LA VALIDITÉ DES RÉSULTATS EN ANALYSE DES DONNÉES

par

Ludovic LEBART

SOMMAIRE

1. INTRODUCTION

Pour pouvoir apprécier la validité de représentations issues des méthodes d'analyse des données, il est nécessaire de comprendre quelle est la nature de ces résultats : aucune analogie, aucun exemple emprunté à d'autres techniques d'investigation ne permet jusqu'à présent de caractériser de façon pleinement satisfaisante ces méthodes et les représentations auxquelles elles conduisent.

On peut, en première analyse, parler *d'instrument d'observation de la réalité multidimensionnelle* ; définition peu précise qui suscite aussitôt toute une série d'analogies avec les instruments d'observation de « l'infiniment éloigné » que constituent les lunettes astronomiques et les télescopes, de « l'infiniment petit » que permettent d'explorer les divers types de microscopes, de « l'opaque » partiellement révélé par les appareils radiographiques et les scintillographes.

L'apparition et la diffusion de techniques nouvelles se faisant toujours à l'intérieur de groupes particuliers, dans des cadres institutionnels spécifiques, on peut penser que les obstacles qui surgissent, les incompréhensions, les illusions éventuelles vis-à-vis des techniques de statistiques multidimensionnelles sont similaires aux réactions qui accompagnèrent, en d'autres temps, l'apparition de nouveaux instruments d'observation (plus précisément des instruments permettant d'observer des aspects de la réalité jusque-là inexplorés).

L'exemple de l'apparition, puis de la diffusion du microscope dans la communauté scientifique de l'époque est, selon nous, assez riche d'enseignement. Comme le note F. Jacob [17] (1), « *Même quand un instrument vient soudain accroître le pouvoir de résolution des sens, il ne présente jamais que l'application pratique d'une conception abstraite. Le microscope, c'est la réutilisation des théories physiques sur la lumière. Et il ne suffit pas de* voir *un corps jusque-là invisible pour le transformer en objet d'analyse. Quand Leeuwenhoeck contemple pour la première fois une goutte d'eau au microscope, il y trouve un monde inconnu, des formes qui grouillent, des êtres qui vivent ; toute une faune imprévisible que l'instrument, soudain, rend accessible à l'observation. Mais la pensée n'a alors que faire de tout ce monde...* ». Le champ d'observation de l'analyse des données n'est évidemment aucunement comparable à celui des microscopes.

Il n'y aurait cependant que quelques mots à changer pour évoquer l'apparition de certaines techniques d'analyse factorielle et les réactions qu'elles suscitent.

(1) Les chiffres entre crochets renvoient à la bibliographie *in fine*.

On trouve déjà en effet la même perplexité, les mêmes engouements hâtifs et éphémères, les mêmes erreurs et les mêmes ostracismes.

Parées du prestige de l'ordinateur, les cartes bidimensionnelles investissent pratiquement toutes les disciplines ; même lorsqu'elles ont été établies à bon escient et avec discernement, il est rare qu'elles soient accompagnées d'appréciations concernant leur validité et leur portée réelle.

C'est pour l'essentiel à cette phase de validation que sont consacrés les travaux et les réflexions critiques repris dans cet article.

On s'intéressera surtout aux méthodes d'analyse factorielle descriptives (analyse en composantes principales et analyse des correspondances), qui sont les plus utilisées, et selon nous les plus utiles. On étudiera tout d'abord les problèmes techniques de validation, puis les problèmes plus méthodologiques posés par l'utilisation de ces méthodes dans les sciences humaines.

1.1. Les méthodes

Pour l'essentiel, les méthodes d'analyse des données recouvrent les méthodes d'analyse factorielle au sens large et les techniques de classification. Par une ironie terminologique assez largement imposée par l'usage, les méthodes d'analyse factorielle au sens strict (analyse factorielle des psychologues, ou encore, analyse factorielle en facteurs communs et spécifiques) ne font pas partie des techniques d'analyse des données proprement dites, puisqu'elles ne se limitent pas au rôle d'instrument d'observation et stipulent un modèle statistique particulier. (Les factorialistes eux-mêmes étaient cependant assez divisés sur ce sujet ; Cyril Burt et ses disciples ne rejetaient pas systématiquement l'utilisation de la méthode à des fins purement exploratoires).

Les deux principales méthodes d'analyse factorielle de données amorphes ([1]) sont alors l'analyse en composantes principales et l'analyse des correspondances, cette dernière s'appliquant avec succès à une classe de codage sensiblement plus large. Les méthodes de classification automatique sont d'un usage moins courant, bien qu'indispensables pour des opérations spécifiques comme les constructions de nomenclatures par agrégation. Elles peuvent cependant compléter et enrichir les résultats des analyses factorielles. La multiplicité des techniques existantes et l'effervescence qui règne autour de ce domaine nouveau de recherche (plus de 1 000 publications par an selon R. M. Cormack [9]) nous incitent à une certaine prudence méthodologique.

L'utilité des algorithmes d'agrégation autour de centres mobiles [1], [2 : T. 1], [11], est indiscutable lors du traitement de fichiers volumineux ; certaines méthodes de classification ascendante hiérarchique ont également

(1) Nous désignons par recueil de données amorphes un recueil sur lequel on n'a pas d'information *a priori* du type « division des variables ou individus en deux ou plusieurs groupes », « relations d'ordre sur les individus », etc.

fait leurs preuves [2 : T. 1], mais l'utilisateur est perplexe devant la multitude des variantes proposées, au point qu'on peut dire [9] « qu'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... ». A peine nées, les méthodes de classification sont menacées par le développement incontrôlé de métastases théoriques, source de diversion et de confusion. Ce n'est pas un phénomène nouveau en mathématique appliquée ni même en statistique. Il sera de toute façon limité si l'on s'astreint à juger des *applications* d'après les normes et les critères de la discipline d'où sont issues les données.

Cependant, si l'outil qu'est la classification est fondamentalement différent de l'analyse factorielle, le champ d'observation et la méthode de travail sont relativement proches. C'est pourquoi nous ne pensons pas être trop restrictifs en nous limitant, dans les propos qui vont suivre, à des considérations sur les deux méthodes d'analyse factorielle précitées.

1.2. Les matériaux observables

Les domaines auxquels les méthodes d'analyse des données peuvent s'appliquer sont évidemment très variés, mais il est bien connu que le recueil de données soumis à l'analyse doit posséder certaines qualités. Les deux premières qualités sont l'homogénéité et l'exhaustivité [2 : T. A2 n° 2]. L'homogénéité est habituellement comprise comme une homogénéité de texture du tableau analysé (le codage doit permettre une certaine comparabilité entre lignes ou entre colonnes; on ne doit pas mélanger des quantités exprimées en grammes et en mètres : d'où une conversion en codage par classe ou par rang afin d'atteindre cette homogénéité). Cependant, il est important, pour la clarté de l'interprétation, que les matériaux analysés simultanément aient également une certaine homogénéité de substance, ou, si l'on préfère de contenu, respectant ainsi le principe de pertinence recommandé par les linguistes [30] qui consiste à ne retenir dans la masse hétérogène des faits que ceux qui se rapportent à un seul point de vue. Cette condition supplémentaire permet souvent de clarifier et de rendre plus aisées les interprétations : en pratique, cela reviendra à distinguer plusieurs familles de variables, certaines d'entre elles jouant un rôle actif dans la construction des typologies, les autres n'intervenant que comme variables illustratives.

Le critère d'exhaustivité est plus difficile à expliciter; il s'agit plus, en fait, d'un conseil que d'une règle stricte : il implique que toutes les situations ou tous les aspects d'un phénomène soient représentés, sans exiger cependant une représentativité au sens de la théorie des sondages; car l'exhaustivité s'applique aussi bien aux variables qu'aux individus, et l'on serait bien en peine de définir « la population parente » d'où est extrait l'« échantillon » de variables qui nous intéresse.

A ces deux exigences, nous ajouterons une première condition assez évidente, mais parfois oubliée : le tableau de données doit être vaste. En statistique, l'infini est parfois de l'ordre de 30... Il est cependant extrêmement

embarrassant de dire à partir de quelle taille de tableau l'analyse factorielle mérite d'être utilisée car ce choix dépend de la nature même du tableau : une table de contingence 10 × 10 peut être intéressante à analyser, alors qu'un tableau binaire (notes de présence-absence) de même taille ne le sera probablement pas. L'apport des méthodes d'analyse factorielle sera en général d'autant plus important que le tableau analysé est vaste; la visualisation s'avèrera en effet beaucoup plus nécessaire, les résultats seront plus stables, l'interprétation plus riche.

Une seconde condition d'obtention de résultats aisément interprétables est d'exiger du recueil de données qu'il soit amorphe. Il convient en effet d'utiliser tout ce que l'on sait (si cela est possible), pour arriver facilement à en savoir plus. Des exemples de tableaux homogènes (en texture et en substance) et non-amorphes sont fournis par les recueils de données traduisant des évolutions chronologiques [e. g. évolution du prix de l'ensemble des produits alimentaires sur 60 trimestres]. Il existe dans ce cas un facteur prédominant et objectif : le temps; une possibilité de description aussi efficace qu'élémentaire : les diagrammes chronologiques; des techniques numériques élémentaires d'analyse (lissage, désaisonnalisation), enfin des techniques plus élaborées (analyse des spectres et cospectres, etc.).

On voit que, dans ce cas, même en se limitant aux techniques descriptives élémentaires, on peut espérer atteindre des résultats assez fins, ce qui diminue l'urgence du recours à l'analyse des données brutes.

2. LES ÉPREUVES DE VALIDITÉ

Aspects techniques

L'appréciation des résultats d'une analyse factorielle descriptive dépend en partie des valeurs de paramètres d'aide à l'interprétation (contributions absolues et relatives en analyse des correspondances), dont font partie les taux d'inertie (ou pourcentages de variance expliquée), et les valeurs propres elles-mêmes. Elles dépendent également d'épreuves de validité, de calculs de sensibilité ou de stabilité qui ne pourront la plupart du temps être réalisés que par simulation.

On rappellera brièvement tout d'abord que, dans beaucoup d'applications, les taux d'inertie relatifs à un sous-espace ne représentent en aucun cas une part d'information extraite par ce sous-espace, et ceci, contrairement à une idée assez répandue.

Ceci nous amènera à tenter une réflexion sur la nature de l'information qu'utilisent et que produisent les techniques d'analyse des données. Enfin, on étudiera les épreuves de validité proprement dites, qu'elles relèvent de schémas inférentiels classiques ou simplement de calculs de stabilité.

2.1. Les taux d'inertie

Quelques contre-exemples suffisent à mettre en évidence l'inadéquation de ces coefficients pour juger de la qualité d'une représentation, sauf peut-être dans le cas d'analyses de correspondances réalisées sur de vraies tables de contingences.

a) *Le codage disjonctif*

Ce type de codage introduit une « sphéricité artificielle » qui a pour effet de rendre extrêmement faibles les taux d'inertie extraits : si le tableau concerne Q questions totalisant J modalités de réponses, aucun facteur ne peut extraire plus de $100 \times Q/(J - Q)$ % de l'inertie totale.

On sait que, pour deux questions, l'analyse des correspondances d'un tableau disjonctif et celle du tableau de contingence croisant les deux questions donnent les mêmes représentations. Prenons l'exemple, publié ailleurs, du tableau de contingence croisant 373 communes de la région parisienne et 29 catégories d'activités. Le premier facteur explique 50 % de l'inertie totale.

La formule ci-dessus nous montre que l'analyse du tableau disjonctif correspondant ne peut donner un premier facteur expliquant plus de 0,5 % de l'inertie.

Pour une carte sociale identique, à partir d'informations de base identiques, mais présentées de façon différente, le taux d'inertie est plus de 200 fois plus faible.

b) *Matrices associées à certains graphes*

On sait que l'analyse des correspondances de la matrice associée à un graphe planaire reproduit de façon satisfaisante (sauf exceptionnellement pour certains points périphériques) les relations d'adjacence. Ainsi, dans le cas d'un cycle, un calcul analytique direct nous donne l'équation paramétrique d'un cercle dans le plan des deux premiers facteurs. Et pourtant, le taux d'inertie correspondant à ce plan est une fonction décroissante du nombre de sommets du graphe et peut donc être rendu aussi petit que l'on veut. On voit qu'il s'en faut de beaucoup que ceux-ci représentent une part d'information.

c) *Influence du choix des variables*

Il s'agit ici du problème classique de la complétion d'un tableau par des lignes (ou des colonnes) de « bruit blanc ». Bien entendu, la part d'inertie expliquée par les facteurs décroît, alors que ceux-ci sont inchangés.

Il en est ainsi lorsque le nombre potentiel des variables est très grand (par exemple : présence d'espèces animales ou végétales rares dans des relevés écologiques). En principe, une certaine discipline dans le choix du recueil des données (critères d'homogénéité, d'exhaustivité) doit permettre d'éviter ces inconvénients. Mais d'une part le statisticien n'a pas toujours la maîtrise de la collecte des données ni une connaissance suffisante du domaine

d'application, et d'autre part, ces critères sont eux-mêmes trop qualitatifs et trop généraux pour définir de façon rigoureuse un tableau optimal, parmi tous les tableaux potentiels. Ici encore, les taux d'inertie seront donc beaucoup plus sensibles que les facteurs aux différentes modalités pratiques qui précèdent ou accompagnent une analyse. Certains praticiens, afin d'obtenir des taux d'inertie « honorables », n'hésitent pas à recommencer une analyse en supprimant les variables ou observations contribuant peu aux premiers facteurs ; cette manipulation opportuniste n'est pas sans fondement : il est exact que les taux d'inertie finaux représentent mieux la « part d'information » expliquée par les facteurs ; cette démarche relève, selon nous, d'une déformation qui est peut-être héritée de la théorie de la régression multiple : on juge la qualité globale d'une régression par le coefficient de corrélation multiple, dont le carré représente la part de variance expliquée par la liaison exprimée par le modèle.

Or un coefficient de corrélation multiple *ne peut qu'augmenter* si l'on ajoute des variables explicatives (exogènes), fussent-elles des nombres au hasard. Précisément dans le cas de complétion du tableau par des lignes de « bruit », le taux d'inertie des *premiers facteurs* d'une analyse en composantes principales, par exemple, ne peut en général que diminuer. Ces propriétés qui découlent de considérations géométriques élémentaires font que le coefficient de corrélation multiple 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. Un coefficient de corrélation multiple voisin de 1 ne signifie pas forcément que la régression est de bonne qualité (une seule des variables exogènes peut être colinéaire à la variable endogène), alors qu'un taux d'inertie voisin de 100 % pour un sous-espace signifie nécessairement que ce sous-espace permet une excellente représentation de l'ensemble des variables et individus analysés. A l'inverse, comme le montrent en particulier les quatre situations caractéristiques que nous venons d'examiner, des taux d'inertie faibles peuvent laisser espérer des représentations de bonne qualité. Il est cependant fâcheux que dans les deux cas, l'on parle de « part de variance expliquée » pour désigner des choses si différentes.

2.2. Information et forme

a) *Ambiguïté du concept d'information*

La plupart des mathématiciens, physiciens ou statisticiens qui ont élaboré ou simplement étudié la théorie de l'information (fondée principalement par Cl. Shannon) ont été conscients des limites de cette théorie qui ne s'intéresse qu'à la conservation, la transmission, la transformation de l'information, sans aborder le problème (évidemment complexe) de la valeur pratique de l'information. L. Brillouin [6], lors d'un survol des problèmes « dépassant la théorie actuelle » rappelle à propos des machines à calculer, que « le mécanisme de calcul n'ajoute aucune information nouvelle, il ne fait que la reproduire dans un langage différent et probablement avec des pertes.

Mais le calcul, comme le décodage ou la traduction, ajoute certainement à la valeur pratique de l'information et la rend pleine de signification pour l'utilisateur... tout critère relatif à la valeur aura pour corollaire une évaluation de l'information reçue. Ceci revient à sélectionner l'information suivant une certaine loi de mérite. Certaines fractions de l'information pourront être considérées comme sans valeur et seront écartées. Ce problème présente une grande ressemblance avec celui du filtrage. Un filtre est caractérisé par deux importantes propriétés : il est irréversible et il diminue la quantité totale d'information. Par analogie, nous pouvons énoncer le résultat suivant : la valeur relative de l'information pour un utilisateur donné est au plus égale à l'information absolue. Nous avons là un point de départ dont la validité est générale et le problème se pose de définir les différentes méthodes de filtrage que l'on peut utiliser pour évaluer la valeur relative ».

L'analogie avec le filtre est très parlante pour nous; il est assez tentant de considérer un programme d'analyse des données comme un filtre très particulier qui permettrait d'isoler un certain signal (par exemple un réseau d'interrelations que l'on peut difficilement imputer au hasard) d'un bruit de fond. Cependant, l'analogie ne peut aller très loin, car le signal, dans notre cas, n'a que peu de chose à voir avec un message, et nous ne sommes pas dans la situation d'un processus de communication. Nous devons, pour espérer aller plus loin dans l'étude de l'information en analyse des données, tenir compte de la dérive du concept en étudiant notamment, d'après R. Thom [37], les glissements sémantiques du mot « Information ». Nous ne ferons en fait que résumer brièvement l'argumentation de cet auteur qui explique l'instabilité sémantique de ce mot par la complexité des relations que recouvre le concept : « il y a sous-jacent à l'idée d'information — l'existence d'un demandeur à qui apprendre cette information est profitable; et d'un donneur qui donne, en général de son plein gré, l'information au demandeur. Dans tous les emplois du mot où l'on est dans l'incapacité d'identifier ces quatre éléments : demandeur-demande-donneur-avantage apporté au demandeur par la connaissance de l'information, on doit suspecter dans l'emploi du mot une certaine malhonnêteté... ». On peut alors distinguer plusieurs sens selon les mutilations que subit le schéma; l'information aux sens judiciaire et journalistique, où il y a effacement du demandeur, et même parfois de la demande; information au sens des techniques de publicité, où « l'abus est cette fois flagrant : l'information est imposée au destinataire qui ne l'a pas demandée, et son contenu est notoirement plus avantageux au donneur qu'au destinataire »; l'information au sens de la théorie de Shannon où pratiquement tout est effacé à l'exception du canal de transmission entre le demandeur et le donneur. En biologie, des expressions telles que « information génétique » permettent de masquer une partie de l'ignorance de certains mécanismes « ... tout en apportant par la connotation d'intentionnalité du mot information une caution implicite au finalisme qui sous-tend toute pensée biologique. En ce sens, l'information, c'est la forme obscure de la causalité ».

Nous nous rapprochons de nos préoccupations avec le sens du mot information dans la phrase suivante : « l'information fournie par les rayons X après traversée du corps du patient », pour laquelle, selon R. Thom, « ... il n'y a ambiguïté ni sur le donneur (le corps du patient) ni sur le demandeur-récepteur (l'observateur) ni sur la finalité globale du processus (interprétation diagnostique). La seule ambiguïté concerne le contenu signifié du mot information que... je réduirais à la forme du message reçu sur la plaque sensible, et interprété par l'observateur ». Ceci fait dire à l'auteur que l'on gagnerait beaucoup à remplacer, dans certains cas, le mot information par le mot *forme*. L'analogie entre un appareil radiographique et les instruments d'observation que constituent les techniques d'analyse factorielle est grande. Nous l'avons développée ailleurs en comparant notamment les deux obstacles à l'observation directe constitués par l'opacité des tissus dans un cas, le caractère multidimensionnel des données dans l'autre. Ce caractère multidimensionnel ne fait qu'ajouter à l'inadaptation de la notion classique d'information comme le souligne plus loin R. Thom : « En réalité, le rêve de la théorie de l'information a été d'établir un algorithme permettant d'évaluer la complexité d'un système, en même temps que son degré d'organisation, l'écart qu'il présente par rapport à une structure *totalement désordonnée*. Un tel algorithme n'existe que pour les morphologies de dimension 1 : les suites finies de lettres extraites d'un alphabet fini. Pour les morphologies naturelles — pluridimensionnelles — un tel algorithme n'existe pas et les concepts mêmes de complexité, d'ordre, d'organisation, de désordre n'y sont pas définis ».

Il est maintenant bien connu des statisticiens que les généralisations multidimensionnelles ne se font pas à n'importe quel prix : le problème de la séparation de populations qui se pose souvent à propos de reconnaissance des formes, est relativement simple dans le cas unidimensionnel (un point sépare deux demi-droites). Dans le cas d'échantillons multidimensionnels, on est obligé de spécifier la forme des cloisons (hyperplans, hypersurfaces de degré donné, etc.) et par conséquent d'introduire des hypothèses supplémentaires qu'un traitement combinatoire ou en apparence non-paramétrique ne doit pas camoufler. Cependant, s'il n'existe pas d'algorithme universel d'évaluation de la complexité ou de l'organisation d'un recueil de données multidimensionnelles, il existe de nombreux algorithmes particuliers — comme ceux que nous utilisons en analyse des données — qui permettent de mettre en évidence certains aspects de cette organisation : un nuage est-il approximativement hypersphérique ? Est-il concentré sur une sous-variété particulière ? Est-il formé de deux sous-nuages que l'on peut considérer comme distincts ? Existe-t-il des axes principaux d'inertie correspondant à des moments prédominants ? Ces questions très partielles ne peuvent être posées qu'après avoir donné de l'objet observé une première représentation sous forme de nuage dans un espace euclidien de grande dimension, étape préliminaire qui suppose une approximation, ou peut-être dans certains cas, toute une théorie.

Le statisticien qui dépouille un fichier d'enquête socio-économique, même aidé d'efficaces instruments de visualisation, est en fait souvent submergé par l'information pourtant partielle qu'il en extrait (ou par les formes qu'il reconnaît ou qu'il découvre, si l'on préfère cette terminologie). Il a surtout besoin, c'est du moins notre propos et notre ambition de l'aider dans cette tâche, de mieux comprendre et d'évaluer le peu qu'il voit.

Nous retiendrons surtout de ces considérations sur la notion d'information qu'il sera parfois utile, en analyse des données, de parler de forme plutôt que d'information (ce qui nous conduira à vérifier la *stabilité d'une forme* et non pas à mesurer une *quantité d'information*, pour valider une représentation); ces techniques fonctionnent un peu comme des filtres (très généraux) : elles tentent d'accroître la valeur pratique de l'information, au prix d'une perte d'information brute qui peut être considérable. Dans cette optique, il apparaît peu pertinent de considérer la quantité d'information brute initiale comme une mesure de référence.

b) *Recherche d'une description ou d'une relation?*

Nous voudrions insister ici sur certaines particularités des analyses factorielles qui les distinguent fondamentalement des techniques de régression et des méthodes apparentées (analyses canoniques, analyses discriminantes) (1).

On sait que si l'on cherche la meilleure relation linéaire liant p variables, pour lesquelles nous disposons de n observations, la solution (l'hyperplan de régression orthogonale) est fournie par le vecteur de la matrice des covariances expérimentales de ces p variables correspondant à la plus petite valeur propre. La régression linéaire classique est un cas limite de régression orthogonale, dans une métrique-limite qui accorderait un poids infiniment grand à une des variables (*cf.* par exemple [29]). Inversement, on peut dire que si l'on cherche la relation linéaire la plus « mauvaise »... c'est-à-dire la plus dispersée sur l'ensemble des observations, on obtient la première composante principale, qui est le vecteur propre de la matrice des covariances correspondant à la plus grande valeur propre.

Les coefficients de cette première composante sont les cosinus directeurs de l'axe principal d'inertie du nuage des n observations dans $\mathbb{R}^p$. Ils nous fournissent également la meilleure description à une dimension des proximités entre variables (préciser plus le sens de « meilleure » et de « proximité » exigerait un exposé théorique pour lequel nous renvoyons aux manuels existants).

Le bas du spectre nous fournit donc des *relations* linéaires permettant éventuellement de *prévoir*, pour un individu, la valeur d'une variable connaissant les valeurs de toutes les autres. Nous allons voir que ces

(1) Le fait que l'analyse des correspondances des tableaux de contingence puisse être considérée comme un cas particulier d'analyse discriminante (et donc d'analyse canonique) appliquée à des codages spéciaux n'est pas en contradiction avec la différence fondamentale dont nous allons parler, précisément parce que ces codages spéciaux (disjonctifs) induisent des propriétés très particulières.

relations font intervenir des coefficients qui sont *instables* et *vulnérables*, et d'autant plus que les variables sont nombreuses.

Au contraire, le haut du spectre nous fournit des combinaisons linéaires qui nous permettent de discriminer les individus et qui décrivent les associations entre variables. Les coefficients en sont généralement stables et le sont d'autant plus que les variables sont nombreuses.

c) *Stabilité le long du spectre*

Supposons que l'on ait réalisé l'analyse factorielle d'un tableau X de valeurs numériques quelconques, donnant lieu à une séquence de valeurs propres distinctes.

Les premiers facteurs de l'analyse seront beaucoup plus stables, vis-à-vis de petites perturbations apportées au tableau X, que des facteurs correspondant aux plus petites valeurs propres. Ce résultat bien connu de la théorie de la perturbation [19] [38] n'était pas ignoré non plus des statisticiens.

Nous pouvons citer le théorème établi par J. C. Deville, dont la démonstration est facilement accessible [10], et qui considère de façon plus générale un opérateur compact hermitien positif C d'un espace de E. Hilbert (de tels opérateurs sont classiquement munis de la norme

$$\|C\| = \sup_{\|x\|=1} \|Cx\|,$$

les k premières valeurs propres de C sont supposées simples. A la valeur propre λ_i correspond le vecteur propre unitaire x_i.

On note

$$\Delta = \inf_{i=1,\ldots k} (\lambda_i - \lambda_{i+1}),$$

Le résultat est le suivant : si C' désigne un opérateur compact hermitien positif de E, tel que $\|C - C'\| \leqslant \varepsilon < \Delta/2$, on a alors $|\lambda_i - \lambda'_i| < \varepsilon/2$ pour $i \leqslant k$; de plus, à chaque λ'_i (valeur propre de C') on peut associer un vecteur propre unitaire x'_i de C' tel que

$$\|x_i - x'_i\| \leqslant 2\sqrt{2}\,\varepsilon/\Delta$$

Ainsi, la perturbation subie par le vecteur propre sera évidemment d'autant plus importante que la matrice d'inertie sera modifiée (coefficient ε), mais sera surtout sensible à l'écart Δ entre deux valeurs propres consécutives qui devra être le plus grand possible. L'écart Δ ne sera grand en général que dans le haut du spectre, dont le résultat ci-dessus montre qu'il correspondra à des sous-espaces stables (*cf.* également [13]).

Les contre-exemples ne manquent pas pour mettre en évidence l'instabilité des vecteurs propres correspondant à des valeurs propres dont l'écart absolu est faible, et donc l'instabilité des coefficients intervenant dans une relation (valeurs propres très faibles) approximativement vérifiée par les variables.

En fait, la recherche de relations n'est justifiée que si un problème de prévision se pose à propos d'un petit nombre de variables. Qu'il s'agisse de régression linéaire ou de régression orthogonale, les variables sont *mises en concurrence*, ce qui hypothèque lourdement l'interprétation des résultats, sensibles aux colinéarités et aux substitutions.

Puisque l'analogie entre « analyse factorielle » et « filtres » a été évoquée au paragraphe précédent, on peut déjà préciser comment ce filtre agit : il isole la partie la plus stable du réseau d'interrelations existant entre les lignes et les colonnes d'un tableau de valeurs numériques. Bien entendu, la définition des interrelations dépend de la nature du tableau et de la technique particulière d'analyse factorielle employée (similitudes entre profils pour l'analyse des correspondances, liaisons linéaires entre variables pour l'analyse en composantes principales); la définition de la stabilité dépend de la classe de perturbations que l'utilisateur entend associer à son recueil de données.

2.3. Principes et mise en œuvre des épreuves de validité

Les calculs par simulations peuvent être utilisés dans les épreuves de validation en analyse des données, de deux façons radicalement différentes : pour valider une hypothèse (en général hypothèse d'indépendance), ou pour vérifier la stabilité d'un résultat [2, chap. B.9].

a) *Validation de l'hypothèse d'indépendance*

Dans ce cas, les simulations ne sont qu'un substitut de techniques statistiques classiques; celles-ci ne peuvent pas toujours s'appliquer, puisque les hypothèses donnant lieu à des formulations analytiques et à des tabulations ne forment qu'un éventail restreint. On simule alors la loi du (ou des) paramètre(s) en générant plusieurs pseudo-réalisations de ce paramètre, que l'on compare à la valeur réellement observée. Un programme-test permet alors de remédier à l'absence de table et d'estimer, de façon grossière en général, des seuils de signification.

S'il ne s'agit pas d'analyse de correspondances sur vraies tables de contingence, pour laquelle des tables approchées existent (*cf.* ci-dessous), on procèdera à une vingtaine de simulations pour atteindre l'équivalent d'un seuil de 5 % (qui n'est évidemment pas comparable à un seuil exact comme celui que fourniraient des tables, du point de vue des risques encourus). La détermination simultanée de plusieurs seuils pose quelques problèmes : la loi jointe des valeurs propres et des pourcentages d'inertie n'est évidemment pas connue lorsque l'hypothèse d'indépendance des lignes et des colonnes du tableau n'est pas vérifiée.

Ces procédures en général coûteuses ne permettent donc pas de formuler des conclusions de façon rigoureuse. Elles sont cependant le seul recours lorsqu'il s'agit de juger la signification d'un paramètre issu de calculs complexes.

— *Le cas de l'analyse des correspondances*

L'hypothèse d'indépendance totale des lignes et des colonnes d'un tableau est beaucoup trop générale et sévère pour être réaliste. Dans les recueils de données issues des sciences humaines, elle est pratiquement impossible à observer, dès que les effectifs mis en jeu deviennent importants. Il est beaucoup plus utile, nous le verrons, de tester en fait la stabilité d'une configuration, en simulant des données fictives entachées d'erreur ou de fluctuations suggérées par la nature du problème traité et la qualité des mesures.

Une situation favorable est fournie par les tableaux de contingence, pour lesquels l'inertie est riche de signification. Bien que l'hypothèse d'indépendance constitue un cas limite sans grande portée pratique, il nous a paru utile de mettre à la disposition des utilisateurs les « garde-fous » que constituent les seuils de signification des valeurs propres et des taux d'inertie, calculés précisément dans le cas où les données analysées ne reflèteraient qu'un « bruit de fond » imputable à des fluctuations d'échantillonnage.

La loi des valeurs propres issues de l'analyse des correspondances a donné lieu à maintes publications erronées. Ainsi, dans le traité de statistiques de M. G. Kendall et A. Stuart [20], les valeurs propres sont supposées suivre des lois du χ^2, comme l'inertie totale.

H. O. Lancaster [22] a réfuté ce résultat en montrant que l'espérance mathématique de la première valeur propre est toujours supérieure aux résultats préconisés par les assertions de M. G. Kendall et A. Stuart.

En fait, nous avons montré [25], [27] que la loi cherchée est étroitement liée à celle des valeurs propres des matrices distribuées selon la loi de Wishart, dont la densité de probabilité a été explicitée en 1939 par Fisher [14].

L'intégration de cette densité assez complexe a donné lieu à plusieurs publications parmi lesquelles nous citerons celle de P. R. Krishnaiah et V. B. Waikar (1971) [21], qui s'inspirent des travaux de physiciens tels que M. L. Mehta [31].

Une table des seuils de signification de la plus grande et de la plus petite valeur propre a été publiée dès 1968 [8] pour des tableaux dont la plus petite dimension n'excède pas 10, puis en 1970 [33].

Nous avons publié des tables approchées pour tous les tableaux dont les dimensions n'excèdent pas 50 × 100, relatives aux cinq premières valeurs propres et aux taux d'inertie correspondants (estimations des moyennes, écart-types, et seuil 0,05 unilatéral de ces quantités). Les figures 1 et 2 donnent des abaques résumant une partie des résultats pour les deux premières valeurs propres. On lit, par exemple, sur la figure 1 que, pour un tableau (10 × 10), la première valeur propre peut atteindre 45 % de l'inertie dans 5 % des cas, dans l'hypothèse d'indépendance des lignes et des colonnes de la table (la loi des taux ne dépend pas de l'effectif total de la table).

FIGURE 1

Seuil (0,05 unilatéral) du pourcentage d'inertie de la plus grande valeur propre

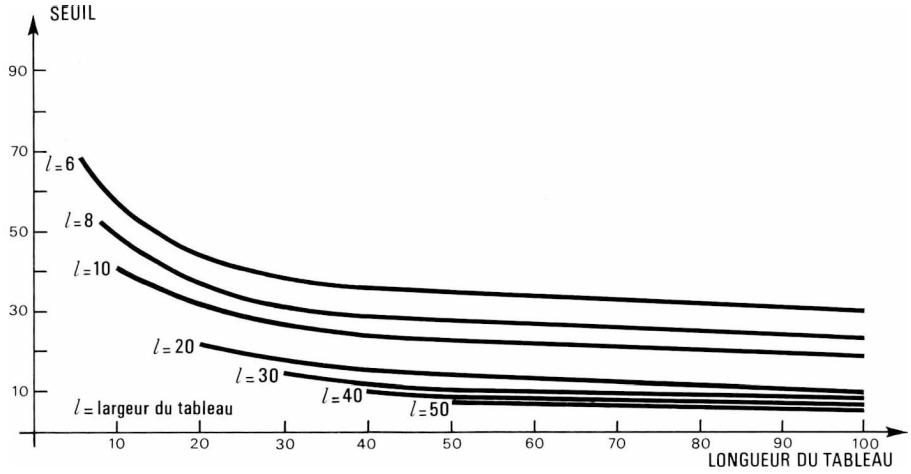
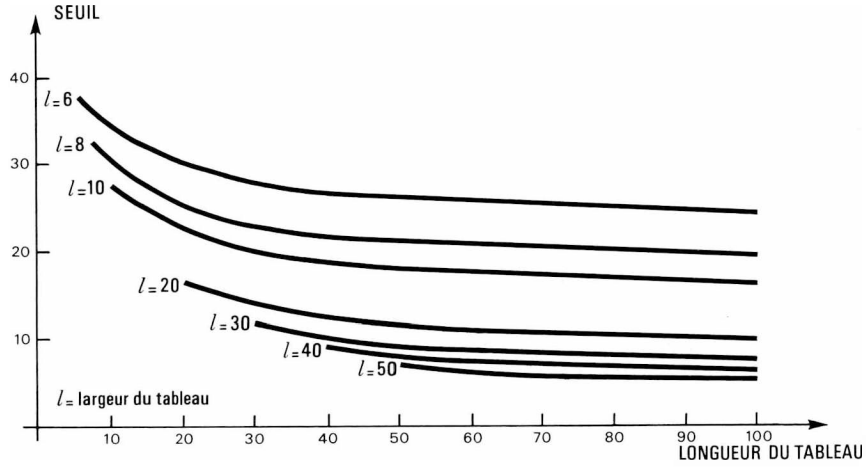


FIGURE 2

Seuil (0,05 unilatéral) du pourcentage d'inertie de la deuxième valeur propre



b) *Stabilité des formes*

Il s'agit ici d'une application beaucoup plus pertinente et beaucoup moins coûteuse des procédures de simulation. Elle consiste simplement en une vérification de la stabilité des configurations obtenues à la suite de modifications du tableau initial. Une seule simulation peut suffire dans certains cas.

Avant d'examiner les différents types de perturbations que subiront les données soumises à l'analyse, il convient de rappeler quels sont les différents facteurs qui peuvent conditionner la qualité des résultats d'une analyse factorielle.

Nous en distinguerons quatre :

- les erreurs de mesure ;
- le choix et le poids des variables ;
- le codage des variables ;
- le choix des individus ou observations, les aléas d'échantillonnage.

Chacune de ces quatre sources de perturbation initiale donnera lieu à des modifications du tableau de données, qui ne devront pas affecter les configurations supposées stables.

La nature et l'amplitude des modifications, l'appréciation de la stabilité des résultats se font lors de « colloques singuliers » réunissant l'utilisateur et le statisticien.

1° *Les erreurs de mesure :* l'ordre de grandeur de ces erreurs, ainsi que leur distribution approximative dans la population doivent être spécifiées par l'utilisateur en fonction de sa propre connaissance du domaine d'étude. Dans le cas classique des réponses ordonnées du type : « pas du tout d'accord, pas très d'accord, assez d'accord, tout à fait d'accord », on peut supposer par exemple que l'individu enquêté a une chance sur deux d'avoir exprimé exactement ce qu'il ressentait, une chance sur quatre (sauf aux extrémités) de répondre en fait une modalité immédiatement contiguë. Les programmes permettront en général de simuler une grande variété de situations intraduisibles analytiquement.

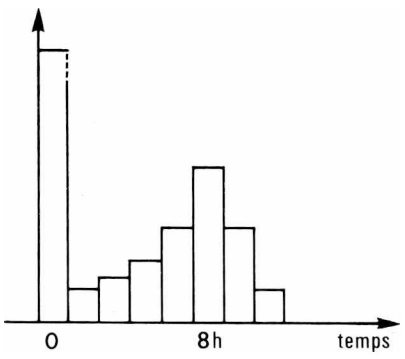
2° *Le choix et le poids des variables :* ce problème se pose lorsque le statisticien a la possibilité d'échantillonner dans l'espace des variables, ce qui n'est pas toujours le cas. Les classiques critères d'homogénéité et d'exhaustivité ne fournissent qu'un cadre général. On pourra donc effectuer des « ponctions aléatoires » dans l'ensemble des variables, afin d'éprouver la sensibilité des résultats vis-à-vis de la composition de cet ensemble. Ce procédé a été utilisé pour vérifier la stabilité de la typologie des départements français décrits par environ 180 variables [24]. Le fait de supprimer 50 d'entre elles ne modifiait pratiquement pas la typologie plane des départements répartis d'abord suivant un axe d'urbanisation, puis un axe de latitude. Le problème du poids des variables se pose surtout en analyse en composantes principales (ou en analyse des correspondances s'il s'agit de tableaux de notes ou de mesures, et non de comptages). On peut par exemple, pour éprouver une éventuelle invariance, imposer à chaque variable une variance comprise entre 1 et 2, en procédant par exemple à un tirage de série pseudo-aléatoire uniforme entre ces deux valeurs, et diagonaliser la matrice des covariances ainsi construite.

3° *Le codage des variables :* nous désignons ici par codage la transformation préliminaire et contrôlée des données brutes qui précède l'analyse multidimensionnelle. Il s'agit selon nous d'une opération qui est fondamentalement empirique (ce qui n'implique évidemment pas que l'école philosophique de l'utilisateur soit l'empirisme!) parce que liée indissolublement au contenu signifié de l'information (terminologie de R. Thom). Le codage a, comme l'analyse des données, pour raison d'être l'augmentation de la valeur pratique de l'information. Il ne s'agit pas dans la phase de codage de rendre celle-ci plus facilement assimilable à l'homme, mais plus aisément utilisable par l'algorithme, qu'il s'agisse d'analyse factorielle ou de classification.

Prenons un exemple pratique qui met en évidence ces différents aspects du codage. Cet exemple sera emprunté à l'enquête C.N.A.F.-C.R.E.D.O.C. de 1971 sur les « Besoins et aspirations des familles et des jeunes », dont un premier compte-rendu a été publié par N. Tabard [34]. Une grille socio-administrative a été construite, à partir de variables telles que « profession », « revenus », « nombre d'enfants », « âge », etc. Le mode de construction et d'utilisation de cette grille est précisé en [26]. Le problème de codage que nous allons examiner concerne une variable supplémentaire (ou illustrative). Il se poserait de la même façon pour une variable participant effectivement à l'analyse. Cette variable est « le temps de travail à l'extérieur » de la mère de famille, mesuré pour 2 003 personnes. L'histogramme de cette variable est fortement bimodal puisque, dans cet échantillon particulier, près de la moitié des femmes ne travaille pas à l'extérieur.

Le problème du codage se résume ici à un choix de partitions, le nombre de classes et les limites de classes restant à fixer. Il est clair que la variable numérique « temps de travail » est inutilisable telle quelle en analyse factorielle. Ces calculs ne faisant intervenir que les moments d'ordre 2 sont surtout intéressants pour les variables ayant des distributions marginales unimodales et relativement symétriques, à défaut de variables normales; ces calculs privilégient également les relations linéaires entre variables, ce qui est une contrainte difficilement acceptable.

Esquisse de l'histogramme

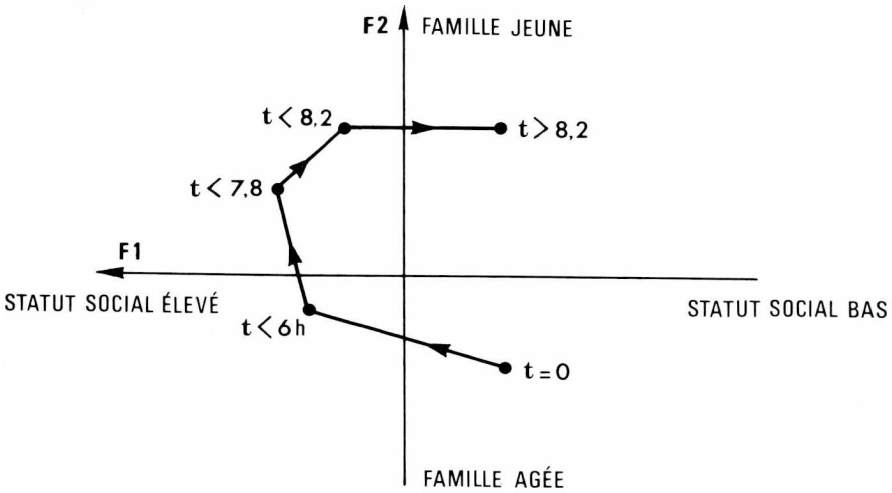


Il paraît *a priori* souhaitable d'isoler à l'intérieur d'une même classe les femmes au foyer ($t = 0$), d'isoler également les femmes ayant un temps de travail exceptionnellement long. Aucun algorithme aveugle, à notre connaissance, ne peut produire une partition qui satisfasse ces deux exigences à partir du seul histogramme. Un algorithme tel que celui de W. D. Fisher [15]

qui fournit pourtant des partitions en k classes optimales exactes ne pourra même pas, si l'on demande moins de cinq classes, isoler exactement dans une classe les femmes au foyer.

Nous avons, en pratique, construit cinq classes à partir d'un histogramme détaillé, et obtenu le résultat suivant, dans le plan des deux premiers facteurs.

Trajectoire des temps de travail dans le plan F1, F2



La liaison avec le premier facteur est non linéaire, et n'aurait pas été mise en évidence si le temps de travail avait été positionné comme une variable numérique, sans division en classes, ou même si l'on s'était contenté de deux classes. C'est donc au prix d'une perte d'information brute (division en classes) avant toute analyse, et à partir de considérations d'ordre sémantique sur les limites de classe que l'algorithme fournit les indications les plus fines sur le phénomène étudié.

Ainsi le codage n'est donc pas une simple formalité avant l'analyse : il suppose une connaissance simultanée des données (en général à partir de statistiques descriptives élémentaires) de la méthode et de ses exigences. Le codage peut surtout être considéré comme source de perturbation éventuelle des résultats dans le cas des notes, des échelles ou des classements (en analyse des rangs ou des préférences, par exemple). Il est alors utile de vérifier que les configurations obtenues résistent à des changements de variables monotones très déformants (logarithme, exponentiel, etc.) afin de s'assurer que l'ordre des notes est plus important que les propriétés métriques particulières de l'échelle utilisée.

Enfin, dans tous les cas, il est intéressant de mettre en évidence un « codage minimal » qui est la forme de codage la plus fruste susceptible de conserver les configurations observées. Citons-en deux exemples : une analyse factorielle fut réalisée sur un tableau de dépenses individuelles de consommation (d'après l'enquête C.R.E.D.O.C.-U.N.C.A.F. de 1963) et

produisit une certaine typologie des postes de consommation ; cette analyse fut refaite [18] en codant simplement « 1 » les dépenses strictement positives, quels que soient leurs montants et donna alors une typologie des postes très voisine de la précédente. On conçoit que l'interprétation de la première analyse soit modifiée par ce dernier résultat, qui souligne l'importance de l'accès à certains types de consommation, indépendamment de l'intensité de ces consommations.

Un résultat analogue a été établi à propos d'une typologie des activités réalisée à partir de budget-temps, qui n'a pas été bouleversée lorsque les durées positives des activités ont été remplacées par des « 1 », les durées nulles étant toujours codées par des « 0 ».

La simple mention d'une activité (lecture, promenade, soins aux enfants...) jouait donc un rôle déterminant.

4° *Les fluctuations d'échantillonnage :* deux types de calculs de stabilité peuvent être exécutés comme dans le cas *b*) ci-dessus : des modifications des pondérations des individus ou observations, des ponctions ou des fractionnements de l'échantillon. Les deux opérations donnent une idée de la stabilité des résultats. Bien entendu, un échantillon où certains aspects de la population parente ne sont pas du tout représentés ne pourra fournir des résultats extrapolables, même si les configurations obtenues sont stables. Toutefois, les typologies obtenues par analyse factorielle n'exigent pas une représentativité de l'échantillon aussi stricte que les estimations de pourcentages ou de moments d'ordre 1. Cette relative stabilité est surtout un fait d'expérience. On peut néanmoins en donner une explication très partielle en rappelant que les sous-espaces correspondant au haut du spectre sont les plus stables vis-à-vis des éventuelles perturbations de la matrice à diagonaliser ; et que cette matrice elle-même (par exemple matrice des corrélations expérimentales en analyse en composantes principales) est moins sensible que les moments d'ordre 1 (moyennes, pourcentages) aux fluctuations d'échantillonnages.

La grille socio-administrative de l'analyse citée en [26] est à cet égard assez démonstrative. Le plan de sondage particulier de cette enquête destinée à étudier des couches de la population très spécifiques prévoyait notamment une sur-représentation des mères de famille exerçant une activité extérieure ; à chaque famille est ainsi affecté un coefficient de redressement (les disparités de ces coefficients sont considérables). La typologie des variables socio-économiques fait preuve d'une stabilité étonnante, puisqu'elle est la même, que l'échantillon soit redressé ou que l'analyse soit faite directement sur les données brutes. De même qu'il était utile de rechercher un codage « minimal » susceptible de reconstituer les configurations, on peut s'intéresser à la taille minimale de l'échantillon qui conserve la typologie des variables. Ceci permet de hiérarchiser les facteurs à partir de leur « assise » dans l'échantillon (i.e. : l'analyse de 200 individus pris au hasard parmi les 2 000 suffit à isoler l'axe 1 (dit de statut social), alors qu'il en faut au moins 500 pour faire apparaître les axes 1 et 2, etc.).

Dans l'étude déjà citée sur la répartition géographique du système des soins médicaux, le plan des deux premiers facteurs restait pratiquement identique à lui-même, que les départements soient considérés comme des unités statistiques de même poids (aspect géographique privilégié) ou au contraire soient affectés d'un poids proportionnel à leur population (aspect démographique privilégié). Ce résultat est d'autant plus rassurant quant à la stabilité de ces facteurs que ces poids varient presque de 1 à 100.

3. PROBLÈMES MÉTHODOLOGIQUES

3.1. Quelques difficultés

On étudie brièvement ci-dessous les circonstances de l'utilisation effective de l'analyse des données et quelques-unes des critiques que peuvent susciter ces méthodes de la part des utilisateurs dans les disciplines relevant des sciences humaines et aussi de la part des statisticiens familiers de méthodes plus classiques.

On évoque tout d'abord le problème des utilisations inconsidérées, qui relève probablement plus d'une sociologie des institutions scientifiques que de notre discipline, et ne concerne donc pas spécifiquement l'analyse des données.

Le contexte informatique pose ici un problème particulier : l'ordinateur introduit une division du travail qui peut s'accompagner d'une déqualification du statisticien, d'une perte de contact avec le matériel brut.

On rappelle enfin que le mot « hypothèse » peut désigner des notions fort différentes lors des applications statistiques, ce qui est à l'origine de certaines confusions, et l'on souligne l'importance des procédures de codage.

a) Qu'une technique nouvelle donne lieu à des usages inconsidérés n'a rien de surprenant, surtout si le champ d'application de la technique est vaste, si son utilisation fait intervenir cet outil encore mythologique qu'est l'ordinateur. Les techniques n'ayant aucun intérêt pratique risquent évidemment moins d'être galvaudées.

Il arrive malheureusement que la méthodologie statistique ne soit pour un chercheur en sciences sociales, qu'un recours devant une situation désespérée, ou même un refuge où il sera à l'abri des critiques de ses pairs ; les méthodes sont aussi des exutoires ! Comme le déplorait déjà le statisticien britannique Cyril Burt en 1955, « *Beaucoup de chercheurs semblent supposer que l'on peut partir de n'importe quel recueil de données qui vous tombent sous les mains et ils attendent que l'analyse statistique en extraie les facteurs appropriés. Quand elle n'y parvient pas, ils se plaignent des méthodes* » [7].

Les méthodes ne doivent pas être jugées d'après les circonstances sociales de leur utilisation. J. P. Benzecri dans son *Histoire de l'analyse des données* [3] cite à ce propos la phrase de Stuart Mill, « *On peut peser du cuivre et le donner pour or, la balance reste sans reproche* ». Les excès et les erreurs que nous

dénonçons ici sont dus pour une large part à l'inexpérience et aux illusions des utilisateurs en sciences sociales; ceci peut nous laisser espérer qu'il s'agit d'une phase transitoire d'adaptation. Il est d'ailleurs évident que la statistique classique n'a pas été épargnée : les applications licites et pertinentes des techniques relevant du modèle linéaire général (régression et analyse de la variance) ne sont qu'une part infime des applications effectivement réalisées. L'utilisation de l'ordinateur introduit incontestablement une dispersion et des diversions peu favorables à une utilisation critique des outils statistiques, et nécessite par conséquent un surcroît de vigilance de la part de l'utilisateur.

b) *L'usage de l'ordinateur*

On a pu reprocher aux résultats d'être trop facilement acquis : leur interprétation détaillée est négligée par les statisticiens « nantis ». *La disparition de l'effet dissuasif du calcul n'est pas sans inconvénient :* il y a seulement une quinzaine d'années, avant d'aborder l'épreuve du calcul qui le détournait souvent de son champ d'intérêt propre, le statisticien étudiait longuement ses hypothèses, envisageait méticuleusement toutes les conséquences de ses calculs, éventuellement faisait des essais sur maquette (sous-échantillon).

Au cours de la phase de calcul, il était en contact quasi-physique avec les données statistiques; les anomalies ou incohérences pouvaient apparaître de façon progressive; la critique toujours possible au cours de l'exécution de certaines opérations arithmétiques ou des diverses manipulations permettait éventuellement d'infléchir le déroulement des calculs, ou de les abandonner précocement. Mais aussi le temps consacré à l'interprétation des résultats était à la mesure du travail nécessaire à leur établissement. Brièvement, disons que toute une série d'opérations qui n'était pas sans influence sur la qualité scientifique des travaux ne relève plus de l'activité du statisticien.

Le travail des taxinomistes polonais (« Wroclaw Taxonomy », comprenant notamment l'algorithme de recherche de l'arbre de longueur minimale de Florek (1951) redécouvert plus tard par les statisticiens occidentaux) est un exemple d'analyse « manuelle » des données donnant lieu à une connaissance très fine du matériel statistique. Les représentations obtenues sont beaucoup moins riches d'information que celles issues de l'analyse des correspondances, et les défauts et la vulnérabilité de ces algorithmes sont connus. Cependant, le processus d'établissement des résultats est lui-même un mode de connaissance.

Un exemple beaucoup plus actuel est fourni en France par les utilisateurs de la technique manuelle de traitement de tableau de J. Bertin [5] qui préfèrent se contenter d'une représentation très partielle (ne faisant intervenir que des modifications de l'ordre des lignes et des colonnes du tableau) mais dont le processus de formation est contrôlable visuellement et aisément explicable aux utilisateurs les plus variés.

Cependant, ces manipulations manuelles sont limitées aux petits tableaux (les fichiers d'enquêtes les plus courants donnent lieu à des tableaux de l'ordre de (200 × 1 500) qui nécessitent évidemment le recours à l'ordinateur).

Le bilan de l'automatisation est cependant loin d'être négatif; l'effet dissuasif du calcul n'a pas toujours été salutaire : il a parfois entraîné une modélisation excessive, un éloignement des faits et des observations, des théories pléthoriques n'ayant que peu de rapport avec la réalité. Cependant, l'apport le plus fondamental ne concerne évidemment pas la vitesse de résolution des problèmes anciens, mais la possibilité de traiter simultanément des masses d'informations qui seraient dénaturées ou mutilées par une parcellisation; également la possibilité d'expérimenter sur ces traitements.

c) *Quelles hypothèses ?*

L'étude de la validité des images ou des représentations issues d'un instrument d'observation suppose la mise en évidence d'erreurs instrumentales, mais aussi d'erreurs de l'observateur, et enfin d'erreurs ou confusions relatives au matériel observé lui-même.

Dans certaines disciplines, notamment dans les sciences sociales, une utilisation hâtive et inadaptée de l'analyse des données a engendré quelques déceptions à la mesure de certaines illusions : les méthodes n'utiliseraient aucune sorte d'hypothèse, n'auraient aucune exigence en ce qui concerne la préparation des données, et fourniraient néanmoins des images ayant une valeur intrinsèque.

Une confusion concernant les différentes acceptions du mot hypothèse, et les différents types d'hypothèses, pourrait être à l'origine d'un véritable malentendu.

En se limitant au domaine des applications de la statistique, on peut distinguer deux grandes familles d'hypothèses :

1° *Les hypothèses relatives au contenu même de la recherche*, sans lesquelles il n'y aurait même pas de recueil de données (que nous appellerons : hypothèses générales).

Par exemple, dans un modèle économétrique classique : l'hypothèse selon laquelle la consommation de viande d'un ménage dépend du revenu et de la taille de ce ménage.

2° *Les hypothèses nécessaires à l'utilisation d'une technique statistique particulière* (hypothèses techniques). Pour l'exemple précité, pour fixer les idées, hypothèses que les variables exogènes sont mesurées sans erreurs trop importantes, que les résidus vérifient la condition d'homoscedasticité et éventuellement soient distribués suivant une loi normale, etc.

Dans beaucoup d'applications statistiques, l'hypothèse technique la plus importante est l'hypothèse de normalité.

Et quand un statisticien dit qu'il ne fait pas ou peu d'hypothèses, cela signifie en général qu'il va utiliser une procédure robuste ou non-paramétrique, donc qu'il compte s'affranchir de certaines hypothèses techniques.

Les techniques d'analyse de données figurent précisément, en statistique, parmi les techniques qui stipulent les hypothèses techniques les plus faibles.

d) *Codage et étalonnage*

Les méthodes d'analyse factorielle descriptive permettent d'obtenir des images d'un tableau de données. Contrairement aux images de la statistique descriptive élémentaire, ces images ne sont pas d'une lecture évidente, elles méritent la connaissance de *règles d'interprétation* strictes, dont l'expérience montre qu'elles sont en général assez mal connues des utilisateurs.

Ces règles d'interprétation permettent de procéder à des inductions concernant la structure du tableau de données, d'où l'idée de procéder à un *étalonnage* de l'instrument d'observation, en soumettant à l'analyse des tableaux ayant des structures connues, comme les matrices associées à des graphes symétriques. Nous avons ainsi comparé les images obtenues (dans les plans des deux premiers facteurs) en appliquant respectivement l'analyse en composantes principales et l'analyse des correspondances à la matrice associée à un réseau à maille carrée. Un travail également empirique de H. Lebras [28] s'adresse, semble-t-il, à ceux qui entretiennent de dangereuses illusions sur la capacité qu'auraient certaines méthodes d'analyse de mettre en évidence des structures abstraites indépendamment du codage et de la « mise en tableau ».

Si l'on change le codage, les images changent (en général) *mais les règles d'interprétation changent aussi*, le seul invariant est le couple (image, règle). C'est pourquoi nous estimons risqué de diffuser dans des revues non spécialisées des plans factoriels sans les accompagner de minutieuses recommandations à l'intention des lecteurs.

Il existe cependant des codages meilleurs que d'autres : ce sont ceux qui conduisent aux règles d'interprétation les plus simples et les plus claires.

Dans le cas des réseaux à mailles carrées, J. P. Benzécri a montré qu'un calcul analytique pouvait éviter le recours à l'ordinateur [2, T. II-B, n° 10, sur l'analyse de la correspondance définie par un graphe, p. 256]. En pratique, il suffit de choisir un graphe plus simple (par exemple *le cycle*) pour étudier *analytiquement* l'effet des variations de codage de la relation binaire sur les résultats.

En fait, le codage est une préparation du matériel à observer nécessitant obligatoirement la collaboration du statisticien et de l'utilisateur, dont les exigences peuvent être contradictoires. L'une des premières fonctions du codage est de donner un *sens* aux distances entre lignes et colonnes du tableau *avant toute analyse*, de façon à permettre une induction aisée au moment de la lecture des résultats. Cette exigence limite considérablement l'éventail des codages possibles.

3.2. Nouvelles possibilités offertes par l'analyse des données

L'utilisation licite et pertinente des méthodes d'analyse des données aux sciences humaines a un impact qu'il est encore difficile d'apprécier. On peut, sans prétendre être exhaustif, distinguer plusieurs types de contributions que nous allons expliciter et commenter. Ces conséquences heureuses de l'utilisation des méthodes ont été observées lors d'applications au domaine socio-

économique; il n'est donc pas impossible que des apports décisifs remarqués à propos d'autres domaines (psychologie, histoire, etc.) viennent s'ajouter à ceux que nous avons relevés.

Nous distinguerons, de façon peut-être artificielle, des *apports d'ordre technique* (gain de productivité dans certaines phases des dépouillements d'enquête, possibilités nouvelles d'éprouver la cohérence d'informations complexes, de détecter des erreurs de natures diverses, de construire des indices synthétiques), et des apports fondamentaux que nous regroupons dans le paragraphe intitulé : *De nouvelles voies méthodologiques.*

Nous évoquerons les perspectives offertes par l'extension des champs d'observation, les possibilités de découverte de faits ou phénomènes jusque-là inaccessibles, enfin l'enrichissement conceptuel qui résulte des possibilités de visualisation.

3.2.1. *Apports d'ordre technique*

a) *Gain de productivité*

Pour certaines phases des dépouillements d'enquêtes, l'analyse des données est devenue un outil selon nous indispensable : ces techniques permettent de procéder à un ordonnancement rationnel des tâches dans le temps, d'éviter certains piétinements, de contrôler par des représentations visuelles la plupart des étapes de travail, enfin d'accéder à des informations autrefois inaccessibles. Elles permettent également de procéder à des tests de cohérence et à des détections d'erreurs, mais ce dernier point fera l'objet du paragraphe ci-dessous. L'ensemble des opérations implique en fait simultanément un gain de productivité et un gain dans la qualité des résultats. L'aide apportée par l'analyse des données (en fait par une variante de l'analyse des correspondances) est résumée dans l'article cité en [26].

Il s'agit ici d'application d'analyse des données, non pas à un travail de recherche, mais à une activité presque routinière pour un statisticien qui dispose maintenant d'outils à la mesure des recueils de données qu'il a pour charge d'analyser.

b) *Épreuves de cohérence des données et détection d'erreurs*

La détection des valeurs aberrantes ou erronées est une péripétie familière aux statisticiens qui utilisent les techniques d'analyses factorielles.

Toujours dans le domaine des dépouillements de fichiers d'enquêtes socio-économiques, on peut procéder à une *véritable critique des données* en faisant apparaître des « variables techniques » sur les « grilles socio-administratives » [26]. De même que l'on procède à une illustration de cette grille par des caractéristiques des ménages relatifs au contenu même de l'enquête (ménages ayant répondu positivement à telle question, ménages appartenant à une association particulière, à une certaine tranche de revenu, etc.), on peut aussi illustrer cette grille par des variables caractérisant la fabrication de l'information (ménages enquêtés par une même personne, heure de l'interview, éventuellement appréciations ou remarques de l'enquêteur codées convenablement, etc.).

On peut également procéder à une analyse critique du questionnaire lui-même, en étudiant les positions des « non-réponses » à certaines questions par rapport aux positions des réponses effectivement exprimées [35]. Enfin, à propos de n'importe quel recueil de données, on peut détecter des erreurs de mesure ou des anomalies tenant à la méthode de mesure.

Ainsi un récent travail sur les notations de 1 à 20 (analyse de tableaux de contingence croisant les 20 notes et les matières ou objets auxquels ces notes s'appliquaient, lors de deux recueils concernant respectivement des matières scolaires et des appréciations sur des qualités de magasins [32]) a montré que sur le premier facteur, l'ordre des notes était respecté, avec dans les deux cas une interversion, celle des notes 18 et 19. Ceci permet de conjecturer l'existence d'un biais inhérent à l'acte de notation, imputable sans doute à une perception particulière de l'échelle des notes.

Un autre exemple est fourni par un travail de géologie [16], qui a consisté, en particulier, en l'analyse d'un ensemble d'échantillons de sables caractérisés par la proportion des différents minéraux lourds qu'ils contenaient. Le premier axe factoriel extrait a classé ces minéraux dans l'ordre de leurs densités, avec cependant une exception, qui révèle probablement une confusion (due à une ressemblance malheureuse) dans la détermination (optique) des différents minéraux.

Il s'agit dans ce dernier cas d'une anomalie qui aurait été indécelable sans un traitement global des matériaux disponibles.

c) *Construction d'indices synthétiques*

Le premier facteur est la combinaison linéaire des variables ayant variance maximale, et donc le meilleur indice discriminatoire entre individus. Dans beaucoup d'applications, cet indice synthétique a un grand pouvoir descriptif et possède une interprétation intéressante : c'est par exemple le facteur général d'aptitude de l'analyse des psychologues; d'autres facteurs peuvent également avoir des interprétations intéressantes et constituer des variables artificielles. On a pu ainsi constituer un indicateur socio-culturel [34] à partir de variables telles que : profession, niveau d'instruction, profession des ascendants, branche d'activité, etc., ayant un grand pouvoir explicatif de certaines attitudes.

De façon analogue, a été construit un indicateur de la forme de la mortalité [23] à partir des composantes des profils de mortalité départementale qui s'est révélé fortement corrélé au taux net de mortalité départementale.

3.2.2. *De nouvelles voies méthodologiques*

a) *Accès à de nouveaux champs d'observation*

La probabilité de traiter simultanément de nombreuses informations permet de procéder à des recueils de données ne *réduisant pas a priori le champ de l'observable.* Les statisticiens ont pendant longtemps été préoccupés par les observations nombreuses de variables elles-mêmes peu nombreuses, qu'il

s'agisse de valider un modèle particulier de dépendance, ou simplement de tester une hypothèse. *L'infini* du statisticien concerne presque toujours la dimension répétitive « individu ou observation » et exceptionnellement le nombre des variables. (La convergence de certaines estimations en analyse factorielle, lorsque le nombre des variables augmente indéfiniment, a été évoquée par Hotelling dès 1933, puis par d'autres factorialistes). L'obstacle du calcul étant levé, on peut maintenant procéder à un *échantillonnage sur deux dimensions* que l'analyse des correspondances, destinée initialement aux tables de contingence, traite d'ailleurs de façon symétrique. Cette possibilité d'exploration de la dimension « variable » est une innovation dont les premières conséquences ne sont qu'entrevues.

Traitant de l'économie, J. P. Benzécri [réf. 4] doute que les approches réductionnistes (explication du complexe par le simple) y soient possibles comme en physique car dans cette discipline encore, comme dans d'autres branches des sciences humaines, « *l'ordre du composé vaut plus que les propriétés élémentaires des composants* ».

Ceci permet de définir, sinon un champ d'observation précis, du moins une certaine optique, un certain regard sur le « composé », le « complexe », mots vagues qui se préciseront lors de chaque application, car ces entités se matérialiseront sous forme de « corpus » ou recueil de données ayant certaines qualités (homogénéité et exhaustivité, principalement, *cf.* supra). Définir les limites d'un champ d'observation nous procure le même embarras que définir l'exhaustivité : à vrai dire les limites n'ont pas besoin d'être précisées, pourvu qu'elles ne soient pas un obstacle à la compréhension du phénomène à un certain niveau.

L'analyse des données apparaît ainsi comme une méthode de description préalable à une approche « compréhensive ou structurale », par opposition à l'approche « explicative ou réductionniste » (terminologie de R. Thom (1) [36]).

Nous verrons plus bas que l'analyse d'un corpus dans sa totalité apporte des informations originales. De nombreuses réponses à des questions d'opinions relatives à un même thème permettent de déceler et d'analyser les contradictions, les incompréhensions, les réticences des enquêtés.

Les analyses des villes de France caractérisées par leurs profils socio-professionnels ou par leur profil de mortalité [23] permettent de dégager des axes stables et font apparaître des phénomènes que l'on aurait pu croire étrangers au corpus. Les résultats qui seront évoqués au paragraphe *b*) ci-dessous sont précisément à imputer au traitement *simultané* de nombreuses variables.

(1) « L'approche structurale considère la morphologie empirique telle qu'elle apparaît, et elle ne cherchera pas — à moins d'y être contrainte — à introduire une théorie causale externe au champ empirique donné, ni, non plus, à considérer des « atomes » appartenant à un niveau d'organisation plus petit ».

b) *Mise en évidence de phénomènes ou de faits méconnus*

Arrive-t-il que l'on découvre des phénomènes alors qu'on ne s'y attendait pas lors d'une analyse? La méthode de dépouillement d'enquête évoquée au paragraphe 1 nous a montré que l'on pouvait trouver plus vite et de façon plus systématique des résultats quand même accessibles par des méthodes plus classiques au prix d'un effort considérable et souvent dirimant. Mais il arrive que la seule description d'un corpus dégage des résultats parfois surprenants. Nous choisirons deux exemples à titre d'illustration schématique en renvoyant le lecteur aux travaux cités pour plus de détails.

1° *Le profil de mortalité des villes* [23 (pages 134-150)]

Une typologie des villes de plus de 50 000 habitants et des zones départementales complémentaires a été faite en 1969; chacune de ces villes ou départements était caractérisée par 17 causes de décès relatifs à toutes les classes d'âge. Il s'agissait donc de données extrêmement grossières, d'une précision toute relative. Les données concernaient la période 1962-1965. (L'analyse a été faite pour chaque année, chaque ville était décrite par son profil de mortalité dont les composantes sont les parts de chacune des causes de décès dans le total des décès; la population n'intervenant pas.) La typologie obtenue s'est révélée stable au cours du temps.

Le plan des deux premiers facteurs permet de voir les villes se séparer des zones départementales complémentaires et des regroupements régionaux. Mais même à l'intérieur de la région parisienne, on observe une séparation entre Boulogne, Neuilly, Asnières, Saint-Maur, Vincennes, Rueil, Versailles, d'une part, Saint-Ouen, Saint-Denis, Aubervilliers et d'autres communes ouvrières d'autre part.

Sans faire de commentaires plus approfondis, nous dirons que les résultats sont aussi fins que les données étaient grossières; mais le tableau était vaste, et, pour caractériser une ville, 17 informations fragiles peuvent permettre d'observer deux ou trois dimensions stables.

2° *Contradiction et approbation dans les questionnaires*

Un ensemble de réponses à un questionnaire d'opinion (17 questions totalisant 60 modalités de réponses, posées à 2 003 enquêtés [35]) décrivant les attitudes face au travail féminin a été soumis à une analyse des correspondances. Le troisième facteur extrait devait réserver une surprise : alors que le premier facteur discriminait entre les attitudes favorables et défavorables au travail féminin, le second opposait les réponses modérées aux deux types de réponses extrêmes; le troisième opposait les réponses positives aux réponses négatives, indépendamment du contenu des questions. Les individus enquêtés les plus discriminés par ce dernier axe sont donc des individus qui se contredisent, soit pour approuver systématiquement, soit au contraire (mais ils sont moins nombreux) pour répondre négativement.

La mise en évidence d'un tel phénomène est incontestablement d'une grande importance pour comprendre la validité des enquêtes d'opinion, puisqu'elle révèle notamment une absence de symétrie entre les réponses

positives et négatives. La technique d'analyse a fait ici apparaître la *non-neutralité des données*. On se reportera, pour plus d'information, à l'article de N. Tabard.

A l'issue de ces deux exemples, il convient néanmoins de rappeler que ce n'est pas toujours la phase de description qui est la plus riche d'enseignement, mais au contraire la phase d'illustration de la représentation obtenue par des données n'ayant pas participé à l'analyse.

c) *Nouveaux matériaux et nouveaux concepts*

La présentation sous forme de graphiques des résultats (graphique dont les règles de lecture sont extrêmement délicates, contrairement à ce que leur caractère suggestif laisserait prévoir) constitue une innovation méthodologique en soi. En effet, le langage usuel, de par son caractère linéaire et séquentiel, permet d'exprimer facilement des liaisons non-symétriques telles que les *implications*. Une telle relation causale est plus aisée à traduire qu'une relation de *covariation*; ceci explique en partie pourquoi les chercheurs sont parfois fascinés (et rendus perplexes) par les figures bidimensionnelles issues de l'analyse factorielle. Une description en langage clair de l'ensemble des covariations observées est une tâche décourageante. Que l'on songe au nombre de phrases nécessaires à la description des cartes factorielles si l'on désire pouvoir reconstituer, même très grossièrement, l'allure générale des figures (sans faire intervenir, évidemment, les valeurs numériques des coordonnées...). Par où commencer la description, puisqu'il n'existe pas de *relation d'ordre* naturelle dans un espace à deux dimensions? L'exercice de la description mérite d'être fait. On s'aperçoit très rapidement que l'agencement des diverses variables peut être décrit de façon plus *économique* en faisant intervenir des catégories d'un ordre différent de celui des variables (statut social, cycle de vie, ruralité, aisance, dénuement, privilèges, etc., dans le cas de données socio-économiques) et la *description* devient progressivement *interprétation*. (En sciences humaines, un simple résumé contient presque toujours une part d'interprétation). Pour résumer une partie de la typologie de l'exemple 1 ci-dessus, nous pouvons par exemple dire : « *à l'intérieur de la région parisienne, qui se distingue elle-même du reste du pays, on peut distinguer deux zones relativement connexes : les agglomérations à caractère plutôt résidentiel et les agglomérations plutôt ouvrières, qui occupent une position plus excentrée* ». Il va de soi qu'il ne s'agit pas d'une *explication* du phénomène observé (relation entre forme de la mortalité et situation géographique) mais d'une assertion interprétative, puisque certains éléments extérieurs au corpus ont été nommés.

L'image bidimensionnelle antérieure au choix d'un langage de description, avec le système de connotation que véhicule celui-ci constitue selon nous un document de travail, un objet de communication d'une grande densité.

Le chercheur en sciences humaines qui désire observer peut maintenant rassembler ses données sur une base plus large. Il dispose de plus d'une *méthode* pour analyser ses données, qui, parce qu'elle est systématique, suppose un apprentissage, permet des comparaisons, facilite la communication.

Il peut espérer observer des faits qui échappent à tout examen visuel direct des données et dispose, sous forme de cartes munies de leurs règles de lecture, de nouveaux matériaux, pas plus neutres que ses données, mais pas moins, s'il maîtrise l'instrument ; plus neutres en tout cas que toute description en langage clair, et cependant plus proches de la pensée que les données brutes.

RÉFÉRENCES BIBLIOGRAPHIQUES

[1] BALL (G. H.) et HALL (D. J.), *A Clustering Technique for Summarizing Multivariate Data*, Behavioral Sciences n° 12, 1967, p. 153-155.

[2] BENZECRI (J. P.), *L'analyse des données*, Tome 1 : La taxinomie, Tome 2 : L'analyse des correspondances, Dunod, 1973.

[3] BENZECRI (J. P.), *Histoire et Préhistoire de l'analyse des données*, Les Cahiers de l'analyse des données, n° 1 à 4, Dunod, 1976.

[4] BENZECRI (J. P.), *La place de l'a priori*, Encyclopedia Universalis (Organum), 1974.

[5] BERTIN (J.), *Article Graphique (représentation)*, Encyclopedia Universalis, 1973.

[6] BRILLOUIN (L.), *La science et la théorie de l'information*, Masson, Paris, 1959.

[7] BURT (C.), *L'analyse factorielle, Méthodes et résultats*, Colloques internationaux du C.N.R.S. (1955) : « L'analyse factorielle et ses applications », 1955.

[8] CHOUDARY HANUMARA (R.) et THOMPSON (W. A.), *Percentage point of the extreme roots of the Wishart Matrix*. Biometrika, 55, 1968, p. 505-512.

[9] CORMACK (R. M.), *A Review of Classification*, Journal of the Royal Statistical Society, Serie A, vol. 134, part. 3, 1971.

[10] DEVILLE (J. C.), *Méthodes statistiques et numériques de l'analyse harmonique*, Annales de l'INSEE, n° 15, 1974.

[11] DIDAY (E.), *La méthode des nuées dynamiques*, R.S.A., vol. 19, n° 2, 1971.

[12] DREYFUS (J.), *Implications ou neutralité des méthodes statistiques appliquées aux sciences humaines*, Rapport C.R.E.D.O.C.-D.G.R.S.T., 1975, 96 p.

[13] ESCOFIER (B.) et LEROUX (B.), *Étude de trois problèmes de stabilité en analyse factorielle*, Publication de l'I.S.U.P., vol. XXI, 1972.

[14] FISHER (R. A.), *The Sampling Distribution of some Statistics obtained from non linear Equations*, Ann. Eugen 7, 1939, p. 179-188.

[15] FISHER (W. D.), *On grouping for maximum homogeneity*, Journal of the American Statistical Association, n° 53, 1958.

[16] HEIN (P.), *Méthodes statistiques nouvelles et sédimentologie*, Thèse de 3e cycle, Géologie quantitative, Université de Paris VI, 1974.

[17] JACOB (F.), *La logique du vivant*, Gallimard, 1970.

[18] JOUSSELLIN (B.), *Les choix de consommation et les budgets des ménages*, Consommation, n° 1, 1972.

[19] KATO (T.), *Perturbation Theory for Linear Operator*, Springer Verlag, Berlin, 1966.

[20] KENDALL (M. G.) et STUART (A.), *The Advanced Theory of Statistics*, vol. 2, Griffin, Londres, 1961.

[21] KRISHNAIAH (P. R.) et WAIKAR (V. B.), *Exact joint Distribution of any few ordered roots of a class of Random Matrices*, Journal of Multi. An. 1-3, 1971, p. 308-315.

[22] LANCASTER (H. O.), *Canonical Correlation and Partition of* χ^2, Quart. J. Maths. 14, 1963, p. 220-224.

[23] LEBART (L.), *Recherche sur le coût de protection de la vie humaine dans le domaine médical*, Rapport D.G.R.S.T.-C.R.E.D.O.C., 1970, 162 p.

[24] LEBART (L.), TONNELLIER (F.) et SANDIER (S.), *Aspects géographiques du système des soins médicaux*, Consommation, n° 4, 1974.

[25] LEBART (L.), *Validité des résultats en analyse des données*, Rapport C.R.E.D.O.C.-D.G.R.S.T., 1975, 158 p.

[26] LEBART (L.), *L'orientation du dépouillement de certaines enquêtes par l'analyse des correspondances multiples*, Consommation, n° 2, 1975.

[27] LEBART (L.), *The significancy of eigenvalues issued from correspondence analysis*, Proc. in comput. stat. Physica Verlag, Vienne, 1976.

[28] LEBRAS (H.), *Vingt analyses multivariées d'une structure connue*, Mathématique et Sciences humaines, n° 47, 1974, p. 37-55.

[29] MALINVAUD (E.), *Méthodes statistiques de l'économétrie*, Dunod, 1964.

[30] MARTINET (A.), *Eléments de linguistique générale*, A. Colin, 1960.

[31] MEHTA (M. L.), *Random Matrices and the Statistical Theory of Energy Levels*, Academic Press, New York, 1967.

[32] PAGES (J.), *Notation et classement : deux méthodes de recueil de données*, Étude critique à partir d'un échantillon restreint, Consommation, n° 4, 1975.

[33] PILLAI (K. C. S.) et CHANG (T. C.), *An Approximation to the c.d.f. of the largest root of a covariance matrix*, Journal of the Institute of Statist. Math., 1970, p. 115-124.

[34] TABARD (N.), *Besoins et aspirations des familles et des jeunes*, Coll. Études C.A.F., n° 16, 1974, 514 p.

[35] TABARD (N.), *Refus et approbation systématique dans les enquêtes par sondages*, Consommation, n° 4, 1975.

[36] THOM (R.), *La science, malgré tout...*, Encyclopedia Universalis (Organum), 1974.

[37] THOM (R.), *Modèles mathématiques de la morphogenèse*, 10/18, n° 887, 1974.

[38] WILKINSON (J. H.), *The Algebraic eigenvalue Problem*, Clarendon Press, Oxford, 1965.