LES CAHIERS DE L'ANALYSE DES DONNÉES

F. J. GALLEGO

Codage flou en analyse des correspondances

Les cahiers de l'analyse des données, tome 7, n° 4 (1982), p. 413-430

<http://www.numdam.org/item?id=CAD_1982__7_4_413_0>

Les Cahiers de l'Analyse des Données
Vol. VII - 1982 - *n°* 4 - *p.* 413-430

CODAGE FLOU EN ANALYSE DES CORRESPONDANCES

[COD. FLOU]

par F. J. Gallego (1)

1 *Rappel d'Analyses des Correspondances Multiples (A.C.M.)*

Plusieurs approches peuvent être faites de l'A.C.M. (cf. 1, 2, 4, et 7). La donnée d'une telle analyse est celle d'une population I distribuée en classes d'après un certain ensemble Q de critères q ayant chacun un ensemble Jq de modalités ; ce qui donne lieu à un tableau de contingence K à Card Q entrées : $J_{1\alpha}$.. $J_{q\alpha}$...

Dans le cas où l'on connaît le détail de la population I, il est classique d'effectuer l'analyse des correspondances du tableau logique Z qui croise les individus avec l'ensemble $J = \cup\{Jq \mid q \in Q\}$ des modalités de tous les critères.

$$z_{ij} = \begin{cases} 1 \text{ si } i \in j \text{ (i.e. rentre dans la modalité j)} \\ 0 \text{ sinon} \end{cases}$$

La même description des modalités fournie par Z peut être obtenue aussi à partir du tableau de Burt B_{JJ}, qui juxtapose tous les tableaux de contingence binaires Jq × Jq' à deux entrées q, q' possibles :

$$B_{jj'} = \text{card}(j \cap j')$$

(où j et j' sont considérés comme des sous-ensembles de I : e.g. : j = ensemble des individus rentrant dans la modalité j). On notera que le tableau de Burt ne tient manifestement compte que des marges du tableau multiple K : en sorte qu'en toute rigueur il s'agit non d'une A.C.M., mais de l'analyse d'un tableau binaire associé à la correspondance multiple.

Pour la construction de Z, les variables sont supposées qualitatives. On peut toutefois rendre qualitative une variable quantitative à l'aide d'une partition de la droite réelle en un ensemble J_q d'intervalles, dont chacun détermine une modalité.

2 *Discontinuité et perte d'information du codage disjonctif*

L'analyse des tableaux décrits a des avantages remarquables : souplesse de la méthode, qui permet de traiter ensemble des variables qualitatives et quantitatives, et capacité de décrire des rapports non linéaires entre variables quantitatives. Pourtant, deux objections peuvent être faites au codage disjonctif complet par découpage en intervalles : d'un côté, il y a une certaine perte d'information quand une valeur de la variable q est remplacée par

(1) *Docteur 3° cycle. Professeur de statistique à la Faculté des Sciences de Valladolid (Espagne).*

l'appartenance à un intervalle j ; on ne peut plus reconstituer exactement les données d'origine. La discontinuité du codage est le deuxième inconvénient et le plus grave, surtout quand on s'intéresse à la description de l'ensemble I ; en effet, si b_k est la borne qui sépare deux intervalles, une distance est artificiellement créée entre des points proches qui sont d'un côté et de l'autre de b_k ; distance qui est en général du même ordre de grandeur que celle qui sépare après le codage les plus petites des plus grandes valeurs de q en tant que quantitative.

2.1 *Une alternative au découpage en intervalles :* Pour échapper à cette discontinuité, l'idée vient de faire un passage progressif d'une modalité à l'autre en faisant qu'une valeur proche de la frontière soit partagée par les deux modalités. On peut définir un codage flou pour chaque partition floue de la droite réelle, mais la plupart n'ont aucun intérêt vis-à-vis de l'interprétation (cf. 8). Parmi ceux qui débouchent sur des résultats lisibles après une A.C.M. se trouve le codage semi-linéaire (fig. 1), dont les modalités sont déterminées par des fonctions d'appartenance qui valent 1 dans un point de référence et décroissent linéairement d'un côté et de l'autre.

2.2 *Un codage continu et injectif :* Le codage employé dans la suite modifie le semi-linéaire et le rend injectif par l'élimination d'intervalles à valeur d'appartenance constante. Ce codage nécessite le choix de J_q-2 points de référence (où J_q est le nombre de modalités souhaité pour la variable q), soit $r_2 \ldots r_{Jq-1}$. Chacune des modalités intermédiaires doit être interprétée comme "ce qui est autour de r_j", alors que la première et la dernière, z_1 et z_{Jq} peuvent être considérées associées à $-\infty$ et à $+\infty$. Pour appliquer cette version du codage, on doit avoir un nombre de modalités $J_q \geq 4$, bien que l'adaptation au cas de deux ou trois modalités ne pose pas de difficultés insurmontables.

Les fonctions d'appartenance aux J_q modalités sont :

$$z_1(x) = \begin{cases} 1-\exp(-(r_2-x)/(r_3-r_2)) & x \leq r_2 \\ 0 & x > r_2 \end{cases}$$

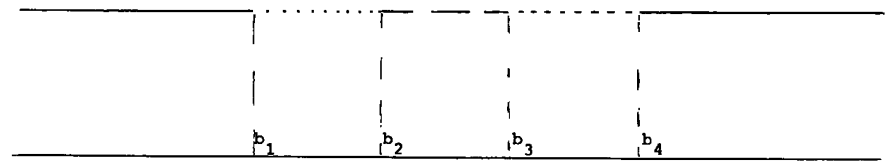
$$z_2(x) = \begin{cases} \exp(-(r_2-x)/(r_3-r_2)) & x \leq r_2 \\ (r_3-x)/(r_3-r_2) & r_2 \leq x \leq r_3 \\ 0 & x > r_3 \end{cases}$$

$$\vdots \qquad (1)$$

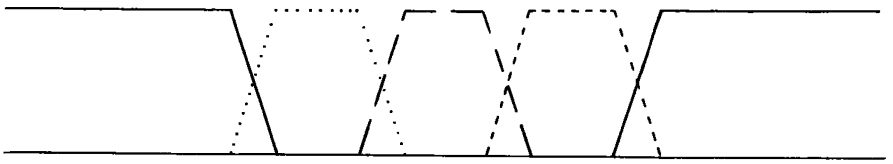
$$z_k(x) = \begin{cases} 0 & x \leq r_{k-1} \\ (x-r_{k-1})/(r_k-r_{k-1}) & r_{k-1} < x \leq r_k \\ (r_{k+1}-x)/(r_{k+1}-r_k) & r_k < x \leq r_{k+1} \\ 0 & x > r_{k+1} \end{cases}$$

$$\vdots$$

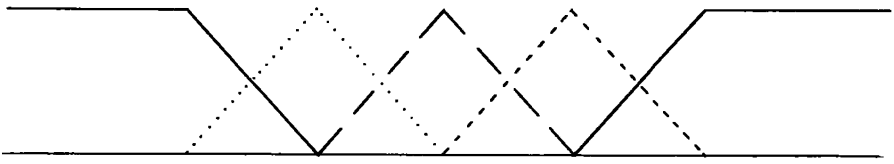
$$z_{Jq}(x) = \begin{cases} 1-\exp(-(x-r_{Jq-1})/(r_{Jq-1}-r_{Jq-2})) & x \geq r_{Jq-1} \\ 0 & x < r_{Jq-1} \end{cases}$$



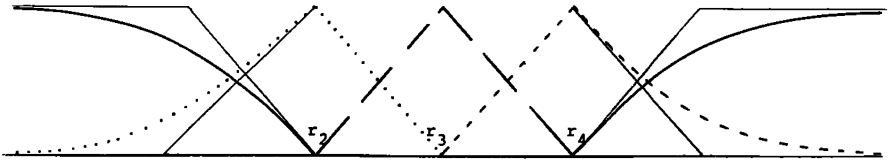
a) Codage disjonctif complet



b) Codage trapézoidal



c) Codage semilinéaire



d) Codage semilinéaire modifié

Figure 1 : Fonctions d'appartenance pour plusieurs découpages d'une variable quantitative en cinq modalités.

Le choix des exponentielles pour les modalités extrêmes est quelque peu arbitraire, et nous ne voyons pas d'inconvénient *a priori*, à employer pour Z_1 n'importe quelle fonction qui soit strictement décroissante entre $-\infty$ et r_2, approche asymptotiquement 1 quand $x \downarrow -\infty$ et atteigne le 0 quand $x = r_2$.

Le tableau résultant du codage est du type :

I \ J			
i	0 t 1-t 0	.....	s 1-s 0 0

Si $r_k \leq x_q(i) \leq r_{k+1}$, les valeurs d'appartenance $Z_k(i)$ et $Z_{k+1}(i)$ sont les masses qu'il faudra placer en r_k et r_{k+1} pour que leur barycentre soit $x_q(i)$.

3 *Analyse des correspondances sous codage flou*

L'étude de l'analyse des correspondances sous codage flou peut être entreprise sous plusieurs points de vue ; on ne s'occupera ci-après que de quelques propriétés élémentaires et des résultats obtenus dans des exemples d'application.

3.1 Quelques propriétés : Il est facile de vérifier que le codage flou conserve les deux propriétés suivantes de l'analyse des tableaux disjonctifs :

- Les modalités de chaque variable forment un sous-nuage dont le barycentre coïncide avec le barycentre global.

- L'ensemble des points-modalités engendre un sous-espace de dimension inférieure ou égale à J-Q+1 (qui est donc le nombre maximum de valeurs propres non nulles à part le facteur trivial).

L'inertie totale et les contributions des modalités de chaque variable à cette inertie sont toujours inférieures aux quantités calculées dans le cas du tableau 0-1 associé au codage disjonctif. En effet, la contribution des modalités de la variable q à l'inertie vaut :

$$C(q) = Q^{-1} \Sigma\{\Sigma\{z^2(i,j)/m(j) \mid i \in I\} - (m(j)/n) \mid j \in Jq\} \leq Q^{-1}(Jq-1)$$

puisque $z^2(i,j) \leq z(i,j)$

avec $m(j) = \Sigma\{z(i,j) \mid i \in I\}$

en conséquence, l'inertie totale vaut :

$$\Sigma\{C(q) \mid q \in Q\} \leq (J/Q) - 1$$

3.2 L'interprétation des graphiques plans : Le codage disjonctif complet d'une variable quantitative est une application de la droite réelle sur les sommets d'un simplexe de $R^J q^{-1}$; l'image de la droite réelle par un codage flou affectant un individu à deux classes contiguës est la ligne polygonale constituée par une suite d'arêtes de ce simplexe. Dans le cas du codage donné par les expressions (1), cette application est injective (fig. 2).

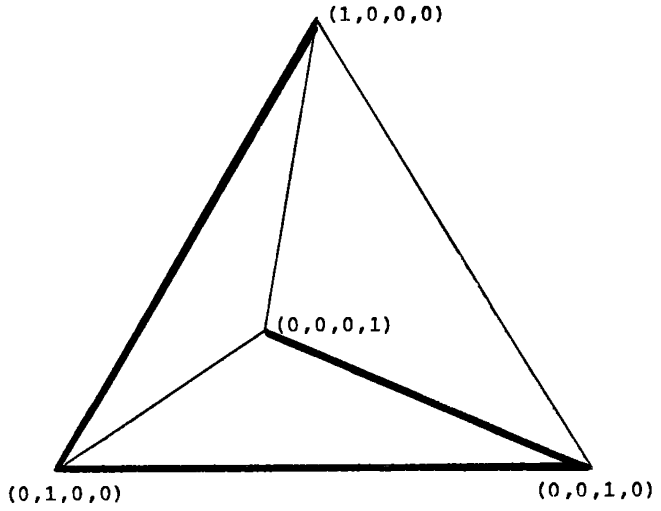


Figure 2 : Le nuage résultant du codage disjonctif en quatre classes est constitué par les sommets d'un tétraèdre , alors que les points qui résultent du codage flou se trouvent sur les arêtes en traits gros.

Après avoir fait une A.C.M., il est fréquent de dessiner sur les plans croisant deux axes factoriels la ligne polygonale qui joint de façon ordonnée les modalités, et donne une idée d'ensemble du comportement de la variable. Dans le cas du codage disjonctif les points intermédiaires des segments qui joignent les points-modalité n'ont aucune signification précise, mais avec le codage donné par (1), chaque point entre la k-ième et la k+1-ème modalité représente une valeur entre r_k et r_{k+1} . Ceci trouve une expression plus précise dans la formule de transition ou formule barycentrique, qui est dans le cas disjonctif :

$$F_\alpha(i) = \lambda_\alpha^{-1/2} \frac{1}{Q} \Sigma\{G_\alpha(j) \mid Z(i,j) = 1\}$$

et qui devient dans le cas du codage flou :

$$F(i) = \lambda_\alpha^{-1/2} \frac{1}{Q} \Sigma\{\gamma_\alpha(i,q) \mid q \in Q\}$$

où $\gamma_\alpha(i,q) = G_\alpha(j)\, Z(i,j) + G_\alpha(j+1) Z(i,j+1)$

j et j+1 étant les deux modalités de la variable q auxquelles appartient i. $\gamma_\alpha(i,q)$ est donc la coordonnée factorielle du point de la suite d'arêtes qui est l'image par le codage de $x_q(i)$. La relation barycentrique se conserve, à ceci près que, chaque individu n'étant pas affecté à une seule modalité, il doit être considéré comme appartenant à une "modalité mixte" intermédiaire entre deux modalités pures.

3.3 *Test sur les modalités supplémentaires* : On sait que sous l'hypothèse nulle d'indépendance d'une modalité supplémentaire j avec les variables principales, sa coordonnée factorielle $G_\alpha(j)$ vérifie cf. 7, p. 137) :

$$E(G_\alpha(j)) = 0$$

$$Var(G_\alpha(j)) = 1/n_j$$

où n_j est le nombre d'individus qui appartiennent à la modalité j. Si Card(I) est sufisamment grand, $G_\alpha(j)$ suit approximativement une loi normale, et le seuil à partir duquel (à un certain niveau de confiance) la modalité j est à regarder comme significativement écartée de l'origine sur cet axe, en découle immédiatement.

Ce test se reproduit pour le cas d'un codage flou, l'expression de la variance étant :

$$Var(G_\alpha(j)) = \Sigma\{z^2(i,j)/m^2(j) \mid i \in I\}$$

ce qui nous mène au test correspondant pour le codage flou.

3.4 *Mise en oeuvre* : Un programme de codage flou avec les fonctions d'appartenance décrites en (1) est donné en (5) , ainsi que son mode d'emploi. Le programme prévoit une option manuelle de choix des points de référence $r_2 \ldots\ldots r_{k+1}$ et une option automatique avec laquelle les points de référence sont calculés de façon que les modalités obtenues soient approximativement équipondérées.

On peut trouver en (9) la description du programme STEKMA qui peut réaliser des analyses de correspondances sous codage flou.

4 *Comparaison des résultats avec les codages disjonctif et flou*

Guilonneau et Roux (cf. 6) présentent une application du codage trapézoïdal (fig. 1) et comparent le résultat avec celui d'une analyse en composantes principales pondérée, mais ne font pas la comparaison avec celui d'une A.C.M. avec codage disjonctif.

Le Foll (cf. 8) a appliqué un codage semi linéaire à un tableau de mesures de pollution des eaux du Bassin Parisien, qui comportait 1452 observations de 50 variables éclatées en 220 modalités. Les dix premiers facteurs sont étudiés et comparés avec ceux qui sont obtenus avec le codage disjonctif. La seule différence remarquable constatée pour ces données est la régularisation par le codage flou des trajectoires formées par les points-modalité de chaque variable.

On présente ci-après deux applications du codage semi-linéaire modifié, donné par les expressions (1) sur des tableaux de taille plus restreinte sur lesquels le codage disjonctif a été aussi appliqué. Outre l'objectif propre à chaque étude, on voulait comparer les comportements de l'A.C.M. avec chacun des deux codages.

4.1 *Une application aux Comptes Trimestriels de l'I.N.S.E.E.* : Plusieurs tableaux ont été analysés avec les deux codages dans le cours d'une étude qui portait sur un ensemble de séries trimestrielles macroéconomiques élaborées d'après le modèle METRIC (Modèle Economique Trimestriel de la Conjoncture). Nous ne présenterons pas le détail des résultats des analyses, qui peut être consulté en (5) , et nous nous bornerons à signaler comment étaient constitués les tableaux qui ont conduit aux résultats les plus intéressants et quelles ont été les différences principales constatées entre les deux codages.

Les variables étaient des taux de variation relative par trimestres de quantités macroéconomiques (PIB, Activité, Pouvoir d'achat

des salaires et des prestations sociales, Prix à la consommation, Balance commerciale, Epargne des ménages, Intérêt du Marché monétaire , Pression fiscale, Bénéfice des Sociétés, Chômage, Difficultés de Trésorerie des entreprises, Compétitivité industrielle, et Prix des Importations), ainsi que les contributions à la Croissance de la consommation, Investissement, Importations, Exportations, Stocks et Dépenses de l'Administration en Biens et Services.

Les analyses présentées ici sont appariées, chaque paire étant constituée de deux analyses faites sur le même tableau avec des codages flou et disjonctif, de façon que dans les deux cas les modalités correspondent approximativement aux mêmes zones de la droite réelle.

Les quatre paires considérées sont :

1) 12 taux entre le deuxième trimestre en 1963 et le deuxième de 1979. Les bornes des modalités (points de référence pour le découpage flou) sont choisies après observation des histogrammes ; dans les cas où ceux-ci montrent un aspect multimodal, les points de référence sont choisis au milieu des "bosses", qui sont inclues à leur tour dans une seule modalité disjonctive. Les modalités extrêmes qui en résultent ont souvent des poids faibles.

2) Même tableau découpé avec un critère d'équipondération approximative des modalités.

3) 21 taux entre le deuxième trimestre de 1963 et le quatrième de 1978 découpés avec un critère d'équipondération.

4) Même codage sur les 21 taux lissés par :

$$LX(i) = (X(i-1) + 2\,X(i) + X(i+1))/4$$

Dans tous les cas, chaque variable a été découpée en six modalités.

4.2 *Un problème médical :* Les deux codages, disjonctif et flou, ont été essayés sur des tableaux de taille plus restreinte concernant une expérience réalisée par M.C. Boffa au CNTS*et à l'hôpital St Jacques. Le problème était ici de caractériser les réactions d'un groupe d'individus sains à des granules imprégnées d'une solution infinitésimale de venin de *Naja nigricollis* en vue de son utilisation en médecine homéopathique.

L'expérience a été faite en double aveugle sur une population de vingt individus, dont dix, choisis au hasard, ont été traités, et dix ont reçu une dose de placebo. Pendant toute la durée de l'expérience, ni les sujets ni le médecin qui suivait leur évolution ne connaissaient le résultat de ce tirage au hasard.

Un nombre irrégulier de prises de sang a été effectué avant le traitement et après chacune des deux phases qu'il comprenait. Après un certain nombre d'essais, on n'a retenu pour la description des effets qu'un tableau de vingt lignes et seize colonnes (individus et variables biologiques déterminées dans chaque prise de sang) qui contient les variations relatives après la première des deux phases du traitement, les observations correspondantes à la deuxième manquant dans un trop grand nombre de cas pour que les résultats de la description soient fiables.

Comme dans le tableau des séries trimestrielles, il s'agit de variables continues qui permettent d'appliquer le codage flou sans

* *Centre National de Transfusion Sanguine.*

restrictions dues à des valeurs ayant une signification particulière qui obligerait à faire des modalités en 0-1. (*).

Toutes les variables ont été découpées en quatre modalités avec un critère d'équipondération approximative, aussi bien dans le cas disjonctif que dans le cas flou.

4.3 *Inertie du nuage et des facteurs* : Pour le codage disjonctif, l'inertie totale du tableau Z ne dépend que du nombre de modalités et du nombre des variables :

$$\text{Inertie} = (J/Q) - 1$$

on a vu que cette quantité est le maximum des valeurs possibles pour le codage flou ; il est facile de voir que, plus les valeurs après le codage tendent à être proches de 1 et de 0, plus l'inertie est grande, et plus les valeurs non nulles de Z sont proches de 1/2, plus l'inertie est petite. Cette observation nous suggère de tester une éventuelle tendance des valeurs avant le codage, à être concentrées autour des points de référence $r_2 \ldots r_5$ (rappelons qu'il y a ici 6 modalités). Sous des hypothèses d'équidistribution des valeurs d'appartenance non nulles dans l'intervalle [0,1] et d'équipondération des modalités, l'espérance de l'inertie vaut $((2J/(3Q)) - 1$, et sa variance $36/(45\ \text{Card}\, I)$. Les valeurs obtenues dans toutes les analyses où un critère d'équipondération a été employé sont proches de l'espérance calculée et le test suggéré ne permet en aucun cas de rejeter l'hypothèse d'équidistribution.

Le codage flou donne lieu à un nuage plus allongé que celui issu du codage disjonctif, ce qui se comprend bien en regardant la figure 2 ; et les pourcentages d'inertie des premiers axes sont plus grands pour le codage flou. La différence des pourcentages est beaucoup plus nette et plus régulière dans les analyses faites sur les données des comptes trimestriels que sur les données médicales. On donne à titre indicatif les pourcentages des cinq premiers facteurs dans chacune des analyses considérées.

	Taux Trimestriels								Naja	
	1		2		3		4			
	dis	flou	dis	flou	dis	flou	dis	flou	dis	flou
λ_1	319	249	328	271	257	204	293	262	331	217
% cumulé	6.4	9.1	6.6	8.6	5.1	6.4	5.9	8.2	11.5	12.0
λ_2	304	215	298	227	254	193	281	244	317	212
% cumulé	12.5	16.9	12.5	15.8	10.2	12.4	11.5	15.9	22.5	23.7
λ_3	273	178	289	195	217	172	243	184	269	185
% cumulé	17.9	23.4	18.3	22.0	14.6	17.8	16.4	21.6	31.9	34.0
λ_4	268	156	257	176	199	146	219	162	232	162
% cumulé	23.3	29.1	23.4	27.6	18.5	22.3	20.7	26.7	40.0	43.0
λ_5	244	133	245	157	198	139	199	149	217	129
% cumulé	28.2	33.9	28.3	32.6	22.5	26.6	24.7	31.4	47.5	50.1
Trace	5	2.74	5	3.15	5	3.21	5	3.19	2.87	1.81

Tab. 1 : Valeurs propres (en millièmes) et pourcentages cumulés d'inertie des cinq premiers facteurs de chacune des analyses sous codage flou et disjonctif.

(*) Imaginons une variable du type : "taux d'alcool dans le sang" ; il s'agit d'une variable quantitative, mais la valeur zéro a une signification suffisamment différente du reste pour qu'on en fasse une modalité non floue.

L'importance de cette constatation reste néanmoins peu claire , car il est bien connu que les pourcentages d'inertie sont très différents dans une même A.C.M. selon qu'elle est faite sur un tableau logique Z ou sur le tableau de Burt associé.

4.4 *Ecart-type des contributions des variables aux facteurs :* Quand il s'agit d'interpréter un facteur, on peut désirer choisir un sous-ensemble de variables qui contribuent le plus à la formation de chaque facteur. Ces contributions sont calculées par addition de celles de leurs modalités, car toutes les modalités d'une variable doivent être interprétées ensemble. Une question se pose alors : combien de variables doit-on choisir ? Il s'agit de déterminer un seuil de contribution à partir duquel on considère que la variable est importante dans ce facteur. Il n'y a pas pour le moment de critère objectif pour guider ce choix, qui est d'autant plus difficile que les valeurs des contributions sont plus proches les unes des autres. Il est donc intéressant d'observer si les contributions des variables aux facteurs sont dispersées par l'emploi d'un certain codage. Dans ce cas l'écart-type a été utilisé comme mesure de dispersion, et il est nettement plus fort avec le codage flou qu'avec le disjonctif dans les analyses faites sur les données des comptes trimestriels ; cette caractéristique est beaucoup moins claire sur les données de naja. On peut être tenté de donner des résultats plus précis sur ce point en faisant des tests sur les écarts-types, mais les tests classiques seraient trop forcés puisqu'une hypothèse d'indépendance n'est nullement admissible.

Facteur	Comptes Trimestriels								Naja	
	1		2		3		4			
	dis	flou	dis	flou	dis	flou	dis	flou	dis	flou
1	55.4	66.7	41.4	65.7	27.2	39.2	22.7	35.3	42.6	43.3
2	35.8	47.5	26.6	43.0	24.3	34.1	29.6	31.8	28.8	44.8
3	28.4	45.3	37.2	66.0	26.2	28.7	25.1	26.9	36.2	33.6
4	49.4	41.6	50.0	46.3	26.4	21.4	20.5	34.5	44.7	32.5
5	50.6	40.3	33.0	58.6	27.1	36.1	20.6	20.9	33.1	38.8

Tab. 2 : Ecarts-types des contributions relatives des variables aux cinq premiers facteurs (en millièmes).

4.5 *Composition des facteurs et aspect des graphiques plans :* Les observations qui viennent d'être énumérées ne peuvent justifier que très partiellement l'emploi du codage flou. La question qui se pose avant tout est : est-ce que le codage flou permet de déceler des caractéristiques des données qui n'auraient pas pu être trouvées par le codage disjonctif? Autrement dit : est-ce que les rapports non-linéaires sont mieux décrits par le codage flou que par le disjonctif ?

Un premier avantage, qui est sans doute général, est celui de la régularisation des trajectoires associées aux variables, due à la diminution de la distance entre modalités contiguës, qui rend plus agréable l'observation des plans, les rapports entre les variables devenant plus clairs.

La constitution des facteurs, mesurée par les contributions des variables a été nettement différente avec les deux codages, et cette

différence est d'autant plus grande que la taille du tableau est petite. Dans le § 5, cette variation sera précisée un peu plus en termes de corrélations.

On ne présentera pas les résultats de toutes les analyses dont on a parlé, mais on en choisira deux paires, une concernant chacun des deux problèmes traités.

4.5.1 Commentaires aux analyses relatives aux comptes trimestriels :

Les figures 3 à 6 représentent les plans 1-2 et 1-3 de la paire d'analyses n° 1 sur les comptes trimestriels. On constate que le PIB et l'Investissement, qui apparaissent avec des contributions importantes sur les deux premiers axes de l'analyse disjonctive sont nettement plus ajustées au premier axe que dans le cas flou. La Consommation est déplacée du plan 2-3 au plan 1-2, les Dépenses de l'Administration le sont du plan 2-3 au deuxième facteur, les Prestations sociales du troisième au deuxième facteur, etc. .

On trouve d'une façon assez générale ce phénomène de plus forte attraction vers les premiers axes du codage flou, ce qui n'est pas étonnant, vu que leur pourcentage d'inertie a augmenté.

L'augmentation de l'écart-type des contributions est associée au fait que chacune des variables apparaisse sur un plus petit nombre de facteurs. Ainsi, en regardant la figure 4, on peut baptiser le premier axe d'axe "de croissance", alors que le deuxième est associé à l'inflation. Il est beaucoup plus difficile de baptiser les axes de l'analyse disjonctive.

Quel que soit l'intérêt de ces constatations, la question fondamentale reste : est-ce que l'on trouve avec le codage flou des phénomènes qui seraient restés inaperçus avec le codage disjonctif ? Dans le problème des comptes trimestriels, les séries n'ont pas été observées directement, mais calculées à partir d'observations auxiliaires à l'aide d'un modèle qui présuppose certaines relations qui constituent l'essentiel de ce qui est trouvé dans les analyses. Il s'agit d'un ensemble de données bien connues où l'intérêt de l'application de l'A.C.M. peut être contestable du point de vue économique, mais qui sert fort bien pour tester une méthode parce qu'on sait à peu près ce qu'on devrait trouver comme résultat.

La différence constatée entre les deux codages réside surtout dans la clarté et la facilité de lecture, associées au codage flou, bien qu'un phénomène inattendu soit apparu avec ce codage flou après avoir échappé au disjonctif : dans le plan 1-3 , on observe un effet Guttman affectant le PIB, Investissement (INV), Importations (IMP), et Balance Commerciale (BAL), alors que la trajectoire des Exportations (EXP) est associée au troisième axe (fig. 6), ce qui suggère une association des faibles valeurs de EXP avec les niveaux intermédiaires de la croissance. Le nuage des trimestres projeté sur les plans EXP-IMP et EXP-PIB présente des formes plus ou moins paraboliques. Les questions qui se posent alors sortent du domaine de l'analyse des données qui aurait déjà accompli son rôle en lançant une hypothèse.

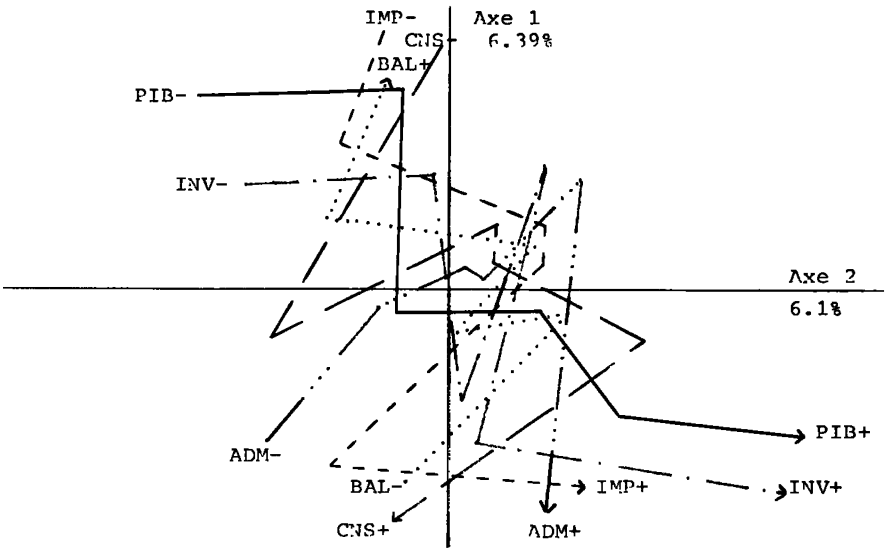


Fig 3: ACM sous codage disjonctif de 12 taux d'évolution trimestrielle. Variables plus contribuantes au plan 1-2

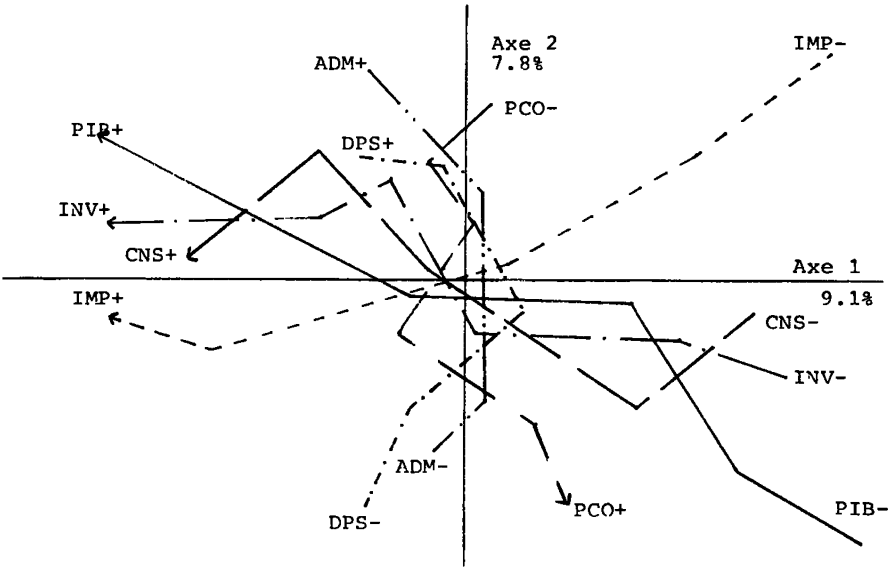


Fig. 4: ACM du même tableau sous codage flou

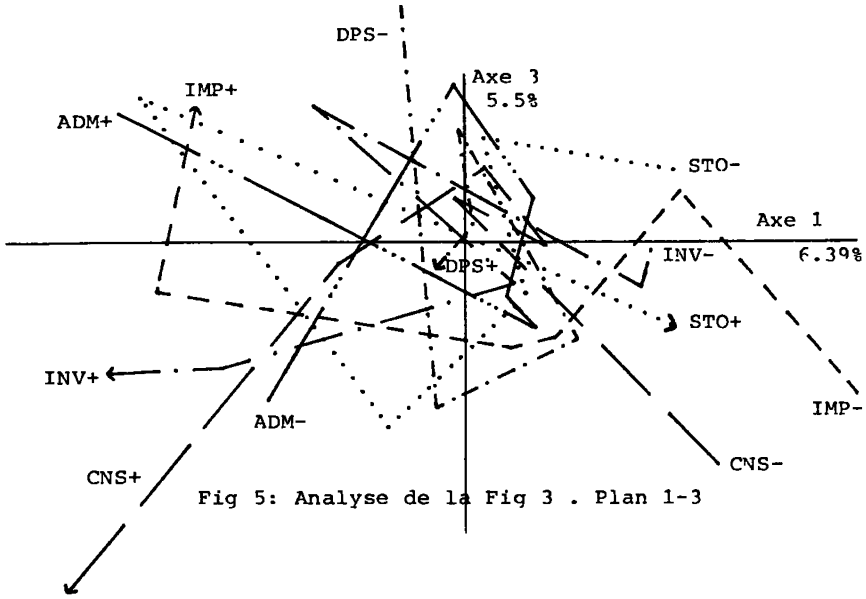


Fig 5: Analyse de la Fig 3 . Plan 1-3

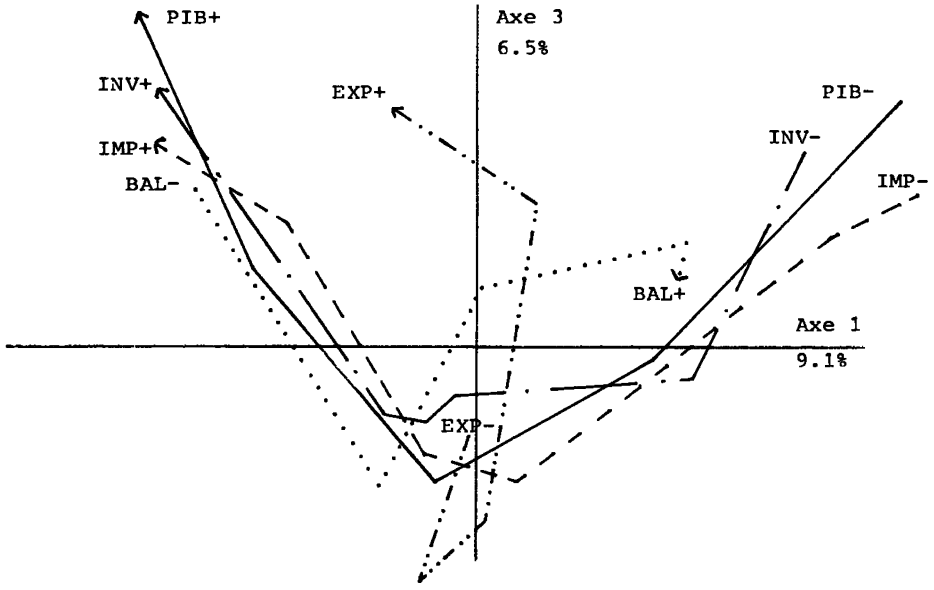


Fig. 6: ACM sous codage flou. Plan 1-3

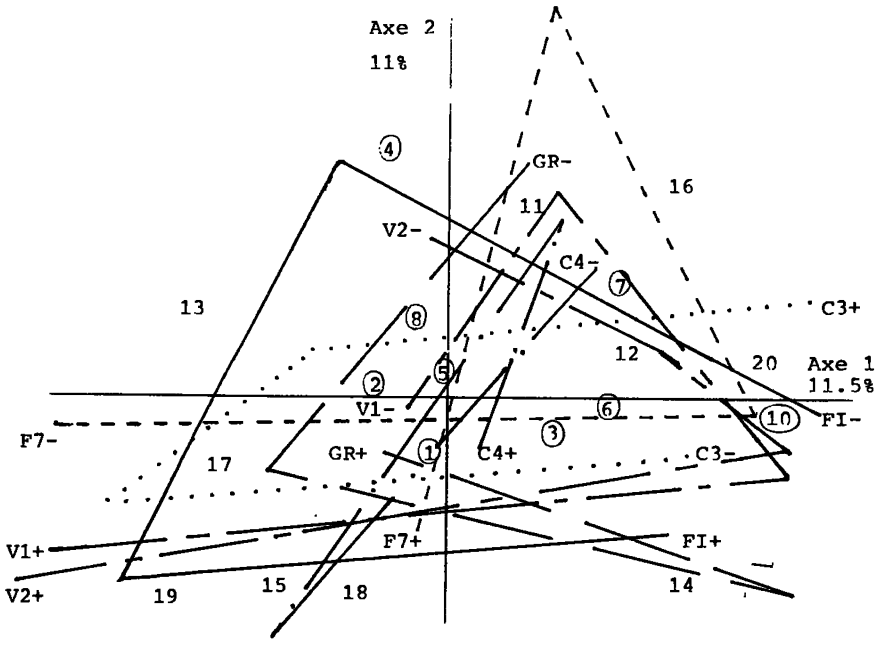


Fig. 7: Expérience "venin Naja". A.C.M. sous codage disjonctif des variations relatives. Plan 1-2

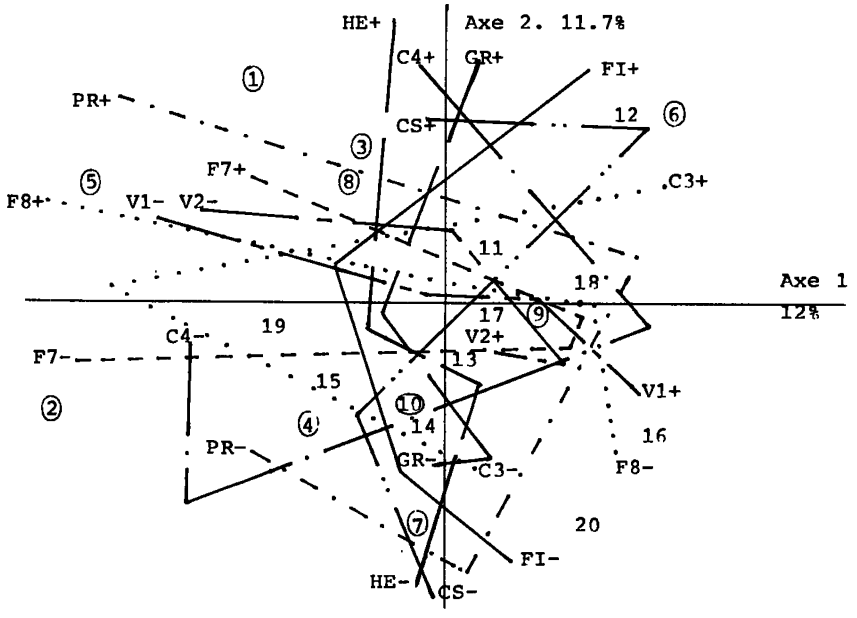


Fig. 8: A.C.M. sur le même tableau avec codage flou. Les sujets traités sont numérotés de 1 à 10, et les témoins de 11 à 20.

4.5.2 *Commentaires sur le problème médical :* Avant l'A.C.M. plusieurs études avaient été faites sur les données, qui avait fourni des renseignements qu'on peut résumer ainsi :

Des tests fishériens de permutation sur les variations, sur les valeurs absolues des variations, et sur un indice d'importance globale de la réaction de chaque individu face au traitement permettent de dire :

- Les individus traités réagissent plus violemment que les placebos (probabilité critique $<10^{-5}$), réaction qui est observable en particulier pour les variables : facteur C4 du complément sérique ($p<10^{-5}$), facteur VIII de coagulation ($p\simeq0.008$), proportion de neutrophiles ($p\simeq0.015$), et protéines dotées dans le sérum ($p\simeq0.045$).

- Pour trois variables les variations se produisent dans le même sens pour l'ensemble des traités : diminution de la vitesse de sédimentation première heure ($p\simeq0.02$) et deuxième heure ($p\simeq0.02$) , et augmentation de FVIII ($p\simeq0.006$).

Les résultats d'une analyse en composantes principales sont concordants avec ceux des tests : les témoins apparaissent concentrés autour du point qui représente un individu hypothétique sur lequel toutes les variations auraient été nulles, et les traités sont plus éloignés, mais les directions de cet éloignement sont disparates. Les réactions des individus sont donc qualitativement diverses, et dépendent de leut typologie. Le problème est maintenant de classer et caractériser les types de réaction.

Le problème étant clairement non-linéaire, on a voulu décrire la situation par une A.C.M. ; vu la taille de la population, chaque variable a été éclatée en quatre modalités avec un critère d'équipondération approximative comme on l'a déjà dit.

L'A.C.M. sous codage disjonctif, dont le plan 1-2 est représenté dans la figure 7, donne des trajectoires-variable qui se bouclent souvent sur elles-mêmes, et offre un aspect peu lisible, en partie à cause des grandes distances que le codage a produites entre valeurs qui étaient proches. Ces distances artificielles ont dispersé les individus témoins (numérotés de 11 à 20), ce qui n'est pas cohérent avec les résultats antérieurs.

Le codage flou a rapproché les modalités contiguës entre elles, et le plan 1-2 de l'A.C.M. offre l'aspect de la figure 8, où les deux populations sont assez bien discriminées sur le premier axe. Les trajectoires associées aux variables sont plus régulières et trois d'entre elles suivent le premier axe, il s'agit de V1 et V2 (vitesse de sédimentation première et deuxième heures), et FVIII (facteur VIII de coagulation), précisément celles qui discriminaient les deux sous-populations dans le test des variations homogènes (sur les variations signées). Les variables GR (globules rouges), HE (hématocrite), CS (complément sérique), et FI (fibrinogène) longent plutôt le deuxième axe, alors que C4, PR (proaccélérine), et FVII décrivent des trajectoires "paraboliques" avec des modalités intermédiaires (faibles variations) associées aux sujets placebos, et les fortes variations d'un sens et de l'autre apparaissent du côté des traités, ce qui est cohérent avec les résultats des tests sur les valeurs absolues des variations . C3 semble subir l'effet contraire (fortes variations associées plutôt à certains témoins) bien que cette hypothèse ne soit pas appuyée par les résultats des tests.

Une classification ascendante hiérarchique de tous les individus avec la métrique du χ^2 sur le tableau des modalités floues, et un critère d'agrégation de minimisation de l'inertie perdue dans chaque pas, donne le découpage décrit dans le tableau ci-dessous.

sujets dans la classe	Rho-2	fortes contributions	modalités dominantes
1,3,8	8.79	CS(1.41), F7(1.27) F8(0.94), C4(0.73) PR(0.67), GR(0.67) HE(0.62)	CS++, F7++ F8++, C4++ PR++, GR++ HE++
6,11,12,16	6.47	C3(1.27), SE(0.96) PR(0.55), C4(0.54)	C3++, SE+ PR+ , C4+
9,13,17,18	5.35	V1(1.05), 2(0.83) CS(0.59)	V1++, V2++ CS+,++
14,15	13.94	HE(2.24), F2(1.86) C4(1.68), CS(1.49) F7(1.41), PR(1.08)	HE+ , F2++ C4- , CS- F7++, PR+
7,10,20	8.08	PR(1.83), CS(1.47) FI(0.85)	PR- , CS-- FI--
2,4,5,19	6.13	F7(0.78), C4(0.77) V1(0.62)	F7--, C4-,-- V1-,--

Tab. 3 : Description des classes obtenues par découpage de la C.A.H. sur le tableau des modalités floues. Les quatre modalités de chaque variable sont notées : --, -, +, ++ (fortes et faibles diminutions et augmentations). Les sujets traités sont numérotés de 1 à 10 et les témoins de 11 à 20. ρ-2 est le carré de la distance au barycentre global ; son objet n'est pas de mesurer l'excentricité de la classe, qui n'a pas d'intérêt dans ce cas, mais d'évaluer l'importance des contributions des variables.

5 *Le tableau de Burt associé à un codage flou*

Il a déjà été dit au § 1 que l'analyse des correspondances du tableau logique disjonctif Z_d est équivalente à celle du tableau de Burt $B_d = {}^t Z_d \, Z_d$ quant à la description de l'ensemble des modalités J . Il est facile de vérifier que l'équivalence est aussi vraie pour l'ensemble I des observations si Z_d est ajouté en supplémentaire à B_d (cf. 1, pp 311 sqq ; voir aussi [*BIN. BURT*], C.A.D. Vol. II, n° 1 , pp 55 -71). Parallèlement, si Z_f est le tableau logique issu d'un codage flou, son A.C. est équivalente à celle de $B_f = {}^t Z_f \, Z_f$ avec Z_f ajouté en supplémentaire (*).

(*) Les indices d et f sont introduits là où il y a une possibilité de confusion entre tableaux issus du codage disjonctif ou flou.

B'_f n'est pourtant pas le tableau de Burt associé à Z_f, au moins si l'on admet sa définition comme juxtaposition des tableaux des marges binaires du tableau de contingence K à Q entrées.

$$K(j_1 \ldots j_Q) = \Sigma\{Z(i,j_1) \times \ldots \times Z(i,j_Q) \mid i \in I\}$$

les sous-tableaux qui se trouvent à la diagonale de B_f *étant des tableaux diagonaux contenant les marges simples de K.* (cf. 4).

Ce tableau est trouvé pareillement si l'on considère chaque individu composé de 2^Q morceaux ayant chacun un poids

$m = Z(i,j_1) \times \ldots \times Z(i,j_Q)$, appartenant à la case $(j_1 \ldots j_Q)$, ce qui nous place dans un schéma de codage disjonctif avec une pondération particulière, et l'on calcule le tableau de Burt B_f correspondant. B_f peut encore être calculé à partir de B'_f en rapportant à la diagonale la masse hors de la diagonale des blocs B'_{qq} qui se trouvent dans la diagonale de B'_f.

On a maintenant intérêt à comparer trois résultats : ceux des analyses de correspondances de $Z_d(B_d)$, $Z_f(B'_f)$, et B_f, ce que nous ferons par les corrélations des coordonnées factorielles des observations I, puisque l'ensemble des modalités J est différent au moins en Z_d et Z_f et les corrélations de leurs coordonnées factorielles n'ont pas de sens. Dans le cas de B_f ces coordonnées sont obtenues en ajoutant Z_f comme supplémentaire.

L'essai a été fait sur le tableau issu de l'expérience "venin Naja", et les corrélations obtenues ont été :

Facteurs Z_d \ Facteurs Z_f	1	2	3	4	5	6	7
1	391	-43	-210	-767	272	28	127
2	-22	-123	-794	205	43	272	41
3	35	-875	294	46	-49	117	-47
4	-94	129	116	290	675	72	169
5	759	-81	99	269	180	82	4
6	40	207	258	129	266	246	-288
7	-253	-180	-93	-225	317	-223	-644

Tab. 4 : Corrélations des facteurs calculés sur I dans les A. C. de Z_d *et* Z_f *(en millièmes).*

Facteurs Z_d	Facteurs B_f 1	2	3	4	5	6	7
1	454	-84	-212	-736	240	89	34
2	38	-75	-777	285	26	310	23
3	-95	-890	233	35	-65	106	-28
4	-114	139	147	229	689	175	91
5	709	-166	209	321	171	137	10
6	35	193	307	92	172	314	-311
7	-234	-141	-144	-199	254	-144	-733

Tab. 5 : Corrélations entre les premiers facteurs calculés sur I dans les A.C. de Z_d et B_f (en millièmes).

Facteurs Z_f	Facteurs B_f 1	2	3	4	5	6	7
1	986	-116	126	78	0	8	30
2	106	991	76	-13	-20	-13	13
3	-110	-56	985	-96	-23	-41	31
4	-67	13	77	989	57	13	68
5	3	13	19	-40	971	179	-158
6	-9	3	27	-3	-167	966	21
7	-11	-3	-23	-54	146	9	981

Tab. 6 : Corrélations entre les premiers facteurs calculés sur I dans les A.C. de B_f et Z_f (en millièmes).

On constate dans ces tableaux :

- Les A.C. de B_f et $B'_f(Z_f)$ donnent pratiquement le même résultat. En effet, les différences entre les trajectoires associées aux variables dans les deux analyses sont négligeables et les différences d'interprétation sont inexistantes.

- Chacun des cinq premiers facteurs de l'A.C. de B_d correspond à peu près à un facteur de l'A.C. de B_f (ou de $B_{f'}$), mais il y a un mélange important et l'ordre est complètement bouleversé.

On signalera finalement les valeurs propres, traces et contributions des premiers facteurs qui ont été obtenues dans les A.C. de B_d, B_f, et B'_f toujours pour l'expérience "venin Naja".

On y voit la fragilité des constatations faites au § 4.3 à propos des pourcentages d'inertie. Les traces dans le cas des tableaux de Burt ont d'ailleurs une interprétation plus claire que dans le cas des tableaux logiques (cf. 4 pp 148 sqq).

	1		2		3		4		5		trace
	v.p.	%	v.p.	%	v.p.	%	v.p.	%	v.p.	%	
B_d	109	18.7	101	36.0	72	48.4	54	57.6	47	65.7	0.583
B'_f	47	19.3	45	37.8	34	51.9	26	62.7	17	69.6	0.244
B_f	55	15.6	52	30.4	40	41.9	32	51.1	21	57.0	0.349

Tab. 7 : Valeurs propres (en millièmes), pourcentages cumulés d'inertie des premiers facteurs, et traces des l'A.C. des tableaux de Burt issus des données "venin Naja".

6 *Bibliographie*

(1) Benzécri J.P., Bastin Ch., Bourgarit Ch., Cazes P. - Pratique de l'Analyse des Données - Vol. 2 - Dunod - 1980.

(2) Benzécri J.P. - Sur l'analyse des tableaux binaires associés à une correspondance multiple - C.A.D. Vol. 2, n° 1-1977.

(3) Cazes P. - Etude de quelques propriétés extrémales des facteurs issus d'un sous-tableau d'un tableau de Burt - C.A.D. Vol. 2, n° 2 - 1977.

(4) Cazes P. - L'analyse de certains tableaux rectangulaires décomposés en blocs - C.A.D. Vol. 5 n°s 2 et 4 - 1980.

(5) Gallego F.J. - Un codage flou pour l'Analyse des Correspondances. Analyse des données des comptes trimestriels. - thèse de 3° cycle - Université Pierre et Marie Curie (Paris VI) - 1980.

(6) Guitonneau G.G., Roux M. - Sur la taxinomie du genre Erodium. - C.A.D. - Vol. 2 n° 1 - 1977.

(7) Lebart L., Morineau A., Tabard N. - Techniques de la description statistique - Dunod - 1977.

(8- Le Foll Y. - Sur les propriétés de l'Analyse des correspondances pour certaines formes complètes de données - Thèse 3° cycle - Université Pierre et Marie Curie (Paris VI) - 1979.

(9) Maïti D. - Programme d'homogénéisation et d'analyse d'un tableau de données hétérogènes - C.A.D. Vol. 4 n° 4 - 1979.