REVUE DE STATISTIQUE APPLIQUÉE

H. CAUSSINUS
A. DE FALGUEROLLES

Tableaux carrés : modélisation et méthodes factorielles

Revue de statistique appliquée, tome 35, nº 3 (1987), p. 35-52
<http://www.numdam.org/item?id=RSA_1987__35_3_35_0>

© Société française de statistique, 1987, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

Rev. Statistique Appliquée, 1987, **XXXV** (3), 35-52

TABLEAUX CARRÉS : MODÉLISATION ET MÉTHODES FACTORIELLES

H. CAUSSINUS et A. de FALGUEROLLES

Laboratoire de Statistique et Probabilités, UA CNRS 745
Université Paul Sabatier,
118, route de Narbonne, 31062 Toulouse Cedex, France

RÉSUMÉ

Dans cet article, la complémentarité des méthodes factorielles et de modélisation est étudiée pour des tableaux carrés. Les modèles considérés sont la quasi-indépendance et la quasi-symétrie (cas additif et cas multiplicatif). Les méthodes de l'analyse des données sont utilisées aussi bien pour explorer la structure des résidus que pour produire des représentations graphiques des paramètres issus de la modélisation.

Mots clés : *Tableaux carrés, Quasi-indépendance, Quasi-symétrie, Représentations graphiques, Classification.*

SUMMARY

In this article, the complementarity of factorial and modelling methods is investigated in the context of square tables. The models which are under consideration are quasi-independence and quasi-symmetry (additive and multiplicative cases). Data analysis methods are used either to investigate the structure of residuals or to produce graphical representations for the parameters introduced by modelling.

1. Introduction

Nous considérons ici une démarche spécifique à l'analyse descriptive des tableaux carrés. Les développements que nous donnons en nous appuyant sur deux exemples cherchent à montrer la complémentarité des diverses approches statistiques dans un contexte où celle-ci nous semble particulièrement frappante. On verra en effet que la *modélisation* fournit des paramètres (estimés) dont une étude descriptive se fait très bien par les *techniques de « l'analyse des données »* (classification, représentation euclidienne). D'autre part, les résidus obtenus par rapport au modèle choisi sont utilement analysés par une *technique exploratoire* proche de l'analyse des correspondances.

Ce dernier point appelle un commentaire d'ordre général. Si un modèle était pour nous soit exact soit à rejeter, il est clair que, par définition, les résidus ne traduiraient que des bruits sans intérêt; leur analyse n'aurait pas alors grand sens sauf, peut-être, pour confirmer le modèle lorsqu'elle ne révèlerait justement aucune structure de ces résidus (on en verra des exemples ci-dessous). Mais nous ne croyons pas qu'un modèle soit utile seulement s'il est exact (H. CAUSSINUS (1985)). En d'autres termes, le modèle peut n'être qu'un modèle virtuel (J.R. BARRA (1985)) : il est destiné à mettre en évidence et à quantifier les traits les

plus importants des données selon la conception la plus communément admise par les statisticiens britanniques (voir par exemple L.S.P.-C.A.S. (1986), paragraphe 1.2.). Dans ces conditions, une fois nos idées a priori les plus nettes exprimées dans la construction même du modèle, il est intéressant de tenter une analyse exploratoire des résidus pour chercher ce que les données pourraient bien encore recéler...

2. Rappels et généralités

Par tableau carré $\ell \times \ell$ nous entendons un tableau de nombres structurellement carré, c'est-à-dire dans lequel les modalités des lignes et celles des colonnes sont identiques; il peut comporter ou non des éléments diagonaux, l'absence de diagonale perturbant souvent les analyses classiques (voir par exemple B. BURTSCHY (1984)).

2.1. Quasi-indépendance

La notion de quasi-indépendance (ou si l'on veut de quasi absence d'interaction) peut être définie comme la condition d'existence de réels m, a (i), b (j), tels que, x (i, j) désignant le terme général du tableau considéré :

$$x(i,j) = m + a(i) + b(j) \qquad i,j = 1,\dots,\ell \qquad i \neq j \qquad 2.1.a$$

ou bien

$$x(i,j) = m \cdot a(i) \cdot b(j) \qquad i,j = 1,\dots,\ell \qquad i \neq j \qquad 2.1.b$$

avec, dans ce second cas, $m, a(i), b(j) > 0$

pour lesquels on imposera les contraintes suivantes (conditions d'identifiabilité) :

— cas additif

$$\sum_i a(i) = \sum_j b(j) = 0$$

— cas multiplicatif

$$\sum_i a(i) = \sum_j b(j) = 1$$

Les contraintes (2.1.a) ou (2.1.b) ne concernent pas les termes diagonaux du tableau qui peuvent être absents ou négligés dans l'analyse proposée.

2.2. Quasi-symétrie

La quasi-symétrie peut être définie comme la condition d'existence de réels (positifs dans 2.2.b) m, a (i), b (j), s (i, j) = s (j, i) tels que :

$$x(i,j) = m + a(i) + b(j) + s(i,j) \qquad i,j = 1,\dots,\ell, \qquad 2.2.a$$

ou

$$x(i,j) = m \cdot a(i) \cdot b(j) \cdot s(i,j) \qquad i,j = 1,\dots,\ell, \qquad 2.2.b$$

pour lesquels on imposera les contraintes suivantes (conditions d'identifiabilité) :

— cas additif

$$\sum_i a(i) = \sum_j b(j) = \sum_{\substack{j \\ j \neq i}} s(i, j) = \sum_{\substack{i \\ i \neq j}} s(i, j) = 0$$

— cas multiplicatif

$$\sum_i a(i) = \sum_j b(j) = \sum_{\substack{j \\ j \neq i}} s(i, j) = \sum_{\substack{i \\ i \neq j}} s(i, j) = 1$$

Là encore les termes diagonaux peuvent être absents ou négligés.

Remarquons tout d'abord que d'autres formulations équivalentes d'un tableau quasi-symétrique sont encore possibles, les paramètres en jeu étant dans chaque cas définis de façon unique (cf. pour le cas multiplicatif H. CAUSSINUS (1976)). Cette flexibilité des écritures possibles d'un tableau quasi-symétrique peut être un handicap dans certains cas, mais elle peut aussi s'avérer intéressante dans la mesure où elle permet une adaptation à la problématique des données considérées. Notons aussi l'utilisation d'un modèle de quasi-symétrie pour l'analyse des matrices de confusion (modèle de LUCE : cf. W. HEISER (1985)).

Enfin on rappelle que les modèles multiplicatifs de quasi-indépendance et de quasi-symétrie sont souvent appelés log-linéaires pour des raisons évidentes.

2.3. Analyse descriptive de tableaux carrés

En général, le modèle log-linéaire (b) est préférable pour des fréquences, le modèle linéaire (a) pour des mesures, notes,... Ces modèles peuvent se généraliser de plusieurs façons pour des tableaux carrés stratifiés (cf. H. CAUSSINUS et C. THELOT (1976), H. CAUSSINUS et A. de FALGUEROLLES (1986)).

Bien entendu, une condition telle que (a) ou (b) n'est pas en général vérifiée par un tableau de données, mais correspond parfois à une approximation très satisfaisante.

Lorsque c'est le cas, l'approche du tableau des données par un tableau vérifiant (a) ou (b) permet d'introduire des quantités souvent interprétables. En particulier, s (i, j) ou une fonction croissante de s (i, j) s'interprète comme un indice de similarité entre les modalités i et j; de même une fonction décroissante de s (i, j) peut s'interpréter comme un « indice de dissemblance » de i et j. Des méthodes de positionnement multidimensionnel permettent d'obtenir une représentation graphique des modalités cohérente avec les valeurs de ces indices. Des méthodes de classification automatique peuvent aussi être appliquées.

Si le modèle ne s'ajuste qu'assez imparfaitement aux données, l'étude des écarts à ce « modèle virtuel » peut être effectuée par des méthodes factorielles parfois présentées comme des variantes de l'analyse factorielle des correspondances (cf. D. DOMENGES et M. VOLLE (1980), B. ESCOFIER (1984) et P. van der HEIJDEN (1985)).

La mise en œuvre de ces idées sera maintenant concrétisée sur deux exemples.

3. « Sociomatrice de Thomas »

3.1. Données, problème

Nous considérons ici les données de la sociomatrice de Thomas publiées dans F. CAILLIEZ et J.P. PAGES (1976) et étudiées dans cet ouvrage (pages 294 à 302) par des méthodes usuelles d'Analyse en Composantes Principales. Elles sont constituées des notes, entre 0 et 20, attribuées par les 24 élèves d'une classe à chacun de leurs camarades. Il s'agit donc d'un tableau carré sans éléments diagonaux.

L'objectif essentiel de l'analyse statistique est de réaliser un bilan des affinités exprimées par ces notes.

Rappelons d'abord la méthode d'analyse proposée par F. CAILLIEZ et J.P. PAGES. Les données diagonales, structurellement absentes, sont posées égales à 20. Sont alors effectuées deux A.C.P. centrées réduites; celle des notes données et celle des notes reçues.

Malgré son intérêt et sa simplicité cette démarche peut susciter certaines réserves. La première, un peu formelle, concerne le caractère arbitraire de la valeur affectée à la diagonale ou plutôt le fait même de fabriquer une donnée là où, structurellement, il n'y en a pas. La seconde, plus fondamentale, concerne l'objectif même de l'analyse : dans l'A.C.P. la proximité de deux individus dépend très lourdement de la similarité des jugements qu'ils portent sur (ou qui sont portés sur eux par) les 22 autres, alors que le jugement qu'ils portent l'un sur l'autre n'intervient qu'assez faiblement. Dans la démarche que nous proposons, et qui pourrait au moins être complémentaire, nous cherchons d'abord à voir s'il y a une réciprocité de ces deux jugements puis à utiliser celle-ci pour mesurer très simplement les affinités individuelles.

3.2. Modèles

On considère ici des modèles additifs pour les notes; la note $x(i, j)$ attribuée par l'élève i ($i = 1, \ldots, \ell$) à l'élève j ($j = 1, \ldots, \ell$ et $j \neq i$) s'écrit :

— pour le modèle de quasi-indépendance

$$x(i, j) = m + a(i) + b(j) + e_1(i, j)$$

— pour le modèle de quasi-symétrie

$$x(i, j) = m + a(i) + b(j) + s(i, j) + e_2(i, j)$$

où m apparaît comme la note moyenne, $a(i)$ l'effet du « notant » i, $b(j)$ l'effet du « noté » j et $s(i, j) = s(j, i)$ la ressemblance « intrinsèque » entre l'élève i et l'élève j. Enfin les $e_1(i, j)$ (ou $e_2(i, j)$) sont des termes aléatoires d'erreur supposés indépendants et de même loi.

Il est assez facile de vérifier que les estimateurs de moindres carrés des paramètres ont, avec des notations évidentes, les expressions suivantes :

$$\hat{m} = \frac{1}{\ell(\ell - 1)} x(+, +)$$

$$\hat{a}(i) = \frac{(\ell - 1)\, x(i, +) + x(+, i) - x(+, +)}{\ell(\ell - 2)}$$

$$\hat{b}(j) = \frac{(\ell - 1)\, x(+, j) + x(j, +) - x(+, +)}{\ell(\ell - 2)}$$

$$\hat{s}(i, j) = \frac{\hat{e}_1(i, j) + \hat{e}_1(j, i)}{2}$$

avec :

$$\hat{e}_1(i, j) = x(i, j) - \hat{m} - \hat{a}(i) - \hat{b}(j)$$

On notera enfin :

$$\hat{e}_2(i, j) = x(i, j) - \hat{m} - \hat{a}(i) - \hat{b}(j) - \hat{s}(i, j)$$

3.3. Propriété spécifique des résidus

Considérons les tableaux carrés E_1, E_2 et S des résidus du modèle de quasi-indépendance, des résidus du modèle de quasi-symétrie et des paramètres de ressemblance s (i, j). Nous affectons la valeur 0 aux termes diagonaux absents (ils n'interviennent pas dans l'analyse mais cela permet d'alléger certaines écritures). La matrice E_1 est la somme de deux matrices l'une symétrique (le tableau S) et l'autre antisymétrique (le tableau E_2). On vérifie aisément

$$\sum_i \sum_j \hat{e}_1(i, j)^2 = \sum_i \sum_j s(i, j)^2 + \sum_i \sum_j \hat{e}_2(i, j)^2 .$$

En suivant J.C. GOWER (1977) et Y. ESCOUFIER et A. GRORUD (1980), les matrices S et E_2 peuvent être analysées par des méthodes factorielles spécifiques exploitant la symétrie de l'une et l'antisymétrie de l'autre.

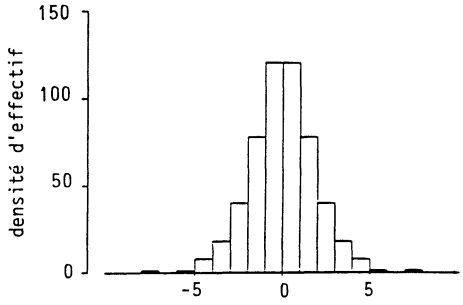
3.4. Ajustement et analyse des résidus

Le tableau suivant rend compte de la qualité globale de l'ajustement des modèles considérés en 3.2.

Modèle	Nombre de paramètres indépendants	ddℓ	Somme des carrés des résidus
m	1	551	10 583
m + a (i) + b (j)	47	505	7 120
m + a (i) + b (j) + s (i, j)	299	253	1 756

L'étude de la répartition des 23 × 24 = 552 résidus du modèle de quasi-symétrie ainsi que celle des résidus ventilés par élève noté (cf. Fig. 1) indique un bon ajustement, encore que l'écart type estimé soit grand (2,6 environ) et que la symétrie de l'histogramme de la figure 1a soit artificielle car inhérente au modèle choisi (le tableau E_2 est antisymétrique). Les résidus les plus grands 7,33 et 5,73

a) histogramme des 24 x 23 résidus $\hat{e}_2(i,j)$



b) répartition des résidus par élève noté ("box-plots" parallèles)

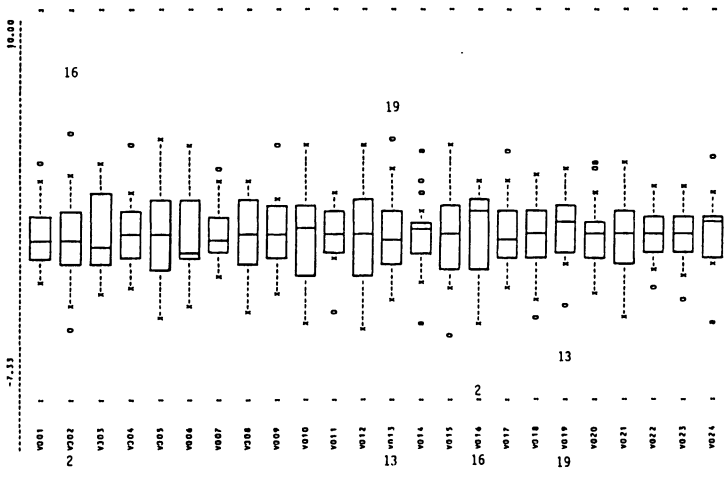
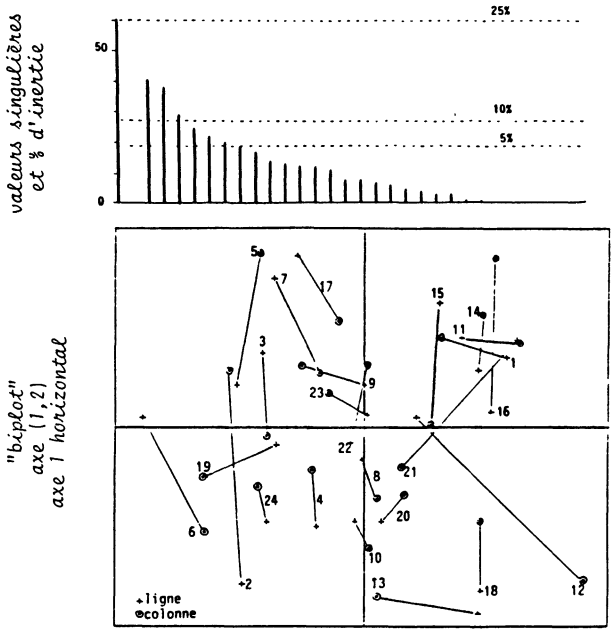


FIGURE 1
Sociomatrice. Répartition des résidus du modèle de quasi-symétrie.

(resp. −7,33 et −5,73) sont observés pour l'élève 2 noté par 16 et 13 noté par 19 (resp. 16 noté par 2 et 19 noté par 13).

Les résultats de la décomposition en valeurs singulières des matrices des résidus E_1 et E_2 des deux modèles ajustés et les représentations en dimension 2 (représentation euclidienne du type « biplot » de K. GABRIEL (1971) pour les résidus de la quasi-indépendance, et représentations non euclidienne du type de celle proposée par J. GOWER (1977) pour les résidus antisymétriques de la quasi-symétrie) sont reportés en fig. 2. Rappelons ici que les pourcentages d'inertie sont proportionnels aux carrés des valeurs singulières.

a) modèle de quasi-indépendance



b) modèle de quasi-symétrie

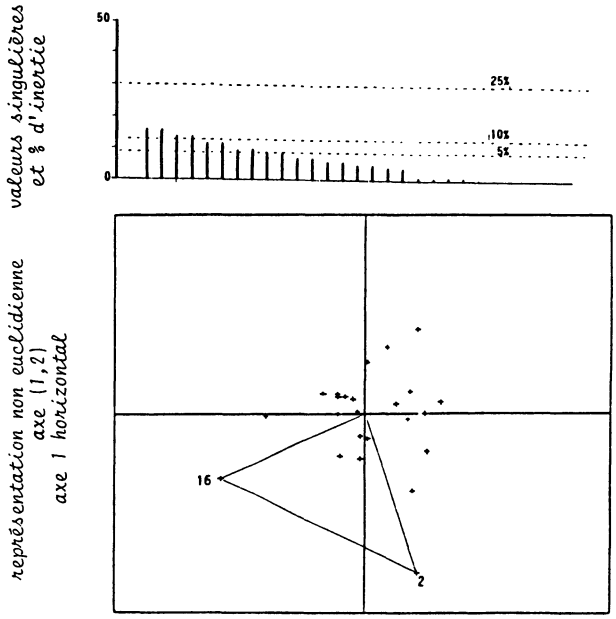


FIGURE 2
Sociomatrice. Analyse factorielle des résidus.

Le « biplot » de E_1 suggère une certaine symétrie et le test de Fisher de quasi-indépendance contre quasi-symétrie conduit à rejeter la quasi-indépendance ($F = 3{,}067$ pour $\nu_1 = 252$, $\nu_2 = 253$). Enfin la décomposition en valeur singulière de E_2 ne met en évidence aucune structure en dehors de l'éventuelle inadéquation du modèle pour le couple (16,2).

Le modèle de quasi-symétrie additive semble donc globalement acceptable.

3.5. Etude des paramètres

En représentant chaque élève par un point de coordonnées ($\hat{a}$ (i), $\hat{b}$ (i)) on fait clairement apparaître les effets individuels : « effet notant » et « effet noté » (cf. Fig. 3).

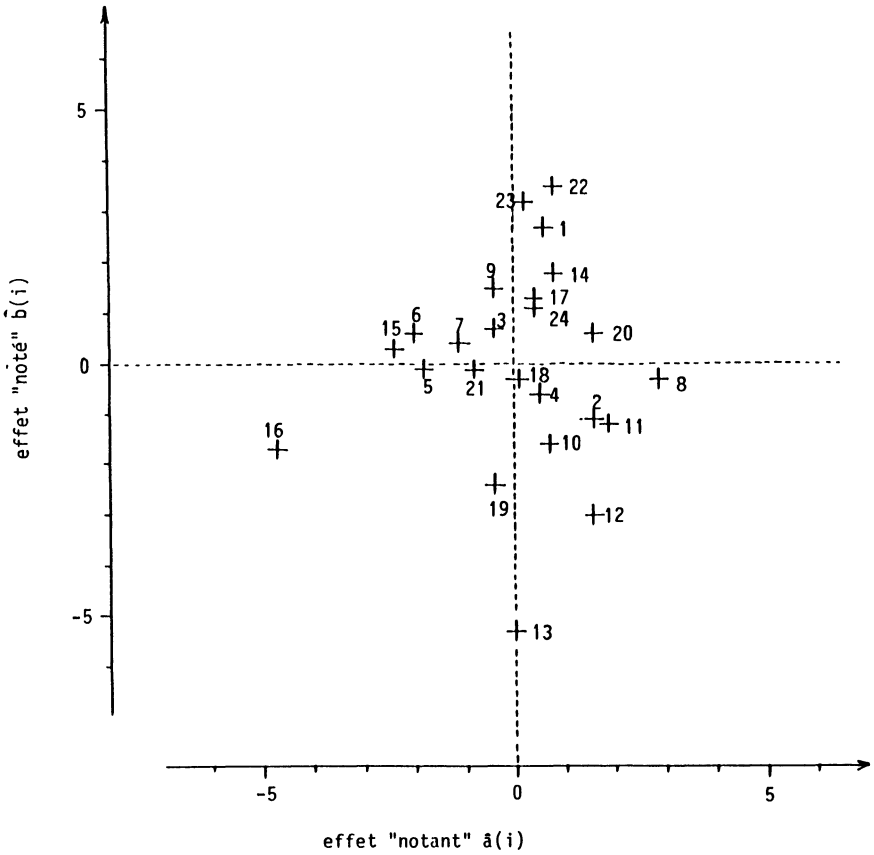
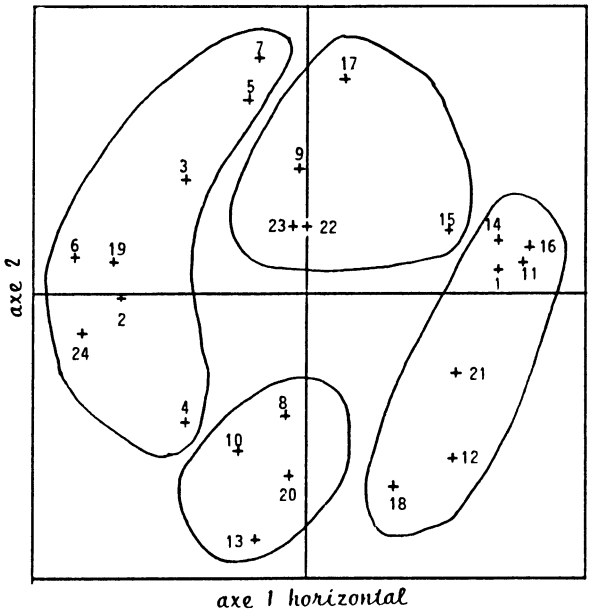


FIGURE 3
Sociomatrice. Effets individuels.



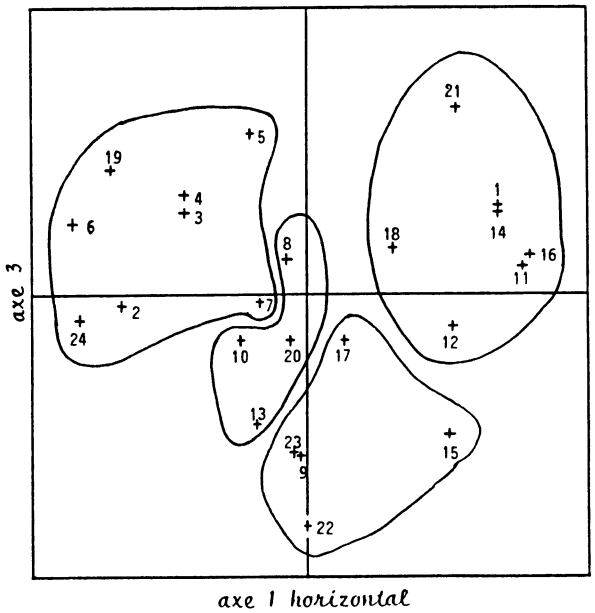


FIGURE 4
Sociomatrice. Affinités entre élèves.

Plus délicate est l'interprétation statistique des $\hat{s}(i, j)$ qui reflètent les similarités intrinsèques entre élèves. L'indice s introduit dans le modèle étant à valeurs dans **R**, une transformation préalable de cet indice doit être envisagée pour pouvoir appliquer des méthodes usuelles de classification automatique ou de positionnement multidimensionnel. Deux méthodes nous sont apparues potentiellement intéressantes. La première consiste à associer à la paire (i, j), j > i, son rang dans la préordonnance induite par $\hat{s}$; après division par $\ell(\ell - 1)/2$ on obtient un indice de ressemblance à valeurs dans]0,1[. La seconde consiste à poser :

$$\hat{\sigma}(i, j) = \begin{cases} F\left(\dfrac{\hat{s}(i, j)}{\lambda}\right) & i \neq j \\ 1 & i = j \end{cases}$$

où λ est un paramètre d'échelle et F la fonction de répartition d'une loi de probabilité simple, par exemple la loi normale réduite (les $\hat{s}(i, j)$ sont centrés par construction); pour le choix de λ on pourra par exemple retenir l'écart-type empirique des $\hat{s}(i, j)$: c'est le point de vue adopté ici.

Il est alors aisé (cf. J.P. BENZECRI (1973) et J. GOWER et P.G.N. DIGBY (1981)) de construire une image euclidienne des élèves en sorte que la distance usuelle entre deux élèves soit approximativement égale à $\sqrt{2(1 - \hat{\sigma}(i, j))}$. Les représentations graphiques données en figure 4 correspondent aux coordonnées associées aux trois plus grandes valeurs propres de la matrice Σ des $\hat{\sigma}(i, j)$ préalablement transformée par double centrage. Dans une certaine mesure, cette représentation peut être considérée comme une forme épurée du biplot des résidus de la quasi-indépendance (Fig. 2).

Par ailleurs, l'étude du tableau des $\hat{\sigma}(i, j)$ par une méthode de classification automatique permet de dégager des groupes homogènes d'élèves. En utilisant la méthode de classification proposée par A. de FALGUEROLLES (1977) (méthode qui fournit des classifications localement optimales) on a obtenu en particulier la répartition suivante des élèves 4 classes : (1,11,12,14,16,18,21), (2,3,4,5,6,7,19,24), (8,10,13,20), (9,15,17,22,23). Ces classes ont été matérialisées sur la figure 4 afin de souligner la cohérence avec la représentation euclidienne.

4. « Mobilité sociale »

4.1. Données, problème

Nous reprenons ici les données de mobilité sociale étudiées par P. van der HEIJDEN (1985) dans son extension de l'Analyse des Correspondances à un modèle de quasi-indépendance : 17 professions sont considérées, et le tableau des données croise les modalités de la profession du père avec celles du premier emploi du fils. Les fréquences observées sont notées x (i, j), i et j variant de 1 à 17.

A première vue il semble que cette table puisse être classiquement étudiée par l'analyse factorielle des correspondances. Cependant, comme le note P. van der HEIJDEN (1985), les résultats de celle-ci sont fortement influencés par les comportements associés aux modalités 16 et 17 du tableau.

En fait, cette analyse est très influencée par les forts effectifs diagonaux (ou « adjacents » à la diagonale), ce qui suggère de distinguer deux types de comportements, immobilité et mobilité, et de les étudier séparément. D'un point de vue pratique cela revient à considérer la diagonale et les marges du tableau pour l'immobilité et le tableau sans sa diagonale pour la mobilité.

Sur l'interprétation sociologique de tableaux analogues observés en France on pourra se reporter à l'ouvrage de C. THELOT (1982).

4.2. « Immobilité sociale »

Comme on l'a déjà vu, le tableau précédent traduit une immobilité sociale variable selon les catégories socio-professionnelles. On peut chercher à quantifier cette immobilité. Par exemple, pour chaque catégorie socio-professionnelle, on peut tenter de repérer la position de l'immobilité réelle par rapport aux deux configurations extrêmes : immobilité dans l'hypothèse de « mobilité parfaite » et dans l'hypothèse d'« immobilité maximale » (cf. R. POHL et J. SOLEILHAVOUP (1982)).

On introduit alors des coefficients α_i tels que :

$$x(i, i) = \left(\frac{x(i, +) \times (+, i)}{x(+, +)}\right)^{\alpha_i} [\inf(x(i, +), x(+, i))]^{1-\alpha_i}$$

avec des notations évidentes.

Ces coefficients présentent ici les valeurs suivantes :

i	1	2	3	4	5	6	7	8	9	10	11
α_i	0.41	0.48	0.63	0.63	0.37	0.77	0.68	0.64	0.67	0.50	0.68

i	12	13	14	15	16	17
α_i	0.64	0.62	0.53	0.67	0.12	0.39

Le phénomène d'immobilité est donc particulièrement marqué pour la profession 16, puis pour les professions 5, 17, 1, ...

4.3. Modèles pour la mobilité

On s'intéresse maintenant à la table de mobilité hors diagonale et on considère pour ce tableau les deux modèles log-linéaires présentés en 2.1 et 2.2. Les effectifs théoriques $\hat{x}(i, j)$ des transitions ont alors pour expression :

dans le cas du modèle de quasi-indépendance

$$\hat{x}(i, j) = m_1 . a_1(i) . b_1(j) \qquad i \neq j$$

dans le cas du modèle de quasi-symétrie

$$\hat{x}(i, j) = m_2 . a_2(i) . b_2(j) . s(i, j) \qquad i \neq j$$

où les a (i) et les b (j) sont des effets de marge et, dans le cas de la quasi-symétrie, les s (i, j) = s (j, i) sont des indices de mobilité hors effets de marge.

Evaluée par les statistiques usuelles (chi-deux et rapport de vraisemblance) la qualité de l'ajustement de ces modèles est ici :

modèle	chi-deux	rapport de vraisemblance	degrés de liberté
quasi-indépendance	4725	4023	239
quasi-symétrie	174	175	120

Celà conduit à analyser la mobilité soit par un modèle de quasi-indépendance *complété* par une étude des résidus de ce modèle (« biplot » des résidus standardisés par exemple) soit par un modèle de quasi-symétrie.

4.4. Les résultats de la quasi-indépendance

En associant à chaque profession i (i = 1, ..., 17) un point de coordonnées (â (i), $\hat{b}$ (i)) dans un repère cartésien, on peut mettre en évidence les professions qui en première approximation sont plus spécifiquement observées chez les fils ($\hat{b}$ (i) > â (i)) ou chez les pères (â (i) > $\hat{b}$ (i)) (cf. Fig. 6).

La décomposition en valeurs singulières de la matrice des résidus standardisés indique que les deux (resp. trois) premières valeurs singulières rendent compte de 87 % (resp. 90 %) du χ^2 (cf. Fig. 5). Des représentations graphiques associées (plan 1,2) et plan (2,3), il ressort essentiellement que la mobilité associée aux modalités 16 (ligne) et 17 (colonne) est très différente de celle des autres modalités.

On notera que, dans son étude, P. van der HEIJDEN (1985) avait considéré un modèle de quasi-indépendance avec la contrainte supplémentaire $\hat{x}$ (i, j) = x (i, j) pour la cellule (16,17).

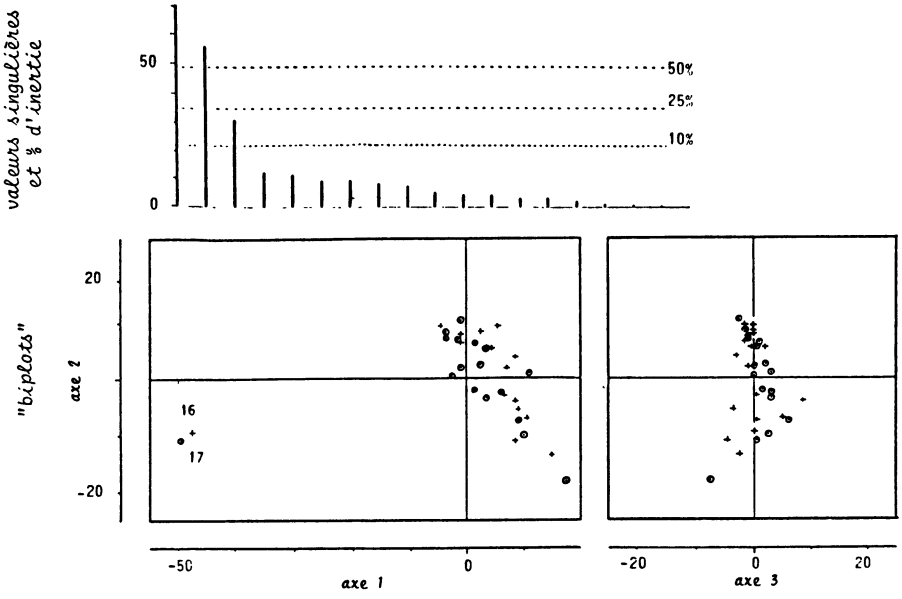
4.5. Les résultats de la quasi-symétrie

Comme pour le modèle de quasi-indépendance on peut représenter graphiquement et interpréter les couples (â (i), $\hat{b}$ (i)). Il apparait qu'ils diffèrent peu de ceux obtenus pour le modèle de quasi-indépendance (cf. Fig. 6).

Le modèle s'ajustant de façon satisfaisante, la décomposition en valeurs singulières des résidus standardisés ne donne aucune représentation graphique facilement interprétable (cf. Fig. 5).

Enfin on peut chercher à interpréter les paramètres s (i, j) et à donner une représentation graphique de leurs estimations ŝ (i, j). Compte tenu des contraintes d'identifiabilité $\sum_i \{\hat{s}(i,j) \mid i \neq j\} = \sum_j \{\hat{s}(i,j) \mid j \neq i\} = 1$, la matrice S des ŝ (i, j) (avec ŝ (i, j) = 1) est semi-définie positive : c'est la somme d'une matrice bistochastique et de la matrice identité. Il est alors aisé (cf. J.P. BENZECRI (1973) et J. GOWER et P.G.N. DIGBY (1981)) de construire une image euclidienne des professions i (i = 1, ..., 17) en sorte que la distance usuelle entre deux professions soit exactement égale à $\sqrt{2(1 - \hat{s}(i,j))}$.

a) modèle de quasi-indépendance



b) modèle de quasi-symétrie

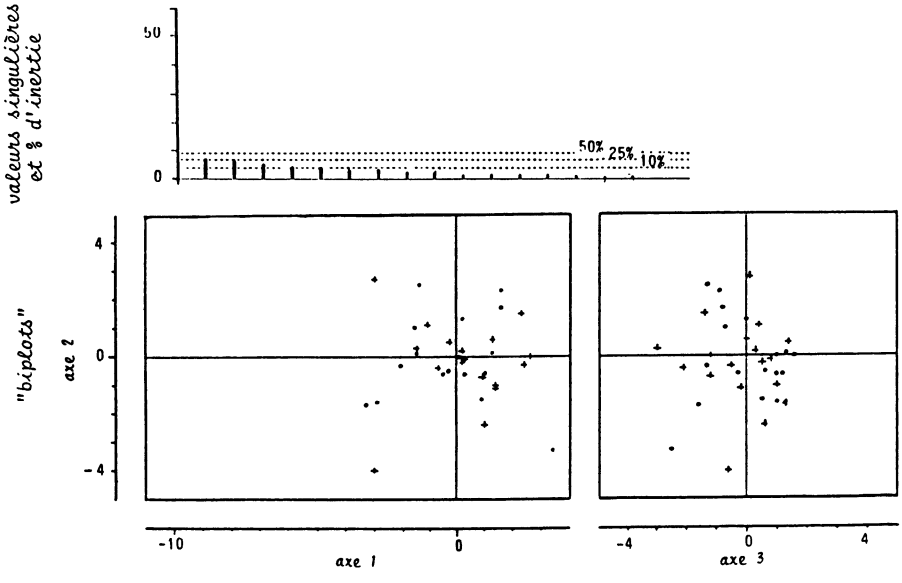
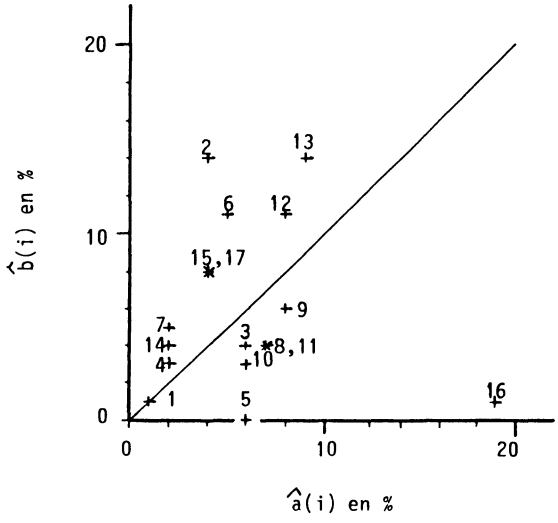


FIGURE 5
Mobilité. Analyse factorielle des résidus (résidus standardisés).

a) modèle de quasi-indépendance



b) modèle de quasi-symétrie

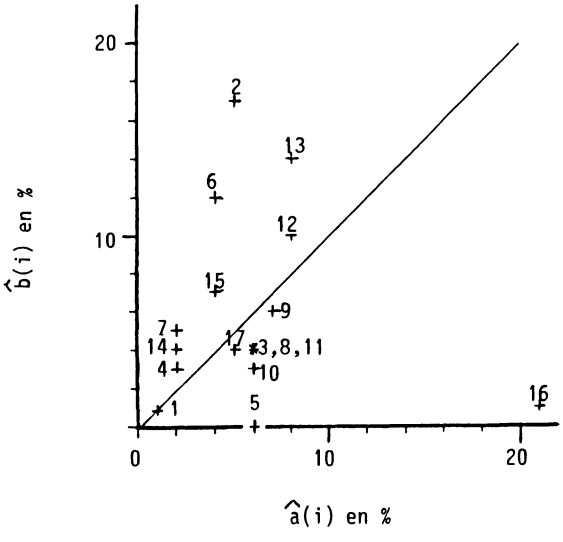


FIGURE 6
Mobilité. Effets individuels.

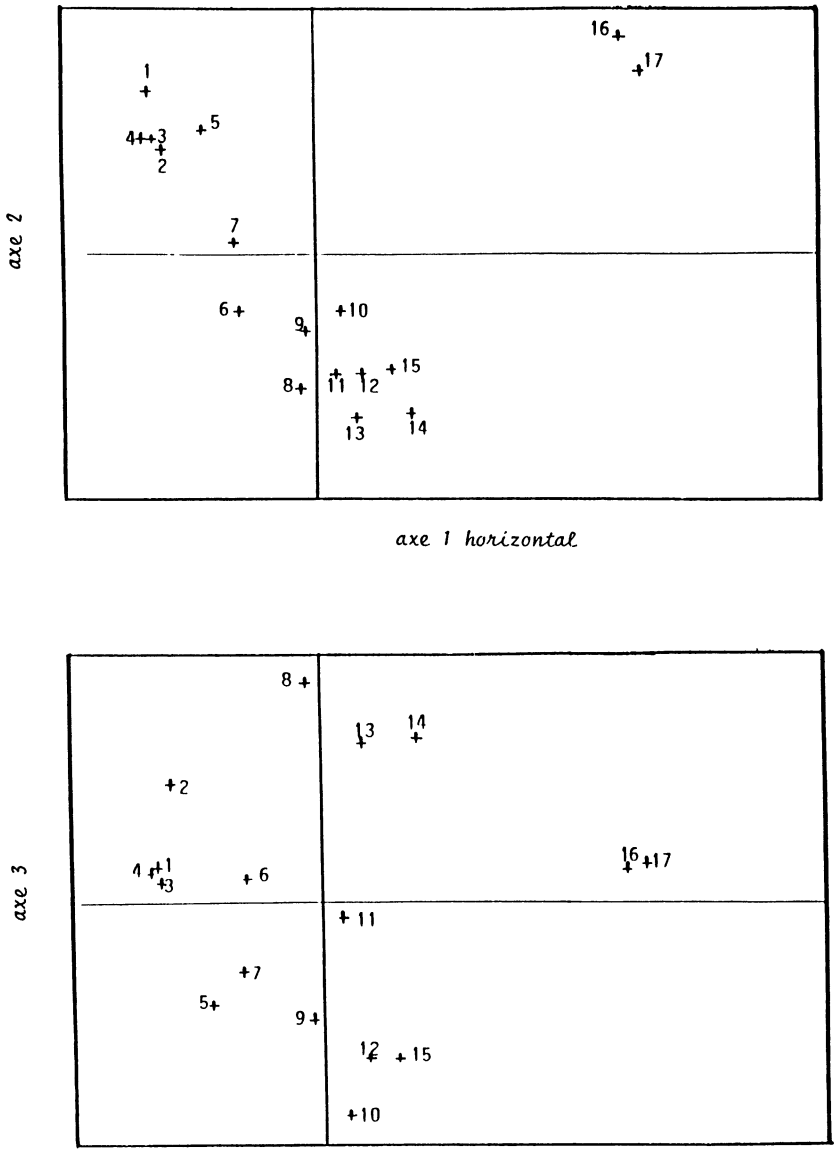


FIGURE 7
Mobilité. Proximités des professions.

Des représentations graphiques sont obtenues en ne considérant que les coordonnées correspondant aux plus grandes valeurs propres de la matrice S, préalablement transformée par double centrage (cf. Fig. 7).

D'autres méthodes de positionnement multidimensionnel pourraient évidemment être utilisées.

5. Conclusions générales et perspectives

L'origine de ce travail est dans une étude comparative des méthodes de modélisation et des méthodes exploratoires dites d'analyse des données. A priori, nous penchions clairement pour leur complémentarité, mais ce sentiment a été encore renforcé par le traitement des exemples ci-dessus ainsi que de quelques autres dont certains rapportés dans CAUSSINUS et FALGUEROLLES (1986) (voir aussi BRADU et GABRIEL (1978)). La modélisation est utile pour tenir compte de la problématique précise du cas étudié. Les méthodes d'analyse multidimensionnelle peuvent *ensuite* intervenir

— pour mieux décrire quelques paramètres pertinents issus de la modélisation

et

— pour l'analyse exploratoire de la partie résiduelle que le modèle a été éventuellement impuissant à prendre en compte.

Cette étude montre donc en particulier que l'articulation des diverses méthodes statistiques n'est pas immuable. Dans les deux exemples ci-dessus la modélisation nous a paru pouvoir intervenir d'abord. Dans bien d'autres cas, l'analyse exploratoire semble au contraire première par nécessité matérielle (voir par exemple A. BACCINI *et al.*, 1987) ou encore intervient « itérativement » (voir par exemple K. WORSLEY (1987)).

A un niveau plus technique, la démarche retenue pour traiter ces exemples met en œuvre des méthodes qui, considérées séparément, ont des propriétés d'optimalité bien connues. Cependant l'articulation emboitée qui en est faite pose ici le problème de leur optimalité jointe. Les réponses à cette question pourraient en particulier influer sur le choix d'algorithmes d'estimation des paramètres et de positionnement multidimensionnel. Ce type de question est abordé par K. GABRIEL (1978) pour le cas linéaire.

Nous voudrions enfin insister sur un point qui semble important concernant la notion même de modèle. Lorsqu'un modèle est soumis à un test (à l'intérieur d'un modèle de référence considéré comme convenable a priori) il s'agit de prendre une décision sur son acceptation ou son rejet. Dans l'esprit de beaucoup d'auteurs, surtout anglo-saxons, l'idée fondamentale est de ne pas « surparamétrer » afin que l'analyse, à force d'être probante, ne devienne pas trop bavarde : c'est assez frappant par exemple quand on considère l'esprit dans lequel L. Goodman a développé son modèle d'analyse des tables de contingence et le compare à l'Analyse des Correspondances (voir L. GOODMAN (1986), et la discussion qui suit). Dans les exemples que nous avons traités ici, et plus particulièrement dans le second, aucun modèle simple ne résistera à un test

d'hypothèse appliqué dans des conditions raisonnables. Nous sommes donc obligés de « sous paramétrer » (à moins de prendre un modèle saturé sans utilité) afin de faire apparaître quelques traits marquants qu'on peut a priori contrôler (quantifier, ...). Mais nous savons que les résidus peuvent encore renfermer une information significative qu'il convient d'explorer : nous avons ici étudié une voie qui semble de plus en plus retenue par nombre de statisticiens.

Bibliographie

A. BACCINI, J.R. MATHIEU et A.M. MONDOT (1987). — Comparaison, sur un exemple, d'analyses des correspondances multiples et de modélisations. *Revue de Statistique Appliquée.* Vol. 35, n° 3, 21-34.

J.R. BARRA (1985). — Méthodes statistiques en psychiatrie — Modèles virtuels. « In Model Choice », *Proceedings of the 4th Franco-Belgian Meeting of Statisticians.* Publications des Facultés Universitaires Saint-Louis, Bruxelles.

J.P. BENZECRI (1973). — *L'analyse des données, tome II : l'analyse des correspondances.* Dunod.

D. BRADU et K.R. GABRIEL (1978). — The biplot as a diagnostic tool for models of two way tables. *Technometrics,* Vol. 20, N° 1, pp. 47-68.

B. BURTSCHY (1984). — Analyse factorielle des matrices d'échanges. *In Data Analysis and Informatics 3* (E. Diday et al. eds), North-Holland, pp. 447-464.

F. CAILLIEZ et J.-P. PAGES (1976). — *Introduction à l'analyse des données.* Paris : SMASH, 9, rue Duban, 75016 Paris.

H. CAUSSINUS (1976). — Quelques points de vue sur l'analyse des tableaux d'échanges. *Ann. de l'INSEE,* N° 22-23, pp. 83-100.

H. CAUSSINUS (1985). — Quelques réflexions sur la part des modèles probabilistes en analyse des données. *4e Journées Internationales d'Analyse des Données et Informatique. Versailles.* (Paru dans *Data Analysis and Informatics 4,* North-Holland, Amsterdam, 151-165).

H. CAUSSINUS et A. de FALGUEROLLES (1986). — Modèle de quasi-symétrie et analyse descriptive de tableaux carrés. In comparaison et évaluation des approches française et britannique de l'analyse de données complexes. *Publications du laboratoire de Statistique et Probabilités.* Université Paul Sabatier, N° 02-86 pp. 79-95.

H. CAUSSINUS et C. THELOT (1976). — Note complémentaire sur l'analyse statistique des migrations. *Ann. de l'INSEE,* N° 22-23, pp. 135-145.

A.G. CONSTANTINE et J.C. GOWER (1978). — Graphical representation of asymmetric matrices. *Appl. Statist.,* Vol. 27, N° 3, pp. 297-304.

D. DOMENGES et M. VOLLE (1980). — L'analyse factorielle sphérique. *In Data Analysis and Informatics 2,* North-Holland, Amsterdam, pp. 253-262.

B. ESCOFIER (1984). — Analyse factorielle en référence à un modèle : application à l'analyse des tableaux d'échange. *Revue de Statistique Appliquée,* Vol. 34, n° 4, pp. 25-36.

Y. ESCOUFIER et A. GRORUD (1980). — Analyse factorielle des matrices carrées non symétriques. *In Data Analysis and Informatics 2,* North-Holland, Amsterdam, pp. 263-273.

A. de FALGUEROLLES (1977). — Classification automatique : un critère et des algorithmes d'échange. *Séminaires INRIA,* pp. 29-40.

K.R. GABRIEL (1971). — The biplot graphic display of matrices with application to principal component analysis. *Biometrika,* Vol. 58, n° 3, pp. 453-467.

K.R. GABRIEL (1978). — Least squares approximation of matrices by additive and multiplicative models. *J.R. Statist. Soc. B,* Vol. 40, n° 2, pp. 186-196.

L.A. GOODMAN (1986). — Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables, with disussion. *International Statistical Review,* Vol. 54, n° 3, pp. 243-309.

J.C. GOWER (1977). — The analysis of asymmetry and orthogonality. *In Recent Developments in Statistics* (J. BARRA *et al.,* eds). North-Holland, pp. 109-123.

J.C. GOWER et P.G.N. DIGBY (1981). — Expressing complex relationships in two dimensions. In Interpreting Multivariate Data (V. Barnett, ed.). Proceedings of the Conference *Looking at multivariate data;* Wiley, pp. 119-146.

P.G.M. van der HEIJDEN (1985). — Transition matrices, model fitting and correspondence analysis. *4e Journées Internationales d'Analyse de Données et Informatique,* Versailles. (Paru dans *Data Analysis and Informatics 4,* North-Holland, Amsterdam, 221-226).

W.J. HEISER (1985). — On the selection of a stimulus set with prescribed structure from empirical confusion frequencies. *Internal Report RR-85-08. Dep. of Data Theory,* University of Leiden.

L.S.P.-C.A.S. (1986). — Comparaison et évaluation des approches française et britannique de l'analyse de données complexes. Rapport de contrat (ATP C.N.R.S. « Comparaisons Internationales »). *Laboratoire de Statistique et Probabilités,* Université Paul Sabatier, Toulouse.

R. POHL et J. SOLEIHAVOUP (1982). — La transmission du statut social sur deux ou trois générations. *Economie et Statistique,* n° 144, pp. 25-42.

C. THELOT (1982). — *Tel, père, tel fils ?.* Dunod.

K. WORSLEY (1987). — Un exemple d'identification d'un modèle log-linéaire grâce à une analyse des correspondances. *Revue de Statistique Appliquée,* Vol. 35, n° 3, pp. 13-20.