	ALC1	F301 F401 F601 F>61 HIN1 HEM1 HCS1 HOV1 HCM1 HSP1 HEC1 H>61
	ALC2	F302 F402 F602 F>62 HOV2 HCM2 H>62
	ALC3	HIN2 HEM2 HCS2 HOV3 HEC2 HSP2
	ALC4	F403 F603 HEM3 HOV4 HCM3 HEC3 HSP3 H>63
	ALC5	HOV5 HCM4 HEC4 HSP4 H>64

6.5. Les analyses réalisées.

6.5.1. Résumé des premières analyses.

Ce tableau de contingence k_{IJ} croisant les ensembles I et J, a tout d'abord été soumis à l'analyse des correspondances qui a généré un plan principal, représentant 86% de l'inertie totale (dont 72% pour le premier axe), et principalement expliqué par l'âge et le sexe : les plus âgés se détachent des plus jeunes en prenant les valeurs les plus importantes pour l'ensemble des marqueurs biologiques, et les femmes se distinguent des hommes par des valeurs plus faibles et plus particulièrement pour l'acide urique.

Ainsi, l'âge et le sexe étant des variables prépondérantes sur la consommation d'alcool, (ce que l'on peut vérifier aussi par d'autres analyses), pour faire apparaître les liaisons entre consommation d'alcool et marqueurs biologiques, une analyse factorielle intra-classe du tableau k_{IJ}, mettant en jeu la partition P_1 (P_1 = AGE × SEXE) a été réalisée.

Cette analyse, qui a pour but de minimiser l'influence de l'âge et du sexe, génère un plan principal totalisant 75% de Intra(P_1) (c'est-à-dire 17% de l'inertie totale de k_{IJ}) dont 59% pour le premier axe. Ce dernier s'interprète comme un axe d'accroissement de la consommation d'alcool, les variables biologiques la marquant le mieux étant GGT et TOTAL LEGO. 3 zones se dégagent alors nettement, dans le plan (1,2).

Celles

- des consommations faibles ($F_1 > 0, F_2 > 0$)
- des consommations moyennes ($F_1 < 0, F_2 > 0$)
- des consommations importantes ($F_1 < 0, F_2 < 0$).

6.5.2. Analyses des interactions entre P_1 et P_2 (Graphiques 1 et 2).

Il s'agit de l'analyse (présentée au § 4.3.) avec $k = 2$ partitions ($\mathcal{P}_1$ et $\mathcal{P}_2$), affectées de poids égaux à 1, k_{IJ} étant le tableau de contingence initial.

Cette analyse génère un plan (1,2) représentant 83% de l'interaction $\mathcal{A}(\mathcal{P}_1, \mathcal{P}_2)$ (c'est-à-dire 8% de l'inertie de k_{IJ}), dont 77% pour le premier axe.

A l'aide des indices INR, (voir (1) en bas de page), calculés par le programme d'analyse des correspondances, et en choisissant les plus faibles d'entre eux (INR $\simeq$ 1% de $\mathcal{A}(\mathcal{P}_1, \mathcal{P}_2)$), on voit apparaître une zone (voir graphiques 1 et 2) où il y a quasi-additivité. Ce qui s'écrit, avec les notations employées dans cet article :

$$\begin{cases} i \in I \\ f_J^i \simeq f_J^{p_1^i} + f_J^{p_2^i} - f_J \end{cases} \qquad \begin{cases} j \in J \\ f_I^j \simeq P_{F_1}(f_I^j) + P_{F_2}(f_I^j) - f_I \end{cases}$$

Il apparaît, ainsi, que les modalités de I correspondant aux consommations moyennes d'alcool et les modalités des marqueurs (donc, de J) traduisant des valeurs moyennes sont celles pour lesquelles il y a quasi-additivité.

---

(1) INR = indice représentant les contributions de chaque $i \in I$ (ou $j \in J$) à l'inertie totale du tableau analysé, rapportées à cette inertie.

D'autre part, le tableau analysé kr se définit comme suit :

$$kr(i,j) = \left[\frac{k(p_1^i,j)k(i)}{k(p_1^i)} + \frac{k(p_2^i,j)k(i)}{k(p_2^i)}\right] - k(i,j)$$

et représente la différence entre la valeur que l'on aurait sous l'hypothèse de l'additivité des effets de l'âge et du sexe d'une part et de la consommation d'alcool d'autre part, et la valeur observée $k(i,j)$. Dans le cadre de cet exemple, $kr(i,j)$ est toujours positif.

Autrement dit, les formules de transition (§ 4.5 d) et e)) s'interprètent facilement.

Ainsi, une absence croissante d'additivité va se traduire par des valeurs $k(i,j)$ de plus en plus faibles devant celles que l'on aurait sous cette hypothèse d'additivité. Les représentations obtenues (graphiques 1 et 2) mettent alors en évidence les 2 résultats suivants :

- Pour les hommes, plus les individus sont consommateurs d'alcool, moins il y a d'additivité ce qui est surtout observé pour les modalités traduisant les valeurs élevées des marqueurs GGT et TOTAL LEGO.
- A l'inverse, chez les femmes, plus la consommation est faible, moins il y a additivité ce qui est, en particulier, remarqué pour les modalités traduisant les valeurs faibles du marqueur Acide Urique.

6.5.3. Classifications des modalités de I (Graphiques 3 et 4).

La première classification est celle réalisée sur I, à partir du tableau k_{IJ}, en utilisant le critère classique de Ward. Cette première classification ascendante hiérarchique génère des classes marquées principalement par l'âge et le sexe (graphique 3).

On obtient en effet une partition de I en 5 classes :

- Deux d'entre elles regroupent, pour l'une (classe 71) les femmes jeunes et pour l'autre les femmes âgées (classe 69).
- Les hommes, quant à eux, se répartissent au sein des autres classes : la classe 70 regroupant les jeunes (avec pour la tranche d'âge [0,30[tous les niveaux de consommation), la classe 68 les hommes âgés (avec tous les niveaux de consommation pour la tranche d'âge > 60), reste la classe 12 ne comportant que le seul élément HIN1 dont le profil se rapproche de celui des femmes jeunes (classe 70).

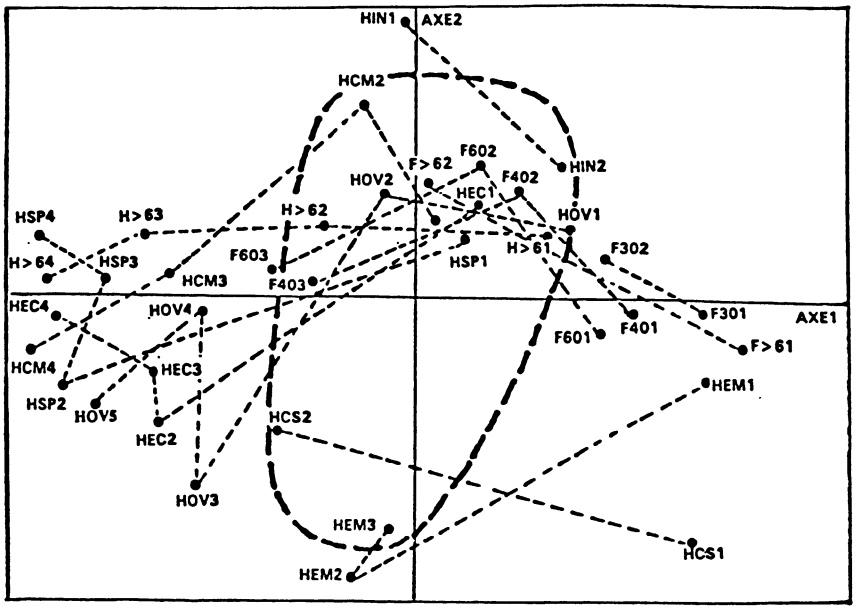
La seconde classification des éléments de I est une classification sous contraintes réalisée à partir des 2 partitions données a priori : $\mathcal{P}_1$ et $\mathcal{P}_0$ (partition de I réduite à la seule classe I). Les 3 partitions $\mathcal{P}_1, \mathcal{P}_0, \mathcal{P}^{(n)}$ (cette dernière étant pour le niveau n de la hiérarchie, celle constituée des sommets) seront munies du même poids. Ainsi, le critère d'agrégation employé (§ 5.3) serait le critère de Ward modifié suivant :

$$\nu(n) = \frac{4}{9}\{[\mathcal{A}(\mathcal{P}_1,\mathcal{P}^{(n)}) - \mathcal{A}(\mathcal{P}_1,\mathcal{P}^{(n-1)})] + \text{Ward}(n)\}$$

et a donc pour but de minimiser l'influence de l'âge et du sexe.

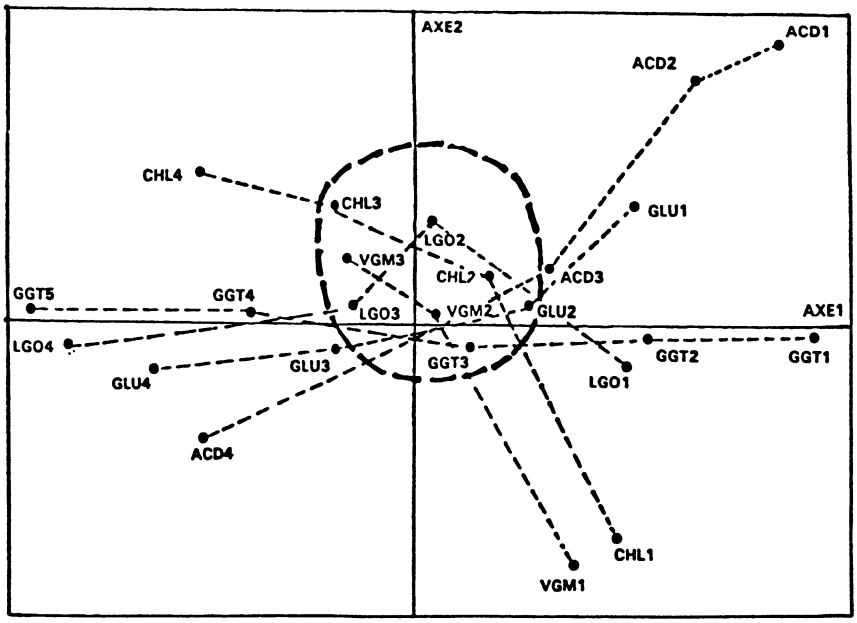
GRAPHIQUE 1

Analyse des interactions (P1, P2) Projection dans le plan (1,2) des éléments de I, issus du tableau $kr_{IJ}(\tau_1 = 77\%, \tau_2 = 6\%)$

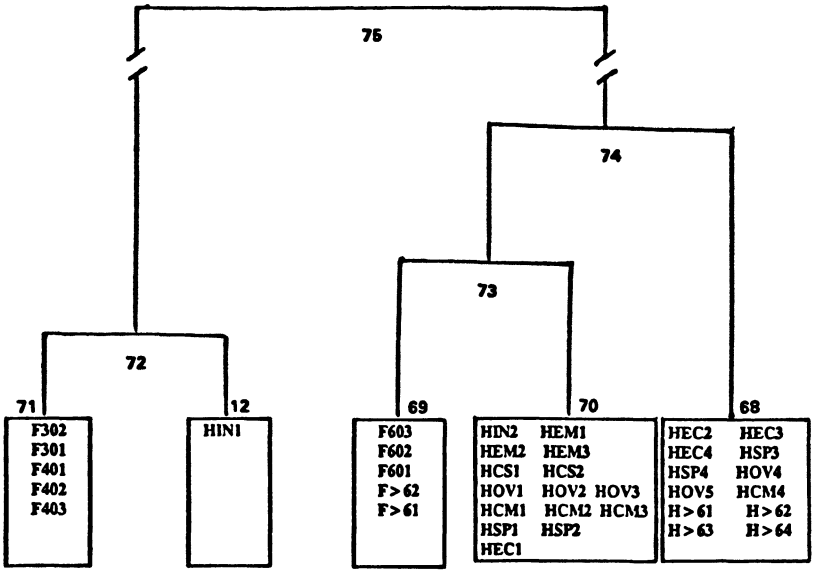


GRAPHIQUE 2

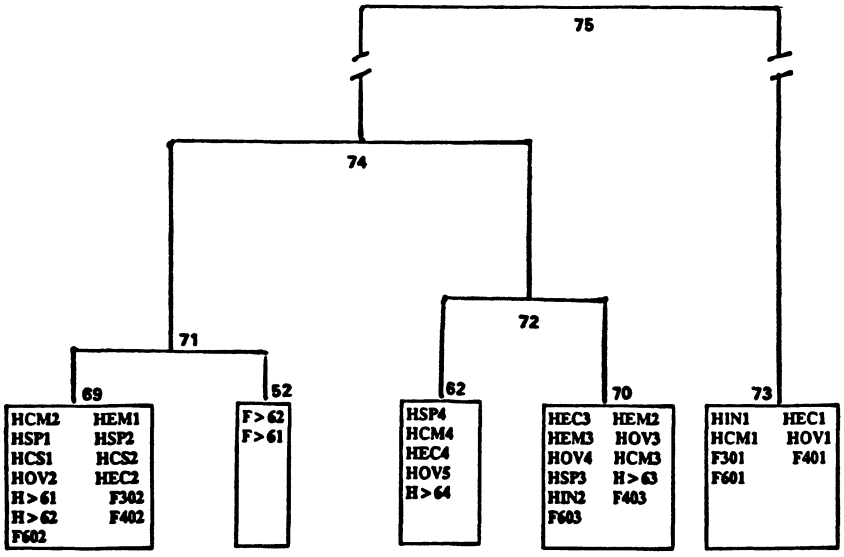
Analyse des interactions (P1, P2). Projection dans le plan (1,2) des éléments de J, issus du tableau $kr_{IJ}(\tau_1 = 77\%, \tau_2 = 6\%)$.



GRAPHIQUE 3
Classification des éléments de I – Critère classique de Ward



GRAPHIQUE 4
Classification des éléments de I sous contraintes



La classification hiérarchique obtenue génère une partition en 5 classes s'expliquant cette fois par la consommation d'alcool. On obtient, en effet, les classes 73, 69, 70, 62 regroupent respectivement les non-consommateurs, les consommateurs d'alcool moyens, importants et très importants. Seules, les femmes d'âge supérieur à 60 ans, formant la classe 52, se regroupent avec la classe des consommateurs moyens. Pour les autres classes (73, 69, 70, 62) seule la consommation d'alcool les explique et l'on y trouve regroupées, dans chacune de ces classes, les tranches d'âge et les 2 sexes correspondant à un type de consommation donné.

Annexes

1 - Démonstration de 3.1. Après avoir mis le terme :

$\frac{2}{r^2}\sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j < P_{F_i}(x) - P_{F_j}(x), P_{F_i}(y) - P_{F_j}(y) >$ sous la forme

$$\frac{1}{2r^2}\sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j \langle (2P_{F_i} - I - A)(x) - (2P_{F_j} - I - A)(x),$$
$$(2P_{F_i} - I - A)(y) - (2P_{F_j} - I - A)(y)\rangle.$$

La formule du lemme ci-dessus s'obtient, après avoir remarqué que :

- d'une part :

$$\frac{1}{2r^2}\sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j \langle (2P_{F_i} - I - A)(x), (2P_{F_i} - I - A)(y)\rangle$$
$$= \frac{1}{2}\langle x, y\rangle - \frac{1}{2}\langle A(x), A(y)\rangle$$

(En effet, $\forall i \in [1,k]$: $\langle (2P_{F_i} - I)(x), (2P_{F_i} - I)(y)\rangle = \langle x, y\rangle$ et

$$\frac{1}{2r^2}\sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j \langle (2P_{F_i} - I)(x), A(y)\rangle$$
$$= \frac{1}{2r}\sum_{j=1}^{k} a_j \langle A(x), A(y)\rangle = \frac{1}{2}\langle A(x), A(y)\rangle)$$

- d'autre part :

$$\frac{1}{2r^2}\sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j \langle (2P_{F_i} - I - A)(x), (2P_{F_j} - I - A)(y)\rangle = 0.$$

2 - Démonstration de 3.2.1. Cette décomposition provient du lemme de décomposition (§ 3.1.), en prenant dans $E = \mathbf{R}^{\text{card I}}$ muni de la métrique du CHI2, $x = y = f_i^j - f_I$.

On obtient, en effet, par application de la propriété § 2.2.2. b) et c) que :

$$\forall \ell \in [1,k] : \quad P_{F_\ell}(x) = \left(\frac{k(p_\ell^i, j)k(i)}{k(p_\ell^i)k(j)} \right)_{i \in I} - f_I$$

On en déduit, alors, A étant l'opérateur $\frac{2}{r} \sum_{\ell=1}^{k} a_\ell \cdot P_{F_\ell} - I$, que :

$$A(x) = \frac{2}{r} \sum_{\ell=1}^{k} a_\ell \left(\frac{k(p_\ell^i, j)k(i)}{k(p_\ell^i)k(j)} \right)_{i \in I} - f_I^j - f_I$$

Le lemme permet donc d'écrire que l'inertie In de k_{IJ} se décompose comme suit :

$$In = \sum_{j \in J} f_j \| f_I^j - f_I \|^2 = \sum_{j \in J} f_j \| \frac{2}{r} \sum_{\ell=1}^{k} a_\ell \cdot \left(\frac{k(p_\ell^i, j)k(i)}{k(p_\ell^i)k(j)} \right)_{i \in I} - f_I^j - f_I \|^2 +$$
$$\sum_{j \in J} f_j \left\{ \frac{2}{r^2} \sum_{\ell=1}^{k} \sum_{\ell'=1}^{k} a_\ell \cdot a_{\ell'} \| \left(\frac{k(p_\ell^i, j)k(i)}{k(p_\ell^i)k(j)} \right)_{i \in I} - \left(\frac{k(p_{\ell'}^i, j)k(i)}{k(p_{\ell'}^i)k(j)} \right)_{i \in I} \|^2 \right\}$$

D'où l'on déduit facilement la formule de cette propriété (§ 3.2.1.).

3 - Démonstration de 4.5. Nous nous limiterons à celle du a).

Sachant que l'on a, par définition, l'égalité : (en utilisant les notations § 4.2.) $f_{rJ}^i = \frac{2}{r} \sum_{\ell=1}^{k} a_\ell \cdot f_J^{p_\ell^i} - f_J^i$, nous en déduisons que :

$$Fr_\alpha(i) = \frac{2}{r} \sum_{\ell=1}^{k} a_\ell \cdot F_\alpha(p_\ell^i) - F_\alpha(i).$$

En conséquence :

$$\lambda_\alpha = \sum_{i \in I} f_i \cdot Fr_\alpha^2(i) = \sum_{i \in I} f_i [\frac{2}{r} \cdot \sum_{\ell=1}^{k} a_\ell F_\alpha(p_\ell^i) - F_\alpha(i)]^2$$
$$= \sum_{i \in I} \frac{1}{f_i} [\frac{2}{r} \sum_{\ell=1}^{k} a_\ell F_\alpha(p_\ell^i) \cdot f_i - F_\alpha(i) f_i]^2$$

ce qui montre que $\lambda_\alpha = \|A(F_\alpha(i)f_i)_{i\in I}\|^2$ où A est l'opérateur $\frac{2}{r}\sum_{\ell=1}^{k} a_\ell P_{F_\ell} - I$

D'où, finalement, en utilisant le lemme de décomposition (§ 3.1), il apparaît que :

$$\sum_{i\in I} f_i F_\alpha^2(i) = \sum_{i\in I} f_i F r_\alpha^2(i) + \frac{2}{r^2}\sum_{\ell=1}^{k}\sum_{\ell'=1}^{k} a_\ell a_{\ell'}\Big(\sum_{i\in I} f_i (F_\alpha(p_\ell^i) - F_\alpha(p_{\ell'}^i))^2\Big).$$

ce qui démontre la formule a).

4 - Démonstration de 5.1.2. $\delta(r,s)$ peut s'écrire sous la forme : $\delta(r,s) = \delta_1(r,s) - \delta_2(r,s)$ avec :

$$\delta_1(r,s) = \frac{f_r \cdot f_s}{f_r + f_s}\|(\overline{f_J^r} - f_J^r) - (\overline{f_J^s} - f_J^s)\|^2$$
$$\delta_2(r,s) = \frac{f_r \cdot f_s}{f_r + f_s}\|\overline{f_J^r} - \overline{f_J^s}\|^2$$

δ_1 (resp. δ_2) représente le critère de Ward pour le nuage $\mathcal{N}_1(I)$ (resp. $\mathcal{N}_2(I)$) associé au tableau k_{IJ}^1 (resp. k_{IJ}^2) suivant :

$$k^1(i,j) = k(i,j) - \frac{k(p_2^i,j)k(i)}{k(p_2^i)} + \frac{k(i)k(j)}{k}$$
$$k^2(i,j) = \frac{k(p_2^i,j)k(i)}{k(p_2^i)}$$

les 2 tableaux k_{IJ}^1 et k_{IJ}^2 ayant même marges que le tableau initial k_{IJ}, δ_1 et δ_2, et par conséquent δ vont vérifier l'égalité demandée.

Bibliographie

BENER A. (1982) *Décomposition des interactions dans une correspondance multiple.* Cahiers de l'Analyse des Données 7 n° 1.

BENZECRI J.-P. (1983) *Analyse de l'inertie intra-classe par l'analyse d'un tableau de correspondance.* Cahiers de l'Analyse des Données 8, n° 3.

CAZES P. (1986) *Correspondances entre deux ensembles et partitions de ces deux ensembles.* Cahiers de l'Analyse des Données, 11, n° 3.

CAZES P., CHESSEL D., DOLEDEC S. (1988) *L'analyse des correspondances internes d'un tableau partitionné. Son usage en hydrobiologie.* Revue de Statistique Appliquée XXXVI n° 1.

DENIMAL J.-J. (1992) *Analyse factorielle des interactions entre k partitions prises 2 à 2.* Publications IRMA, Lille, Volume 27.

DENIMAL J.-J. (1991) *Essai de modélisation de la consommation d'alcool dans une population.* Article (H.C.E.I.A.) supplément scientifique et technique

(CRESGE - Centre d'Examens de Santé de l'Institut Pasteur de Lille). Journal d'Alcoologie.

MOREAU J. (1990) *Analyse factorielle et relationnelle de données structurées.* Centre Vaudois de recherches pédagogiques n° 91103, Lausanne.

SABATIER (1987) *Analyse factorielle de données structurées et métriques.* S.A.D. Volume 12, n° 3.