LA CLASSIFICATION CROISEE

Gérard GOVAERT

L.R.I.M. Université de Metz, Ile du Saulcy 57045 Metz
INRIA-Lorraine, BP 239 54506 Vandoeuvre les Nancy Cedex

Résumé :

Contrairement à la plupart des méthodes de classification, et comme le font les méthodes factorielles, la classification croisée a pour objectif de traiter simultanément les deux ensembles qui sont mis en correspondance dans le tableau de données à analyser. Nous rappelons ici le principe général de cette approche qui englobe plusieurs algorithmes suivant le type de tableaux de données envisagés (tableaux de contingence, tableaux de variables quantitatives, questionnaire, tableaux binaires), puis nous précisons quelques uns de ces algorithmes en détaillant surtout leur utilisation à l'aide d'exemples.

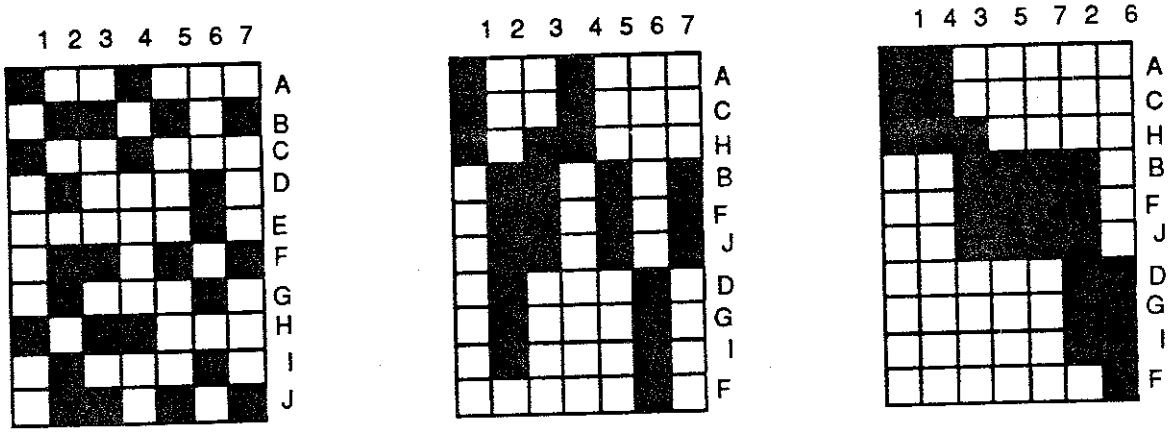
Mots-clés : classification, classification simultanée, tableau de contingence, tableau binaire, questionnaire.

1. INTRODUCTION

Les **méthodes factorielles** (analyse en composantes principales, analyse des correspondances, analyse des correspondances multiples) ont tous pour objectif la simplification d'un tableau rectangulaire de données défini sur deux ensemble I et J. Une des particularités essentielles de ces méthodes est d'obtenir des résultats **simultanément** sur les deux ensembles. Il apparait donc naturel d'avoir la même démarche lorsqu'on cherche à faire de la classification sur un tableau rectangulaire de données. Or, la plupart des méthodes de classification ont une approche tout à fait dissymétrique en ne faisant porter la structure recherchée que sur un seul des deux ensembles, privilégiant ainsi l'un d'entre eux. Au contraire, la **classification croisée** (Govaert 1977, 1983 et 1984), qui englobent plusieurs méthodes suivant le type de tableaux de données envisagés (tableaux de contingence, tableaux de variables quantitatives, tableaux binaires, tableaux de variables qualitatives) a comme objectif principal de fournir **simultanément** une partition de chacun des deux ensembles.

Cette démarche n'est pas nouvelle. Fisher (1969) pose le problème de recherche simultanée de deux partitions de manière matricielle ; il définit un critère à optimiser, mais ne propose aucune méthode pour résoudre ce problème. Anderberg (1973) cite parmi la liste des problèmes posés par la classification celui du choix de l'ensemble à classifier. Il considère tout aussi raisonnable de classifier les variables que les individus et il suggère même une approche itérative dans laquelle la classification est faite alternativement sur les individus et les variables jusqu'à ce que les classifications obtenues soient "mutuellement harmonieuses". Dans le cas d'un tableau de contingence, Benzecri (1973) définit deux mesures d'information associées à un tableau de contingence. L'une d'entre elles est le χ^2 de contingence, critère que nous reprendrons dans le cas des tableaux de contingence. Toledano et Brousse (1977) posent un problème analogue mais en recherchant cette fois une hiérarchie sur chacun des deux ensembles. Bock (1979) montre l'intérêt de la classification simultanée et donne plusieurs exemples de problèmes pour lesquels une bonne solution est fournie par une classification simultanée. Il propose alors un certain nombre de méthodes de classification simultanée qui reposent à chaque fois sur des modèles.

La classification croisée peut être aussi rapprochée des méthodes de **classification directe** (Hartigan, 1972 et 1975), qui essaient d'organiser directement le tableau initial ou des méthodes manuelles de (Bertin, 1977) dont le principe est de simplifier le tableau en permutant lignes et colonnes.



Réorganisation par permutation des lignes et des colonnes

On peut enfin indiquer que, dans de nombreux cas, bien que les classifications des deux ensembles soient obtenues séparément par des méthodes habituelles, l'analyse des résultats est faite en essayant de prendre en compte simultanément les deux types de résultats. On peut citer par exemple Maroy et Peneau (1972), Jambu (1976) et Lerman et Leredde (1977).

Avant d'entrer plus dans le détail des algorithmes, nous pouvons préciser notre objectif sur un petit exemple :

Soit le tableau des données suivant :

	1	2	3	4	5	6	7	8	9	10
a	1	0	1	0	1	0	0	1	0	1
b	0	1	0	1	0	1	1	0	1	0
c	1	0	0	0	0	0	0	1	1	0
d	1	0	1	0	0	0	0	1	0	0
e	0	1	0	1	0	1	1	0	1	0
f	0	1	0	0	0	1	1	0	1	0
g	0	1	0	0	0	0	0	1	0	1
h	1	0	1	0	1	1	0	1	1	1
i	1	0	0	1	0	0	0	0	0	1
j	0	1	0	1	0	0	1	0	0	0

L'ensemble des lignes correspond à un ensemble de 10 micro-ordinateurs et l'ensemble des colonnes à 10 propriétés que ces micro-ordinateurs peuvent posséder (valeur 1) ou ne pas posséder (valeur 0). On cherche à résumer ce tableau de la manière suivante : classer les deux ensembles de telle sorte que les classes de micro-ordinateurs soient bien caractérisées par les classes de propriétés et, vice versa, que les classes de propriétés soient bien caractérisées par les classes de micro-ordinateurs.
Dans notre exemple, si on considère les classifications {{a,d,h},{b,e,f,j},{c,g,i}} et {{1,3,5,8,10},{2,4,6,7,9} on obtient en réordonnant les lignes et les colonnes suivant ces deux partitions le tableau suivant :

		1					2				
		1	3	5	8	10	2	4	6	7	9
A	a	1	1	1	1	1	0	0	0	0	0
	d	1	1	0	1	0	0	0	0	0	0
	h	1	1	1	1	1	0	0	1	0	1
B	b	0	0	0	0	0	1	1	1	1	1
	e	0	0	0	0	0	1	1	1	1	1
	f	0	0	0	0	0	1	0	1	1	1
	j	0	0	0	0	0	1	1	0	1	0
C	c	1	0	0	1	0	0	0	0	0	1
	g	0	0	0	1	1	1	0	0	0	0
	i	1	0	0	0	1	0	1	0	0	0

Le tableau initial peut alors être résumé par le tableau croisant les 2 partitions :

	1	2
A	1	0
B	0	1
C	0	0

A, B et C correspondent aux 3 classes de micro-ordinateurs, 1 et 2 aux 2 classes de propriétés. On a associé à chaque case la valeur 0 ou 1 qui était majoritaire. Ce tableau permet de voir que les micro-ordinateurs de la classe A possèdent en général les propriétés 1 mais pas 2, ceux de la classe B les propriétés 2 mais pas 1 et ceux de la classe C aucune propriété. On a résumé ainsi le tableau de 100 valeurs par un tableau de 6 valeurs.

Sur l'exemple précédent, on voit apparaître la notion de tableau **résumé**, ayant la **même structure** que le tableau initial (ici binaire). Nous avons retenu cette caractéristique essentielle pour tous les types de tableaux envisagés. En outre, à chaque fois, nous partons de la définition d'une mesure d'information associée à chaque type de tableau pour définir un **critère numérique** de classification : il s'agit de minimiser la perte d'information obtenue lorsqu'on passe du tableau initial au tableau résumé. Il reste ensuite à trouver le couple de partitions optimisant ce critère. Pour répondre à ces problèmes, une famille d'algorithmes de classification croisée adaptés à la plupart des tableaux de données que l'on peut rencontrer a été développée :

- CROKI2 pour les tableaux de contingence,
- CROMUL pour les tableaux de variables qualitatives,
- CROEUC pour les tableaux de variables quantitatives,
- CROBIN pour les tableaux binaires.

Dans ce travail, après avoir décrit le schéma général de tous les algorithmes de classification croisée, nous étudierons trois des ces quatre algorithmes. Nous donnerons à chaque fois succintement les critères optimisés et les algorithmes, mais nous nous intéresserons surtout à l'utilisation de ces méthodes en décrivant de manière détaillée des exemples d'applications.

2. PRINCIPE GENERAL

I et J étant les deux ensembles sur lesquels est défini le tableau de données , à chaque fois, le problème est de déterminer le couple de partitions de I et J qui minimisera ou maximisera suivant le cas une fonction réelle W appelée critère, mesurant la qualité d'un couple de partitions et dépendant du type de tableau de données.

L'algorithme proposé pour résoudre ce problème définit une suite (P^n,Q^n) de couples de partitions à partir d'un couple initial (P^o,Q^o) en appliquant le mécanisme suivant : on fixe une des partitions et on cherche alors la meilleure partition de l'autre ensemble, puis on fixe cette fois la partition que l'on vient de trouver et l'on cherche la meilleure partition du premier ensemble. On recommence ces 2 étapes jusqu'à la convergence.

On obtient ainsi un algorithme itératif. Pour trouver à chaque étape la meilleure partition, on fera appel à une méthode des Nuées Dynamiques (Diday 1971, Diday et al. 1980, Celeux et al. 1989). Cette méthode étant elle aussi itérative, on aura deux niveaux d'itérations dans les algorithmes proposés . On obtiendra à chaque fois un algorithme qui convergera vers un optimum local.

3. TABLEAUX DE CONTINGENCE

3.1 INTRODUCTION

Il s'agit ici d'effectuer une classification simultanée portant sur des tableaux de contingence définis sur deux ensembles I et J ou, de manière plus générale, sur des tableaux possèdant des propriétés équivalentes. En suivant le principe énoncé dans l'introduction, le tableau résumé associé aux deux partitions est donc aussi un tableau de contingence. La mesure d'information à conserver que nous avons choisie est le χ^2 de contingence. On peut alors poser le problème de manière précise :

Disposant d'un tableau de contingence défini sur I et J, il s'agit de trouver une partition P de I en K classes et une partition Q de J en M classes telles que le χ^2 de contingence du nouveau tableau de contingence construit en regroupant les lignes et les colonnes suivant les partitions P et Q soit maximum.

3.2 NOTATIONS

On notera $T = (n_{ij})$ le tableau de contingence initial défini sur les 2 ensembles I et J de cardinal n et p.

On utilise alors la terminologie habituelle :

- s est la somme des éléments du tableau $\left(\sum_{i \in I} \sum_{j \in J} n_{ij} \right)$,
- F est le tableau des fréquences $\{f_{ij} = \frac{n_{ij}}{s}, i \in I, j \in J\}$,
  (f_{ij} est une estimation de la probabilité qu'un individu présente simultanément la modalité i et la modalité j),
- $f_{i.}$ et $f_{.j}$ sont les fréquences marginales :
$$\forall i \in I \; f_{i.} = \sum_{j \in J} f_{ij} \quad \text{et} \quad \forall j \in J \; f_{.j} = \sum_{i \in I} f_{ij}$$
- $f_I = (f_{1.}, \ldots f_{i.}, \ldots f_{n.})$ et $f_J = (f_{.1}, \ldots f_{.j}, \ldots f_{.p})$ sont les lois marginales définies sur I et J,
- $f_j^i = \frac{f_{ij}}{f_{i.}}$ et $f_i^j = \frac{f_{ij}}{f_{.j}}$ sont les fréquences conditionnelles,
- $f_J^i = (f_1^i, \ldots f_j^i, \ldots f_p^i)$ et $f_I^j = (f_1^j, \ldots f_i^j, \ldots f_n^j)$ sont les lois conditionnelles appelées aussi profils.

On peut alors définir la dépendance entre I et J par le χ^2 de contingence :

$$\chi^2(I,J) = s \sum_{i \in I} \sum_{j \in J} \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}}$$

Cette quantité représente l'écart entre les fréquences théoriques $f_i f_j$ que l'on aurait, s'il y avait indépendance entre les deux ensembles I et J, et les fréquences observées f_{ij}. Le χ^2 sera nul en cas d'indépendance entre I et J et il sera grand lorsque les liens entre I et J seront forts. Ce χ^2 nous servira à mesurer l'**information** apportée par un tableau de contingence.

3.3 χ^2 DE CONTINGENCE ASSOCIE A UN COUPLE DE PARTITIONS

Si $P = (P_1, \ldots P_K)$ est une partition de I en K classes et $Q = (Q_1, \ldots Q_M)$ est une partition de J en M classes, on peut définir un nouveau tableau de contingence en faisant la somme des éléments du tableau de contingence initial correspondant à chaque couple de classes (P_k, Q_m). Ce tableau, noté T(P,Q), est donc défini par :

$$T(k,m) = \sum_{i \in P_k} \sum_{j \in Q_m} n_{ij} \qquad \forall k = 1, \ldots K \text{ et } \forall m = 1, \ldots M$$

Comme pour le tableau initial T(I,J), on peut mesurer l'information apportée par ce nouveau tableau de contingence T(P,Q) à l'aide du χ^2 de contingence qui lui est associé et que l'on notera χ^2 (P,Q). C'est cette quantité que nous choisissons pour mesurer la **qualité** des deux partitions P et Q des ensembles I et J.

3.4 LE PROBLEME D'OPTIMISATION

On peut montrer la relation $\chi^2(I,J) \geq \chi^2(P,Q)$ (1)

Le regroupement des éléments de chaque classe conduit donc à une perte d'information. Chercher à minimiser cette perte d'information revient à chercher les partitions P et Q maximisant le χ^2 de contingence du tableau qui leur est associé, rendant ainsi maximum la dépendance entre la partition P et la partition Q. Le problème que l'on va donc essayer de résoudre est la recherche simultanée de deux partitions P et Q maximisant le χ^2 de contingence du tableau associé. Pour justifier l'intérêt de ce problème, reprenons un exemple de Benzécri (1973) dans lequel le tableau de contingence est défini sur un ensemble I de communes et sur un ensemble J de professions. t_{ij} représente le nombre de personnes pratiquant la profession j dans la commune i :

"Une classification des communes par une partition P sera d'autant meilleure que la connaissance de la classe d'une commune i nous apportera plus d'information sur la part des différentes professions j dans les activités de cette commune. De même une classification Q de l'ensemble des professions qui se pratiquent sera d'autant meilleure qu'elle groupera ensemble les professions qui se pratiquent le plus dans les mêmes communes : ainsi la connaissance de la classe d'une profession j nous fera approximativement connaître la répartition entre les communes de ceux qui pratiquent j".

Signalons que, dans le cas idéal où les profils en ligne et en colonne sont égaux à l'intérieur de chaque classe de P et Q, il n'y a pas de perte d'information.

Remarque sur le nombre de classes :

Le problème que l'on vient de définir n'a de sens que pour des nombres de classes en ligne et en colonne fixés. En effet, si ces deux valeurs K et M sont quelconques, il est clair que le

meilleur couple correspond aux partitions triviales où chaque élément de I et J forment à eux seuls une classe. On a alors $\chi^2(P,Q) = \chi^2(I,J)$.

3.5 RECHERCHE D'UN PARTITION OPTIMISANT LE χ^2

Afin de définir plus précisément notre algorithme, il est nécessaire de rappeler qu'une variante des Nuées Dynamiques peut optimiser le critère du χ^2 :

A partir d'un tableau de contingence défini sur I et J, l'algorithme des centres mobiles appliqué au nuage N(I) des profils f_j^i munis des poids f_i avec comme noyaux les centres de gravité et en prenant la métrique du χ^2 de matrice D_{1/f_J}, permet d'obtenir une partition de I en K classes optimisant le χ^2 de contingence. En effet, le critère optimisé s'écrit :

$$W(P) = \sum_{k=1}^{K} \sum_{i \in P_k} f_i \, d^2(i, G(P_k))$$

où $P = (P_1, \ldots P_K)$, d est la distance D_{1/f_J} et $G(P_k)$ est le centre de gravité de P_k et il aisé de montrer la relation :

$$\chi^2(I,J) = sW(P) + \chi^2(P,J) \qquad (2)$$

sW(P) représente l'information perdue en effectuant les regroupements de la partition P, tandis que $\chi^2(P,J)$ correspond à l'information conservée. En conséquence, puisque la quantité $\chi^2(I,J)$ ne dépend pas de la partition P, la recherche de la partition minimisant le critère W(P) est équivalente à la recherche de la partition P maximisant $\chi^2(P,J)$. La méthode des Nuées Dynamiques décrite précédemment maximise donc le χ^2 de contingence du tableau (P,J). Notons que c'est à partir de cette relation (2) que l'on peut facilement démontrer la relation (1) utilisée dans un paragraphe précédent.

3.6 L'ALGORITHME CROKI2

On construit une suite (P^n,Q^n) à partir d'un couple de départ quelconque (P^0,Q^0) telle que la suite des valeurs du χ^2 associées soit croissante.

Construction de P^{n+1} à partir de (P^n,Q^n)

Soit le tableau de contingence $T(I,Q^n)$ défini par

$$T(i,k) = \sum_{j \in Q_k^n} n_{ij} \quad \text{où} \quad Q^n = (Q_1^n, \ldots Q_L^n)$$

On obtient la partition P^{n+1} en appliquant l'algorithme précédent au tableau $T(I,Q^n)$ en considérant que les objets à classer sont les éléments de I, les variables les classes de Q^n et en prenant comme partition initiale la partition P^n.

Construction de Q^{n+1} à partir de (P^{n+1},Q^n)

Le principe est le même mais on travaille cette fois sur le tableau $T(P^{n+1},J)$ défini par :

$$T(k,j) = \sum_{i \in P_k^{n+1}} n_{ij}$$

Les objets à classer sont cette fois les éléments de J et les variables les K classes de P^{n+1} et on part de la partition Q^n pour obtenir la partition Q^{n+1}.

A partir des propriétés de convergence de la méthode des Nuées Dynamiques, on peut montrer des propriétés équivalentes de convergence pour l'algorithme CROKI2 (Govaert 1983).

3.7 COMPARAISON AVEC LES METHODES HABITUELLES

En effectuant la comparaison entre les résultats de l'algorithme CROKI2 et ceux de l'algorithme des centres mobiles utilisant la distance du χ^2 appliquée séparément sur les deux ensembles sur un certain nombre d'exemples, nous avons toujours obtenu le résultat suivant :

$$\chi^2(P',J) \geq \chi^2(P,J)$$
$$\chi^2(I,Q') \geq \chi^2(I,Q)$$
$$\chi^2(P,Q) \geq \chi^2(P',Q'),$$

où P et Q sont les partitions fournies par l'algorithme CROKI2 et P' et Q' sont les partitions obtenues séparément.

Bien que les partitions P' et Q' soient meilleures séparément que les partitions P et Q, le couple formé par ces dernières est meilleur que le couple formé par les premières. Ce résultat pratique justifie l'utilisation d'un algorithme de classification simultanée.

Il resterait à montrer un résultat équivalent pour les optima globaux. Dans ce cas, bien entendu, on est certain que le couple (P,Q) est le meilleur, par définition, mais on peut se demander si le couple (P',Q'), formé des meilleures partitions sur I et sur J, n'est pas le même couple.

3.8 LE PROGRAMME CROKI2

Il correspond à l'algorithme CROKI2 que l'on vient de décrire, auquel on a rajouté un certain nombre d'éléments concernant l'utilisation des différents tirages, le problème des classes vides et l'édition d'indicateurs succeptibles d'aider l'utilisateur pour étudier les résultats obtenus.

Entrées du programme

- tableau de contingence
- nombre de classes demandées en ligne et en colonne
- nombre de tirages de départ
- divers options d'impression de résultats

Sorties du programme

- valeurs des critères obtenus pour chaque tirage
- description du meilleur résultat :
  - les deux partitions
  - tableaux complémentaires permettant d'analyser les résultats.

La liste complète de ces indicateurs sera fournie plus loin.

3.8.1 Tirages de départ

Comme pour toutes les méthodes convergeant vers un optimum local, les résultats obtenus dépendent des partitions initiales. L'algorithme est donc appliqué plusieurs fois à partir de partitions initiales tirées au hasard. A partir de là, plusieurs possibilités sont envisageables, comme par exemple, la construction des formes fortes qui consistent à rechercher les groupements stables pour tous les tirages. Dans le programme CROKI2 comme dans les suivants, nous avons préféré nous en tenir à l'objectif initial, c'est-à-dire optimiser le critère, aussi après avoir effectué plusieurs tirages, le programme ne retient que le meilleur couple de partitions, c'est-à-dire le couple qui fournit la meilleure valeur du critère.

Dans le but d'obtenir des valeurs du critère comparables, l'utilisateur peut refuser d'obtenir des classes vides. Le programme arrête alors le tirage dès qu'une classe devient vide et recommence un autre tirage. Remarquons qu'en pratique, les classes vides n'apparaissent qu'à la première itération. L'option retenue n'est donc pas trop coûteuse. Il arrive que le nombre de classes demandées soit tel que l'on obtienne toujours des classes vides. L'option retenue précédemment rend alors impossible l'obtention de résultats. Un contrôle est effectué et signale cette information à l'utilisateur. Il peut alors relancer le programme en autorisant cette fois les classes vides.

3.8.2 Les sorties

En dehors de la liste des classes du meilleur couple de partitions obtenues un certain nombre de valeurs sont fournies :

- la valeur du χ^2 du tableau initial
- la valeur du χ^2 du tableau (P,Q)
- le pourcentage d'inertie (ou d'information) conservée, c'est-à-dire $100 \frac{\chi^2(P,Q)}{\chi^2(I,J)}$
- le tableau de contingence initial réordonné en ligne et en colonne suivant les classes de 2 partitions. (option)
- le tableau des $\frac{f_{ij}}{f_i . f_j}$ réordonné suivant les deux partitions permet d'analyser plus finement l'homogénéité des classes. En effet le problème que l'on cherche à résoudre revient à obtenir à l'intérieur des classes (k,l) des valeurs $\frac{f_{ij}}{f_i . f_j}$ voisines . (option)
- le tableau de contingence (P,Q) qui permet d'étudier les liens entre les 2 partitions.
- le tableau des $\frac{f_{kl}}{f_k . f_l}$ donne pour chaque couple (k,l) la moyenne des valeurs f_{ij}, la pondération du couple (i,j) étant $f_i f_j$

- les tableaux (P,J) et (I,Q) ainsi que les tableaux des profils correspondants $\frac{f_{kj}}{f_k f_j}$ et $\frac{f_{il}}{f_i f_l}$. Ces tableaux permettent d'analyser séparément les 2 partitions P et Q. (option)

3.8.3 Indices d'interprétation d'un couple de partitions

Pour définir des indices spécifiques à la classification croisée, il est intéressant de considérer que l'ensemble des valeurs $\frac{f_{ij}}{f_i f_j}$ munies des poids $f_i f_j$ forme un nuage **N**(IxJ) de **R**. Le centre de gravité de ce nuage est alors la valeur 1. Si on munit de plus **R** de la métrique euclidienne identité, le χ^2 du tableau initial représente l'inertie de ce nuage. A chaque couple de classes (P_k,Q_m) on peut associer la classe de **N**(IxJ), notée (k,m) formée par les valeurs f_{ij} obtenues lorsque i et j décrivent les classe P_k et Q_m. On peut alors définir $T_{k,m}$, inertie de la classe (k,m) par rapport au centre de gravité de **N**(IxJ), soit :

$$T_{km} = \sum_{i \in P_k, j \in Q_m} f_i f_j \left(\frac{f_{ij}}{f_i f_j} - 1\right)^2$$

T_{kl} représente donc la part apportée par les couples (i,j) de la classe (k,m) à la valeur du χ^2 du tableau initial. On définit l'inertie conservée par le centre de gravité de la classe (k,m) :

$$B_{km} = f_k f_m \left(\frac{f_{km}}{f_k f_m} - 1\right)^2$$

Si W_{km} est l'inertie de la classe (k,m), on a par le théorème de Huygens $T_{km} = B_{km} + W_{km}$, en outre :

$$\chi^2(I,J) = T = \sum_{k,m} T_{km} \quad \text{et} \quad \chi^2(P,Q) = B = \sum_{k,m} B_{km}$$

Ces relations permettent de définir :

- le tableau des valeurs $\frac{B_{km}}{B}$ représentant la part d'inertie conservée par la classe sur l'inertie totale. Ces quantités indiquent l'importance de la classe.

- le tableau des valeurs $\frac{B_{km}}{T_{km}}$ représentant la part d'inertie conservée par la classe par rapport à l'inertie initiale des points de la classe. Cette quantité indique la qualité de représentation d'une classe. La valeur obtenue comprise entre 0 et 1 sera d'autant plus grande que la classe sera homogène.

3.9 EXEMPLES D'APPLICATION

3.9.1 Les données

Nous présentons ici pour illustrer la méthode les résultats obtenus sur un petit jeu de données qui portent sur la comparaison de budgets-temps (Jambu 1976). On dispose du tableau $T = \{n_{ij} / i \in I, j \in J\}$ où n_{ij} représente le nombre d'heures passées à exercer l'activité j

par la population i durant une période déterminée. L'ensemble I est constitué 28 types de population caracterisés par le sexe, le pays, l'activité professionnelle et le mariage. Dans les identificateurs des lignes, la signification des lettres est la suivante :

- H : homme, F : femme
- A : actif, NA : non actif, M : marié, C : célibataire
- US ou U : USA , WE ou W : pays de l'ouest, ES ou E : pays de l'est, YO ou Y : Yougoslavie.

L'ensemble J est constitué de 10 classes d'activités :

- ***1 Travail professionnel
- ***2 Occupations dues ou liées au travail professionnel
- ***3 Travail ménager
- ***4 Occupation liées au travail des enfants
- ***5 Courses ou empletes ménagères
- ***6 Toilette, soins personnels
- ***7 Repas
- ***8 Sommeil
- ***9 Télévision
- **10 Autres loisirs

Le tableau de données est le suivant :

HAUS	610	140	60	10	120	95	115	760	175	315
FAUS	475	90	250	30	140	120	100	775	115	305
FNAU	10	0	495	110	170	110	130	785	160	430
HMUS	615	141	65	10	115	90	115	765	180	305
FMUS	179	29	421	87	161	112	119	776	143	373
HCUS	585	115	50	0	150	105	100	760	150	385
FCUS	482	94	196	18	141	130	96	775	132	336
HAWE	652	100	95	7	57	85	150	807	115	330
FAWE	510	70	307	30	80	95	142	815	87	262
FNAW	20	7	567	87	112	90	180	842	125	367
HMWE	655	97	97	10	52	85	152	807	122	320
FMWE	168	22	529	69	102	83	174	825	119	392
HCWE	642	105	72	0	62	77	140	812	100	387
FCWE	389	34	262	14	92	97	147	848	84	392
HAYO	650	140	120	15	85	90	105	760	70	365
FAYO	560	105	375	45	90	90	95	745	60	235
FNAY	10	10	710	55	145	85	130	815	60	380
HMYO	650	145	112	15	85	90	105	760	80	357
FMYO	260	52	576	59	116	85	117	775	65	295
HCYO	615	125	95	0	115	90	85	760	40	475
FCYO	413	89	318	23	112	96	102	774	45	409
HAES	650	142	122	22	76	94	100	764	96	334
FAES	578	106	338	42	106	94	52	752	64	228
FNAE	24	8	594	72	158	92	128	840	86	398
HMES	652	133	134	22	68	94	102	762	122	310
FMES	434	77	431	60	117	88	105	770	73	229
HCES	627	148	68	0	88	92	86	770	58	463
FCES	433	86	296	21	128	102	94	758	58	379

3.9.2 Sorties

L'ordre dans lequel nous décrivons les résultats ne correspond pas nécessairement à l'ordre dans lequel ils sont imprimés (qui est essentiellement choisi pour des raisons d'efficacité de programmation) mais un ordre plus logique pour analyser les résultats. Dans cet exemple, nous avons demandé 5 classes en ligne, 3 en colonne et 10 tirages au hasard.

```
VAIEUR DU KI2 DU IABLEAU INIIIAI   9658
```

Qualité du résultat et description des partitions obtenues

```
VALEUR DU KI2 OBIENU   8048

POURCENIAGE DU KI2 CONSERVE          83.33
```

```
        PARTITION EN LIGNES

CLASSE  1 :      3 ELEMENTS: FMUS FMWE FMYO
CLASSE  2 :      3 ELEMENTS: FAYO FAES FMES
CLASSE  3 :      6 ELEMENTS: FAUS FCUS FAWE FCWE FCYO FCES
CLASSE  4 :      4 ELEMENTS: FNAU FNAW FNAY FNAE
CLASSE  5 :     12 ELEMENTS: HAUS HMUS HCUS HAWE HMWE HCWE HAYO HMYO HCYO HAES HMES HCES

        PARTITION EN COLONNES

CLASSE  1 :      2 ELEMENTS: ***3 ***4
CLASSE  2 :      2 ELEMENTS: ***1 ***2
CLASSE  3 :      6 ELEMENTS: ***5 ***6 ***7 ***8 ***9 **10
```

Le pourcentage d'inertie expliquée par le couple de partitions obtenu est sur ce petit exemple très bon puisque 83% de l'information a été conservée.

Dans toute la suite, un certain nombre de tableaux de contingence peuvent être imprimés suivant les options. En pratique, plutôt que d'étudier directement les résultats sur le tableau de contingence n_{ij}, il est préférable de le faire sur le tableau des valeurs normées correspondantes $\frac{f_{ij}}{f_i f_j}$. En effet, cette fois le tableau est plus "lisible". L'objectif du programme étant de regrouper les valeurs $\frac{f_{ij}}{f_i f_j}$ voisines, les valeurs de ces tableaux normés sont maintenant comparables. En réalité, nous avons imprimé $1000 \frac{f_{ij}}{f_i f_j}$ pour des raisons de clarté. L'apport de chacune de ces valeurs au χ^2, c'est-à-dire à l'information sera d'autant plus grand que ces valeurs seront éloignées de 1. Par conséquent, sur nos tableaux, il faudra comparer ces valeurs à la valeur centrale 1000. Les tableaux de contingence non normés sont aussi imprimés, mais, étant moins intéressants, nous ne les avons pas repris dans ce texte.

Tableau résumé associé aux 2 partitions

Il s'agit du tableau de contingence obtenu en regroupant les lignes et les colonnes suivant les 2 partitions P et Q. Rappelons que l'objectif du programme est de maximiser l'information conservée par ce tableau, c'est-à-dire son χ^2. Ce tableau permet de caractériser mutuellement les deux partitions obtenues.

TABLEAU DES (FKM/FK*FM)*1000

	1	2	3
1	1846	437	1024
2	1395	1168	863
3	953	993	1011
4	2165	41	1096
5	322	1423	989

Les valeurs les plus intéressantes sont les valeurs éloignées de la moyenne 1000. Nous avons mis en évidence ces valeurs en les soulignant. On peut remarquer que les classes 3 en lignes et en colonnes sont toujours moyennes et donc peu caractéristiques. Il faudrait détailler un peu plus ces classes. Par contre, les autres classes en lignes sont très marquées. On peut retenir les associations suivantes :

classes en lignes	classe en colonnes
4	très faible pour 2 et très forte pour 1
1	faible pour 2 et très forte pour 1
5	forte sur 2 et faible sur 1
2	assez forte sur 1

Contribution de chaque classe et couple de classes

Le tableau suivant représente la part apportée par chaque couple de classe au χ^2 de contingence conservé par le tableau (P,Q). En marge, est reportée la contribution des classes en ligne et en colonne.

TABLEAU DES POURCENTAGES D'INERTIE (BKM/B)

	1	2	3	
1	8	6	0	15
2	2	1	1	3
3	0	0	0	0
4	21	24	1	46
5	21	14	0	36
	52	46	2	

Remarquons que les classes correspondant à des valeurs proches de la valeur moyenne 1000 sont des classes de faible inertie et par conséquent interviennent moins dans la valeur du critère obtenu (tableau des $\frac{B_{km}}{B}$) ; c'est le cas des deux classes 3.

Homogénéité des classes

Dans ce tableau, est reporté le pourcentage entre la part apportée par un couple de classe ou une classe dans le χ^2 final et la part apportée par tous les éléments de ces classes dans le χ^2 initial. Si la classe est parfaitement homogène on obtient 100.

POURCENTAGE D'INERTIE CONSERVEE PAR CLASSE (BKM/IKM)

	1	2	3	
1	91.	96.	3.	87.
2	87.	56.	44.	62.
3	6.	0.	1.	2.
4	92.	100.	28.	93.
5	97.	95.	0.	82.
	92.	96.	12.	83.

Tableau initial réordonné suivant les partitions obtenues

Cette sortie optionnelle n'est possible que si la taille du tableau n'est pas trop grande. Elle permet d'étudier les partitions directement sur les données initiales.

TABLEAU DES (FIJ/FIFJ)*1000 REORDONNES

	***3	***4	***1	***2	***5	***6	***7	***8	***9	**10
FMUS	1517	2607	398	336	1479	1179	1018	988	1436	1069
FMWE	1843	1998	361	246	905	844	1439	1015	1155	1085
FMYO	2076	1768	579	603	1065	894	1001	986	652	845
FAYO	1352	1348	1247	1218	826	947	813	948	602	673
FAES	1239	1280	1309	1250	990	1006	452	973	653	664
FMES	1564	1810	973	899	1082	932	904	987	738	660
FAUS	901	899	1058	1044	1286	1263	856	986	1155	874
FCUS	706	539	1074	1090	1295	1368	821	986	1325	963
FAWE	1107	899	1137	812	735	1000	1216	1038	874	751
FCWE	961	426	881	401	860	1038	1280	1098	858	1143
FCYO	1155	694	927	1040	1037	1018	880	993	455	1181
FCES	1087	641	983	1016	1198	1094	820	983	593	1107
FNAU	1784	3296	22	0	1562	1158	1112	999	1606	1232
FNAW	2046	2610	44	81	1030	948	1542	1073	1257	1053
FNAY	2559	1648	22	116	1332	894	1112	1037	602	1089
FNAE	2141	2157	53	92	1451	968	1095	1069	863	1140
HAUS	216	299	1359	1624	1102	1000	984	967	1757	902
HMUS	234	299	1369	1635	1056	947	984	973	1807	873
HCUS	180	0	1303	1334	1378	1105	856	967	1506	1103
HAWE	342	209	1454	1161	524	895	1285	1028	1155	946
HMWE	350	300	1461	1126	478	895	1302	1028	1226	918
HCWE	259	0	1432	1219	570	811	1200	1035	1005	1110
HAYO	432	449	1448	1624	781	947	898	967	703	1046
HMYO	403	449	1449	1683	781	947	899	968	803	1023
HCYO	342	0	1370	1450	1056	947	727	967	401	1361
HAES	439	659	1448	1647	698	989	856	972	964	957
HMES	483	659	1453	1543	625	990	873	970	1225	888
HCES	245	0	1397	1717	808	968	736	980	582	1327

On peut constater que les valeurs sont assez homogènes à l'intérieur de chaque couple de classes et proches de la valeur correspondante du tableau résumé qui en est d'ailleurs le centre de gravité.

Si ce tableau réordonné T(I,J) ne peut être sorti, des sortie intermédiaires entre ce tableau complet et le tableau résumé T(P,Q) peuvent être obtenues : il s'agit des tableaux T(P,J) et T(I,Q). Le premier permet d'analyser plus finement le comportement des classes de la partition P sur toutes les colonnes initiales alors que la seconde permet d'analyser de manière symétrique la partition Q. Nous donnons ici ces deux tableaux, bien que dans notre notre petit exemple, ils n'apportent pas grand chose puisque nous avons pu sortir le tableau initial réordonné.

Description de la partition des lignes suivant les colonnes

Si on note (P,J) le tableau de contingence obtenu en regroupant les lignes suivant la partition P, le programme imprime le tableau (P,J) et éventuellement le tableau normé correspondant en ordonnant les colonnes suivant la partition Q. Cette sortie permet d'analyser la partition des colonnes après avoir effectué le regroupement des lignes. Le nombre de colonnes étant en général beaucoup plus grand que le nombre de classes en ligne, ce sont les tableaux transposés qui sont sortis. Cette sortie n'est envisageable que si le nombre de ligne n'est pas trop grand.

TABLEAU DES (FKJ/FK*FJ)*1000 SORII SOUS FORME TRANSPOSEE

	1	2	3	4	5
***3	1813	1385	986	2133	327
***4	2123	1480	684	2428	277
***1	445	1176	1011	35	1412
***2	393	1122	902	72	1480
***5	1147	966	1069	1344	821
***6	971	961	1131	992	953
***7	1156	724	978	1216	967
***8	997	969	1014	1045	985
***9	1082	664	878	1082	1095
**10	1001	666	1002	1128	1038

Description de la partition des colonnes suivant les lignes

Il s'agit d'une sortie analogue à la précédente. Il suffit d'échanger le rôle des lignes et des colonnes.

TABLEAU DES (FIM/FI*FM)*1000

	1	2	3
FMUS	1634	388	1083
FMWE	1860	343	1054
FMYO	2043	583	934
FAYO	1351	1243	846
FAES	1243	1300	847
FMES	1590	961	895
FAUS	901	1056	1000
FCUS	688	1076	1035
FAWE	1085	1085	953
FCWE	903	804	1086
FCYO	1106	945	997
FCES	1039	988	995
FNAU	1947	18	1148
FNAW	2107	50	1105
FNAY	2462	37	1039
FNAE	2143	59	1094
HAUS	225	1402	1016
HMUS	241	1412	1009
HCUS	160	1308	1061
HAWE	328	1406	994
HMWE	344	1407	990
HCWE	232	1398	1016
HAYO	434	1476	948
HMYO	408	1486	950
HCYO	305	1383	1006
HAES	463	1480	941
HMES	502	1468	938
HCES	218	1448	1001

3.10 LIENS AVEC L'ANALYSE DES CORRESPONDANCES

La classification croisée d'un tableau de contingence s'applique aux mêmes types de données que l'analyse des correspondances et utilise le même critère : le χ^2 de contingence. On peut aller plus loin dans cette comparaison et montrer que les deux problèmes sont proches : la classification croisée de tableau de contingence peut être considérée comme une analyse des correspondances **sous contraintes** (Govaert 1983).

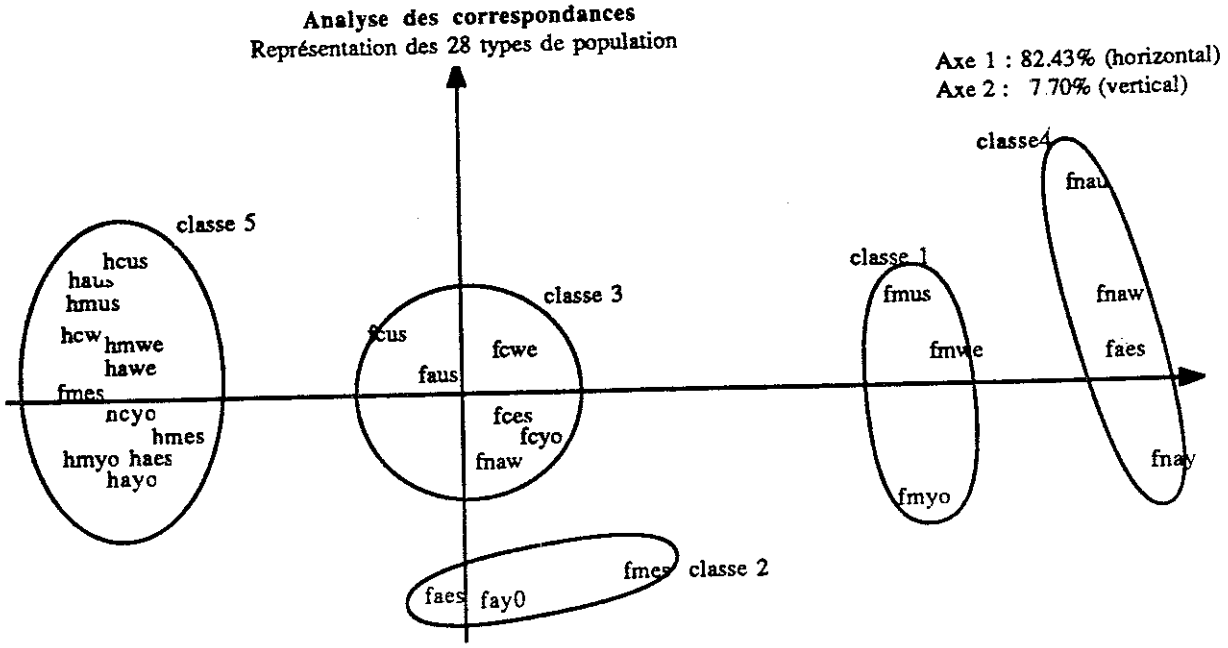
Pour illustrer ce lien, nous avons appliqué l'analyse des correspondances à l'exemple précédent. Le pourcentage d'inertie expliquée par le premier plan factoriel s'élève à 90%. Nous n'avons pas utilisé la représentation simultanée afin de pouvoir reporter de manière claire les classes obtenues par CROKI2. Sur cet exemple très simple, on retrouve sans difficulté les conclusions que nous avions obtenues à partir de la classification croisée:

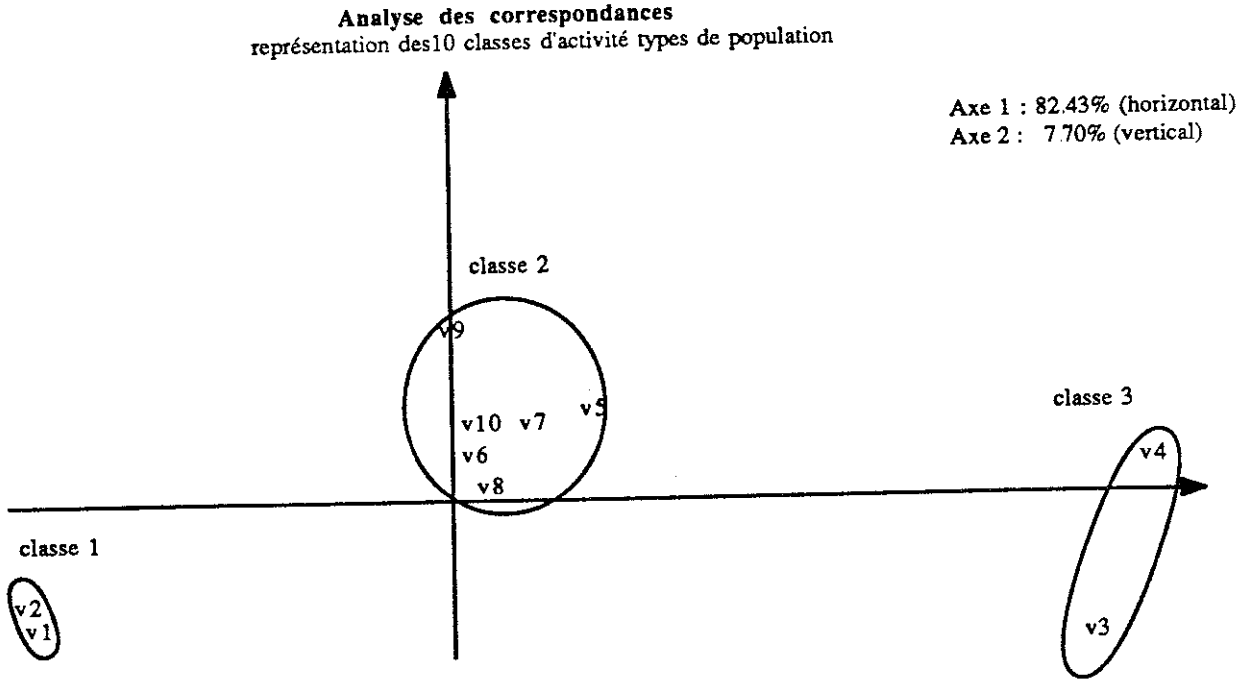
Lignes:

classes 4 et 5 très opposées
classe 1 ayant la même tendance que la classe 4 mais un peu moins forte.
classe 3 moyenne
classe 2 située un peu à part.

Colonnes:

classe 1 et 2 très opposées
classe 3 moyenne


Analyse des correspondances
Représentation des 28 types de population
Axe 1 : 82.43% (horizontal)
Axe 2 : 7.70% (vertical)
classe 5
hcus
hmus
hcw
hmwe
hawe
fmes
ncyo
hmes
hmyo haes
hayo
classe 3
fcus
faus
fcwe
fces
fcyo
fnaw
classe 1
fmus
fmwe
fmyo
classe4
fnau
fnaw
faes
fnay
fmes
classe 2
faes
fay0


Analyse des correspondances
représentation des10 classes d'activité types de population
Axe 1 : 82.43% (horizontal)
Axe 2 : 7.70% (vertical)
classe 2
v9
v10
v7
v5
v6
v8
classe 3
v4
v3
classe 1
v2
v1

4. TABLEAUX DE VARIABLES QUALITATIVES

4.1 L'ALGORITHME CROMUL

Une version modifiée du programme CROKI2, appelée CROMUL, destinée à l'analyse de tableaux de variables qualitatives ou questionnaires a été mise au point. Comme pour l'analyse des correspondances multiples (Lebart, Morineau, Tabard, 1977), on peut considérer le questionnaire, après l'avoir transformé en un tableau disjonctif complet, comme un tableau de contingence. On peut alors lui appliquer la méthode CROKI2 qui permet la recherche simultanée d'une partition des individus et d'une partition des modalités. L'inconvénient de cette approche est la nécessité de disposer du tableau disjonctif complet dont la taille peut vite devenir trop grande. Pour cette raison, la version CROMUL a été écrite. Cette version nécessite simplement en entrée le tableau individus x variables qualitatives. Les tableaux de contingence intermédiaires sont directement recalculés à partir de celui-ci. Les caractéristiques (entrées et sorties) du programme CROMUL sont les même que celles du programme CROKI2.

Dans cette approche, on se place exactement dans les mêmes nuages et avec les mêmes métriques que dans l'analyse des correspondances multiples, mais cette fois, l'objectif n'est pas de rechercher des axes mais des classes sur chacun des deux ensembles. Ce parallèle peut d'ailleurs être poursuivi en considérant que la recherche d'une partition est équivalent à la recherche d'axes orthonormés sous contrainte (Govaert 1983). Nous étudions maintenant un exemple d'application de cette méthode.

4.2 UN EXEMPLE D'APPLICATION

4.2.1 Les données

fourrure	poil : poils ras	
	poi2 : poils longs	
comportement prédateur	com1 : diurne	
	com2 : diurne et nocturne	
	com3 : nocturne	
forme des oreilles	ore1 : rondes et arrondies	
	ore2 : en pointe	
os hyoïde	lar1 : présence	
	lar2 : absence	
taille du garrot	tai1 : < 50 cm	
	tai2 : entre 50 cm et 70 cm	
	tai3 : taille > 70 cm	
poids de l'animal	pod1 : < 10 kg	
	pod2 : entre 10 kg et 80 kg	
	pod3 : > 80 kg	
longueur du corps	lon1 : < 80 cm	
	lon2 : entre 80 cm et 150 cm	
	lon3 : > 150 cm	
longueur de la queue	que1 : petite	
	que2 : moyenne	
	que3 : longue	
canines développées	den1 : très développées	
	den2 : peu développées	
type de proie	pro1 : grosse	
	pro2 : grosse ou petite	
	pro3 : petite	
monte ou non aux arbres	arb1 : monte aux arbres	
	arb2 : ne monte pas	
chasse à courre	cha1 : oui	
	cha2 : non	

1	1.	1.	1.	2.	3.	3.	3.	2.	2.	1.	1.	2.
2	1.	3.	1.	2.	3.	3.	3.	2.	2.	1.	1.	1.
3	1.	3.	1.	2.	3.	3.	2.	1.	2.	1.	2.	1.
4	1.	3.	1.	2.	3.	3.	2.	2.	2.	2.	2.	1.
5	2.	1.	1.	2.	2.	2.	2.	3.	2.	2.	2.	1.
6	1.	1.	1.	1.	3.	2.	2.	3.	1.	2.	1.	2.
7	1.	2.	1.	1.	2.	3.	2.	3.	2.	2.	2.	1.
8	1.	3.	1.	2.	2.	2.	2.	3.	2.	3.	2.	1.
9	1.	1.	2.	1.	2.	2.	2.	1.	1.	3.	2.	2.
10	1.	2.	1.	1.	2.	2.	2.	2.	1.	3.	2.	1.
11	2.	2.	2.	1.	2.	2.	2.	1.	2.	2.	2.	1.
12	1.	2.	2.	1.	2.	2.	1.	1.	1.	3.	2.	2.
13	1.	2.	1.	1.	1.	1.	2.	2.	1.	3.	1.	1.
14	1.	2.	1.	1.	1.	2.	2.	3.	1.	3.	2.	1.
15	2.	3.	2.	1.	1.	2.	1.	2.	1.	3.	2.	1.
16	1.	3.	1.	1.	1.	1.	1.	1.	1.	3.	2.	1.
17	1.	3.	1.	1.	1.	1.	1.	2.	1.	3.	2.	1.
18	2.	2.	1.	1.	1.	1.	1.	2.	1.	3.	1.	1.
19	1.	3.	1.	1.	1.	1.	1.	2.	1.	3.	2.	2.
20	1.	2.	2.	1.	1.	1.	1.	1.	1.	3.	2.	1.
21	1.	3.	1.	1.	1.	1.	1.	2.	1.	3.	2.	1.
22	1.	2.	1.	1.	1.	1.	1.	2.	1.	3.	2.	1.
23	2.	3.	1.	1.	1.	1.	1.	1.	1.	3.	2.	1.
24	1.	3.	1.	1.	1.	1.	1.	2.	1.	3.	2.	1.
25	1.	2.	1.	1.	1.	1.	1.	1.	1.	3.	2.	2.
26	2.	3.	1.	1.	1.	1.	1.	1.	1.	3.	2.	1.
27	1.	3.	1.	1.	1.	1.	1.	3.	1.	3.	2.	1.
28	1.	3.	1.	1.	1.	1.	1.	2.	1.	3.	2.	1.
29	1.	3.	1.	1.	1.	1.	1.	2.	1.	3.	2.	1.
30	2.	3.	1.	1.	1.	1.	2.	2.	1.	2.	2.	1.

Comme pour le programme CROKI2, nous avons pris un exemple limité en taille afin de pouvoir détailler les résultats. Il s'agit d'un tableau portant sur 30 félins mesurés par les 12 variables qualitatives.

4.2.2 Les Sorties

```
VALEUR DU KI2 DU TABLEAU INITIAL    540
```

Tout les résultats qui vont suivre correspondent uniquement au meilleur tirage.

Qualité du résultat et description des partitions obtenues

```
VALEUR DU KI2 OBTENU    229

POURCENTAGE DU KI2 CONSERVE          42.57

PARTITION DES INDIVIDUS

CLASSE  1 :     2 ELEMENTS:  1   2
CLASSE  2 :     5 ELEMENTS:  3   4   5   7   8
CLASSE  3 :     1 ELEMENTS:  6
CLASSE  4 :     2 ELEMENTS: 14  30
CLASSE  5 :     6 ELEMENTS:  9  10  11  12  20  25
CLASSE  6 :    14 ELEMENTS: 13  15  16  17  18  19  21  22  23  24  26  27  28  29

PARTITION DES MODALITES

CLASSE  1 :     6 ELEMENTS: lar1 tai1 pod1 lon1 den1 pro3
CLASSE  2 :     3 ELEMENTS: com1 tai3 arb1
CLASSE  3 :     3 ELEMENTS: lon2 que3 pro2
CLASSE  4 :     7 ELEMENTS: poi1 poi2 com3 ore1 que2 arb2 cha1
CLASSE  5 :     6 ELEMENTS: com2 ore2 tai2 pod2 que1 cha2
CLASSE  6 :     5 ELEMENTS: lar2 pod3 lon3 den2 pro1
```

Tableau de contingence associé aux 2 partitions

Il s'agit du tableau de contingence obtenu en regroupant les individus et les modalités suivant les 2 partitions P et Q. Rappelons que l'objectif du programme est que ce tableau conserve le maximum d'information (au sens du χ^2) du tableau disjonctif complet initial. Ce tableau permet d'étudier les liens entre les 2 partitions. Nous ne donnons que le tableau normé correspondant :

TABLEAU DES (FKM/FK*FM)*1000

	1	2	3	4	5	6
1	0	5357	0	888	326	6250
2	102	1285	2749	1066	913	3250
3	512	6428	3749	444	1304	0
4	1025	0	2499	1111	652	0
5	982	357	833	666	2717	208
6	1501	306	178	1158	419	0

Nous avons souligné les valeurs les plus intéressantes de ce tableau. Elles mettent en évidence les relations qui existent entre les deux partitions : en tenant compte de la taille des classes, on voit par exemple apparaître une classe de félins (classe 6) importante caractérisée par la classe 1 de variables, il s'agit des petits félins.

Contribution de chaque classe et couple de classes

Le tableau suivant représente la part apportée par chaque couple de classe au χ^2 de contingence du tableau (P,Q). En marge, est reportée la contribution des classes en ligne et en colonne.

TABLEAU DES POURCENTAGES D'INERTIE (BKM/B)

	1	2	3	4	5	6	
1	3.	8.	1.	0.	1.	19.	32.
2	7.	0.	5.	0.	0.	9.	21.
3	0.	6.	3.	1.	0.	0.	10.
4	0.	0.	2.	0.	0.	1.	3.
5	0.	1.	0.	1.	12.	1.	15.
6	6.	1.	3.	1.	3.	5.	19.
	17.	16.	14.	3.	16.	35.	

Homogénéité des classes

Dans ce tableau, est reporté le pourcentage entre la part apportée par un couple de classe ou une classe dans le χ^2 final et la part apportée par tous les éléments de ces classes dans le χ^2 initial. Si la classe est parfaitement homogène on obtient 100.

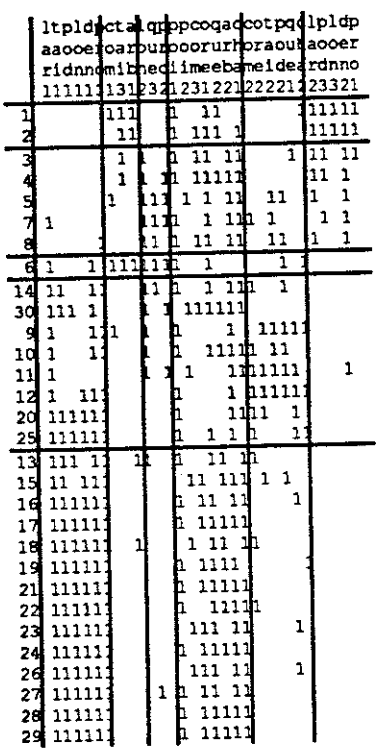
POURCENTAGE D'INERTIE CONSERVEE PAR CLASSE (BKM/IKM)

	1	2	3	4	5	6	
1	100.	79.	100.	2.	23.	72.	70.
2	86.	1.	50.	1.	0.	45.	37.
3	37.	98.	83.	46.	2.	100.	69.
4	0.	100.	42.	2.	7.	100.	26.
5	0.	14.	1.	17.	45.	46.	30.
6	71.	22.	51.	5.	20.	100.	34.
	61.	49.	47.	8.	27.	64.	43.

Tableau disjonctif réordonné suivant les partitions obtenues

Cette sortie, optionnelle, permet d'étudier les partitions directement sur les données initiales. Cette sortie n'est envisageable que si le nombre de lignes et le nombre de colonnes ne sont pas trop importants.

TABLEAU DISJONCTIF COMPLET REORDONNE

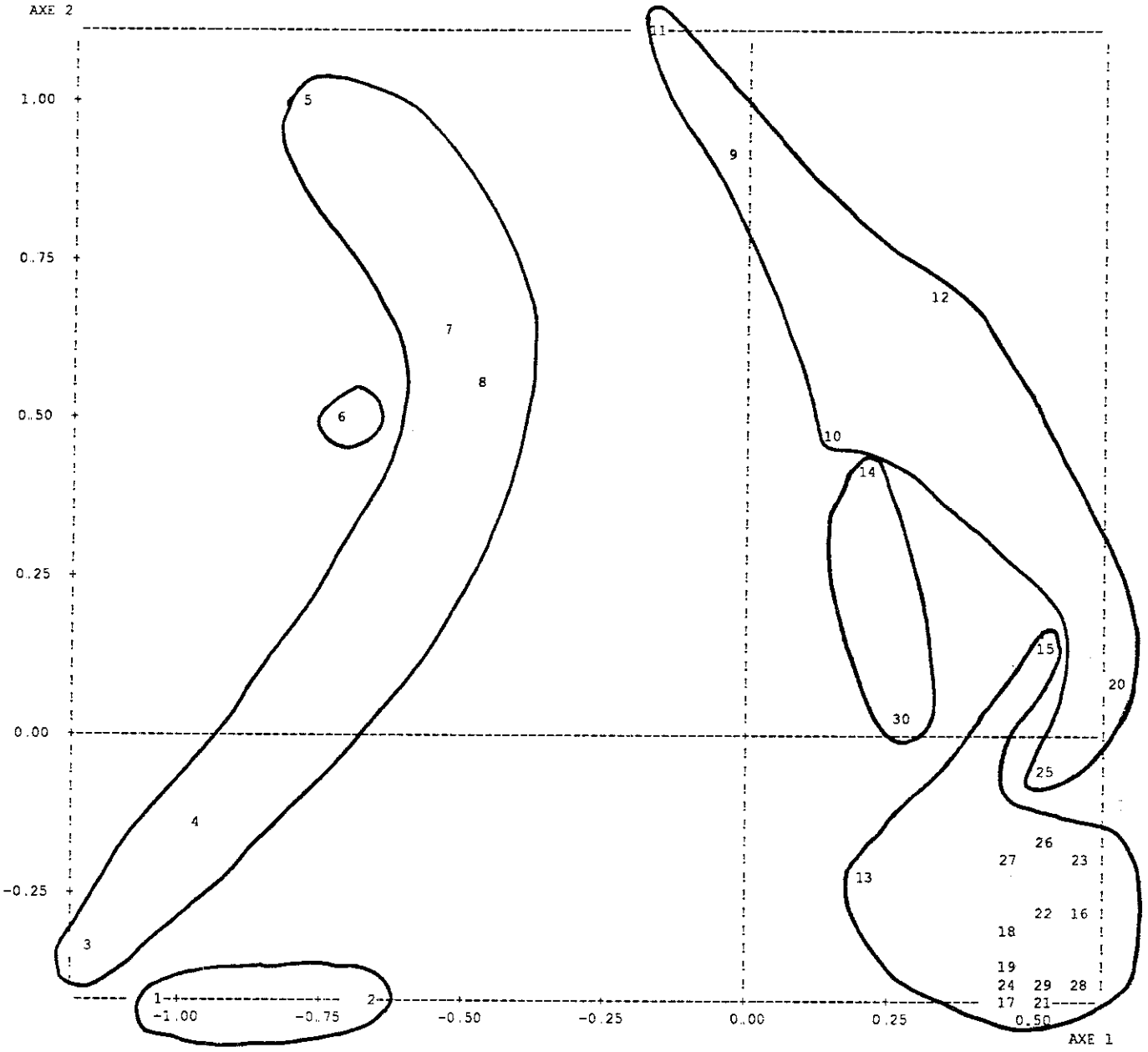


Sur ce tableau, on retrouve par exemple le lien important entre la classe 6 des félins et la classes 1 des modalités.

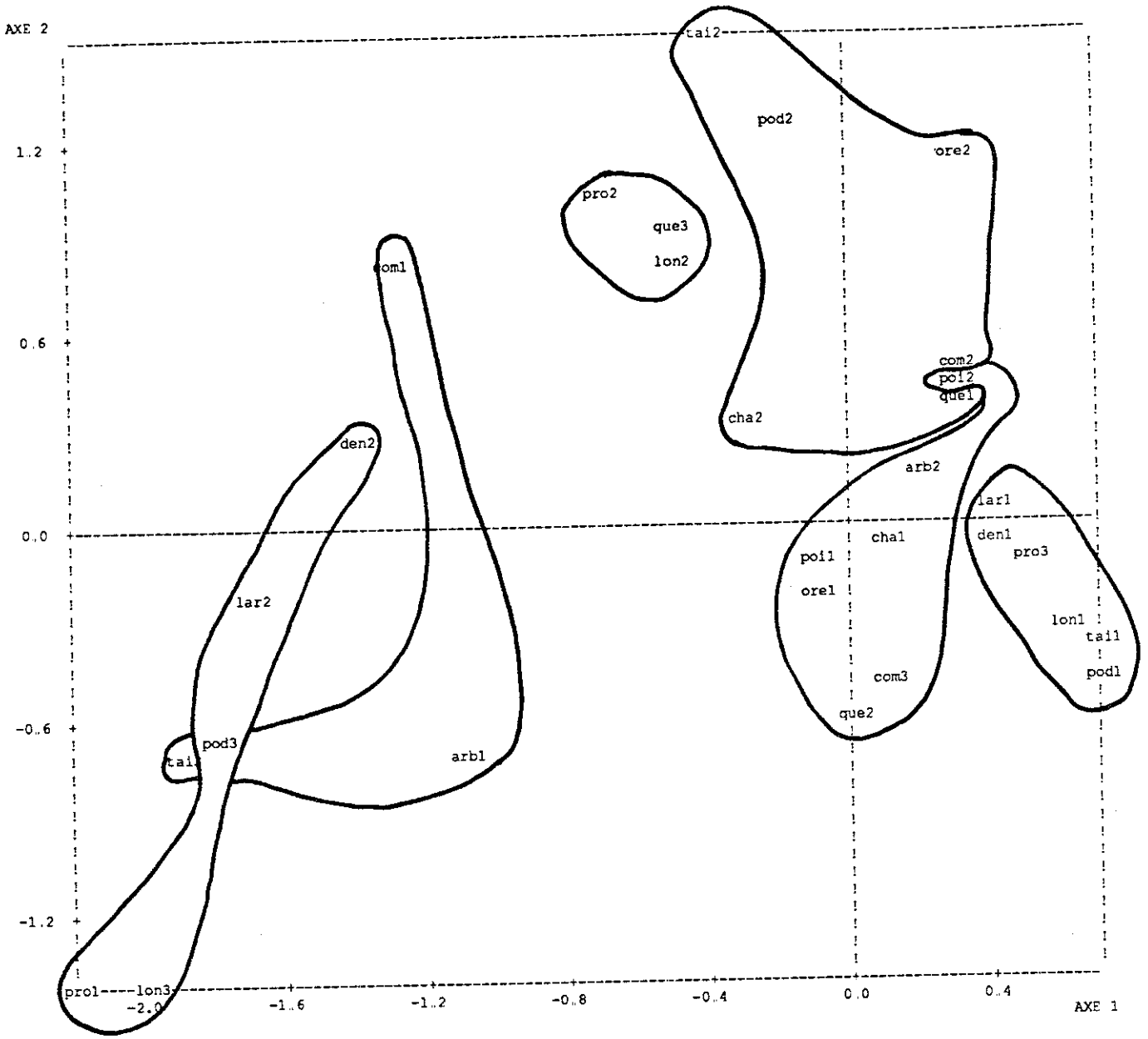
Comme pour l'algorithme CROKI2, il est possible, si le tableau précédent ne peut être sorti en raison de sa taille, d'étudier séparément une des deux partitions P et Q grâce aux tableaux de contingence T(I,Q) et T(P,J). Le premier, par exemple, permet d'analyser la partition des modalités après avoir effectué le regroupement des individus. Nous ne décrirons pas ici ces sortie optionnelles.

4.2.3 Comparaison avec l'analyse des correspondances multiples

Comme pour l'algorithme CROKI2, les résultats obtenus avec CROMUL sont à étudier en liaison avec l'analyse des correspondances multiples puisque les deux approches travaillent sur les mêmes nuages de points avec les mêmes métriques.
Nous donnons ici la représentation obtenues dans le premier plan factoriel.



Représentation factorielle des félins



Représentation factorielle des modalités

5. TABLEAUX BINAIRES

5.1 INTRODUCTION

On se place cette fois exactement dans les conditions du petit exemple décrit dans l'introduction. Le tableau résumé est cette fois un tableau binaire (à chaque couple de classes est donc associée une valeur "idéale") et le critère va mesurer l'écart entre ce tableau idéal et les données initiales. Il est facile de voir que ceci revient à essayer d'obtenir, en réordonnant les lignes et les colonnes du tableau initial suivant les deux partitions, des blocs homogènes de 1 ou de 0.

5.2 LE PROBLEME

Pour justifier et préciser la recherche de partitions simultanées d'un tableau binaire, on peut reprendre les propositions faites par Lerman à propos de la notion de classe polythétique (Lerman 1981). Il rappelle d'abord la notion de classe polythétique selon Becker : "une classe polythétique G d'une classification naturelle se réfère à un sous-ensemble B d'attributs tel que :

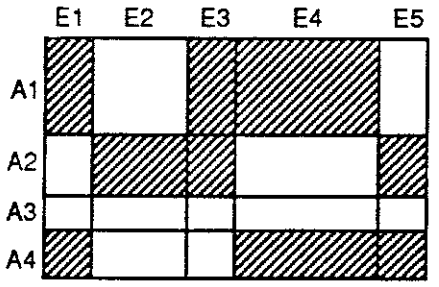
a) Chaque élément de la classe possède une proportion importante d'attributs de B.

b) Chaque attribut de B est présent dans une proportion importante

c) Il n'y a pas nécessairement un attribut qui soit possédé par tous les éléments de G".

Lerman généralise alors cette notion : *"la situation plus générale d'une bonne classification sur E avec une bonne classification sur A conformément à la définition approximative de Becker est celle où à chacune des classes E_i de la partition $(E_1 \ldots E_k)$ correspond une réunion B de classes de la partition $(A_1 \ldots A_k)$ sur A, à laquelle la classe E_i se réfère. Réciproquement à toute classe A_j correspond une réunion G de classes E_i à laquelle la classe A_j se réfère"*.

La situation ainsi décrite peut être représentée par le tableau binaire réordonné suivant les partitions. On suppose que les zones hachurées correspondent à des zones de forte densité de 1 et que les zones non hachurées correspondent à des zones de forte densité de 0.



Représentation simultanée des partitions

L'objectif étant de trouver des blocs homogènes, c'est-à-dire remplis soit de 1, soit de 0, on associe à chaque couple (k,m) de classes une valeur binaire idéale (1 ou 0). On obtient ainsi un tableau binaire que l'on appelle noyau. On cherche alors à minimiser le nombre de fois où la

valeur associée à un couple (i,j) est différente de la valeur idéale associée au couple de classes auquel appartient (i,j). Cette quantité représente l'écart entre le tableau initial et le tableau idéal.

Exemple : soit le tableau binaire

		J							
		1	1	0	1	1	1	0	0
		0	0	1	0	0	1	1	0	0
I		1	1	1	1	1	0	1	0	1
		0	0	0	1	1	0	1	1	1
		0	1	0	1	1	1	1	1	1

Soit le couple de partitions ({1,2,3},{4,5}) pour I et ({1,2,3},{4,5,6,7}{8,9}) pour J.

Soient les valeurs idéales suivantes

1	1	0
0	1	1

Alors l'écart entre le tableau initial et le tableau idéal est égal à 9.

5.3 LE CRITERE

On note :

- (x_i^j) le tableau binaire initial défini sur les deux ensembles I et J de cardinal n et p,
- P et Q les partitions en K classes et M classes des deux ensembles I et J et
- **L** l'ensemble des tableaux binaires à K lignes et M colonnes ; **L** représente l'ensemble des noyaux :

$$L \in \mathbf{L} \Leftrightarrow L = (a_k^m) \quad \text{où} \quad a_k^m \in \{0,1\} \quad \forall\, k=1, K \text{ et } \forall\, m=1, M.$$

Le critère que l'on cherche à minimiser s'écrit alors :

$$W(P,Q,L) = \sum_k \sum_m \sum_{i \in P_k} \sum_{j \in Q_m} | x_i^j - a_k^m |$$

5.4 L'ALGORITHME CROBIN

Plusieurs algorithmes sont possibles pour obtenir un optimum local du critère. Nous avons retenu l'algorithme suivant :

Partant d'un élément (P,Q,L) initial, on fixe Q et on cherche à améliorer P et L, puis on fixe P et on cherche à améliorer Q et L. On construit ainsi une suite (P^n,Q^n,L^n) qui fait décroître le critère W. L'algorithme est donc construit à partir de deux étapes intermédiaires que nous allons maitenant préciser.

Soit P et Q un couple de partitions et L un noyau. Fixons Q et cherchons à améliorer la partition P et le noyau L, c'est-à-dire cherchons une partition P' et un noyau L' telle que

$$W(P,Q,L) > W(P',Q,L').$$

On peut écrire

$$W(P,Q,L) = \sum_k \sum_{i \in P_k} \underbrace{\left(\sum_m \sum_{j \in Q_m} | x_i^j - a_k^m | \right)}_{A}$$

Si on note $y_i^m = \sum_{j \in Q_m} x_i^j$ et $q_m = Card(Q_m)$

$$\text{on a } A = \begin{cases} y_i^m & \text{si} \quad a_k^m = 0 \\ q_m - y_i^m & \text{si} \quad a_k^m = 1 \end{cases}$$

On a

$$A = |y_i^m - q_m a_k^m| \quad \text{et} \quad W(P,Q,L) = \sum_k \sum_{i \in P_k} \sum_m |y_i^m - q_m a_k^m|$$

Il est alors aisé de montrer que l'algorithme des Nuées Dynamiques défini sur l'ensemble de n éléments et L variables associé au tableau (y_i^m) muni de la distance L^1 avec des noyaux de la forme $(q_1 a_k^1, \ldots, q_L a_k^L)$ où $a_k^m \in \{0,1\}$ (chacune des composantes du noyau ne peut être que le minimum ou le maximum atteint par le regroupement des colonnes du tableau initial, c'est-à-dire les valeurs 0 ou q_m puisque les valeurs initiales étaient 0 ou 1 et qu'il n'y a que q_m colonnes réunies pour former la classe Q_m) fournit une solution à notre problème.

On peut évidemment, de façon similaire, à partir d'une partition L fixée améliorer la partition Q et le noyau L.

5.5 UN EXEMPLE D'APPLICATION

5.5.1 Les données

Dans cet exemple (Leredde, Perrin 1980), il s'agit d'étudier un ensemble de plaques-boucles de ceintures damasquinées du nord-est de la France s'échelonnant entre la fin du VI^{ième} et le début du VII^{ième} siècle. L'ensemble est constitué de 59 plaques-boucles sur lesquels ont été observée la présence ou l'absence de 26 critères techniques de fabrication, de forme et de décor.

5.5.2 Les résultats

Nous analysons ici le meilleur résultat obtenu sur 10 tirages initiaux.

Liste des deux partitions

```
valeur du critere obtenu    172
```

Remarquons que cette valeur est bonne puisqu'elle correspond à un taux moyen de désaccord entre un valeur binaire du tableau et la valeur binaire correspondante du noyau de l'ordre de 11%.

partition en ligne

classe	1 :	20 elements:	03	04	09	14	21	22	24	25	26	28	30	31	32	33	34
			36	46	49	52	53										
classe	2 :	19 elements:	05	07	11	15	17	20	23	27	35	47	48	50	51	54	55
			56	57	58	59											
classe	3 :	20 elements:	01	02	06	08	[0	12	13	16	18	19	29	37	38	39	40
			41	42	43	44	45										

partition en colonne

classe	1 :	7 elements:	C16	C23	C25	C32	C36	C37	C42				
classe	2 :	11 elements:	C01	C14	C19	C26	C28	C30	C31	C34	C35	C40	C41
classe	3 :	5 elements:	C17	C29	C33	C38	C39						
classe	4 :	3 elements:	C15	C22	C24								

Le tableau des "valeurs idéales"

Ce tableau permet de caractériser les classes entre elles. Il constitue un résumé du tableau initial.

	1	2	3	4
1	0.	0.	1.	1.
2	0.	0.	0.	1.
3	1.	0.	0.	0.

Ce tableau est extrémement simple à lire et permet une première interprétation grossière des données.

- la classe 2 des variables n'intervient pas dans cette classification,
- la classe 1 des lignes se caractérise par l'absence des propriétés de la classe 1 des colonnes et la présence des propriétés des classes 3 et 4 en colonne,
- la classe 2 des lignes se caractérise par la présence des propriétés de la classe 4 en colonne,
- la classe 3 des lignes se carcatérise, elle, par la présence des propriétés de la classe 1 en colonne.

Le tableau des écarts de chaque couple de classes aux valeurs idéales

Ce dernier tableau permet de mesurer la qualité de représentation de chaque couple de classes. Par exemple, le couple (classe 3 en ligne et classe 4 en colonne) est parfaitement représenté par la valeur 0 (100%), alors que le couple classe 2 en ligne, classe 2 en colonne est moyennement représenté par la valeur 0 puisque seulement 76% des valeurs sont 0. Sur ce petit exemple, la qualité est de toute façon toujours très correcte.

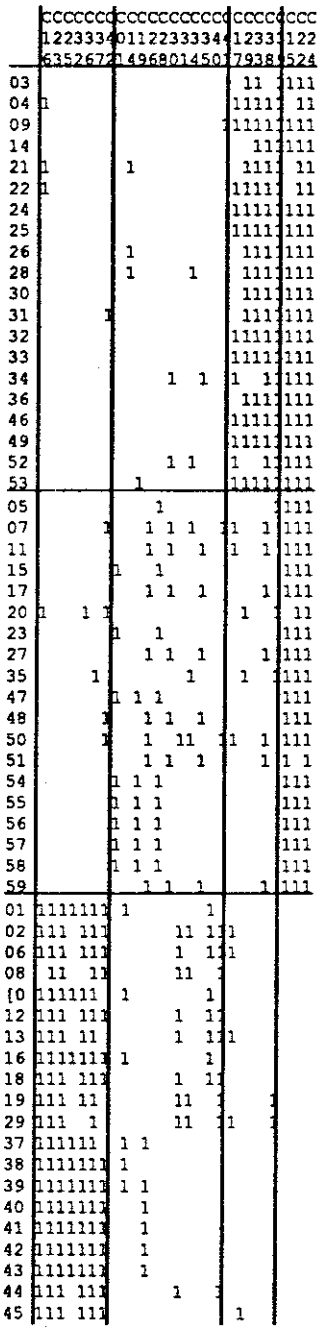
homogeneite par classe

	1	2	3	4
1	97.	95.	86.	95.
2	95.	76.	84.	96.
3	87.	81.	93.	100.

Le tableau binaire réordonné suivant les partitions obtenues en ligne et en colonne

Ce tableau, si sa dimension permet de le sortir permet de mettre en évidence les structures qui pouvaient exister, directement sur les données sans ajouter ni enlever d'information au tableau initial.

tableau initial reordonne



BIBLIOGRAPHIE

Anderberg, M. (1973), *Cluster Analysis for applications*. Academic Press, New-York.
Benzécri, J.-P. (1973), Théorie de l'information et classification d'après un tableau de contingence. *l'Analyse des données*. tome 1, Dunod, Paris.
Bertin, J. (1977), *La graphique et le traitement graphique de l'information*. Flammarion, Paris.

Bock, H. (1979), Simultaneous clustering of objects and variables. *Cours des communautés européennes. Analyse des données et informatique* INRIA, Fontainebleau.
Celeux, G. Diday, E. Govaert, G. Lechevallier, Y. Ralambondrainy, H. (1989), *Classification automatique des données - Environnement statistique et informatique* Dunod, Paris.
Diday, E. (1971), Une nouvelle méthode en classification automatique et reconnaissance des formes. *Revue de Statistiques Appliquées* **19** n° 2.
Diday, E. et Coll. (1980), *Optimisation en classification automatique*. Ed. INRIA, Rocquencourt.
Fisher, W. (1969), *Clustering and aggregation in economics*. The Johns Hopkins Press, Baltimore.
Govaert, G. (1977). Algorithme de classification d'un tableau de contingence. *Premières journées internationales Analyse de Données et Informatique*. INRIA, Versailles.
Govaert, G. (1983),*Classification Croisée*. Thèse d'Etat. Paris 6.
Govaert, G. (1984), Algorithme de classification d'un tableau de contingence. *Data analysis and informatics III*, ed. Diday, North-Holland.
Hartigan, J. (1972), Direct clustering of a data matrix. *Journal of American Statistical Association*.
Hartigan, J. (1975), *Clustering algorithms*. John Wiley & sons, New-York.
Jambu, M. (1976), Sur l'interprétation naturelle d'une classification hiérarchique et d'une analyse des correspondances. *Revue de Statistiques Appliquées* **24** n°2.
Lebart, L. Morineau, A. et Tabard, N. (1977), *Techniques de la description statistique* Dunod, Paris.
Lerman, I.C. et Leredde, H. (1977), La méthode des pôles d'attraction, *Premières journées internationales Analyse de Données et Informatique*. INRIA, Versailles.
Lerman, I.C. (1981), *Classification et analyse ordinale des données*. Dunod, Paris.
Leredde, H. et Perin, P. (1980), Les plaques-boucles mérovingiennes. *Dossiers de l'archéologie* **42**, 83-87.
Maroy, J.P. et Péneau, J.P. (1972), Analyse des données et conception en architecture. *Bulletin IRIA n°13*. Rocquencourt.
Tolédano, J. et Brousse, J. (1977), Une méthode de classification simultanée des lignes et des colonnes. *Premières journées internationales Analyse de Données et Informatique* INRIA, Versailles.