JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

ROBERT GIBRAT

L'analyse des données

Journal de la société statistique de Paris, tome 119, n° 3 (1978), p. 201-228

<http://www.numdam.org/item?id=JSFS_1978__119_3_201_0>

II

COMMUNICATIONS

L'ANALYSE DES DONNÉES (1)

(Première partie)

(Communication faite le 17 novembre 1977 devant les Sociétés de statistique de Paris et de France)

Robert GIBRAT

Président des Sociétés de statistiques de Paris et de France,
membre de l'Institut international de statistique,
ancien président de la Société des ingénieurs civils de France

L'auteur expose les principes de l'analyse des données et en donne les bases mathématiques. Il en étudie des applications aux techniques industrielles. Il montre enfin les abus auxquels peut donner lieu son utilisation.

The author exposes the principles of data analysis and gives its mathematical bases. He studies some applications to industrial technics. He shows at last abuses that may occur when it is used.

« *On peut se réconforter de l'histoire des physiciens qui, dans un semblable dilemme, considéraient que la lumière est une particule les lundi, mercredi et vendredi et une onde les mardi, jeudi et samedi et qui, le dimanche, priaient.* »

CORNFIELD, 1969

« *Donnez un marteau à un enfant et vous verrez que tout lui paraîtra mériter un coup de marteau.* »

A. KAPLAN, 1968

« *La statistique n'est pas seulement un mode de connaissance : elle est un mode d'être.* »

J.-P. BENZECRI, 1974

1. Cette première partie correspond à la conférence faite le 17 novembre 1977 devant les Sociétés de statistique de Paris et de France. Elle a déjà été publiée dans la *Revue des ingénieurs civils de France : Sciences et techniques*, n° 44, sept.-oct. 1977, pp. 20-33, que nous remercions de nous avoir autorisé à la reproduire avec quelques modifications. La 2e partie correspondant à une conférence du 24 mai 1978 devant l'Association des statisticiens universitaires sera publiée ultérieurement.

Introduction

Le 17 juillet 1928 l'après-midi, je me mariai à l'hôtel de ville de Saint-Étienne après avoir le matin même passé, à la faculté de droit de Lyon, mes examens oraux de licence. Ainsi s'ouvrait le chemin vers ma thèse. Soutenue le 15 mai 1931, elle m'apportait le grade de Docteur en droit. Elle avait pour titre : « Les inégalités économiques » et pour sous titre : « Applications aux inégalités de richesse, à la concentration des entreprises, aux populations des villes, aux statistiques des familles, etc., d'une loi nouvelle : la loi de l'effet proportionnel ». J'ai pu, ensuite, l'appliquer à de nombreux phénomènes physiques, par exemple, aux débits des cours d'eau, crues exceptionnelles, etc. (*Revue générale de l'Électricité*, 15 et 29 octobre 1932, pp. 493 à 500 et 525 à 532). Tout récemment, un Américain, Larsen (*cf.* B6) l'a utilisée comme distribution *a priori* des mesures de pollution atmosphérique.

On me permettra d'en reproduire le premier paragraphe : « Les biens sont inégalement répartis entre les hommes, les hommes inégalement répartis entre les entreprises... d'une manière générale, les grandeurs économiques formées de biens et d'hommes comprennent un nombre variable de ces biens ou de ces hommes. Notre but est de définir, puis d'étudier l'inégalité de ces nombres... ».

Après avoir étudié sans moyens artificiels de calcul, plus de six cents répartitions à un paramètre j'y proposai une formule d'ajustement dérivée de la célèbre courbe en cloche (loi de Gauss Laplace) montrant au passage que « les revenus et la fortune étaient distribués entre les hommes avec la même inégalité *depuis près de cinq siècles* et qu'ainsi toutes les conquêtes économiques ou sociales étaient restées sans influence sensible sur elle ». Ma formule s'appliquait remarquablement bien à toutes ces répartitions en n'utilisant que deux ou trois paramètres. J'ai pu, ensuite, en donner une explication *a priori* très simple en consi dérant des variations stochastiques relatives et non des variations absolues, d'où le nom de l'*effet proportionnel.*

Cette thèse était essentiellement une recherche de statistique économique, s'appuyant sur une hypothèse *a priori* dérivée de la loi de Gauss Laplace. Elle était dans la ligne de la statistique mathématique de cette époque, popularisée en France quelques années auparavant par le petit livre de G. Darmois (1928). L'École anglo-saxonne (Karl Pearson — 1857 1936 et R.-A. Fisher — 1890-1962 — en sont, quoique adversaires toute leur vie, les deux plus grands noms) l'a dominée de façon extraordinaire jusqu'au milieu du xx^{e} siècle, mais construite pour la plus grande part sur l'utilisation de lois mathématiques bien définies (principalement loi normale et ses dérivées) et trop confiante dans les hypothèses *a priori*, elle a abordé l'analyse des tableaux de données à plusieurs dimensions de façon jugée peu satisfaisante par un nombre croissant de statisticiens.

Dans l'étude présente, le lecteur me surprendra à brûler en 1977 ce que j'ai adoré en 1930 et il me trouvera, parfois, trop enthousiaste devant l'éclat et l'avenir de ces nouvelles techniques d'analyse, mais prudent et souvent réticent devant des audaces, parfois des abus.

A quoi est dû ce changement si vif : nous énumérerons trois mutations dans la recherche appliquée industrielle qui l'expliquent :

1. L'industrie s'intéresse aujourd'hui non seulement aux sciences dites « exactes » dont les plus importantes sont la physique et la mécanique, mais à l'immense éventail de connaissances que nous appelons sciences « humaines » parce qu'elles prennent l'homme en compte (sciences sociales, biologie et médecine). Elle ne sont plus seulement un objet de

culture générale, mais apportent quotidiennement de nouvelles méthodes de recherches appliquées dans la réalité industrielle. *On ne peut presque jamais dans ces sciences utiliser la méthode expérimentale* classique consistant à construire artificiellement des expériences où intervient un seul paramètre décisif pour le rejet ou l'acceptation d'une hypothèse; on est en face de faits, souvent en très grand nombre, répondant à plusieurs questions à la fois et il faut les analyser en même temps. Il en est de même de plus en plus dans l'industrie, à l'*expérimentation* a succédé dans de nombreux cas l'*observation simultanée;* aujourd'hui la partie nouvelle de la statistique concernant l'analyse des données, conçue dans le cadre des sciences « humaines » pénètre dans les sciences « exactes », prenant ainsi une généralité et une importance insoupçonnées il y a dix ans.

La méthode expérimentale appliquée à une variable avait conduit à admettre implicitement que les distributions statistiques, quel que soit le nombre de variables, suivaient la loi « normale », aussi l'enseignement de la statistique et du calcul des probabilités doit être entièrement revu sous ce double aspect : introduction de l'analyse des données et abandon du caractère systématique de la loi normale.

2. L'arrivée des ordinateurs est un phénomène si banal que nous avons quelques scrupules à en souligner l'importance. Ils existaient certes il y a vingt ans (le langage Fortran date de 1955), mais leur influence sur la recherche industrielle était nulle. Aujourd'hui leur dynamisme est extraordinaire : un ingénieur en informatique est dépassé s'il s'éloigne de son milieu un an seulement : les performances des machines doublent tous les trois ans et demi; aux U. S. A., les groupes de recherches spécialisés prévoient l'usine entièrement automatisée avec ordinateur intégré en l'an 2000, etc. Toute l'industrie ou presque toute calcule aujour d'hui tout ou presque tout sur ordinateur. Rien n'est contrôlé; rien n'est préparé, rien n'est optimisé sans qu'il intervienne. Un véritable raz de marée submerge tout. Mais, pendant longtemps le développement explosif de l'analyse numérique, fille de l'ordinateur, s'est fait avec un tel empirisme que la *crédibilité des modèles mathématiques* en a été considérablement affaiblie et non plus seulement celle de la loi normale.

3. Fort heureusement, depuis quelques années, les mathématiciens purs font un effort considérable en cherchant à établir des fondations solides. Le fait que le langage des ensembles et surtout la topologie des espaces normés jouent un rôle très important dans ces recherches a été un facteur très favorable et explique l'intérêt subit des mathématiciens purs pour ces nouvelles méthodes. L'introduction toute récente des espaces de Soboleff dans la justification des méthodes modernes de calcul, comme celle des éléments finis est un progrès considérable. C'est une grande chance, l'analyse numérique ne pouvant pas continuer à être une science « expérimentale » comme elle l'est encore trop souvent aujourd'hui.

Nous nous proposons, ici, de faire le point sur l'étude des grands ensembles des données numériques. Par exemple, un réseau de surveillance et d'alerte au SO^2 mesure la pollution tous les quarts d'heures en une trentaine de postes; une année apporte donc près d'un million de données à analyser en même temps. Nous aurons, pour cela, à remettre à sa place, à leur vraie place, les hypothèses *a priori* comme la loi normale, à employer à fond les ordinateurs et à utiliser les théories mathématiques sur les matrices et les opérateurs. (Pour les problèmes à plus de deux dimensions ou les problèmes à variables continues, les bases mathématiques manquent encore, un beau sujet pour les jeunes mathématiciens.)

L'étude des structures des ensembles de données est en fait très récente. Cependant, dès le début du xx^e siècle l'analyse factorielle classique avait été développée par les psychologues (Ch. Spearman, 1904) pour « expliquer » des résultats par des facteurs cachés (mémoire,

intelligence, etc.) mais elle supposait un modèle *a priori* et ne concernait jamais, faute de moyens de calcul, que de très petits ensembles. Puis, l'analyse dite en *composantes principales* s'est donnée comme but principal, sans hypothèses particulières *a priori*, de représenter le mieux possible des ensembles de plus en plus grands dans des espaces de petites dimensions. Enfin, l'analyse dite, des *correspondances*, animée par une équipe française, connaît aujourd'hui une vogue extraordinaire; car elle fournit, sans hypothèse *a priori*, des représentations simplifiées appropriées dans un certain sens à l'interprétation. Notre examen portera surtout sur elle, car son succès auprès des jeunes chercheurs, son utilisation de plus en plus fréquente dans l'industrie en particulier pour les problèmes de l'environnement d'une part, les critiques aiguës dont elle fait l'objet souvent de la part de personnalités de premier plan d'autre part, méritent une attention toute particulière.

Toutes les méthodes précédentes et celles qui en ont été dérivées, comme l'analyse des covariances partielles, l'analyse des rangs, l'analyse canonique, l'analyse discriminante, etc., dépendent d'un même corps de résultats mathématiques que nous exposerons rapidement au chapitre II.

On pourra juger de l'importance prise, aujourd'hui, par ces idées en notant que plus de la moitié des thèmes de statistiques présentés pour une thèse portent sur ce sujet et qu'il a existé en 1977 une école d'été de l'analyse des données, sous la responsabilité du professeur Benzecri, en trois stades :

— pré-sélection;

— participation à un des dix ateliers préparatoires (trois journées);

— école proprement dite (dix journées, du 19 au 30 septembre).

Ajoutons, enfin que, parmi bien d'autres, un colloque sur l'analyse des données a été organisé par l'IRIA [1] pour les 7, 8 et 9 septembre 1977. La méthodologie et l'épistémologie de l'analyse des données (rôle et limites) ont été particulièrement examinées dans une séance dont nous avions accepté d'être modérateur.

Nous garderons à l'esprit, ici, les applications industrielles, reportant nos réflexions sur l'application aux sciences humaines à une deuxième partie à paraître.

Les conclusions que nous tirerons du texte actuel ont donc un peu un caractère provisoire.

1. Espoirs et controverses

Nous introduirons le débat par quelques citations aussi caractéristiques que possible :

Écoutons d'abord le créateur de la forme la plus moderne de l'analyse des données, celle des correspondances, le professeur J.-P. Benzecri. Ses premiers travaux (Collège de France, 1963) avaient été dominés par la recherche têtue d'une « méthode inductive d'analyse des données linguistiques ». Il s'opposait à Noam Chomski, qui régnait alors sur la linguistique mathématique, et affirmait « que celle-ci ne peut être inductive, c'est-à-dire s'élever par une méthode explicitement formulée des faits aux lois qui les régissent et qu'elle doit être déductive, c'est-à-dire partant d'axiomes, engendrer des modèles de langues réelles ». Les tableaux de données de la linguistique comporteront, par exemple, des noms suivant les lignes (chien, cheval...), et des verbes suivant les colonnes (courir, hennir, aboyer...), chaque nombre

1. Institut de Recherches sur l'Information et l'Automatisme.

inscrit dans le tableau de « correspondances » (ou de « contingence ») étant le nombre de fois que, dans un texte ou ensemble de textes, le verbe et le nom sont associés. Le professeur Benzecri estime, en 1977, que ses espérances d'une méthode inductive à partir de tels tableaux n'ont pas été déçues, « bien au contraire », et annonce une publication sur l'ensemble des résultats acquis à ce jour en linguistique. Les deux grands principes d'équivalence distributionnelle et de distance du chi², dont nous verrons qu'il a fait la base de son édifice, lui ont été suggérés par ces problèmes particuliers. Par contre, la suite de ses travaux semble avoir été dominée par les problèmes statistiques posés par les autres sciences humaines, les problèmes industriels ou de sciences exactes paraissent ne le préoccuper que depuis peu et encore fort peu.

Le but de l'analyse des correspondances est, pour lui, clair : « traiter simultanément de grands ensembles de faits et les confronter en vue de découvrir l'ordre global ».

Sa méthodologie est impitoyable : « Le modèle doit suivre les données, non l'inverse. Découvrir sans parti pris, sans *a priori*, quels courants de lois traversent l'océan des faits ».

« Asservir la chair des données à l'âme des formules ».

Il s'oppose aux lois *a priori* et en particulier à la loi normale (Gauss-Laplace) surtout en statistique multidimensionnelle. (On pourra trouver notre formulation précise de cette opposition avec la notion toute récente de robustesse au chapitre III).

Les moyens à la disposition du statisticien sont devenus extraordinairement puissants; l'arrivée des ordinateurs et peut-être encore plus les progrès prodigieux de la programmation ont tout transformé, permettant des calculs impensables il y a 20 ans et suggérant même des théorèmes que l'on a ensuite démontrés. Aussi le professeur Benzecri (*cf.* B1) (Tome II, A, n° 1) écrit-il sans indulgence : « Ayant rendu à nos pères l'hommage qui leur est dû, nous sommes pressés de substituer aux algorithmes confus, issus des nécessités du calcul manuel et simultanément justifiés par des mosaïques de théorèmes,... des formules claires et exécutables qui soient les meilleures possibles. Utiliser un ordinateur implique d'abandonner toutes techniques conçues avant l'avènement du calcul automatique ».

Emporté par son enthousiasme, il conclut dans un article sur « la place de l'*a priori* [1] », « Auxiliaire de la synthèse, l'ordinateur est un outil mental : après l'organum d'Aristote et le novum organum conçu par Bacon n'est-il pas le novius organum l'outil le plus nouveau? » Et il résume l'analyse des données par l'analyse des correspondances, « méthode qui, bien mieux que toute autre, nous a permis de découvrir les faits de structure que recèle un tableau de données quel qu'il soit ».

Mais le doute perce cependant chez lui en quelques rares aveux : « Dans quelles limites sommes nous capables d'accomplir ce magnifique programme?... Comment démontrer que, de la manipulation purement mathématique des tableaux des données, sortent des résultats assez significatifs pour que le spécialiste les lise et les accepte comme une évidence? Ne nous vantons pas trop... » ou encore « le trait dont nous dessinons nos conclusions doit être aussi large que la place laissée au doute ».

Il n'a pas convaincu tout le monde, loin de là. Voici deux opinions importantes récentes (1976 et 1977) :

A. Lichnerowicz, remarquable observateur, écrit (*cf.* B6) : « Une donnée n'est jamais donnée et isolée, elle n'a pas de valeur. Les épistémologistes savent depuis longtemps qu'il n'est pas de faits scientifiques bruts, mais seulement des faits scientifiques enserrés dans un modèle scientifique éventuellement modifiable. De même, il n'y a pas de donnée brute, mais

1. Ronéotypé.

saisie de données par une exploitation à travers des modèles variés ». Il ajoute : « Ne nous égarons pas dans le mirage que seraient des données élémentaires, brutes, objectives, indiscutables et gratuites. Au-delà viennent les processus statistiques par lesquels on s'efforce d'extraire, de l'océan des informations, des « données » d'un autre niveau... Or, ces opérations, utiles dans un champ local et pour un objectif déterminé, se révèlent déformantes et dangereuses dès que l'on sort du champ... »

Dans un chapitre d'un article récent (*cf.* B6) intitulé « L'analyse des données n'est pas une panacée », E. Malinvaud écrit :

« Les travaux sur l'analyse des données... auraient dégagé des procédures grâce auxquelles les sciences humaines pourraient progresser suivant une voie purement factuelle. Il semble bien que cet espoir soit le plus souvent déçu lorsqu'on essaie de le matérialiser. Sans doute, le mouvement pour l'analyse des données a-t-il permis de définir des procédures adaptées à des problèmes que l'on n'avait guère étudiés auparavant... Mais pour la recherche en vue de laquelle les méthodes traditionnelles ont été conçues (détermination des lois entre grandeurs quantitatives) il n'existe pas de procédé efficace qui élimine totalement l'a priorisme ».

E. Diday et L. Lebart (*cf.* B5) ont une grande pratique de ces nouvelles techniques statistiques et tous deux leur ont apporté des contributions théoriques fondamentales, or ils précisent, optimistes et dynamiques : « L'analyse des donnés s'affirme de plus en plus comme une discipline autonome, capable de fournir des instruments de classement et d'organisation des informations, sûrs et indépendants. Bénéficiant de l'essor des moyens informatiques, les différentes méthodes d'analyse des données suscitent un intérêt croissant. Toutes n'ont pas la même portée, mais ensemble elles peuvent rénover et stimuler les méthodes de la statistique classique. »

G. Morlat, enfin, dans l'introduction à l'ouvrage de Caillez et Pages (*cf.* B4) que nous aimerions citer tout entière, voit dans cette discussion, non sans raison, un conflit de générations bien qu'il paraisse oublier la nôtre qui est la sienne. Il écrit : « Tout particulièrement au cours des années les plus récentes, on a vu la gent statisticienne se scinder, grosso-modo, en deux classes :

— la première catégorie est celle des statisticiens d'âge moyen qui ont appris et pratiqué la statistique mathématique classique, celle qui prétend formaliser l'induction à la suite des statisticiens anglo saxons, notamment des années 1900 à 1950;

— la seconde classe est formée de gens en général plus jeunes qui ont appris, sous la même étiquette de « statistique », des techniques bien différentes s'appuyant sur un outil mathématique purement algébrique et visant à décrire, réduire, classer des observations multidimensionnelles; ceux-là n'ont cure de l'induction et sont volontiers portés à proclamer que le statisticien doit mettre en œuvre ses techniques d'analyse sans faire aucune hypothèse sur les phénomènes observés. Ils pratiquent "l'analyse des données".

Et il conclut : « L'analyse des données doit rendre service partout où l'on se soucie d'accumuler des observations... Les services rendus montrent bien que l'analyse des données constitue aujourd'hui, et de loin, la partie la plus immédiatement rentable de la statistique... Cela permet, selon les cas, de découvrir dans les phénomènes étudiés des structures directement visibles sur les résultats de l'analyse, alors qu'elles ne l'étaient pas sur les données originelles, ou de retrouver en les précisant des structures que l'on soupçonnait déjà pour telle ou telle raison. »

2. Principaux résultats mathématiques constituant le formulaire de l'analyse des données

Ce bref résumé vise simplement à donner une idée des directions mathématiques qui ont été explorées, il ne peut remplacer la lecture des ouvrages spécialisés. En particulier, nous conseillons vivement au lecteur de lire un ou plusieurs des exemples « pratiques » recensés dans la bibliographie. Faute de place, on ne trouvera dans cet article que quelques exemples très brefs.

A — *Formules communes*

On part d'un tableau reliant deux ensembles U et Q de variables u (p lignes) et q (n colonnes) en faisant correspondre au couple (u, q) un nombre positif $k\ (u, q)$; prenons un exemple cité par le professeur Benzecri : Dans une enquête effectuée en 1971, l'Institut de l'environnement demandait d'associer à une liste de thèmes — haine, amour, vacances, travail, calme... des surfaces colorées en rouge, vert, bleu, mauve ! que l'on présentait aux sujets. Le dépouillement de l'enquête se fait sur un tableau dont les lignes sont les thèmes, haine, amour...; tandis que les colonnes sont des couleurs rouge, vert...

On inscrit par exemple à la croisée de la ligne amour et de la colonne vert le nombre de sujets ayant associé ce thème à cette couleur.

Le tableau peut avoir deux représentations :

— L'une dans un espace vectoriel $\mathbb{R}^n$ avec un nuage de p points correspondant chacun à une ligne; chaque point indice u_0 ayant n coordonnées $k\ (u_0, q)$.

— L'autre dans un espace vectoriel $\mathbb{R}^p$ avec un nuage de n points correspondant donc chacun à une colonne, chaque indice n_0 ayant p coordonnées $k\ (u, q_0)$.

Tout d'abord, on cherche à décrire le plus économiquement possible ces nuages, c'est-à-dire avec un nombre de dimensions aussi faible que possible; puis plus ambitieusement, on cherche à mettre en évidence des structures explicatives. En lisant ce qui suit, on n'oubliera pas que l'on sait traiter en 1977 un tableau de plusieurs milliers de lignes par deux cents colonnes.

On commence évidemment par déterminer une droite passant par l'origine et ajustant au mieux le nuage à étudier, ceci en minimisant la somme des carrés des distances des points à la droite (problème classique depuis longtemps dans la théorie des axes principaux des ellipses ou ellipsoïdes).

Ce calcul conduit à un vecteur support de la droite, dit vecteur propre qui, on le sait, est relatif à une valeur propre. En continuant l'ajustement compte tenu de celui-ci on peut trouver successivement dans $\mathbb{R}^p$ un certain nombre de vecteurs propres et de valeurs propres toutes positives décroissant avec le rang. R étant la matrice du tableau, R' la matrice transposée, V_i les vecteurs propres et λ_i les valeurs propres seront solutions de l'équation (la démonstration est classique) :

$$R'RV_i = \lambda_i V_i \text{ dans l'espace } R^p$$

Le vecteur V est normé par la relation :

$$V'V = 1$$

On aura les formules correspondantes dans le deuxième espace avec RR' au lieu de $R'R$.

Cela revient donc à diagonaliser les matrices $R'R$ ou RR' (R' transport de R). On aura donc les mêmes valeurs propres dans les deux espaces, les deux vecteurs étant liés par les relations :

$$V_i = \frac{1}{\sqrt{\lambda_i}} R'Q_i$$

$$Q_i = \frac{1}{\sqrt{\lambda_i}} RV_i$$

On peut donc considérer un seul espace. Le nombre des valeurs ou vecteurs propres est le plus petit nombre de p ou de n. Les valeurs de λ_i sont les inerties des diverses valeurs propres, etc.

Ces résultats très importants, qui ont permis sous certaines précautions d'utiliser le même espace pour les deux types de variables, furent d'abord reconnus empiriquement à propos de l'analyse de correspondance par une collaboratrice du professeur Benzecri, B. Cordier, puis démontrés dans la thèse de celle-ci.

B — *Analyses diverses*

Elles diffèrent d'abord par le tableau de départ des calculs, ensuite par la norme adoptée pour les distances (ou proximités) des points représentatifs.

1. L'analyse en composantes principales part du tableau :

$$r(u, q) = k(u, q) - \frac{k(u)}{n}$$

dissymétrique par rapport aux indices u et q. Elle est souvent utilisée quand le chercheur estime que le facteur important, pour chaque donnée est son niveau.

2. Pour l'analyse de correspondances, on part du tableau symétrique en (u, q),

$$s_{uq} = \frac{p_{uq} - p_u p_q}{\sqrt{p_u p_q}}$$

avec

$$p_{uq} = \frac{k(u, q)}{k} \qquad p_u = \frac{k(u)}{k} \text{ etc.}$$

k étant le nombre de données :

$$k = \sum_{u, q} k(u, q)$$

Nous verrons plus loin les raisons de cette différence.

3. L'analyse factorielle « classique », beaucoup plus ancienne que les précédentes, part d'un modèle *a priori* pour reconstituer les corrélations existantes entre un certain nombre de variables à l'aide de facteurs « communs et spécifiques », par exemple pour les psychologues l'intelligence et la mémoire. Cette recherche de facteurs donne, sous certaines réserves, des résultats voisins de l'analyse en composantes principales.

4. L'analyse canonique d'Hotelling recherche à représenter les relations entre deux ensembles de variables en recherchant les combinaisons linéaires des variables du premier ensemble qui ont les meilleures corrélations avec des combinaisons linéaires des variables du deuxième groupe.

5. L'analyse discriminante est un cas particulier de la précédente.

C — *Analyse des correspondances*

Le professeur Benzecri (*cf.* B3) dans les Cahiers de l'analyse des données (avril 1977) résume de façon très claire ses principaux résultats et ceux de ses disciples :

« L'analyse des correspondances, telle qu'on la pratique en 1977, ne se borne pas à extraire des facteurs de tout tableau de nombres positifs. Elle donne pour la préparation des données des règles telles que le codage, sous forme disjonctive complète, aide à critiquer la validité des résultats, principalement par des calculs de contribution, fournit des procédés efficaces de discrimination et de régression, se conjugue harmonieusement avec la classification automatique. Ainsi une méthode unique dont le formulaire reste simple est parvenue à s'incorporer des idées et des problèmes nombreux apparus d'abord séparément, certains depuis plusieurs décennies. »

Facteurs, codage, calculs de contribution, classification automatique, quatre mots-clé que nous allons maintenant expliquer. La critique de la validité des résultats sera faite dans le chapitre III. Mais, auparavant, il nous faut exposer la notion ici fondamentale de « distance distributionnelle », dite parfois « distance du chi² ».

1. Distance du Chi²

L'analyse des correspondances cherche à mesurer les proximités de forme entre lignes ou entre colonnes, compte tenu de leurs poids différents. Elle ne mesure donc pas les distances de deux lignes u et u' par la formule euclidienne classique de la somme des carrés des différences de coordonnées :

$$\sum_q \left(\frac{p_{uq}}{p_u} - \frac{p'_{uq}}{p_{u'}}\right)^2$$

qui donnerait trop de poids aux colonnes ayant des effectifs considérables. Mais, se séparant, comme nous l'avons vu, de l'analyse en composantes principales, elle utilise l'expression suivante de la distance dans l'espace des données :

$$d^2(u, u') = \sum_q \frac{1}{p_q}\left(\frac{p_{uq}}{p_u} - \frac{p'_{uq}}{p_{u'}}\right)^2$$

avec l'expression correspondante pour d (q, q').

Le professeur Benzecri montre que ce choix, en apparence arbitraire, est en fait fondamental.

a) Cette distance vérifie ce qu'il a appelé le principe d'équivalence distributionnelle : si on a deux points confondus u_1 et u_2 dans l'espace des Q par exemple et si on considère un seul point u_o avec la somme des poids de u_1 et u_2, les distances entre couples de points des deux espaces U et Q sont échangées; de ce fait, on ne modifie pratiquement pas les

résultats d'une analyse par correspondances si on regroupe deux rubriques très voisines en ajoutant leurs poids, ce qui explique la stabilité inhérente à cette méthode par rapport au codage des données, ce que nous appelons robustesse au chapitre III.

b) Cette distance se relie à une des lois les plus appréciées des chercheurs en statistique mathématique, celle du χ^2 (chi²). Rappelons la brièvement : soient n variables aléatoires indépendantes $x_1, x_2, \ldots, x_n$ chacune suivant la loi normale réduite N (0,1), c'est-à dire à moyenne nulle et variance unité. On pose par définition :

$$\chi^2 = x_1^2 + \ldots + x_n^2$$

χ^2 est donc une variable à n degrés de liberté et par la méthode des fonctions caractéristiques on trouve aisément que χ^2 a pour moyenne n et pour variance $2n$. K. Pearson a donné la formule explicite donnant la probabilité P qu'une valeur de χ^2 soit dépassée, ceci pour diverses valeurs de n. Tous les ouvrages de statistique mathématique en donnent des tables (1).

Soit maintenant dans la généralisation à n variables de la loi du binôme le vecteur X à n composantes $X_1, \ldots, X_n$ dont l'espérance mathématique $E(X)$ a pour composantes

$$E(X_i) = np_i$$

nous cherchons la loi limite de la quantité :

$$t^2 = \sum_i \frac{(X_i - np_i)^2}{np_i}$$

quand n augmente indéfiniment. On démontre qu'elle suit une loi de chi² à n-1 degrés de liberté et ceci permet de tester un ajustement, ce que nous avions à rappeler. On en trouvera un exemple simple dans l'annexe I où la valeur du Chi² dépassant largement celle de la table, l'hypothèse initiale cependant plausible *a priori* doit être rejetée.

Or, dans le tableau k (u, q), si on désirait comparer en utilisant le langage du calcul des probabilités la « fréquence » observée p_{uq} à celle qui résulterait de l'indépendance des lignes et des colonnes soit p $(u) \cdot p$ (q) ont écrirait une distance du χ^2 :

$$t^2 = k \sum_{u,q} \frac{(p_{uq} - p_u p_q)^2}{p_u p_q}$$

t^2 est un χ^2 à $(n\ 1)$ $(p\ 1)$ degrés de liberté (2).

Les tableaux 1 et 2, pages 52 et 53 de l'ouvrage de Lebart, Morineau et Tabard sont très précieux (IV).

2. Facteurs

L'ordinateur calcule à partir du tableau k (u, q) obtenu à partir du tableau des données transformé par l'introduction de la métrique du χ^2 (Chi²), une suite de couples de facteurs (positifs ou négatifs) traditionnellement appelés (F_1, G_1), (F_2, G_2), etc., définis sur l'ensemble des U et l'ensemble des Q. On aura pour le thème U, F_i (u) et pour la couleur Q, G_i (q).

Chaque couple de facteurs correspond à une valeur propre, nombre réel positif compris entre 0 et 1.

1. Pour n grand, supérieur à 30 par exemple, $\sqrt{2\chi^2} - 2\sqrt{2n\ 1}$ suit la loi normale réduite N $(0,1)$.

2. On aurait pu s'attendre à trouver np 1 degrés de liberté, la différence est due au fait que l'on a dû estimer les valeurs théoriques à partir des observations. On la retrouve tout au long de la statistique mathématique, la prise en compte de ce fait ayant été le grand succès de la fin du XIXe siècle et du début du XXe siècle.

On a $\lambda_1 > \lambda_2 > \lambda_3 \ldots$ La somme des λ est la trace T, le rapport $\frac{\lambda_n}{T}$ est dit taux d'inertie extraite par le n^e facteur [1]. Premier résultat à noter : pour l'interprétation, plus les λ sont grands, plus les « contrastes » dans le tableau seront marqués. Ainsi $\lambda = 1$ correspondrait pour deux thèmes de u_1 et u_2 et deux couleurs q_1 et q_2. Mais λ_1 peut valoir 0.01

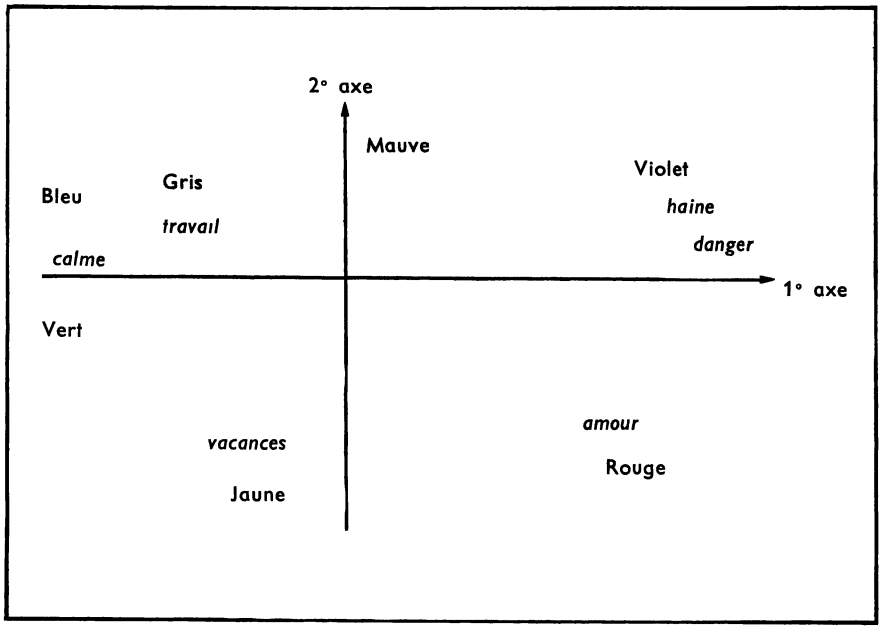

FIG. 1.

et avoir cependant une grande signification. On est donc toujours très tenté de donner aux taux d'inertie des différents axes ou aux valeurs même des λ un grand rôle dans l'interprétation, mais le chapitre III montrera que cela est très délicat.

On démontre les formules :

$$\sum_u k(u) F_\alpha(u) = 0;$$
$$\sum_q k(q) G_\alpha(q) = 0$$
$$\sum_u k(u) F_\alpha(u) F_\beta(u) = k\lambda_\alpha \delta_\alpha^\beta;$$
$$\sum_q k(q) G_\alpha(q) G_\beta(q) = k\lambda_\alpha \delta_\alpha^\beta$$

(δ_α^β symbole de Kronecker = 1 si $\alpha = \beta$, zéro si $\alpha \neq \beta$)

En d'autres termes pour l'ensemble des thèmes munis d'un poids $\frac{k(u)}{k}$ les facteurs ont une moyenne nulle et une variance λ_α; de plus deux facteurs F_α, F_β ont une corrélation égale à zéro. Ils expriment donc des « tendances » indépendantes. On a la même propriété pour l'ensemble des couleurs.

1. Dans une étude faite sur la pollution de Fos, un auteur a calculé les huit premiers taux d'inertie : en %, 21,9; 19,1; 14,9; 12,6; 9,5; 8,7; 7,2; 6,1 et s'émerveille que le total fasse 100 %, or il n'y avait que huit variables, donc il y avait exactement huit axes factoriels et huit seulement qui naturellement contiennent toute l'inertie.

3. Reconstitution des données

Notons maintenant la très importante formule de reconstitution des données :

$$k(u, q) = \frac{k(u)\,k(q)}{k}\left[1 + \sum_{\alpha} \frac{F_\alpha(u)\,G_\alpha(q)}{\sqrt{\lambda_\alpha}}\right]$$

(Le nombre de facteurs est le plus petit diminué de 1 des deux nombres lignes et colonnes, chaque terme a l'ordre de grandeur de $\sqrt{\lambda_\alpha}$ car F et G sont de cet ordre de grandeur.)

La formule limitée à quelques termes sera une formule approchée, le premier s'il était seul correspondrait à l'indépendance des thèmes et des couleurs. On représentera, dans le plan à axes factoriels (i, j), un thème par un point de coordonnées $F_i(u)$, $F_j(u)$ et une couleur par un point $G_i(u)$, $G_j(u)$. La figure 1 donne les deux premiers facteurs. Les facteurs traduisent des « tendances générales » valables pour les thèmes et les couleurs.

L'analyse concluait : « Sur la figure, les couleurs sont rangées dans l'ordre naturel; leurs associations avec les thèmes ne sont pas surprenantes. Il est toutefois très improbable qu'au simple vu des résultats de l'enquête... un psychologue ait été capable de saisir, aussi clairement que sur la figure, ces associations qui aideront par exemple au choix de la couleur d'un emballage ou d'une annonce. »

4. Formules de transition

Signalons encore les formules de transition dites « barycentriques » :

$$F_\alpha(u) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{q} \frac{G_\alpha(q)\,k(u, q)}{k(u)}$$

$$G_\alpha(u) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{u} \frac{F_\alpha(u)\,k(u, q)}{k(q)}$$

La représentation d'un thème sera donc sur les diagrammes, au facteur $\frac{1}{\sqrt{\lambda_\alpha}}$ près, au centre de gravité des couleurs pondérées et inversement : résultats très précieux pour l'interprétation.

5. Contributions absolues et relatives

Deux séries de coefficients peuvent être calculées et apportent souvent une aide à l'interprétation :

a) Les contributions absolues donnent la part prise par une variable donnée dans l'inertie d'un facteur, permettant ainsi de juger de son importance dans la constitution de ce facteur, on a :

$$C_{a_\alpha}(u) = k(u)\,F_\alpha^2(u) \quad \text{avec} \quad \sum_{u=1} C_{a_\alpha}(u) = 1$$

et naturellement :

$$C_{a_\alpha}(q) = k(q)\,G_\alpha^2(q) \quad \text{avec} \quad \sum_{q=1} C_{a_\alpha}(q) = 1$$

La somme des contributions absolues à un même facteur est l'unité.

b) Les contributions relatives donnent une idée des caractéristiques exclusives de ce facteur. On a :

$$C_{r_\alpha}(u) = \frac{\lambda_\alpha F_\alpha^2(u)}{d^2(u)} \qquad \text{avec} \qquad \sum_\alpha C_{r_\alpha}(u) = 1$$

et naturellement le même pour $C_{r_\alpha}(q)$, $d\,(u)$ est la distance dans le nuage d'un élément à l'origine des axes et C_{r_α} représente le pourcentage de d^2 dû à l'axe α. C'est aussi à la fois le carré du cosinus de l'angle avec l'axe α de la droite joignant l'élément au centre et le carré du coefficient de corrélation correspondant, ce qui rend cette notion facile à imaginer dans l'espace.

Tableau I

Partie positive de l'axe								
Direction du vent . .	120°	100°	140°	60°	80°	40°	160°	180°
Poids	50,3	20,0	29,6	5,3	5,8	10,3	39,5	30,9
100 C_r	27,1	16,9	13,0	3,1	2,2	0,5	1,1	0,1
100 C_a	4,6	2,4	1,8	0,3	0,2	0,05	0,1	0,01

Partie négative de l'axe						
Direction du vent	360°	340°	220°	280°	320°	20°
Poids	64,1	59,1	15,6	26,1	33,4	32,37
100 C_r	11,7	7,3	1,0	1,2	1,0	0,9
100 C_a	2,4	1,3	0,1	0,1	0,1	0,1

Tableau II

Partie positive de l'axe	C_r	$\sqrt{C_r}$
Durée pluviométrique	54,27	0,74
Hauteur pluviométrique	23,77	0,49
Nébulosité (8 octas) (ciel complètement couvert)	14,86	0,36
Direction du vent 120°	27,11	0,52
100°	16,88	0,41
140°	13,00	0,36
Température (10° — K°)	20,00	0,44
Partie négative de l'axe		
Nébulosité (0 octe)	39,19	0,63
(1 octe)	29,20	0,54
Direction du vent 360° (mistral)	11,72	0,34
Vitesse du vent 9 m/s	39,19	0,63
10 m/s	29,20	0,63
Humidité inférieure à 50 %	21,03	0,46
50 à 60 %	24,31	0,49

Prenons un exemple :

Dans une étude de pollution, faite par le LECES (Laboratoire d'études et de contrôle de l'environnement sidérurgique) en septembre 1975 pour la région Fos-Etang de Berre, sous la direction de M. Klein, il y avait deux groupes de variables, les concentrations en gaz sulfureux (29 100) et les données météorologiques (60). On a calculé les concentrations pour le premier axe factoriel et pour les directions du vent; le tableau I les donne pour les directions les plus intéressantes ainsi que le poids (moyenne des mesures de concentration pour la direction étudiée). Nous allons y insister un peu, car les ouvrages de la bibliographie en parlent rarement et le mérite du LECES les introduisant est grand.

1. Le total en pour cent des contributions absolues ci-dessus est 9,47 + 4,27 = 13,74. La direction du vent n'est donc pas un paramètre capital pour le premier axe. Mais si on fait le recensement de tous les C_a pour les autres variables en se limitant à 97,43 %, on trouve :

Contributions absolues d'action positives :

— durée pluviométrique	12,4	en %
— hauteur pluviométrique	0,44	
— nébulosité	22,95	
— direction du vent (tableaux ci-dessus)	9,47	
— température	5,07	

Contributions absolues d'action négative :

— humidité relative	15,23
— nébulosité	10,96
— direction du vent	4,27
— température	5,73
— vitesse du vent	10,91

Soit un total de 97,43 sur cent. La différence avec 100 % est peu significative relative par exemple aux directions du vent non examinées. L'eau (nébulosité, pluviométrie, durée et hauteur, humidité relative) apporte 70 %; le vent (direction et vitesse) 24,6 % et la température 10,80 %. Les fortes nébulosités [1] (7 et 8 octas) ont un C_a (partie positive) total de 22,9 % et les faibles (0 et 1 octe) 11,23 % (partie négative). De même, la durée de la pluie a un C_a (partie positive) de 12,4 % et les humidités relatives faibles (inférieures à 60 %) 6,07 % (partie négative). Cinq directions du vent sur 26 ont un rôle. Le groupe 300°, 120°, 140° (positivement), le groupe 360°, 340° (négativement).

Les contributions absolues varient en général avec la variable dans le même sens que les contributions relatives. Nous remarquerons, au passage, que ces termes de contributions ne sont pas très heureux car ils font penser que les deux contributions sont de même nature, ce qui est tout à fait inexact.

2. La somme des contributions relatives, C_r, pour un facteur donné n'est pas une donnée *a priori*, voici (tableau II) les valeurs supérieures à 0,10, nous avons calculé aussi $\sqrt{C_r}$ qui est le cosinus de l'angle de la variable (dans son espace) avec l'axe, ce qui visualise mieux les positions réciproques.

1. La nébulosité se mesure par huitièmes de couverture du ciel.

Sur les 60 variables, 14 seules sont retenues, la droite joignant le point représentatif au centre de gravité du nuage faisant un angle inférieur à 70° ($\cos \varphi \geqslant 0{,}34$), la plus proche est celle de la durée pluviométrique (40°); la plus éloignée, le mistral (70°); les 46 autres variables sont à plus de 70°. Cela donne, sans nécessiter une grande maîtrise de l'espace à 60 dimensions, une image très expressive de la position par rapport à l'axe des différents vecteurs dans le nuage et une très bonne description par rapport au premier axe.

La contribution absolue de la durée pluviométrique à l'inertie de cet axe n'était pas négligeable (un huitième environ); le point représentatif du mistral, au contraire, a à la fois une direction très éloignée de l'axe et une très faible contribution à l'inertie totale (2,41 %). Tout cela est assez net.

6. Codage sous forme disjonctive complète

Nous l'expliquerons sur un exemple (l'analyse des niveaux de condition de vie au Liban en 1959-1960 (*cf.* 384)). Soixante villages avaient été choisis au mieux; à chaque village correspondait une ligne du tableau des données. Cent cinquante cinq questions avaient été posées pour chaque village, relevant de neuf niveaux du standard de vie : sanitaire, économique et technique, domestique, résidentiel, habitat, scolaire, culturel, familial, social, une dixième concernait les « données générales » sorte de résumé. Pour chaque question, on répondait par une note 0 (très mauvais), 1, 2, 3 ou 4 (très bon). Mais, pour assurer une indépendance maximum des résultats vis-à-vis du codage (*cf.* Cahiers 1977, n° 1) la note est dédoublée sur deux colonnes (codage disjonctif) : une première colonne pour chaque question comporte la note x, une seconde la note $4\,x$. Il y a ainsi symétrie, par exemple entre le très mauvais et le très bon. Au total donc, un tableau des données 60×310.

Chaque village a donc en principe deux notes de somme 4, $k\,(u)$ devrait être $155 \times 4 = 620$ pour tout u, mais il y a des questions sans réponses et de même pour q; ainsi pour les villages agricoles, les villages stations de tourisme ne sont pas notés de même que les villages chrétiens ne sont pas utilisés pour la question polygamie.

On notera, avec intérêt, que les quatre premières dimensions donnent pour le problème du Liban des pourcentages d'inertie de :

27,66; 6,09; 4,74; 4,45 au total 42,94 %, ce qui précise la concentration des données dans cet espace réduit, une partie importante des 57 % restants due à l'échantillonnage par exemple est probablement peu significative.

Pour donner au lecteur l'envie de lire le texte de l'étude du Liban, nous indiquons que les quatre premiers facteurs furent interprétés ainsi, grâce à plusieurs diagrammes de couples de facteurs :

— 1 facteur de niveau général;
— 2 opposition entre traditions d'instruction et ouverture au progrès technique;
— 3 état d'équipement dans la société villageoise;
— 4 facteur domestique fondé sur le rôle dévolu à la femme.

7. Classification automatique

Elle cherche à construire des classes « homogènes », ce qui était le problème des premiers botanistes (Linne) : espèces, genre, familles, etc., en partant d'un tableau de données caractérisant chaque plante par des mesures (des « taxons », d'où la « taxonomie » [1]). On

1. Le tome I du professeur Benzecri est intitulé « Taxinomie » (le petit Larousse donne les deux orthographes). En 615 pages, il comporte, après une introduction qu'il relie à l'analyse des données, neuf leçons théoriques, treize applications et deux programmes de calcul (*Cf.* B1 et 3).

peut observer, grâce à la classification automatique, des classements imprévus ou ne pas observer des classements prévus. Le professeur Benzecri est un peu provocateur : « Là où le temps manque pour familiariser l'homme avec son objet et développer l'intuition, une synthèse automatique peut être préconisée ». Aujourd'hui, les chercheurs combinent presque toujours analyse de données par correspondance et classification automatique, cette dernière ayant la préférence quand ils recherchent l'association des variables entre elles (lignes à lignes ou colonnes à colonnes).

On distingue les méthodes par leur procédé d'information : hiérarchie, arbre, partition, typologie, « classes empiétantes » qui relèvent toutes de la même méthodologie : on se donne un indice de ressemblance, par exemple entre l'objet a caractérisé par a_1, a_2, ..., a_n et l'objet b (b_1, b_2, ... b_n) qui aura la forme d'une distance.

Un bon classement réunit les objets qui sont proches et sépare ceux qui sont loin. L'ordinateur permet la construction automatique d'un « arbre » en quelques minutes. On a pu obtenir, en 1975, des partitions fournissant des classes bien écartées pour 40 000 personnes ayant fait l'objet d'une enquête (*cf.* B5).

L'algorithme de classification le plus célèbre aujourd'hui et sans doute un des plus efficaces est celui d'Edwin Diday, de l'IRIA, dit des « nuées dynamiques » (1970). Les méthodes de classification automatique connaissent, dans le monde entier, un succès prodigieux. Chacun a sa méthode. On a compté plus de mille publications par an d'après R.-M. Cormack (*cf.* B6), qui a écrit : « Il est presque plus facile, pour un chercheur, d'inventer à propos d'un recueil de données un nouvel algorithme que de tirer quelque chose de ces données. »

3. Validité des résultats

A de nombreuses reprises, dans les deux tomes du professeur Benzecri, celui-ci a mis en évidence les difficultés que peut entraîner l'interprétation des résultats de l'analyse des correspondances et multiplié les mises en garde. Mais sa marche continue en avant, à notre connaissance, ne lui a pas encore permis d'écrire la synthèse correspondante. On trouvera ci après, un exposé modeste et provisoire de nos premières réflexions.

Ce chapitre pourra être utile à tous ceux qui commandent des études aux spécialistes et désirent pouvoir juger de leur validité, en les aidant à poser des questions.

A — *Stabilité et robustesse*

Nous avons dans l'introduction insisté sur la nécessité de remettre à sa place exacte la fameuse « courbe en cloche » et promis de la préciser en utilisant le concept très récent de robustesse. Lichnerowicz l'ignorait peut-être quand il écrivit (*cf.* B6) :

« A ma connaissance, il n'existe pas d'études suffisantes concernant, pour un modèle choisi, la stabilité des conclusions par rapport à de faibles variations des données. Il n'existe pratiquement pas d'étude sur cette stabilité par rapport à de petites altérations des postulats du modèle. »

Par contre, Malinvaud, dès son ouvrage de 1954 (*cf.* B6), étudie page 76 les tests ou estimateurs robustes, c'est-à-dire « peu sensibles aux hypothèses du modèle ». Il distingue l'insensibilité vis-à-vis des erreurs de données et celle vis-à-vis des variations du modèle.

Plusieurs types de modèles sont particulièrement importants. Il a étudié la robustesse de diverses hypothèses : normalité, hétéroscedasticité (erreurs indépendantes entre elles mais à variances inégales), indépendance par rapport aux variables exogènes, absence de moments du troisième ordre, etc. (1).

Mais ce n'est qu'un début et la théorie en est, chaque jour, de plus en plus élaborée; elle s'applique aujourd'hui à des domaines de plus en plus nombreux de la statistique.

La quarantième Session de l'Institut international de statistique (I. I. S.) (Varsovie 1975) a consacré une séance aux statistiques robustes (tome 1, page 375 et tome 3 page 33). Une des communications, celle de A. Dempster, donne 93 références, on y retrouve l'étude de 1961 de W. James et C. Stein que nous avons pris comme exemple — annexe II — et que le lecteur parcourera avec intérêt s'il est sportif.

La théorie comporte, aujourd'hui, des écoles assez divergentes. En général, le niveau mathématique y est fort élevé, faisant appel par exemple, récemment, à la notion de « capacité » due au mathématicien Choquet.

La loi normale sort de ces études fort malmenée. Nous nous contenterons de citer une remarque en séance de l'I. I. S. de l'Américain, J. Stigler. Il avait rappelé le jugement ironique de M. Poincaré reproduisant G. Lippmann : « Chacun croit à la distribution normale, les mathématiciens parce qu'ils pensent que c'est un fait expérimental, les expérimentateurs parce qu'ils pensent que c'est un théorème de mathématiques. » Stigler écrit : « Aujourd'hui, semble t-il, personne ou presque n'y croit plus » et il continue, dans le style de Lippmann, « les expérimentateurs ne la supposent plus parce que les mathématiciens leur ont dit que ce n'était plus nécessaire et les mathématiciens savent qu'elle n'est plus souvent expérimentalement exacte, surtout quand il y a plus d'une variable; car on sait, aujourd'hui que les queues de distributions sont toujours nettement plus épaisses que celles de la loi normale. La méthode des moindres carrés, comme la loi de Laplace qui lui est souvent reliée, serait totalement non-robuste ».

Cette nouvelle branche de la statistique, la robustesse, vise en fait jusqu'ici les distributions unidimensionnelles. Il serait très précieux qu'elle s'étende à l'analyse des données. A notre connaissance, il n'y a presque rien, excepté par simulation, procédé auquel le professeur Benzecri fait une très brève allusion, mais dont Lebart s'est fait fort utilement une spécialité, comme on va le voir maintenant.

B — *Valeurs propres de taux d'inertie, nombre de dimensions interprétables*

On doit beaucoup sur ces points aux recherches de L. Lebart (*cf.* B6). Contrairement à l'opinion presque générale, il estime : « l'utilisation de ces paramètres n'est pas toujours justifiée en pratique. Leur interprétation est de plus délicate... car ils sont étroitement liés au codage des données ».

Il a donc cherché à donner des « garde-fous » en calculant des seuils de signification de ces paramètres. Il a construit des tableaux des données avec indépendance totale entre lignes et colonnes, les valeurs propres et les taux d'inertie résulteront donc seulement des fluctuations d'échantillonnage et peuvent être considérés comme des seuils. Toutes valeurs des paramètres en dessous, ou seulement voisins, seront donc sans signification. On ne connaît

1. La loi de Pareto, sur l'inégalité des revenus, n'a pas de moments du troisième ordre, mais la loi de l'effet proportionnel (loi logarithmico-normale) qui étudie les mêmes données en a. Leur robustesse sera donc différente.

pas aujourd'hui la loi mathématique des valeurs propres dans l'hypothèse d'indépendance, le plus célèbre traité de statistique classique (Kendall et Stuart, 1961) admettait faussement qu'elles suivaient des lois du chi². En fait, on ne connaît, aujourd'hui, que des approximations (loi de Wishart). L. Lebart a donné, grâce à 20 000 analyses environ de tableaux obtenus par simulation et utilisant chacune k nombres « pseudoaléatoires (1) » (1 000, 5 000, 10 000), 5 abaques et 26 pages de tableaux de seuils. Ceux-ci donnent, pour 170 tableaux de données (6×6 à 50×100), les estimations des moyennes, médianes, écart-types, seuils à 0,05 des cinq premières paires de paramètres, valeurs propres et pourcentages d'inertie. Ils représentent une documentation très précieuse pour l'étude de la validité des résultats. C'est un travail considérable.

1. Taux d'inertie

L. Lebart veut dissuader les chercheurs d'utiliser sans réfléchir, comme ils en ont pris l'habitude, les taux d'inertie pour « noter la qualité d'une représentation » et de les interpréter en « pourcentage d'information ». Il donne 4 exemples assez difficiles à suivre, mais très impressionnants.

Le premier montre comment des « options » de calcul peuvent, pour un même ensemble d'observations et pour une même description finale, modifier totalement des paramètres auxquels on aurait voulu attribuer un « pouvoir explicatif ».

Le deuxième exemple donne des cas où on peut reconstituer totalement une structure donnée à l'avance à partir de résultats dont on peut rendre le total des taux d'inertie aussi petit que l'on veut.

Le troisième a été construit pour montrer à l'évidence que les taux d'inertie reflètent les conditions de la prise des données.

Dans le quatrième, le choix des variables influence le total de l'inertie utilisée. Ceci s'obtient lorsque l'on complète un tableau par des lignes (ou des colonnes) de « bruit blanc ». Le taux d'inertie expliqué par les facteurs décroît, bien que ceux-ci soient inchangés.

On n'hésite pas en pratique, remarque-t-il, pour obtenir des taux « honorables », à recommencer une étude en supprimant des variables ou des observations, alors que des taux d'inertie faibles peuvent donner des représentations de bonne qualité. Le coefficient de corrélation multiple de la statistique classique est une mesure optimiste de la qualité d'une régression, alors que les taux d'inertie mesurent de façon pessimiste la qualité d'une représentation. Tout cela vise à « décourager l'utilisation des taux d'inertie comme unique critère d'évaluation », mais évidemment ne facilite pas la tâche des utilisateurs.

Il a essayé aussi d'approfondir la notion même d'information et les différents concepts réunis sous ce nom. Sa conclusion essentielle est qu'en analyse de données, il vaudrait mieux parler de forme que d'information. La validité d'une représentation serait mieux établie en étudiant la stabilité de la forme plutôt qu'en mesurant une quantité d'information. Tout ceci est à suivre.

2. Nombre de dimensions à interpréter

Le professeur Benzecri (*cf.* B6) (tome IIA, n° 2, § 3) est de plus en plus optimiste dans la découverte de « faits de structures que pourrait révéler un tableau de données ». Au début, lui et ses élèves ne retenaient, pour des raisons techniques (largeur de la feuille d'imprimante),

1. On sait qu'il existe des techniques très puissantes pour obtenir ces nombres.

que cinq dimensions et l'interprétation dépassait rarement le quatrième couple de facteurs. Il était alors très limité par les possibilités de calcul.

La première édition de son traité comporte un algorithme rapide (Golub et Reinsch), mais la deuxième édition (1976) propose le programme IBM Symor, qui peut traiter des tableaux dont le nombre de lignes est aussi grand que l'on veut (plusieurs dizaines de milliers par exemple) et le nombre de colonnes peut atteindre deux cents. Au total donc dix millions de données (200 × 50 000), ce qui peut éviter des distorsions par simplifications. On mesurera ainsi le chemin parcouru en moins de deux ans en notant qu'aux Treizièmes Journées de l'hydraulique, de la Société hydrotechnique de France, les représentants de la Météorologie française (question II, rapport n° 8 — 16-18 septembre 1974) écrivaient pour un tableau de 80 000 données : « Que faire d'une masse d'informations aussi considérable du point de vue informatique, bien qu'encore insuffisante du point de vue météorologique? Il faut condenser, comprimer, trier cette information, sélectionner la partie utile et rejeter le reste, car on ne saurait traiter des matrices d'un tel ordre. » Tout cela a disparu.

Où s'arrêter dans la suite de facteurs? se demande le professeur Benzecri. Pour des événements presque indépendants, le test de chi^2 permettrait, en principe, de fixer le moment où les fluctuations d'échantillonnages deviendraient prépondérantes. Mais ce cas est fort rare. Il lui paraît nécessaire pour décider d'examiner la suite des valeurs propres et des taux d'inertie.

Les conclusions ne seront jamais impératives, il s'agira surtout d'indications. Il donne ainsi quelques informations « d'expérience ». Ainsi on peut avoir un premier facteur très sûr avec un taux d'inertie γ_1 de 10 %, les grandes valeurs ne sont donc pas nécessaires. Des facteurs d'importance très voisine donneront ou non, suivant les cas, des interprétations indépendantes; une valeur λ_1 ou λ_2 de 0,2 peut donner un résultat intéressant; des valeurs propres très faibles ($\lambda_1 = 10^{-3}$, par exemple) ne doivent pas empêcher l'examen et peuvent donner des axes interprétables. Ce serait donc affaire de cas particuliers : ainsi le tableau III donne les valeurs propres et les taux d'inertie pour les dix premiers axes factoriels dans une étude statistique très complète des conditions du développement de la Colombie (un tableau de 45 × 612) :

(Tome II C, § 6.4.4)

Les valeurs propres sont très faibles, le plus grand λ vaut seulement 0,06.

Les dix premiers facteurs n'expliquent que 61 % de l'inertie et l'analyse a été faite jusqu'au sixième facteur : voici l'interprétation :

— facteur de niveau général;
— facteur d'opposition entre niveau technico-économique et niveau spirituel;
— facteur lié à la salubrité;
— facteur de stabilité sociale;
— facteur d'opposition entre religion et irreligion;
— facteur lié à la condition de la femme.

On s'est arrêté là, car au septième facteur une variable à elle-seule représente le quart des poids.

TABLEAU III

1 000 λ	62	25	21	18	14	14	13	12	11	10
1 000 t	189	76	63	54	44	42	39	37	33	31

Au total, l'analyse a ramené près de 30 000 données à six facteurs principaux étudiés en détail sauf le premier, évident, dans quatre graphiques concernant chacun deux facteurs particuliers (2,3) (4,5) (2,5) (2,6).

3. Cas particuliers

Certains cas sont presque pathologiques. Voici par exemple une analyse faite en octobre 1975 par le LECES (Laboratoire d'études et de contrôle de l'environnement sidérurgique) sur la pollution SO^2 dans la région de Fos-Étang de Berre. Un tableau de 3 534 observations a donné en valeurs propres et en taux d'inertie :

TABLEAU IV

1 000 λ_i	859	858	852	849	845	842	840	853	834	833
1 000 t_i	60	60	60	59	59	59	59	59	58	58

Les dix premières valeurs contiennent 59 % de l'inertie, mais ici toutes les « structures factorielles » ont la même importance et les valeurs propres sont très voisines. Le nuage de données se confond presque avec une sphère à dix dimensions.

Seule la direction du vent semblerait avoir une importance et les autres variables météorologiques n'auraient qu'un rôle mineur (pression, température, humidité relative vitesse du vent, nébulosité, hauteurs pluviométriques, durées). Mais est-ce exact?

C — *L'effet Guttmann*

En novembre 1973, J. Verneaux a présenté, dans la troisième partie d'une importante thèse de sciences naturelles sur les cours d'eau de Franche-Comté, une étude d'analyse de données très complète. Il disposait de tableaux de distribution des trente espèces de poissons sur 123 stations le long des différents cours d'eau.

Il essaya, d'abord, différents « tests d'affinité », ensuite plusieurs procédés de classification automatique, puis l'analyse en composantes principales et enfin l'analyse des correspondances, faisant aussi progressivement le tour des algorithmes connus pour aboutir à ceux du professeur Benzecri.

L'analyse en composantes principales se révéla, d'après lui, peu satisfaisante, parfois même aberrante. « Ainsi le poisson-chat, la truite arc-en-ciel, le blageon et la perche-soleil se trouvaient étroitement associés ! »

Un comble ! D'autre part, elle présentait « l'importante lacune de ne pas établir des relations entre espèces et stations ». L'analyse des correspondances fut donc finalement retenue, permettant la mise en évidence de structures c'est-à-dire « d'une organisation logique des rapports entre individus et caractères ».

Elle mit en évidence (*cf.* fig. 2) une distribution des espèces selon une courbe en *U* dans le plan des deux premiers axes factoriels « représentant pour le réseau du Doubs 70 % de l'inertie expliquée totale ». L'auteur estima pouvoir ainsi « distinguer dix niveaux typologiques se succédant le long de la structure dans l'ordre de la succession des stations le long d'un système théorique, des sources à l'embouchure ».

Ces résultats auraient été « vérifiés et complétés par l'examen des travaux relatifs à divers systèmes aquatiques européens » (Seine, Ain, Saône, Cher, Massif Central, Bretagne, Alsace, Vosges, etc.). Leurs conséquences seraient très importantes : notions d'espèces « écologiquement équivalentes », notion d'appartenance écologique. Aussi, l'auteur y insiste dans une étude pour le ministère de l'Environnement qui nous a été aimablement communiquée.

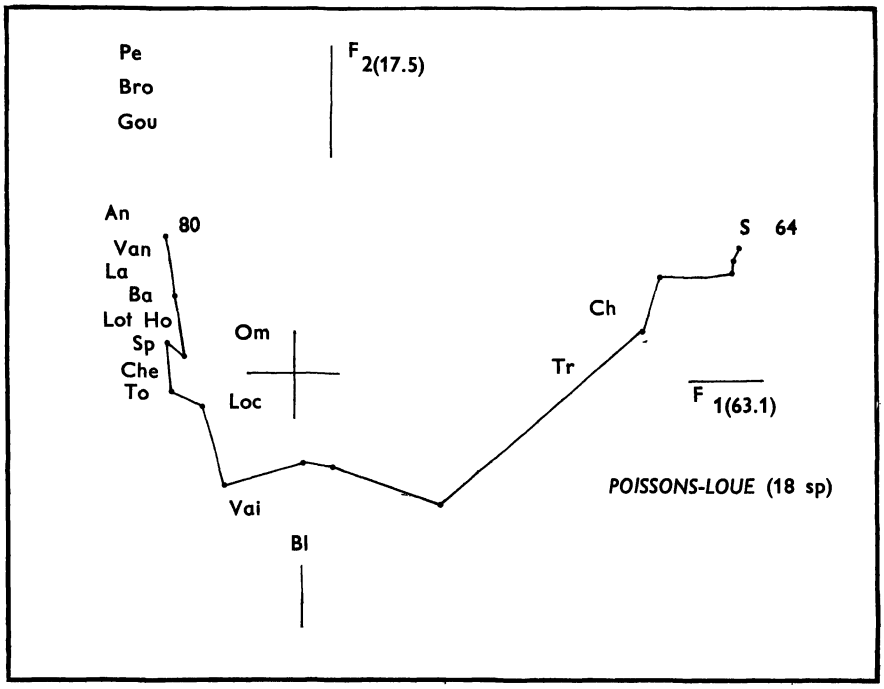

FIG. 2. — Analyse factorielle des correspondances espèces-stations : structure ichtyologique de la Loue dans le plan des deux premiers facteurs d'analyse.

TABLEAU V

Q / I	Oui				Non			
	1	2	3	4	1	2	3	4
Sujets types 1	1	1	1	1				
2		1	1	1	1			
3			1	1	1	1		
4				1	1	1	1	
5					1	1	1	1

Or, cette structure en U se retrouve dans de très nombreuses études de toutes sortes et le professeur Benzecri lui donne le nom d'effet Guttmann. Il importerait donc de savoir si elle apporte, dans ce cas particulier, une structure spécifique au problème ou traduit une structure générale.

Il y a effet Guttmann, pour le professeur Benzecri (*cf.* B1) (II B, n° 7,3) quand « un phénomène fondamentalement unidimensionnel est sousjacent à l'analyse des correspon-

dances ». Il part des scalogrammes introduits par Guttmann dès 1941 (c'est-à-dire des tableaux où l'on peut, par permutation des lignes et colonnes, créer une bande centrale en parallèlogramme), qui a démontré qu'il en résulte « une suite de facteurs qui bien que non corrélés entre eux, n'en sont pas moins tous liés fonctionnellement, le facteur de rang p étant par exemple un polynôme de degré p sur le premier facteur ». Cette démonstration a d'ailleurs fait l'objet d'une question à l'examen du D E A de Statistique en septembre 1971. (Guttmann ne connaissait pas l'ordinateur et travaillait en construisant des modèles (scalogramme, simplexe, circumplex, radex, etc.) recherchant l'effet de plus en plus complexe de leurs structures sur les calculs).

Voici (tableau V) un exemple de scalogramme (*cf.* B3) (Cahiers, § 2.5.3) comportant quatre questions et cinq sujets.

Le calcul exact des facteurs donne, ici, des polynômes de Legendre :

$$F_2 = a\,(2\,F_1^2 - 1); \qquad F^2 = b\,(10\,F_1^3 - 9\,F_1); \text{ etc.}$$

Ce phénomène peut être aussi dans d'autres cas multidimensionnel, les facteurs étant des polynômes suivant les deux premiers facteurs (ou plus). D'autres modèles peuvent donner, à la place des polynômes, une suite de facteurs en sin pt, cos pt, etc.

Dans le plan (F_1, F_2) les points relatifs à la Loue se répartissent, aux erreurs d'échantillonnage près, sur une parabole dont le sommet a une ordonnée négative et dont l'axe est F_2. La courbe en U pourrait donc être due seulement à des particularités de codage. En tout cas, un examen très sérieux, par exemple suivant les méthodes préconisées par Lebart, nous paraîtrait s'imposer avant toute conclusion.

D'autres structures de Guttmann donneraient naissance à des phénomènes du même genre : il faut le savoir et s'en méfier.

En résumé, comme nous l'a écrit Michel Volle, « l'analyse donne des résultats qui ne dépendent pas de l'ordre donné *a priori* aux lignes et colonnes du tableau; si l'on intervertit quelques lignes ou colonnes, cela ne modifie pas le résultat graphique obtenu. L'effet Guttmann permet donc de mettre en évidence cette propriété (cachée jusqu'alors) du tableau : pouvoir se mettre sous forme quasi diagonale, lorsque les lignes et les colonnes sont rangées dans l'ordre des coordonnées des points qui les représentent sur le premier axe ».

Conclusions

Dans sa préface à l'ouvrage de F. Caillez et J. P. Pages (*cf.* B4), G. Morlat résume, avec humour, l'évolution de la statistique dans les cinquante dernières années. Le voici, d'abord, citant De Finetti; pour la période 1930-1950 : « On s'est aperçu, à un certain moment, que la statistique mathématique était bâtie sur du sable (les probabilités *a priori* si discutées) alors on a retiré le sable et la statistique a été bâtie sur rien ! » Puis, pour la période récente caractérisée par l'arrivée de l'analyse des données : « Contrairement à la statistique néo-bayesienne, qui vise à consolider la formalisation de l'induction, en consolidant le « sable » sur lequel s'appuyait la théorie statistique, on peut dire que l'on a retiré, en analyse des données, non seulement le sable mais tout ce qui reposait dessus, c'est-à-dire le modèle probabiliste. Dès lors, que reste-t il de la statistique? Pas grand chose aux yeux des classiques ou des néobayesiens, sinon des procédés descriptifs. »

C'est tout à fait exact à nos yeux, ce ne sont que des algorithmes de description; mais les services rendus aux utilisateurs dès maintenant seraient tels, les applications formeraient un tel « torrent », en un mot l'efficacité de la description serait telle, grâce à l'arrivée des ordinateurs, que G. Morlat lui réserve une place de tout premier plan dans les outils du chercheur : « L'absence de moyens de calcul avait longtemps barré une des voies d'exploration les plus fécondes du réel », mais il conclut de façon inattendue : « pour quelques années encore, priorité à l'analyse des données. »

Pourquoi « quelques années encore »? Parce que, écrit G. Morlat, un jour le chercheur « voudra prévoir les phénomènes (et il lui faudra adopter un modèle probabiliste, ...) ou éclairer des décisions et il se posera des problèmes du type de ceux qu'abordait la statistique mathématique classique ». Retour éternel des disciplines abandonnées puis reprises, écrirons-nous !

Revenons au présent : il semble bien qu'aujourd'hui les nombreux chercheurs qui, chaque année, sortent de l'Université avec un simple D. E. A. et, une fois dans l'industrie, se précipitent avec avidité sur ces algorithmes n'aient pas eu le temps de recevoir une formation suffisante pour en éviter les pièges et obtenir une interprétation correcte des analyses obtenues trop automatiquement. Déjà en 1952, on écrivait sur « The uses et abuses of factor analysis ». Mais écoutons Lebart. Le chapitre III et son étude de : « Validité des résultats en analyse des données » (*cf.* B6) est intitulé : « L'étude critique de l'utilisation des méthodes. » Il y examine successivement quatre types de critiques :

— l'utilisation inconsidérée;
— les résultats sont des évidences;
— l'ordinateur fait perdre le contact avec le matériel brut;
— on ne travaille pas directement sur des structures, mais sur des tableaux.

a) L. Lebart parle des « monomaniaques de la technique dont il faut refréner l'ardeur dévastatrice » et les explique par le recours que les analyses « apportent au chercheur des sciences sociales » qui étoffe ainsi son travail ou par l'activité commerciale d'un statisticien avide de chiffre d'affaires. Pour lui, ces utilisations abusives ou maladroites sont un phénomène social lié à la division du travail, à la parcellisation des tâches, ce phénomène n'épargnerait aucune science.

b) L. Lebart remarque que l'évidence *a posteriori* n'est pas toujours celle *a priori*, ce qui est facile à démontrer en demandant aux utilisateurs d'esquisser à l'avance une certaine synthèse des résultats qu'ils attendent, « dans les sciences humaines, beaucoup d'assertions sont possibles, aussi les résultats sont toujours évidents *a posteriori* ».

c) L'usage de l'ordinateur rendrait le travail trop facile et introduirait des spécialisations regrettables (utilisateurs, statisticiens, informaticiens). Autrefois le statisticien, avant de se lancer dans un calcul fastidieux et coûteux, réfléchissait longuement, faisait des essais et surtout restait en contact avec les données. Il pensait agir sur la suite des calculs et leur interprétation. Aujourd'hui, écrit G. Morlat « L'application des techniques d'analyse de données met en évidence des traits parfois tellement saillants que l'on est tenté d'en rester là, et en effet, le client venu avec des masses de chiffres repart content avec quelques graphiques et des idées pour un bout de temps. »

d) Il s'agit, en fait, du rôle des hypothèses sous-jacentes, qu'elles soient ou non exprimées, car il est toujours difficile de prévoir qu'elles vont introduire ou non des erreurs. On a d'une part les hypothèses indispensables relatives au contenu même de la recherche qui permettent de définir les données à observer et, d'autre part les hypothèses dites de

« commodité » suivant G. Morlat. Ainsi, on admet implicitement que la distribution des « erreurs » (dans les observations ou par l'action de facteurs inconnus) se fait suivant la loi normale ou avec homoscedasticité (existence d'une variance) par exemple.

Leur importance dépendra de la « robustesse » de l'algorithme adopté et l'un des grands avantages de l'analyse par correspondances serait précisément dans le fait qu'elle suppose presque toujours les hypothèses de « commodité » les plus faibles.

L. Lebart insiste enfin, à bon droit, sur la nécessité de respecter très strictement les règles d'interprétation dans la recherche des structures. Il recommande de fréquents « étalonnages » en faisant l'analyse de structures connues, il rappelle la dépendance étroite des résultats envers le « codage ». Si celui-ci change, les règles d'interprétation changent. Il rappelle : « Le codage donne un sens aux distances entre lignes et colonnes avant toute analyse. »

Lors de la discussion (fin 1976) de la communication de E. Malinvaud à la Société de statistique de Paris (*cf.* B6), nous avons exprimé succinctement notre position et nous prendrons comme conclusion de cet article les termes mêmes de notre intervention. « On s'est donc lancé dans l'utilisation des méthodes mises en honneur par le professeur Benzecri. Hélas, sur dix études sur la pollution de la région Fos Berre provenant toutes de thésards 3e cycle ou de bureaux d'études spécialisés, sept ou huit sont fort critiquables. L'un constate des courbes paraboliques dans les résultats, s'y intéresse et essaye de les interpréter. Il oublie qu'il s'agit seulement de l'effet Guttmann. D'autres utilisent des moyennes journalières oubliant que le phénomène est météorologique et à l'échelle de l'heure subit des variations de un à dix. Ne résistent guère que les études de proximité recherchant les profils semblables. On sent que cette nouvelle technologie a été très mal digérée, sans doute trop vite apprise et que le désir marqué de ne pas avoir d'idée *a priori*, de ne pas faire de « modèles » a supprimé dans la plupart des cas tout examen préalable sérieux du phénomène physique. »

Il y a dans ces méthodes l'immense intérêt de pouvoir grâce à l'ordinateur traiter facilement des problèmes autrefois intouchables. Il y a certainement beaucoup à gagner en réduisant au minimum les idées *a priori*, surtout si elles sont basées sur des acceptations irraisonnées de loi normale à plusieurs variables. L'avenir de ces méthodes est très grand. Mais leur utilisation reste très délicate.

Il est très agréable d'avoir devant soi au lieu de dizaines de milliers de données (c'est l'ordre de grandeur habituel), cinq ou six axes principaux détenteurs de presque toute l'information et quelques graphiques plans. Mais leur interprétation est très difficile. Les premières valeurs propres suggèrent des conclusions qui paraissent alors évidentes, les suivantes (3 à 6 par exemple) paraissent de peu de valeur. Il y a des pièges partout.

A ne pas mettre entre toutes les mains. Exiger qu'une réflexion approfondie puisse précéder le choix des données et la mise en route d'un programme. Ne pas se laisser impressionner par la technologie des calculs. Tels sont les conseils à donner aux utilisateurs. Ils paraissent aujourd'hui absolument nécessaires.

Ce serait, cependant, à notre avis, une grave erreur de se refuser à utiliser pleinement l'analyse des données et particulièrement celle par correspondances sous prétexte que cela est difficile.

Elle progressera en marchant comme elle l'a fait depuis une décennie si elle reste lucide et conserve un contact étroit avec la statistique mathématique classique. Il serait particulièrement opportun de commencer à l'enseigner dans les écoles d'ingénieurs où elle est peu connue afin de développer son action sur la recherche industrielle. Elle y gagnera en rigueur et en efficacité.

ANNEXE I

Test du Chi² et autres tests

En voici un exemple très simple, emprunté au petit livre déjà ancien de A. Vessereau (*cf.* B 6) :

53 680 familles de huit enfants ont été classées d'après la répartition des enfants en garçons et filles (tableau VI).

Cela fait 221 023 garçons contre 208 417 filles, soient 515 garçons sur 1 000. Il est naturel d'admettre comme hypothèse de travail que la probabilité d'un garçon est 0,515 et celle d'une fille 0,485 et d'appliquer le schéma de « l'urne », c'est-à-dire la loi binomiale. Les effectifs théoriques du nombre de familles deviennent les suivants (tableau VII) :

Le test du chi² :

$$\Sigma \frac{(n_1 - n_2)^2}{n_2} \text{ donne ici } 92{,}22$$

Il y a huit tirages indépendants, nos propres calculs sur la formule du χ^2 donnent une probabilité pour que le schéma théorique donne une valeur du χ^2 égale ou supérieure à 92,22, extrêmement faible $1{,}6\ 10^{-16}$. (Pour une probabilité de 0,001, $\chi^2 = 26{,}13$.)

Le schéma de l'urne ne peut être une explication et il faut, écrit Vessereau, étudier l'influence de la mortalité infantile, de celle suivant l'âge et le sexe, de la présence de jumeaux, etc.

ANNEXE II

Estimateur de James et Stein

Excellent exemple des surprises que peut apporter une étude peu orthodoxe de statistique mathématique, il permet ainsi de compléter les remarques du Chapitre III sur la robustesse. Ch. Stein, en 1955, démontra *à la surprise générale*, que la moyenne arithmétique d'une série d'observations d'un phénomène n'est pas la meilleure estimation, au sens des moindres carrés, de la vraie valeur de ce phénomène quand on dispose aussi de plus de deux séries de phénomènes non identiques mais de même nature. Avec James, Stein donna, ensuite (1961), l'expression de cette estimation optimum. Soit y_i la moyenne d'une série d'observations, chacune d'écart quadratique σ, $\bar{y}$ la moyenne générale, k le nombre de séries, l'estimation z_i de la vraie valeur d'une série :

$$z_i = \bar{y} + c(y_i - \bar{y})$$

avec

$$c = 1 - \frac{(k-3)\sigma^2}{\Sigma (y_i - \bar{y})^2}$$

est « meilleure » que l'estimation classique y_i. Pour une définition rigoureuse du mot « meilleure », on voudra bien se reporter à la théorie de la décision, par exemple dans l'ouvrage de E. Malinvaud.

Un exemple, pris dans Scientific American (mai 1977), le précise.

Il est relatif au base-ball aux U.S.A. On connaît les résultats d'un certain nombre de « batteurs » après leur 45 premiers essais et on désire une estimation de ce que chacun d'eux fera dans la suite de la saison pour environ 400 autres essais. La méthode classique, depuis Legendre (1806) et GAUSS (1821), consiste à faire pour chacun la moyenne de ses 45 essais connus. Ce n'est une bonne méthode, déclare Stein, que si on a seulement les résultats de 1 ou 2 batteurs. Au-dessus de deux, il faut appliquer la formule ci-dessus avec c. L'étude a été faite sur 18 batteurs dont on connait donc

TABLEAU VI

Nombre de garçons	0	1	2	3	4	5	6	7	8
n_i nombre de familles	215	1 485	5 331	10 649	14 959	11 929	6 678	2 092	342

TABLEAU VII

n_i nombre de familles *théorique*	165	1 402	5 203	11 035	14 627	12 410	6 580	1 994	264

les y_i et le σ : la moyenne générale, était 0,265 et le calcul de c donnait 0,212. Le meilleur batteur avait $y_i = 0{,}40$ le plus mauvais 0,178, mais l'estimation donnait z_i égal à 0,294 pour le premier et 0,247 pour le dernier. Comme on connaît les résultats du reste de la saison (400 essais environ), en prenant ceux-ci pour déterminer la vraie valeur on peut calculer la somme des carrés des écarts des y_i ou des z_i. On trouve respectivement 0,077 et 0.022. L'estimateur James Stein est donc 3,5 fois plus exact que celui de Legendre-Gauss (la moyenne arithmétique). Or, il a réduit l'écart des extrêmes y_i près de 5 fois, ce qui est énorme.

Nous trouvons dans cet estimateur un exemple de plus des transformations profondes qu'apporte l'arrivée de plusieurs dimensions, ici le passage à la quatrième *est décisif*.

BIBLIOGRAPHIE SOMMAIRE

I. L'ouvrage de base, difficile mais irremplaçable, est formé des deux tomes de l'ouvrage du professeur J.-B. Benzecri et de ses 70 collaborateurs :

1. La Taxinomie, 615 pages,
2. L'analyse des correspondances, 619 pages, publiés chez Dunod, 1re édition 1973, 2e édition 1976.

II. On trouvera, ci-après, la liste des études *pratiques* contenues dans ces deux tomes. Seul le spécialiste, possédant une bonne connaissance de l'algèbre linéaire et multilinéaire, lira les études théoriques.

A — *Application de la taxinomie numérique (tome I, Benzecri)*

1. Le genre Myosotis : exemples d'application des méthodes numériques en taxinomie végétale.
2. A propos de quelques méthodes de classification en phyto-sociologie.
3. Compléments à une étude de phyto sociologie alpine.
3. La végétation de l'étage subalpin du bassin supérieur de la Tinée.
4. Un exemple d'analyse factorielle de courbes de croissance (porcins).
5. L'étude de la forme du squelette de la tête chez le Praomys (petit rongeur africain).
6. Sur le genre Cricetomys (rats géants africains).
7. Le genre Cricetomys (II).
8. Sur la craniométrie des Lémuriens.
9. Application de l'analyse des correspondances à l'écologie des Orthoptères.
10. L'écologie des Collemboles.
11. Extrait d'une étude anthropométrique.
12. La construction des nomenclatures d'activités économiques.
13. Les peurs enfantines.

B — Application de l'analyse des correspondances (tome II, Benzecri)

1. Recherche d'un scalogramme par l'analyse factorielle (vases chinois).
2. Une analyse statistique de vocabulaire : les professions de foi des députés élus en 1881.
3. Les marques de cigarettes.
4. Niveaux et conditions de vie au Liban en 1959-1960.
5. *Idem.* Évolution entre 1960 et 1970.
6. Études sur les conditions de développement de la Colombie.
7. La population de l'Ansariin (Tunisie).
8. Les budgets familiaux dans les régions de la C. E. E.
9. Attitude des étudiants du Zaïre vis-à vis des actes attentatoires aux personnes, aux biens et aux mœurs.
10. Sondages sur l'image de la justice au sein de la population française.
11. Extrait d'une analyse des données hydrologiques.

III. Cet ouvrage est prolongé par les Cahiers de l'Analyse des données (Revue trimestrielle, Dunod) revue débutant au premier trimestre 1976 (462 pages en 1976), publiée par le Laboratoire de statistique de l'Université Pierre et Marie Curie et par l'Association pour le développement et la diffusion de l'Analyse des données. Nous recommandons d'y lire J.-B. Benzecri : « Histoire et préhistoire de l'analyse des données », passionnante et passionnée. Voici la liste des études *pratiques* parues dans les neuf premiers numéros.

Quelques exemples tirés des Cahiers de l'analyse des données

• Volume 1, 1976, N° 1.

— P. 61 : Analyse des données sur l'art préhistorique.
— P. 71 : L'état du commerce extérieur en France en 1831.

N° 2

— P. 137 : Tables à l'usage des investisseurs à l'étranger.
— P. 145 : Deux analyses de données granulométriques en géomorphologie, partie 1, p. 161; partie 2 (n° 3), p. 259 : Les scrutins de 1967 à l'Assemblée des Nations Unies.
— P. 197 : Analyse des liens au sein d'un groupe d'enfants.

N° 3

— Partie 1, p. 243; partie 2 (n° 4), p. 367 : Sur l'analyse d'un tableau de notes dédoublé.
— P. 287 : Problème d'implantation de structures dans un atelier.
— P. 319 : L'analyse de correspondance appliquée au marketing : choix d'un nom d'engrais.

N° 4

— P. 381 : Les critiques de cinéma d'après la cote des films publiée par l'hebdomadaire Pariscope.
— P. 401 : L'évolution de la production industrielle française de 1963 à 1975.
— P. 419 : Les leucémies lymphoïdes chroniques, la diversité des cas et leur évolution.

• *Volume 2, 1977, N° 1*

— P. 79 : Sur l'optimisation d'un système d'observation. Application à un réseau de contrôle de la pollution atmosphérique. I. Pollutions et météorologie. Corrélations entre stations.
— P. 97 : Sur la taxinomie du genre Erodium.

N° 2

— P. 173 : II. Pollution et météorologie.
— P. 193 : Archéologie préhistorique.
— P. 215 : Typologie de l'outillage préhistorique en pierre taillée.

N° 3

— P. 273 : Les lycéens du second cycle, comparaison entre filles et garçons.
— P. 293 : Analyse des importations brésiliennes de machines et outils mécaniques.
— P. 303 : Implantation des services d'une société.

N° 4

— P. 412 : L'appréciation des demandes de crédits.
— P. 415 : Aspects pronostiques et thérapeutiques de l'infarctus.
— P. 435 : Taxinomie des micro-mammifères.

• *Volume 3, 1978, N° 1.*

— P. 35 : La distance aux équipements urbains en Languedoc-Roussillon.
— P. 47 : Attitudes des paysans iraniens dans la réforme agraire.
— P. 65 : Le paysage forestier.

IV. On n'oubliera pas, enfin, quatre ouvrages d'enseignement, bien construits et susceptibles de remplacer les cours correspondants absents en général des grandes écoles d'ingénieurs :

— P. Bertier, J. M. Bouroche. Analyse des données multidimensionnelles, P. U. F., 1975.
— L. Lebart, J.-P. Fenelon. Statistique et informatique appliquées, avec son complément d'exercices commentés, Dunod, 3e édition, 1975.
— F. Caillez et J. P. Pages. Introduction à l'analyse des données, Smash. 1976.
— L. Lebart, A. Morineau, N. Tabard. Techniques de la description statistique. Dunod 1977.

V. L'article « Analyse des données » par E. Diday et L. Lebart, dans la revue Recherche, n° 74, janvier 1977, pages 15 à 25, est un exposé élémentaire très bien fait.

Voir aussi le volume de 900 pages : 81 communications du symposium de l'I. R. I. A., Versailles, 7-9 septembre 1977.

VI. Articles cités

— R. M. Cormack. A review of classification, *Journal of the Royal Statistical Society*, Série A. Volume 134, Dark 3, 1971.
— R. Gibrat. Discussion de la communication E. Malinvaud, *Journal de la Société de Statistique de Paris*, Nouvelle série n° 1; 1er trimestre 1977, p. 13.
— M.D. Hill. Correspondance analysis *Appli. Statistics*, 1974, t. 23, p. 340 à 354.
— R. I. Larsen. Un modèle mathématique pour relier les mesures de la qualité de l'air aux normes de celle-ci (traduction LECES).
— L. Lebart. Validité des résultats en analyse de données (recherche financée par la DGRST), novembre 1975, 158 pages.
— A. Lichnerowicz. L'ordinateur dans la société. *Sciences et Techniques*, n° 41, avril 1977, p. 9 à 14.
— E. Malinvaud. Méthodes statistiques de l'économétrie, 1964, un volume, 634 pages, Dunod.
— E. Malinvaud. Les grands échantillons des données individuelles et leur exploitation statistique. *Journal de la Société de Statistique de Paris*, nouvelle série, n° 1, 1er trimestre 1977, p. 2 à 15.
— A. Vesserau. La statistique — un livre. Que Sais-je? n° 281, 2e édition 1955.