La Revue de Modulad

Une histoire de discrétisation

Gilles Celeux, Claudine Robert

INRIA Rocquencourt F78153 Le Chesnay
Université Joseph Fourier, TIMC-IRMA BP 53 X 38041 Grenoble

Résumé

On raconte les aventures de 23 papillons découpés de façon malencontreuse.

Mots-clés : Discrétisation de variables continues.

Beaucoup d'études statistiques portent sur une famille de variables dont certaines sont continues et d'autres irrémédiablement discrètes non ordinales. Un souci de cohérence nous invite et nous incite à transformer toutes les variables afin de se ramener à une seule catégorie ; dès lors, une seule solution : discrétisons les variables continues (et tentons, si cela est possible, de donner un caractère ordinal aux variables discrètes). Dans l'optique de discrétiser les variables, de nombreuses techniques de codage (cf. par exemple Celeux *et al.* 1989) ont été développées, qui permettent de mettre en œuvre l'analyse des correspondances multiples (ACM) sur les variables discrètes ainsi obtenues (cf. Lebart, Morineau et Tabard 1977 ou Escofier et Pages 1988).

Dans un autre domaine, en intelligence artificielle, notamment dans les systèmes experts et dans les réalisations de réseaux de neurones, discrétiser est devenu un réflexe inconditionnel. De plus, argument décisif, la conclusion d'une étude statistique ou celle d'un système expert ou d'un réseau de neurones est de nature discrète. Et puisqu'on en arrivera à du discret, pourquoi ne pas l'introduire d'emblée.... Mais c'est alors que de curieux phénomènes arrivent. Par exemple, dans de nombreux systèmes experts la fièvre (discrétisée bien sûr), la vitesse de sédimentation du sang.... sont considérées comme non suffisamment informatives pour être prises en compte réellement par le système. Pourtant, on prend toujours sa température quand on est malade. Se pourrait-il que, dans certains cas, l'information perdue lors du codage sous forme discrète soit justement celle qui est pertinente?

Pour illustrer ce propos, voici une anecdote. Un soir de l'été 1985, des enfants d'une colonie de vacances ont attrapé 23 papillons. Pour occuper un jour de pluie, on leur a demandé de dessiner ces papillons grandeur nature et, pour cela, de mesurer les longueurs z_1, z_2, z_3 et z_4 indiquées sur la figure 1. Après

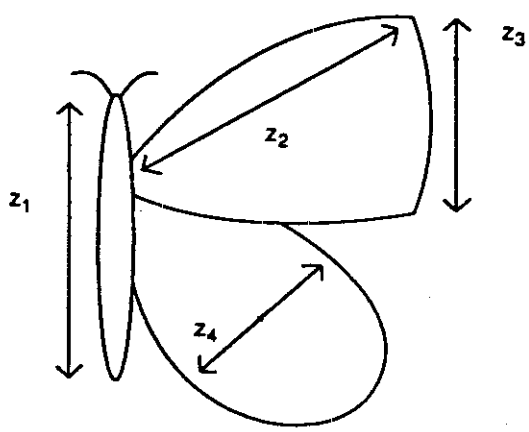

FIG. 1 - *Les 4 longueurs mesurées.*

la séance de mesures, il ne restait, hélas !, plus des papillons que des corps par ci, des antennes par là et des cassures d'ailes. Il ne subsistait en fait que le tableau des mesures (tableau 1), et des indications sur le lieu de capture,

	z_1	z_2	z_3	z_4
1	22	35	24	19
2	24	31	21	22
3	27	36	25	15
4	27	36	24	23
5	21	33	23	18
6	26	35	23	32
7	27	37	26	15
8	22	30	19	20
9	25	33	22	22
10	30	41	28	17
11	24	39	27	21
12	29	39	27	17
13	29	40	27	17
14	28	36	23	24
15	22	36	24	20
16	23	30	20	20
17	28	38	26	16
18	25	34	23	14
19	26	35	24	15
20	23	37	25	20
21	31	42	29	18
22	26	34	22	21
23	24	38	26	21

TAB. 1 - *Les données disponibles.*

le temps ce jour-là..., qui étaient les mêmes pour les 23 papillons. Un moniteur nous a posé la devinette suivante : combien de "sortes" de papillons y avait-il ? Il avait fait des histogrammes (figure 2) et s'étonnait de n'y rien voir (car lui savait combien de sortes de papillons avaient été capturés).

Nous avons réalisé une analyse en composantes principales (ACP) non normées du tableau de données. (Toutes les analyses factorielles ont été réalisées avec le logiciel SPAD.N, cf. Morineau 1991.) La projection sur le premier plan principal (figure 3) permet de soupçonner l'existence de 3 groupes qui est confirmée par l'examen du 3^e axe factoriel (figure 4). La conclusion de cette ACP est claire : il y a 3 groupes de papillons et un papillon à part, le n° 6. Le retour au tableau des données indique une valeur très élevée de la quatrième variable pour cet élément. Connaissant la méthode de recueil des données, disons qu'il doit s'agir d'une erreur de mesure... Pour conforter notre idée de 3 groupes, faisons une ACP qui supprime l'effet taille, c'est-à-dire où les nouvelles variables sont $z_1/z_4, z_2/z_4$ et z_3/z_4. La figure 5 représentant le premier plan principal de cette ACP va dans le même sens : 3 groupes et un cas isolé. Le moniteur de la colonie de vacances a confirmé les trois "sortes" de papillons... Et peut-être est-il maintenant statisticien ?

Essayons maintenant de voir ce qui se passe si on discrétise les variables. Après tout, une sorte de papillon, dans le cadre de nos données numériques, est une forme et autour de cette forme, il y a des variations (et des mesures imprécises) : bref, on a une structure de type *forme* + *bruit*. Comment se transforme cette structure par codage discret ?

Pour découper les 4 variables continues, nous avons utilisé l'algorithme de Fisher (1958) qui permet de trouver les D intervalles (D étant fixé) d'inertie

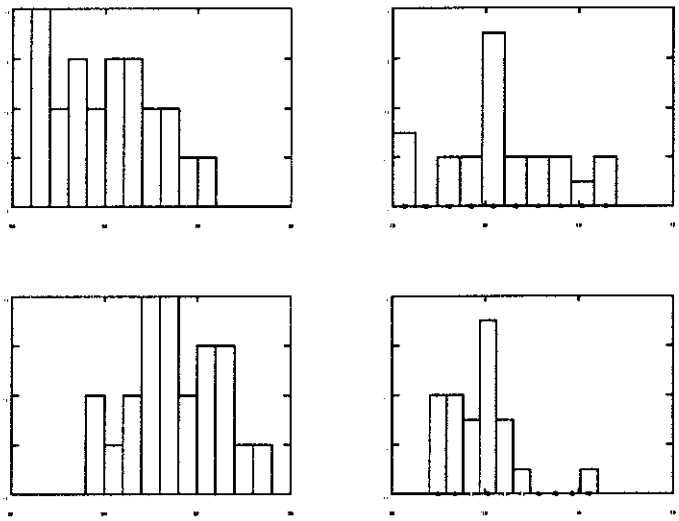

FIG. 2 - *Les histogrammes.*

intraclasse minimum pour les données considérées. Pour les 4 variables, nous avons considéré $D = 3$ intervalles.

Le tableau 2 donne pour chaque variable la valeur des bornes des intervalles et le tableau 3 donne les effectifs des classes associées.

z_1	z_2	z_3	z_4
≤ 24	≤ 34	≤ 23	≤ 17
de 25 à 27	de 35 à 37	de 24 à 25	de 18 à 20
> 27	> 37	> 25	> 20

TAB. 2. *Les bornes des intervalles de discrétisation.*

z_1	z_2	z_3	z_4
9	7	9	8
8	9	6	7
6	7	8	8

TAB. 3. *Les effectifs des intervalles de discrétisation.*

Au vu de ces deux tableaux, on peut penser que d'autres découpages classiques du type découpage à intervalles égaux ou découpage à effectifs égaux conduiraient à des codages peu différents.

Une analyse des correspondances multiples a été faite à partir du codage disjonctif complet associé à notre découpage. Les figures 6 et 7 représentent les projections de cette ACM sur les plans factoriels (1-2) et (1-3). On voit des "groupes" tels {10, 12, 13, 17}, {1, 15, 20}, {5, 8, 16} (qui sont, en fait, des ensembles de points ayant le même vecteur de codage), mais aucune partition ne se dégage. On peut remarquer, sur les figures 8 et 9

où est indiquée l'appartenance des éléments aux groupes (notés A, B et C) dégagés par l'ACP que

- les petits groupes {10, 12, 13, 17}, {1, 15, 20} et {5, 8, 16} sont des sous-groupes de l'ACP, mais que les groupes de l'ACP ne peuvent pas être trouvés par l'ACM. En fait, aucune partition des 23 points ne se dégage des figures 6 et 7.
- le point 6 n'est pas du tout exceptionnel dans le nuage de points. Cela était prévisible : la variable z_4 ayant été partagée en classes d'effectifs comparables, on a gommé la particularité du point 6 (d'ailleurs, que 6 soit inclus dans l'analyse ou mis en point supplémentaire ne change pas l'allure des cartes obtenues par cette ACM).

On peut penser que la situation va s'améliorer si l'on adopte un codage additif où les points tombant dans le premier intervalle sont codés (1, 0, 0) comme dans le codage disjonctif complet, mais où les points tombant dans le deuxième intervalle sont codés (1, 1, 0) et où ceux tombant dans le dernier intervalle sont codés (1, 1, 1). Ainsi, le codage additif conserve une structure ordinale aux données. Effectivement, comme on le voit sur les figures 12 et 13, les points d'un même groupe peuvent à peu près être séparés par des "patates". Mais, comme on peut en juger par l'examen des figures 10 et 11, représentant les premiers plans factoriels de l'ACM sans marquer les groupes, ce codage additif ne conduit pas plus que le précédent à une partition naturelle des éléments.

La perte d'information par les codages discrets précédents pouvait être due au petit nombre de points traités. Aussi, nous avons imaginé que chacun des 23 points était une classe de N points de variables plus finement mesurées. Par exemple, la valeur 21 de la variable z_1 se trouve être la classe [20.5, 21.5[. Le codage discret des 4 variables a ainsi autant de modalités qu'il y a de valeurs distinctes dans le tableau de données. Cela donne 11 modalités pour z_1 et z_3 et 12 modalités pour z_2 et z_4. Mais, on peut constater sur les figures 14 et 15 représentant les plans factoriels (1-2) et (1-3) de l'ACM à partir du codage disjonctif complet, que les groupes sont complétement mélangés et que le point 6 ne ressort pas en dehors du nuage....

Dans l'exemple présenté, la discrétisation a perdu ce qui nous paraît être la structure fondamentale du nuage des 23 points, à savoir l'existence très apparente d'une partition en 3 groupes et un point isolé. Bien sûr, avec ces seules données toutes quantitatives, personne n'aurait eu l'idée de discrétiser les variables. Mais, imaginons que, pour aider à la construction d'un système expert de reconnaissance de certaines espèces de papillons qui se ressemblent tous quant à leurs couleurs, rayures, etc. on ait disposé d'un fichier de données portant les valeurs des variables z_1, z_2, z_3 et z_4 ainsi que d'autres variables continues et qualitatives (heures de vol dans la journée, température du jour de capture, nombre d'excroissances aux ailes inférieures, nombre de rayures -si rayures il y a-, présence de taches et nombre de couleurs sur les ailes, type de végétation du lieu d'habitat....). Alors, dans l'optique système expert, ces variables auraient été recodées sous forme discrète et peut-être,

comme pour le cas de la fièvre cité en introduction, la plupart d'entre elles auraient perdu leur valeur informative.

D'une manière plus générale, notre pratique nous a montré que, dans le domaine de la médecine, pour y voir plus clair dans un fichier de données, il était souvent avantageux de garder les variables dans leur forme originelle le plus longtemps possible (cf. Robert 1991). Plus précisément, dans un premier temps, nous regroupons toutes les variables continues et étudions ce "sous-fichier continu". Soit il ne sort rien de particulier de ces analyses sur les variables continues, et alors, compte tenu de cette vérification, dans la majorité des cas, nous envisageons de discrétiser chaque variable en ayant le souci de guider le choix du codage par des informations a priori ou des informations issues des analyses du sous-fichier continu. Il arrive aussi que nous tentions de regrouper de nombreuses variables qualitatives en plusieurs scores ordinaux : nous réessayons alors des analyses "continues" incorporant ces scores, etc. Par contre, si les analyses des variables continues reflètent clairement une structure des données pertinente pour la problématique envisagée, il convient de ne pas perdre cette information. Pour cela, on peut par exemple discrétiser quelques composantes principales ou discriminantes. Mais bien sûr, chaque domaine étudié et chaque problématique poursuivie amène une pratique particulière de l'analyse des données. Notre objectif est ici de montrer qu'une segmentation systématique en début d'analyse peut être préjudiciable. Bien sûr, cela ne prouve pas qu'il ne soit pas avantageux dans certains domaines de discrétiser les variables continues en tout premier lieu.

Bibliographie

[1] Celeux G., Diday E., Govaert G., Lechevallier Y. et Ralambondrainy H. (1989) *Classification automatique des données*. Dunod, Paris.

[2] Escofier B. et Pagès J. (1988) *Analyses factorielles simples et multiples*. Dunod, Paris.

[3] Fisher W. D. (1958) On grouping for maximum homogeneity. *JASA* **53**, 789-798.

[4] Lebart L., Morineau A. et Tabard N. (1977) *Techniques de la description statistique*. Dunod, Paris.

[5] Morineau A. (1991) SPAD.N logiciel pour l'analyse statistique des données. *La Revue de Modulad* **6**, 27-59.

[6] Robert C. (1991) *Modèles statistiques pour l'intelligence artificielle*. Masson, Paris.

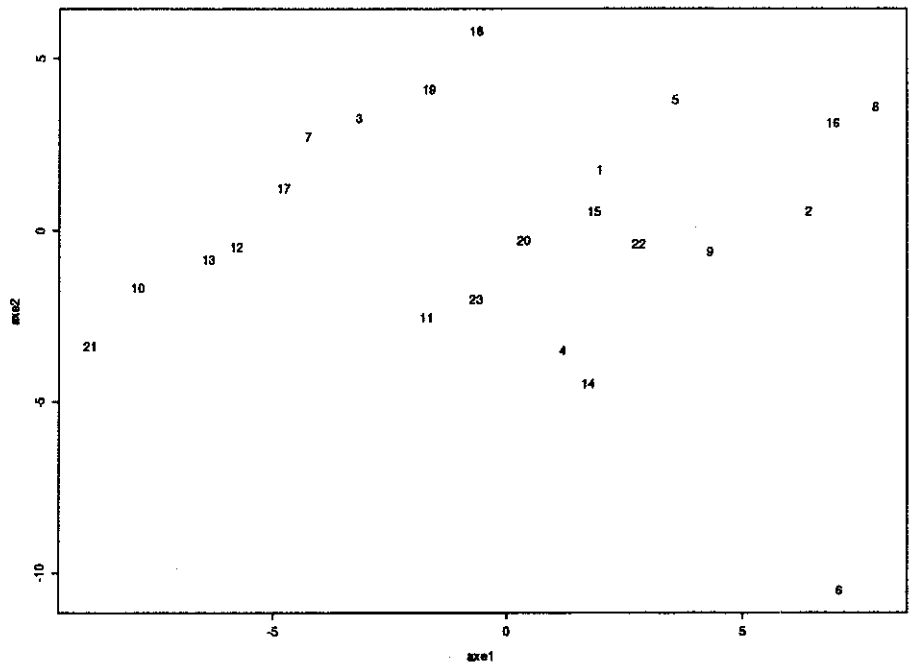

FIG. 3 - *Le plan (1-2) de l'ACP non normée.*

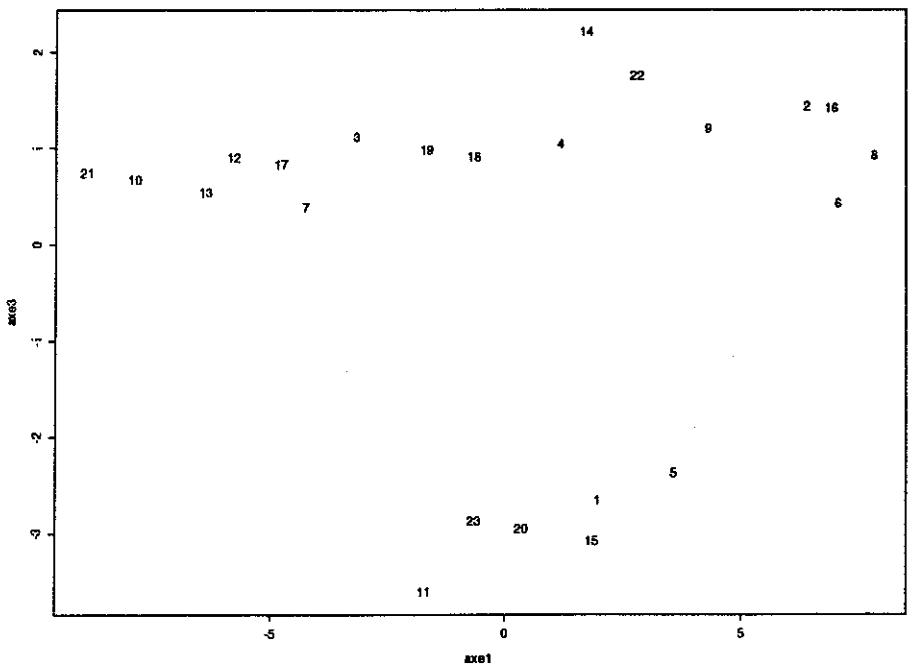

FIG. 4 - *Le plan (1-3) de l'ACP non normée.*

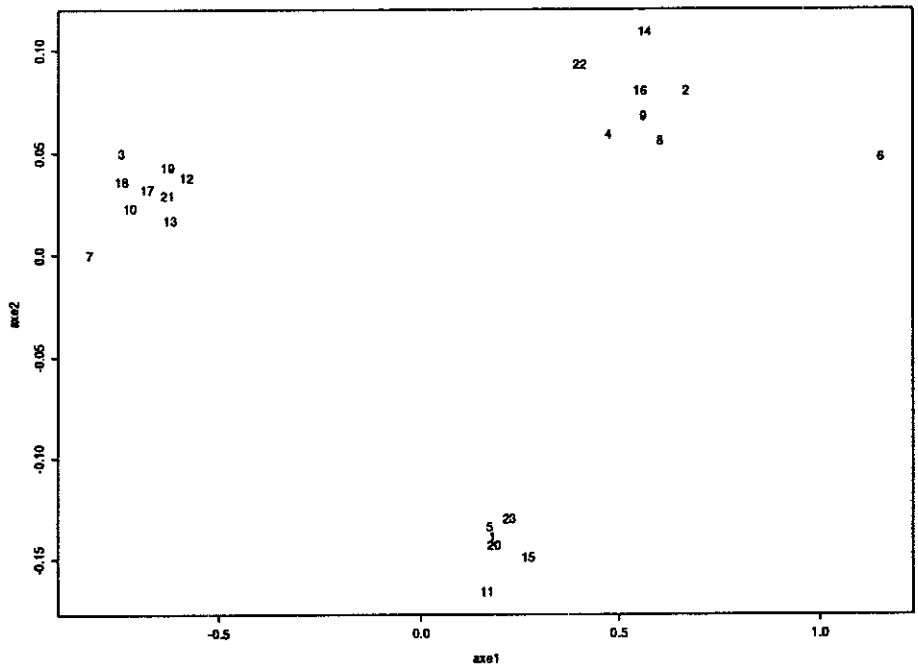

FIG. 5 - *Le plan (1-2) de l'ACP des trois variables transformées.*

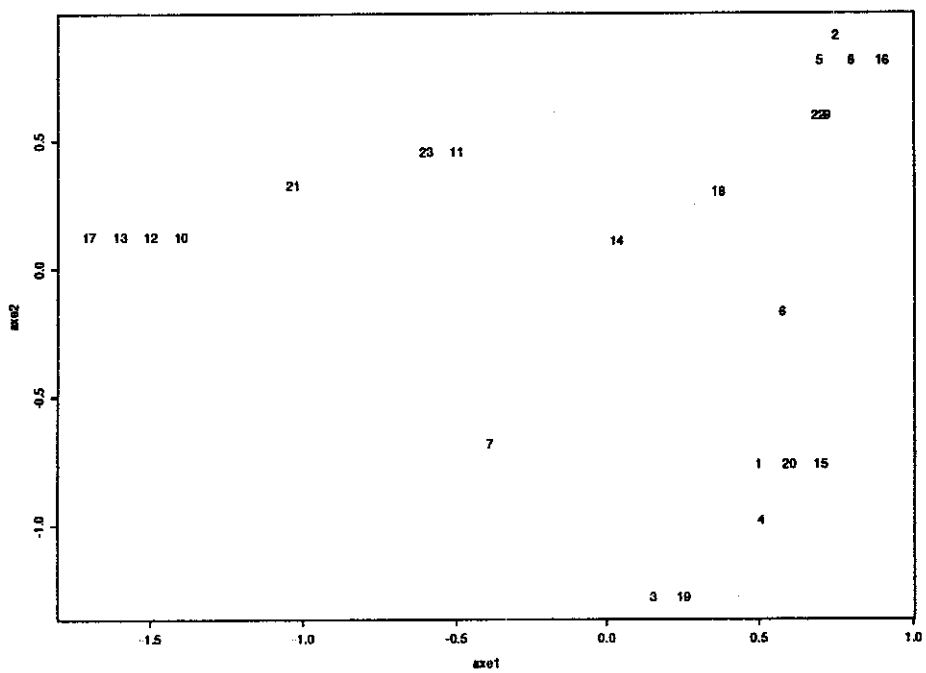

FIG. 6 - *Le plan (1-2) de l'ACM sur le codage disjonctif complet.*

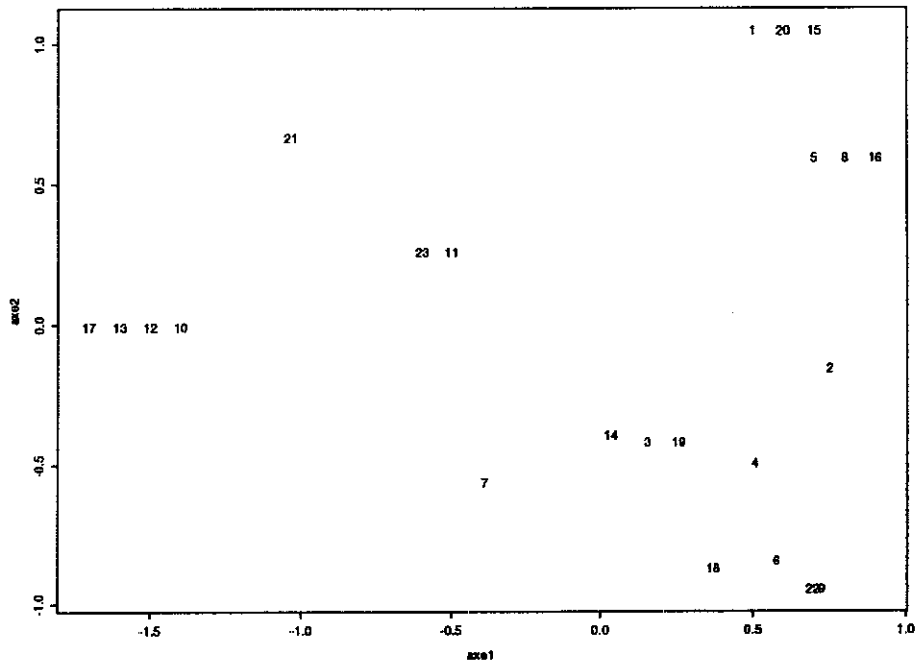

FIG. 7 - *Le plan (1-3) de l'ACM sur le codage disjonctif complet.*

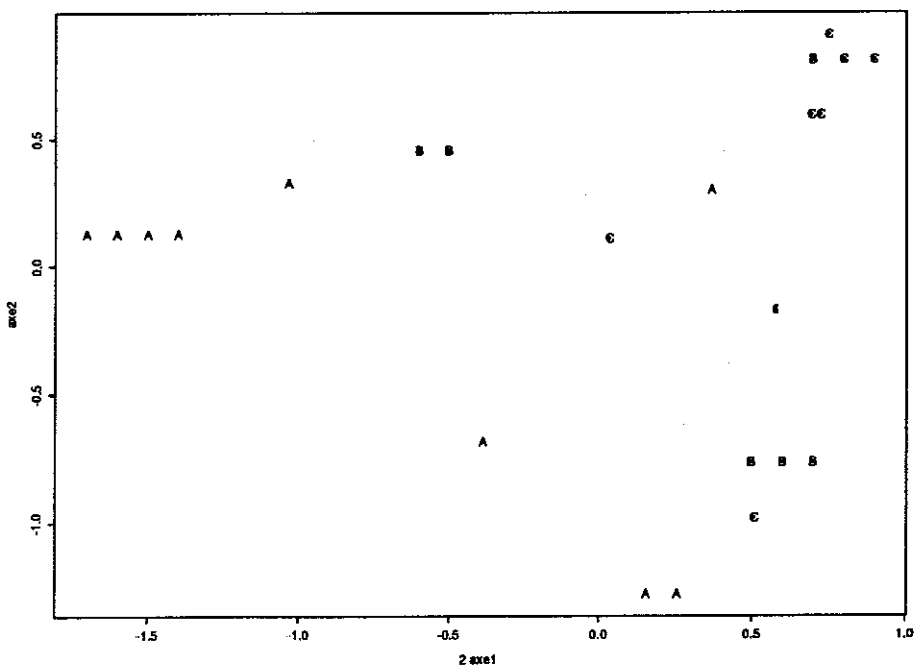

FIG. 8 - *Le plan (1-2) de l'ACM sur le codage disjonctif complet avec marquage des 3 groupes.*

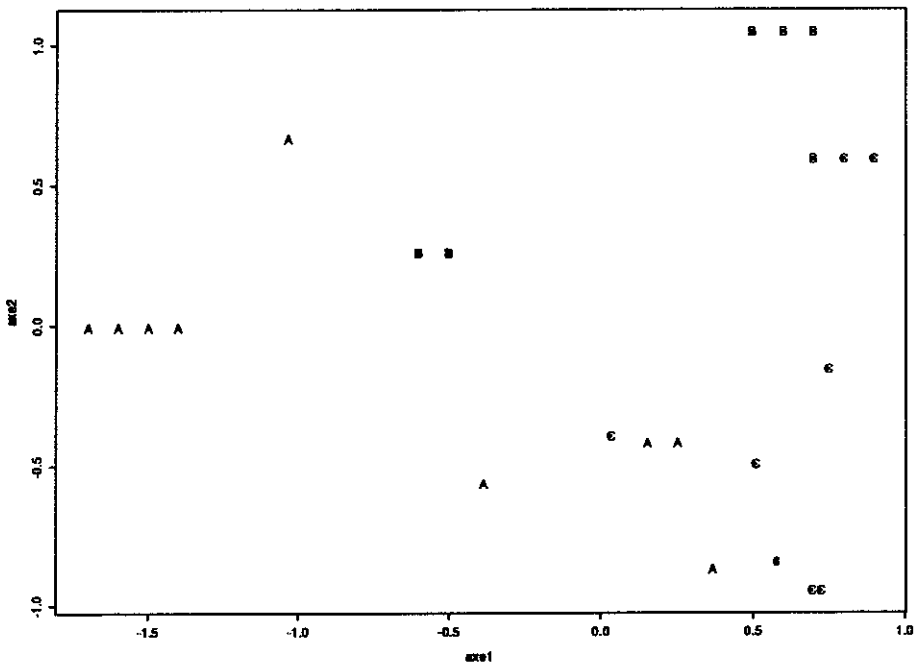

FIG. 9 - *Le plan (1-3) de l'ACM sur le codage disjonctif complet avec marquage des 3 groupes.*

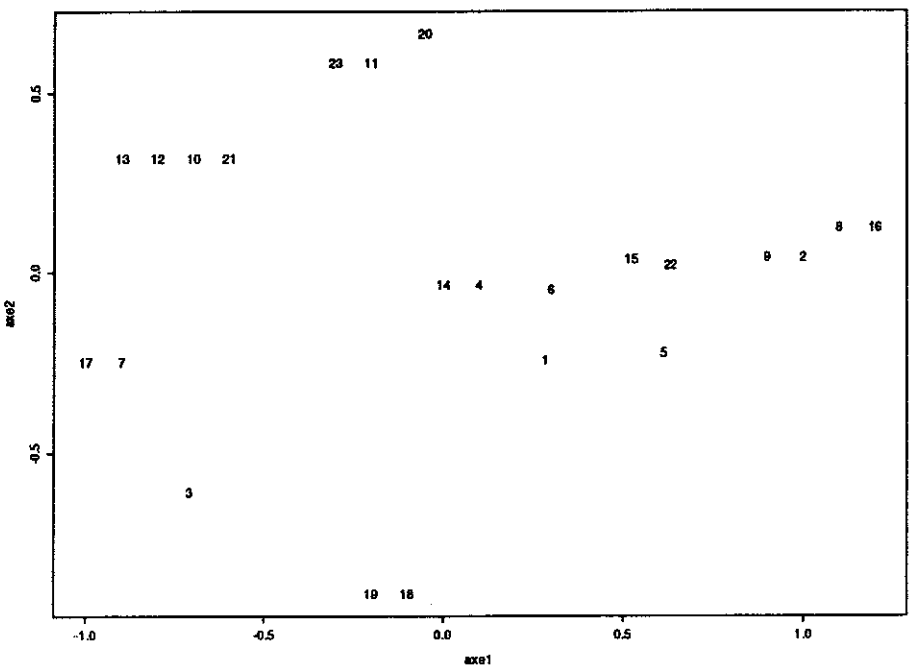

FIG. 10 - *Le plan (1-2) de l'ACM sur le codage additif.*

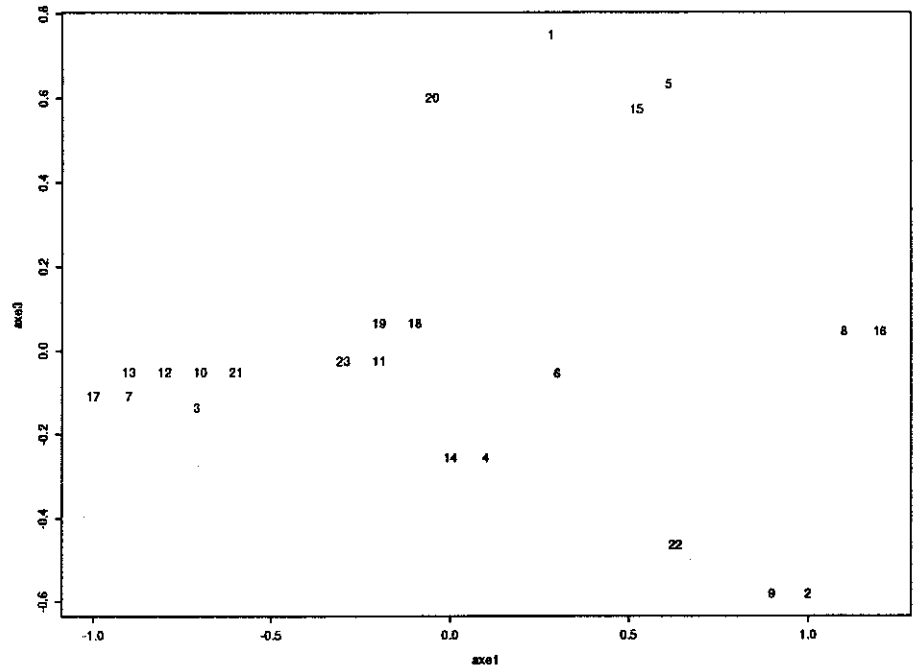

FIG. 11 - *Le plan (1-3) de l'ACM sur le codage additif.*

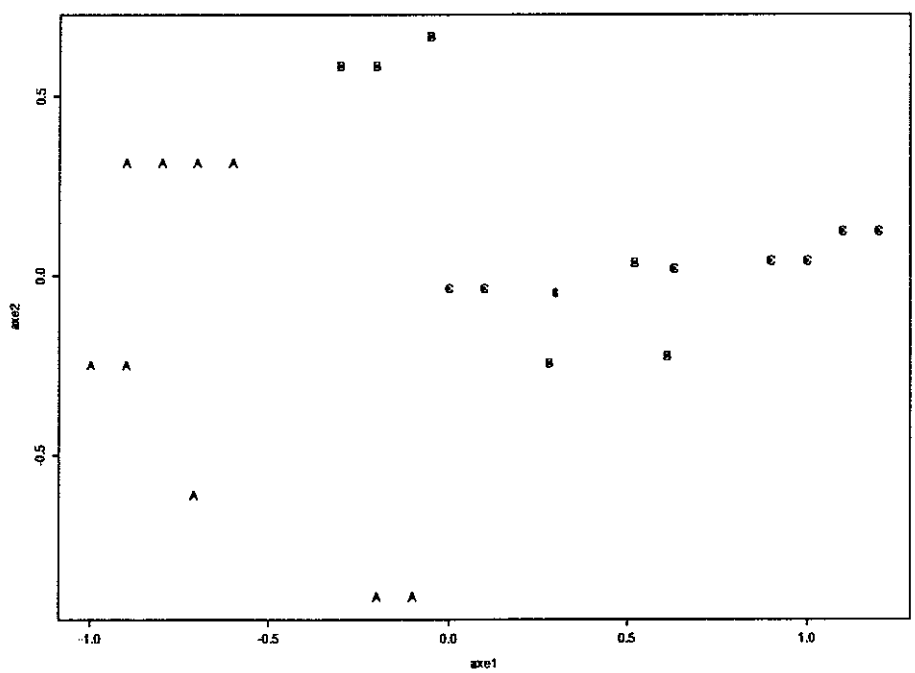

FIG. 12 - *Le plan (1-2) de l'ACM sur le codage additif avec marquage des 3 groupes.*

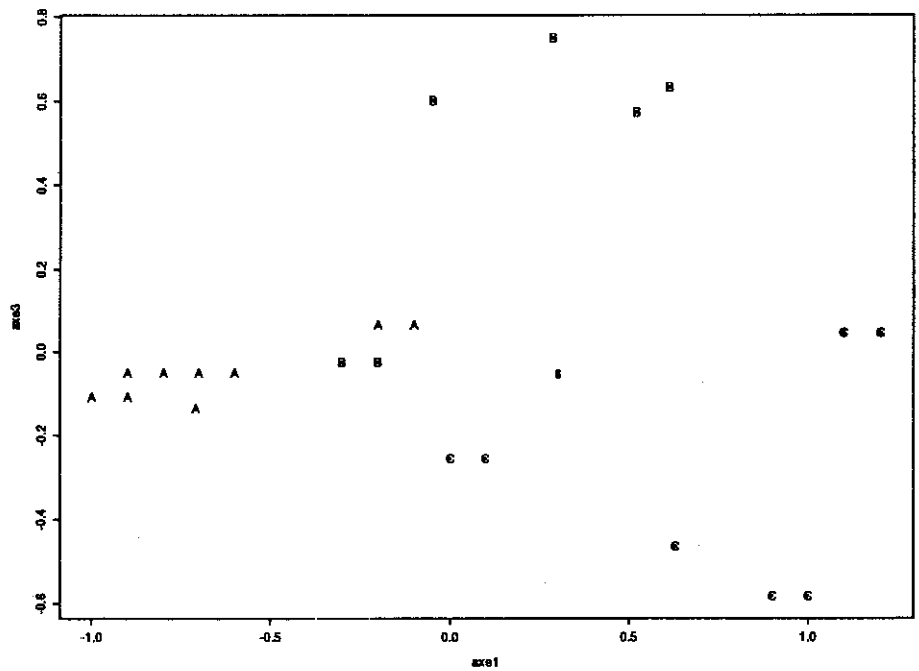

FIG. 13 - *Le plan (1-3) de l'ACM sur le codage additif avec marquage des 3 groupes.*

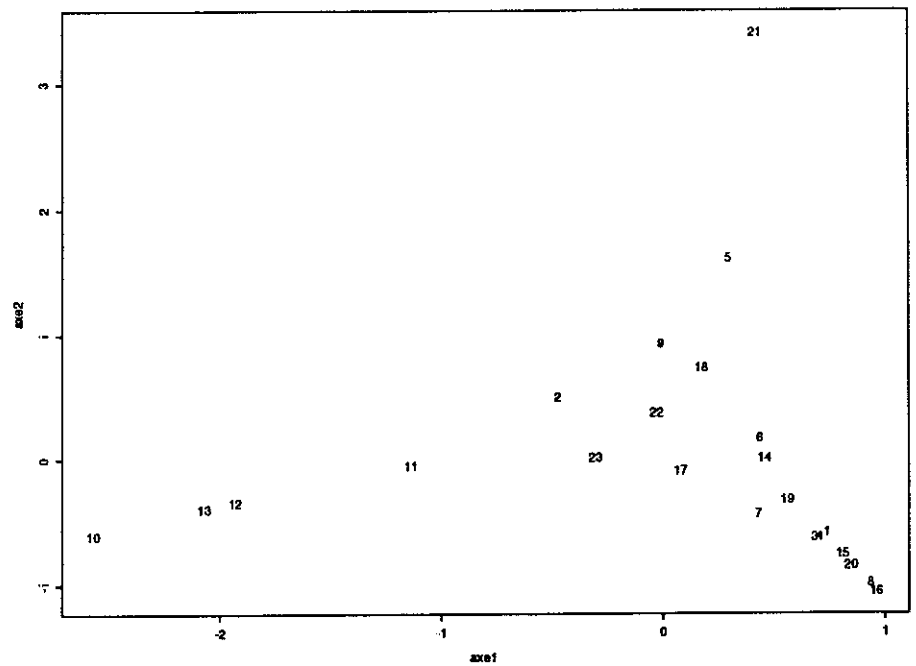

FIG. 14 - *Le plan (1-2) de l'ACM sur les données initiales considérées qualitatives.*

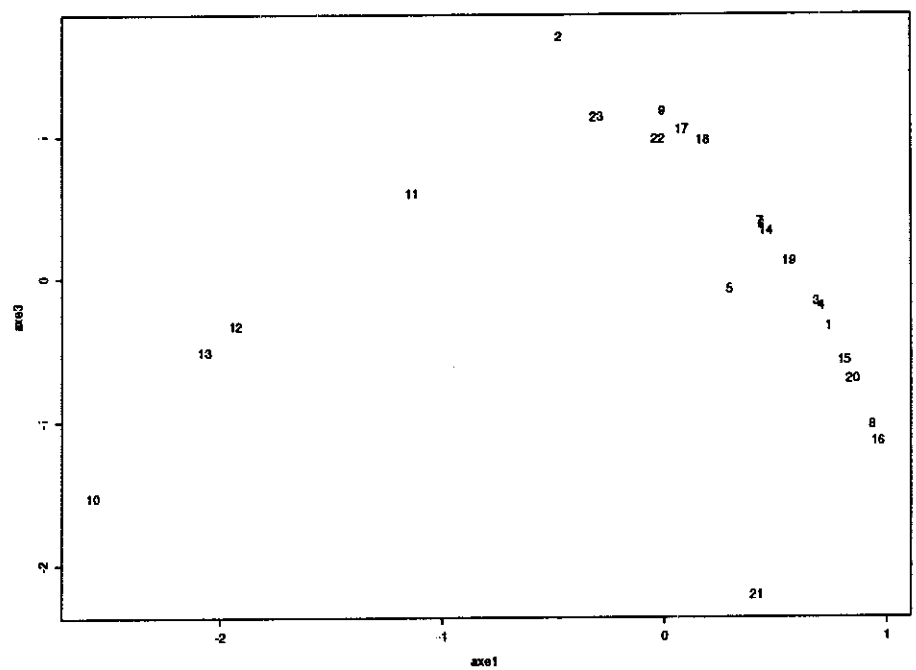

FIG. 15 - *Le plan (1-3) de l'ACM sur les données initiales considérées qualitatives.*