

CAD

**Mémoire reçu : Measures, models and graphical displays in the analysis of cross classified data**

*Les cahiers de l'analyse des données*, tome 16, n° 1 (1991),  
p. 119-126

[http://www.numdam.org/item?id=CAD\\_1991\\_\\_16\\_1\\_119\\_0](http://www.numdam.org/item?id=CAD_1991__16_1_119_0)

© Les cahiers de l'analyse des données, Dunod, 1991, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » 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/>

## MÉMOIRE REÇU: MEASURES, MODELS AND GRAPHICAL DISPLAYS IN THE ANALYSIS OF CROSS CLASSIFIED DATA

[MOD. GRAPH.]

*point de vue de J.-P. BENZÉCRI  
sur un exposé de L. A. GOODMAN*

*Le Professeur GOODMAN nous ayant communiqué pour discussion, avant publication dans le Journal of the American Statistical Association, le compte rendu d'une conférence prononcée par lui à l'invitation de la section des Sciences Sociales de cette Association, il nous a paru opportun de communiquer nos réflexions au public de langue française; tout en préparant, avec l'aimable collaboration du Pr. L. A. GOODMAN, une version anglaise de la présente note destinée au JASA ...*

### 1 Modèles et analyse inductive

Notre propos n'est pas ici de reprendre l'exposé des méthodes que L. A. GOODMAN présente avec une clarté qui satisfera les lecteurs de son mémoire. Il nous suffira de dire que ces méthodes nous paraissent effectuer des constructions assez proches de celles de l'analyse des correspondances; en les complétant toutefois par des épreuves de validité destinées à autoriser l'assimilation des données de base à des schémas simplifiés que l'auteur extrait des résultats de ces constructions. Nous voulons plutôt préciser en quoi notre point de vue diffère de celui du Professeur de Berkeley; en acceptant, faute de mieux, comme pierre de touche, les trois tableaux, de petite taille, qui ont servi dans le mémoire.

L. A. GOODMAN propose d'abord cet aphorisme:

“Beaucoup de chercheurs qui ont analysé des données souscriront à la thèse de statisticiens distingués pour lesquels: de petites tailles d'échantillons se prêtent communément à un ajustement satisfaisant; mais avec de grandes tailles on n'obtient jamais que des conclusions négatives relativement à quelque modèle que ce soit.”

Mais L. A. GOODMAN s'oppose aussitôt à ces chercheurs en affirmant:

“En réfutation de cette thèse nous présentons ici des modèles simples bien ajustés à des données [concernant des ensembles de cas variant de 3.500 à 25.000].”

Avant de passer à l'examen direct des données, il faut préciser ici ce qu'on peut entendre par modèle et par ajustement. Dans les sciences humaines, la notion de modèle est souvent associée à celle de mesure d'une ou plusieurs grandeurs cachées ou latentes. Par exemple, s'il s'agit de lier la profession du fils à celle de son père, le plus simple (sinon le plus judicieux) est de postuler qu'une profession se caractérise par son abscisse sur "l'échelle sociale"; et que l'abscisse de la profession est la même pour le fils que pour le père, à un décalage près qui suit une loi normale dont la variance pourra être arbitrairement fixée à 1.

Un tel modèle comportera pour seuls paramètres les abscisses (définies à un changement d'origine près) des  $n$  professions qu'on distinguera dans la nomenclature adoptée; alors que les données constituent une matrice  $n \times n$ . En effectuant un ajustement, on pourra matérialiser "l'échelle sociale". Mais comme, par exemple, avec 10 professions on a 9 paramètres pour 99 données, il est bien vrai que si le nombre des cas individuels est assez élevé, l'ajustement ne passera pas le test! Il est essentiel de souligner ici que le nombre des cas individuels n'est pas seul en cause; la forte complexité du tableau des données relativement à celle du modèle, est essentielle. Et c'est quant à la faible dimension du tableau de base que nous considérons comme "de petite taille" les exemples choisis par l'auteur pour illustrer son propos; bien que le nombre des cas individuels, soit, comme il l'affirme, considérable.

Nous n'avons, personnellement jamais manifesté d'intérêt pour de tels modèles *a priori*; mais c'est sans doute à ceux-ci que pensent ordinairement les chercheurs que L. A. GOODMAN voudrait rallier à l'analyse inductive des données. Car c'est bien d'elle qu'il s'agit chez lui; même si, répétons-le, ses méthodes diffèrent des nôtres.

On peut, si l'on veut, appeler modèle le résultat d'une analyse inductive; ou, plus précisément, le schéma simplifié extrait d'une telle analyse: comme quand, en analyse des correspondances, on substitue au tableau donné, un tableau reconstitué en tronquant, après le deuxième facteur, une formule de développement qui serait exacte si on avait gardé tous les facteurs. Mais il vaut mieux dire que l'analyse inductive substitue au tableau donné son expression canonique en fonction d'invariants algébriques, tels que vecteurs propres normalisés et valeurs propres.

Le mérite de cette description canonique réside d'abord dans son universalité: sans qu'il soit besoin de concevoir *a priori* un mécanisme latent, que précisément l'analyse inductive est propre à réfuter, on aboutit à une vue synthétique des données. Nous savons que la statistique classique critique une telle synthèse en mettant à l'épreuve l'ajustement aux données initiales du schéma qu'elle peut suggérer de celles-ci. Mais tel n'est pas notre point de vue.

Pour nous, la question n'est pas de savoir si le modèle (ou schéma) passe le test. Nous sommes en effet convaincu, comme les adversaires qu'imagine L. A. GOODMAN, que le modèle ne passera pas. Nous posons plutôt la question dans le sens opposé: le tableau analysé étant significativement distinct du modèle d'indépendance entre lignes et colonnes, jusqu'où avons-nous pu aller dans la construction d'un schéma de ce tableau en termes de facteurs interprétables.

Ici, nous rejoignons, dans leur désir de comprendre, les tenants des modèles *a priori*. L'idéal est d'obtenir *a posteriori* des axes factoriels qui soient comme "l'échelle sociale" hâtivement suggérée ci-dessus; avec, sur celle-ci, l'avantage de proposer des points de vue multiples: l'axe 1 étant, e.g., une échelle de niveau général, l'axe 2 une échelle de différenciation intellectuelle... Même si l'idéal n'est pas atteint, le schéma ne peut être séparé de son interprétation; qui est pour nous le véritable critère. Car, répétons-le, dans toute étude importante, il y a plus de structure valide, au sens statistique usuel du terme, que de structure que nous soyons capable d'extraire sous forme conceptuelle.

Le problème général de la statistique inductive est pour nous de construire des tableaux qui, par des codages et des compléments appropriés, donnent aux données disponibles une forme telle que l'analyse en extraie la réponse aux questions concrètes qu'il est permis de poser. De ce point de vue, il ne nous est jamais apparu utile de recourir à un autre outil que l'analyse des correspondances; outil bien étalonné sur de multiples études; outil qui, par la représentation simultanée des deux ensembles des lignes et colonnes, suggère directement des interprétations; et qui, avec tout le formulaire, d'ailleurs simple, des contributions et corrélations, la formule de reconstitution des données en fonction des facteurs, la formule de transition, le principe d'équivalence distributionnelle ..., offre aussi au chercheur le moyen de critiquer ce qui lui est suggéré.

## 2 Analyse des exemples proposés par A. L. GOODMAN

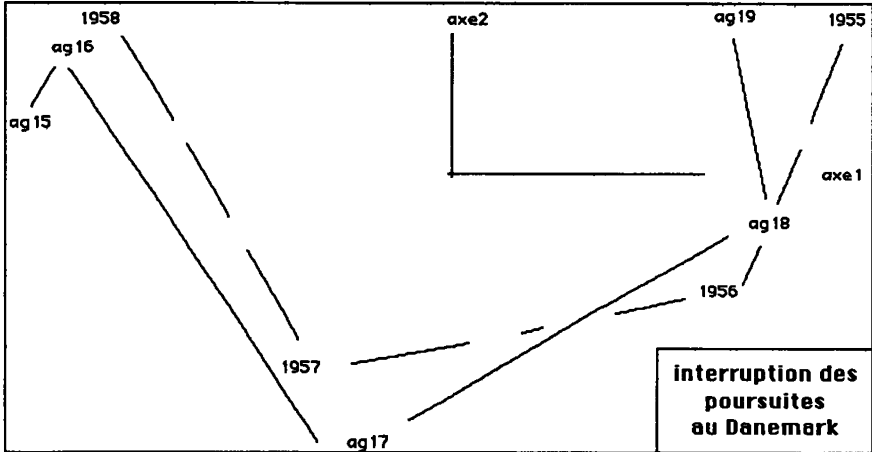
L'auteur propose trois tableaux de données que nous considérerons successivement.

### 2.1 Analyse diachronique de l'interruption des poursuites criminelles contre jeunes hommes au Danemark

L. A. Goodman Tableau A : poursuites criminelles abandonnées

trace :	5.468e-3			
rang :	1	2	3	
lambda :	49	5	0	e-4
taux :	9034	897	69	e-4
cumul :	9034	9931	10000	e-4

Le tableau de base croise l'ensemble des 4 années {1955, 1956, 1957, 1958} avec les 5 âges {ag15, ag16, ag17, ag 18, ag19}.



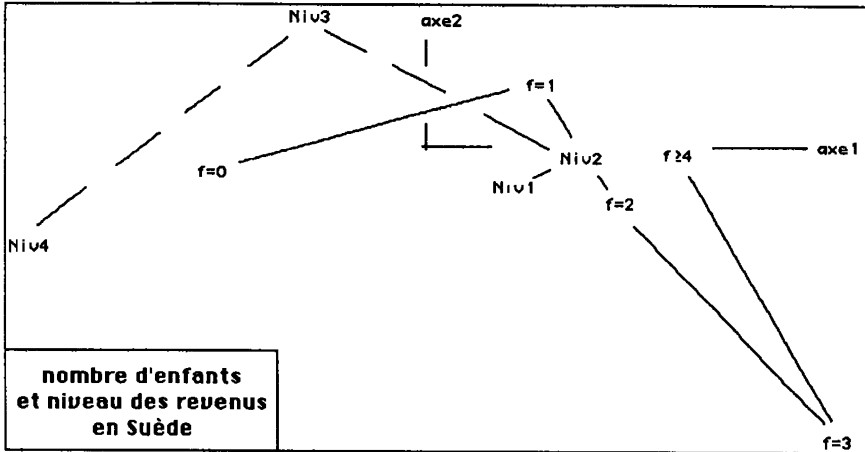
Le plan (1,2) montre un classique effet Guttman: les deux ensembles, se rangent, à peu près dans leur ordre naturel, sur une courbe en croissant: la suite décroissante des âges accompagne la suite croissante des années: ceci suggère un abaissement constant de l'âge de la délinquance au cours de la période étudiée. Mais, pour éprouver cette hypothèse, il faudrait disposer des statistiques de toutes les poursuites interrompues ou non; afin de faire la part de l'indulgence et de la délinquance dans le tableau observé. Il serait assurément souhaitable d'analyser les données réparties par nature des actes incriminés et lieu du crime supposé.

Quant à la validité statistique, sous l'hypothèse que les quelque 7000 cas recensés constituent des événements indépendants; ce que rien ne prouve, indulgence et délinquance pouvant avoir un caractère épidémique; on applique la règle classique, remontant, en substance, à Karl PEARSON: on regarde comme significatifs les facteurs extraits jusqu'à ce que soit inférieur à  $(\text{cardI}-1) \times (\text{cardJ}-1)$  le produit de la somme des valeurs propres restantes par le nombre total  $k$  des cas individuels recensés.

L. A. Goodman Tableau A : poursuites abandonnées

5	ag15	ag16	ag17	ag18	ag19
1955	141	285	320	441	427
1956	144	292	342	441	396
1957	196	380	424	462	427
1958	212	424	399	442	430

Dans le cas présent,  $(\text{cardI}-1) \times (\text{cardJ}-1) = 12$ ; la somme des deux dernières valeurs propres est  $\approx 5 \text{ e-}4$ ;  $k \approx 7000$ ;  $k \times 5 \text{ e-}4 \approx 3,5 < 12$ . On ne retiendra donc comme proprement significatif que l'axe 1. Cependant, le plan (1,2) offre une image plus claire que le tracé d'un seul axe.



### 2.2 Croisement entre nombre d'enfants et revenus pour des familles suédoises

L. A. Goodman Tableau B : nombres d'enfants et revenus

trace :	2.251e-2				
rang :	1	2	3		
lambda :	219	5	1	e-4	
taux :	9724	228	47	e-4	
cumul :	9724	9953	10000	e-4	

Le tableau de base croise l'ensemble des quatre niveaux de revenus {Niv1, Niv2, Niv3, Niv4}, délimités par les bornes 1000, 2000, 3000 kronor, avec les effectifs {f=0, f=1, f=2, f=3, f≥4}, de zéro à 4 enfants et plus. L'échantillon étant de quelque 25.000 familles, la règle rappelée au §2.1 conduit à conserver les axes 1 et 2 car:  $k \times 6 e-4 \approx 15 > 12$ .

L. A. Goodman Tableau B : nombres d'enfants et revenus

5	f=0	f=1	f=2	f=3	f≥4
Niv1	2161	2755	936	225	39
Niv2	3577	5081	1753	419	98
Niv3	2184	2222	640	96	31
Niv4	1636	1052	306	38	14

Quant à l'interprétation, on peut voir la suite croissante des niveaux, de Niv1=Niv2 à Niv4, accompagnée, en sens inverse, par la suite décroissante des effectifs: {f=0, f=1, f=2, ...}. L'effectif tendrait donc à être en raison inverse de la fortune: nous disons "tendrait"; car, même si les résultats sont significatifs, les valeurs sont très faibles: or, en analyse des correspondances, les valeurs propres s'interprètent comme des carrés de coefficients de corrélation.

On s'interrogera, cependant, sur la place de f≥4 au niveau de f=2, alors que f=3 est le plus écarté. Du point de vue de la théorie des tests, on pourrait chercher si le nombre de 14 familles riches ayant 4 enfants ou plus n'est pas en

butte à des erreurs d'échantillonnage telles que l'affinité qu'il crée entre  $f \geq 4$  et Niv4 ne soit sujette à caution. Mais c'est plutôt vers l'interprétation que, selon notre esprit, nous porterons l'attention.

Le nombre d'enfants peut faire l'objet de choix délibérés; ou résulter de l'acceptation religieuse de la fécondité naturelle. Les choix dépendent du genre de vie; des sollicitations de la richesse ou de la peur du lendemain. La religion suit d'autres voies; dont on dit qu'elles sont insondables. On ne s'étonnera pas que d'influences à la fois discrètes et complexes, résulte un tableau nuancé. Plus de nuances apparaîtraient encore si l'on ventilait les données par professions.

Nous ne quitterons pas cette analyse sans considérer la définition des données: s'agit-il de familles achevées, de ménages avancés en âge? mais alors, les revenus correspondent-ils à l'époque où les enfants venaient au foyer, ou y vivaient encore tous? Cette dernière question revêtirait une particulière importance, si, comme on le trouve aujourd'hui, dans certaines enquêtes, le manque de ressources dissuadait des familles d'avoir plusieurs enfants; ou si l'abondance les encourageait à en accueillir trois.

### 2.3 Profession du père et profession du fils en Grande-Bretagne

L. A. Goodman Tableau C : catégorie du père et catégorie du fils

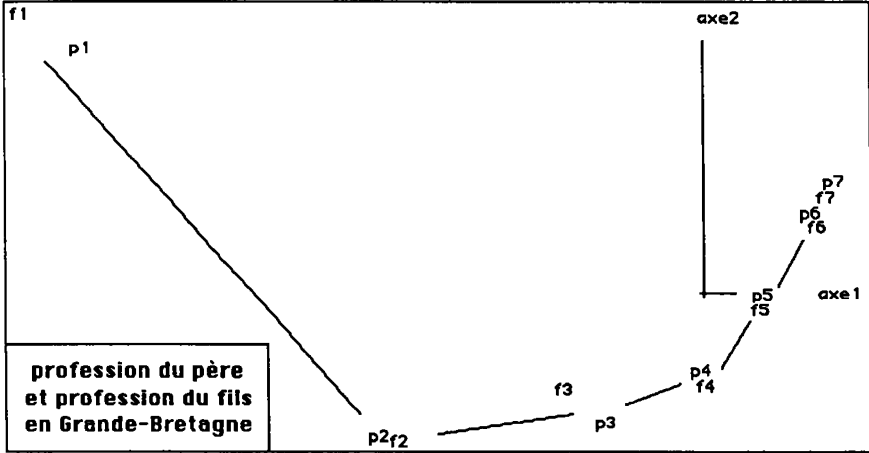
trace :	3.895e-1					
rang :	1	2	3	4	5	6
lambda :	2763	715	273	90	45	8 e-4
taux :	7094	1836	701	231	115	22 e-4
cumul :	7094	8931	9632	9863	9978	10000 e-4

Le tableau croise avec lui-même un ensemble de 7 catégories professionnelles, considérées pour deux générations successives: en ligne, profession du père; en colonne, profession du fils; mais si les ensembles I et J sont identiques, le tableau n'est rigoureusement symétrique.

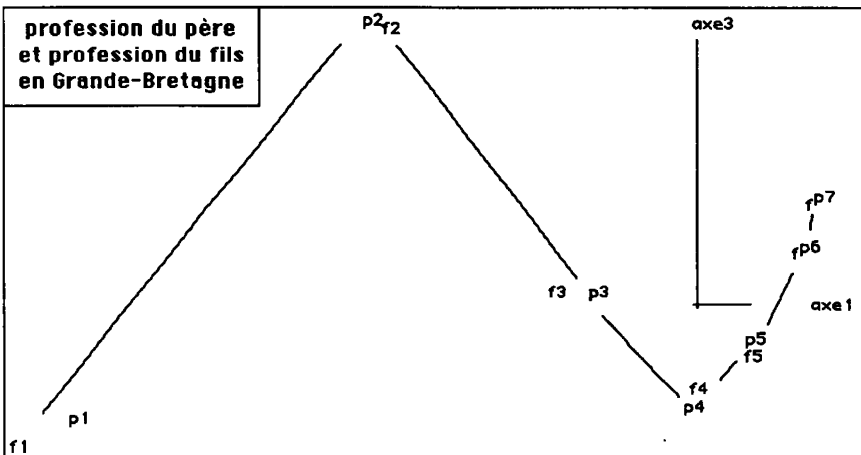
L. A. Goodman Tableau C : catégorie du père et catégorie du fils

7	f1	f2	f3	f4	f5	f6	f7
p1	50	19	26	8	18	6	2
p2	16	40	34	18	31	8	3
p3	12	35	65	66	123	23	21
p4	11	20	58	110	223	64	32
p5	14	36	114	185	715	258	189
p6	0	6	19	40	179	143	71
p7	0	3	14	32	141	91	106

Un statisticien classique écarterait-il l'hypothèse de symétrie? L'effectif total de la ligne p6 apparaît significativement différent de celui de la colonne f6:  $458 \neq 593$ . Plutôt que de conjecturer un accroissement de la catégorie 6, nous préférons, ici comme ailleurs, nous interroger sur les données: d'une génération à une autre, la structure de l'ensemble des professions n'est-elle pas modifiée au point de rendre imprécise l'assimilation des nomenclatures employées pour l'une et pour l'autre?



L'échantillon étant de quelque 3.500 sujets, avec  $\text{cardI} = \text{cardJ} = 7$ , la règle rappelée au §2.1 conduit à conserver les axes 1 à 4 car:  $k \times (90+45+8) e-4 \approx 50 > 36$ . Cependant, l'examen des plans (1,2) et (1,3) montre un effet Guttman poursuivi jusqu'au troisième facteur. Dans le plan (1,2), les couples (fi, pi) sont disposés, suivant leur ordre naturel, sur une courbe d'allure parabolique: le facteur F2 apparaît ainsi comme une fonction du second degré de F1. On a, de même dans le plan (1,3), une courbe en 'S'; qui suggère, pour F3, une expression du 3-ème degré en F1. Le modèle est foncièrement unidimensionnel; mais, comme dans le cas du parallélogramme de Louis GUTTMAN, on approche une fonction en créneau par une combinaison de termes d'allure polynomiale.





### 3 Perspectives et conclusions

De part et d'autre de l'Atlantique, les mêmes algorithmes d'analyse numérique sont appliqués au traitement des données. Les références bibliographiques passent; mais l'esprit de l'analyse des correspondances, familier aux lecteurs des *Cahiers*, n'est pas encore passé. La publication par Marcel Decker d'un "Correspondance Analysis Handbook" pourrait nous mettre sur les flots!

#### Références bibliographiques

##### Tableau du §2.1

U. CHRISTIANSEN & J. STINE: in *Le manuel G. RASH de Statistique Théorique*; (en danois), Copenhagen: Teknisk Forlag; (1969).

E.B. ANDERSEN: *Discrete Statistical Models with Social Science Applications*, Amsterdam: North-Holland Publishing Co; (1980).

##### Tableau du §2.2

H. CRAMER: *Mathematical Methods of Statistics*, Stockholm: Almqvist & Wiksell; (1945).

P. MARTIN-LÖFF: "The Notion of Redundancy and Its Use as a Quantitative Measure of the Discrepancy between a Statistical Hypothesis and a Set of Observational Data," *Scandinavian Journal of Statistics*, 1, pp. 1-28; (1974).

F. ABILG<sup>o</sup>ARD: "Discussion of Paper by P. MARTIN-LÖFF," *Scandinavian Journal of Statistics*, 1, pp. 13-15; (1974).

P. DIACONIS & B. EFRON: "Testing for Independence in a two-Way Table: New Interpretations of the Chi-Square Statistic," *Journal of the American Statistical Association*, 77, pp. 803-815; (1985).

##### Tableau du §2.2

D.V. GLASS, ed.: *Social Mobility in Britain*, Glencoe: Free Press; (1954).

L.A. GOODMAN: "Some Multiplicative Models for the Analysis of Cross-Classified Data," in *Proceedings of the the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vol 1, ed. Le CAM et al., Berkeley: University of California Press, pp. 649-696; (1972).

S.H. HABERMAN: *The Analysis of Frequency Data*, Chicago: University of Chicago Press; (1974).

#### Analyse des Correspondances

J.-P. BENZÉCRI: *Correspondance Analysis Handbook*; translated by T. K. GOPALAN; Marcel DEKKER, inc.; (announced for 1991).