DEUXIEME SECTION

M É T H O D O L O G I E

par Raymond BOUDON

, ... ANALYSE DES DONNEES ET SOCIOLOGIE

•• par Philippe CIBOIS

L.Lebart, A. Morineau, N. Tabard. - **Techniques de la description statistique. -** Paris, Dunod, 1977.

M. Jambu, M.-O. Lebeaux. - **Classification automatique pour l'analyse des données. -** T. I : *Méthodes et algorithmes, t.* II : ` *Logiciels. - Paris,* Dunod, 1978.

L. Lebart, A. Morineau, J: P. Fénelon. - **Traitement des données statistiques. -** Paris, Dunod, 1979.

J.P. Benzécri, F. Benzécri et al. -- **Pratique de l'analyse des données.** Vol. I : Analyse *des correspondances. Exposé élémentaire.*- Paris, Dunod, 1980.

J. M. Bouroche, G. Saporta. - **L'analyse des données. -** Paris, PUF, 1980 (coll. "Que sais-je ?", n° 1854).

Le sociologue qui s'aventure dans la littérature citée est partagé entre plusieurs sentiments : intérêt, affolement, peut-être même agacement.

Intérêt d'abord : voici des gens qui traitent des données sociologiques : morphologie sociale des communes urbaines de la région parisienne (Lebart, 1977), l'origine socioprofessionnelle des étudiants en Grèce (Benzécri, 1980) et beaucoup d'autres exemples de ce style. Les données traitées peuvent être de vastes tableaux de données ou des données d'enquêtes importantes. Le sociologue voit donc son intérêt éveillé : il va peut-être pouvoir faire autre

chose que son éternel test du Khi-deux appliqué à ses innombrables tableaux croisés.

L'affolement vient au moment où l'on examine les textes de près : l'appareil mathématique est envahissant et d'un niveau élevé. Les auteurs nous en avertissent d'ailleurs : « Le public auquel s'adresse ce manuel est censé posséder les éléments de statistique habituellement acquis dans les premières années de l'enseignement supérieur, et un niveau en mathématique comparable à celui de la licence en sciences économiques, ou à celui des écoles d'ingénieurs » (Lebart, 1979, V).

Si l'ouvrage de J: P. et F. Benzécri a pour origine un exposé à des linguistes « de formation purement littéraire » (p. VIII) il n'en comprend pas moins 175 pages d'un exposé mathématique qui, s'il ne présuppose pas de connaissances mathématiques importantes, nécessite cependant une forte motivation pour être parcouru à fond.

Devant les nécessités d'un tel investissement mathématique, le chercheur risque de déclarer forfait et de renoncer à comprendre ce type de méthodes, même s'il les utilise.

L'agacement vient ensuite quand on lit les commentaires des exemples appartenant au champ sociologique : était-il vraiment nécessaire de recourir à une méthode aussi puissante pour découvrir (dans l'exemple de Lebart, 1977, 68) que les communes urbaines de la région parisienne s'opposent selon un critère de ségrégation sociale ? Que l'importance de l'autoconsommation alimentaire dans les budgets familiaux est liée à « un facteur de ruralité exprimant la proximité à la terre, quelle que soit la profession principale du ménage » (Benzécri, 1980, 328). Le premier réflexe du sociologue est de ne voir dans ces techniques que des machines coûteuses destinées à produire des évidences.

Ces réactions sont naturelles, nous voudrions cependant les dépasser et inviter à une lecture plus « professionnelle » des ouvrages cités, c'est-à-dire se poser les questions suivantes : s'il y a littérature abondante, y a-t-il des régularités, y a-t-il une source commune ? Ces méthodes sont-elles liées à un groupe de consommateurs privilégiés ? Ce mouvement existe-t-il à l'étranger ou est-il spécifiquement français voire parisien ? Quelle est la stratégie des acteurs ? Peut-on repérer des thèmes idéologiques ? Chaque sociologue pourrait, selon ses goûts, poser à l'*analyse des données* la question qu'il pose habituellement à la réalité sociale.

Nous voudrions, dans les lignes qui vont suivre, essayer de donner des éléments de réponse à ces questions en situant l'analyse des données dans un contexte plus large et en posant quelques jalons en ce qui concerne le cas français. Ceci fait nous reprendrons sous un nouveau jour les questions initiales.

Jalons historiques

A l'origine de la situation française actuelle se trouvent les travaux de J.-P. Benzécri sur l'analyse factorielle des correspondances. Cette méthode, mise au point dans les années 1962-1965, a voulu introduire en France des points de vue proches de la *data analysis* des Anglo-saxons.

En ce qui concerne ce dernier phénomène, citons Rouanet et Lépine[1] : « L'expression est, au moins depuis une quinzaine d'années, très courante chez les Anglo-saxons, aussi bien chez les théoriciens que chez les usagers de la statistique ; toujours utilisée, outremer, de façon très souple, elle désigne, non pas vraiment un ensemble de techniques, et encore moins une « doctrine établie », mais plutôt « une certaine idée de la statistique », selon laquelle il est légitime *en principe* (même si dans la pratique cela ne va pas toujours sans problèmes) d'examiner les données pour les interpréter, quelles que soient les intentions et les modalités qui ont pu présider à leur recueil, et sans avoir à s'enfermer dans un modèle ou des hypothèses restrictives. Cette conception (...) a dû en fait se constituer et s'affirmer en réaction contre les excès de l'école « décisionniste » naguère dominante, laquelle, selon une déviation certes peu conforme à l'esprit des pères fondateurs de la statistique moderne, en arrivait à ne plus voir dans les données qu'une sorte d'intermédiaire destiné à permettre de prendre mécaniquement une « décision » (celle-ci d'ailleurs en général toute formelle) dont tous les termes (modèle probabiliste, mais aussi, le cas échéant, fonction de coût, probabilités *a priori,* etc.) devaient (ou auraient dû), toujours en principe être posés au départ. »

Les auteurs jugent que chez Benzécri le terme «analyse des données» va au-delà de la *data analysis* car

« - d'une part, le point de vue de l'analyse des données se transforme en un principe radical qui tend à éliminer (plutôt qu'à contrebalancer) tout autre point de vue, notamment le point de vue décisionnel ;

- d'autre part, l'expression même d'« analyse de données » tend à désigner également les méthodes favorites de l'auteur ».

A partir de 1965 la méthode de l'auteur : l'analyse factorielle des correspondances, se répand dans la recherche appliquée (CREDOC), dans les milieux traitant de données économiques et sociales (INSEE), ainsi que dans la recherche en sociologie et en géographie.

1. L H. Rouanet et D. Lépine. A propos de « L'analyse des données » selon Benzécri, *Année psychologique*, 78, 1976. pp. I37-138.

Après la parution en 1973 de *L'analyse des données* (réédité en 1976), ouvrage de référence sur l'analyse des données au sens de l'auteur[2], une revue trimestrielle s'est créée en 1976: *Les Cahiers de l'Analyse des Données* ainsi qu'une collection « Pratique de l'analyse des données » (dont il est rendu compte ici du premier titre), revue et collection dirigées par J.-P. Benzécri et entièrement consacrée au développement de l'analyse des correspondances, des méthodes connexes ainsi qu'à des exemples d'application.

Utilisée dans les bureaux d'étude, puis dans la recherche, l'analyse factorielle des correspondances a atteint des milieux plus vastes, d'abord sous forme de publication de résultats (diverses enquêtes du *Nouvel Observateur)* puis sous forme d'exposés de la méthode *(Le Monde, La Recherche, Pour la Science* et enfin le "Que Sais-je ?" dont il est rendu compte ici)[3].

Du développement rapide de l'usage des techniques de Benzécri nous n'en retiendrons qu'une preuve, c'est l'usage du terme de *mode,* terme qui est très souvent employé pour qualifier le phénomène. En effet si on l'emploie c'est pour marquer qu'il s'agit d'un phénomène qui touche un grand nombre et que cette diffusion est liée à une pression sociale du milieu (la connotation est bien sûr défavorable : chez des scientifiques, il va de soi que les évolutions dans l'emploi des méthodes ne relèvent en rien de la pression sociale !).

Benzécri, pour cette diffusion, n'a pas joué la carte universitaire et par exemple ne fréquente pas les congrès scientifiques sauf ceux organisés par lui et où l'on ne parle que de ses techniques. Par contre il a formé et continue de former un grand nombre d'étudiants qui font la propagande de sa méthode dans les milieux de la recherche appliquée.

Notons cependant qu'un certain nombre de chercheurs en mathématique appliquée, collaborateurs de Benzécri, jouent la carte universitaire et visent à faire reconnaître une "école française" d'analyse de données centrée sur la classification automatique et l'analyse des correspondances. Les ouvrages que certains diffusent, s'ils incluent les méthodes de Benzécri, cherchent à les rattacher à un enseignement statistique très classique (c'est particulièrement le cas de Lebart (1979) dont il est rendu compte ici). Le but de ces chercheurs que nous pouvons désigner sous le terme des "analystes de données" est, éventuellement par des alliances avec des

2. J.-P. Benzécrit et al., *L'analyse des données,* Paris. Dunod, 1973, 2 vol.
3. L. Diday et L. Lebart. L'analyse des données, *La Recherche, .n° 74,* Janvier 1977, pp. 15-25 : J.-M. Bouroche et G. Saporta, L'analyse des données, *Pour la Science,* n° 5, mars 1978, pp. 23-34 : M. Arvonny, Statistiques, sondages et mensonges, *Le Monde,* 29 octobre 1975, p. 17 (Le Monde des Sciences et des Techniques).

chercheurs non collaborateurs de Benzécri, de présenter les caractéristiques d'une « Ecole ».

Sociologie de l'analyse des données

Après ces quelques jalons historiques[4], essayons d'interpréter en sociologue en employant d'abord un registre qui nous semble particulièrement pertinent, celui de la sociologie religieuse.

En effet cette manière de voir nous semble doublement fondée dans une articulation « à étage » entre le mouvement lié à l'analyse factorielle, mouvement créé par un leader de style prophétique, qui dispose d'un corpus de textes de références et de disciples attelés à la propagation du message, et d'autre part pour le contenu lui-même du message délivré par Benzécri.

Revenons sur le premier point : le mouvement est dirigé par un leader de « style prophétique » qui, en particulier par un refus affirmé de sacrifier aux habitudes universitaires ou même simplement vestimentaires, tient à s'affirmer comme "ailleurs", témoin privilégié d'un message que nous étudierons ensuite.

Ce leader a créé un corpus de textes de référence : cette notion de corpus est très explicite : elle se trouve dès l'origine (c'est-à-dire en tête du premier volume qui est l'ouvrage, dit-on, de 70 auteurs, comme la traduction grecque de la Bible par les Septantes) : « Par le présent volume, nous entendons débuter la publication ordonnée de nos recherches statistiques et des travaux de notre laboratoire »[5]. Elle est présente par le biais d'une technique particulière de référence interne : chaque chapitre ou article du corpus est affecté d'un mot ou d'un ensemble de mots en abrégé, entourés de crochets. Par exemple la première leçon du livre de Benzécri dont il est fait ici le compte rendu est synthétisée par la référence [Correspondances : profils]. La présence de ces citations entre crochets peut d'ailleurs être considérée comme un signe d'appartenance à l'univers des collaborateurs directs de Benzécri. Dans les ouvrages étudiés ici, on les trouve dans Benzécri (1980) et dans Jambu (1978), mais non dans les autres qui cependant citent tous Benzécri comme un de leurs inspirateurs.

Dans ce corpus, on trouve aussi une « histoire sainte », c'est-à-dire l'histoire du problème revue par Benzécri[6].

4. Nous renvoyons pour plus de détails au troisième chapitre de notre thèse : « La représentation factorielle des tableaux croisés et des données d'enquête : étude de méthodologie sociologique », Paris, LISH, 1980, chapitre paru dans *Informatique et Sciences* humaines, n°48-47, automne-hiver 1980.

5. Benzécri et al., op. cit., vol. 1. p. V.

6. *J.-P.* Benzécri. Histoire et préhistoire de l'analyse des données, *Les Cahiers de l'Analyse des Données,* 1978, 1-4, et 1977, 1.

Quant aux disciples, ils se caractérisent par une forte liaison personnelle avec leur maître et une incapacité fréquente, car ce sont des disciples, à penser les problèmes qu'ils rencontrent en d'autres termes que ceux qui leur ont été inculqués[7]. *A contrario,* les « analystes de données », dont un certain nombre ont reçu l'enseignement de Benzécri, ont pris vis-à-vis de lui une certaine distance pour pouvoir faire des présentations différentes de l'analyse des données.

Si le mouvement fonctionne selon un mode religieux, le contenu ne l'est que d'une manière indirecte bien que très réelle. Benzécri s'en est d'ailleurs pleinement expliqué dans deux textes : « La place de l'a *priori* »[8], et « Philosophie thomiste et connaissance mathématique de la nature[9]». Sans refaire l'analyse détaillée de ces deux textes[10], on peut les résumer de la façon suivante.

En premier lieu Benzécri pense que les sciences humaines sont menacées par « l'idéalisme », philosophie souvent implicite des chercheurs qui leur fait substituer leurs *a priori* à une observation de la réalité. Pour lutter contre cette manière de faire, il propose sa méthode statistique : « Cet outil nouveau qu'est l'ordinateur électronique peut permettre de substituer à des notions qualitatives du sens commun des quantités définies statistiquement ; en sorte que l'édifice, fondé sur une ample base de faits, s'affranchira, dans sa structure définitive, de l'arbitraire échafaudage des idées *a priori* »[11].

Ce texte est à prendre au sens strict : les « quantités définies statistiquement » existent bien dans l'analyse des correspondances : ce sont les facteurs. Grâce à la recherche des facteurs le chercheur pourra s'affranchir des idées reçues et en particulier, il pourra « rester sourd aux idées que proposent les meneurs à la mode »[12].

Benzécri est très polémique vis-à-vis des sciences humaines (auxquelles il ne veut d'ailleurs pas donner ce titre de sciences)

« A l'idéalisme verbaliste qui menace de vicier notre compréhension d'une science aussi bien fondée que la physique, il semble que plus d'une des disciplines qu'on appelle sciences humaines ait présentement succombé »[13].

7.Cf. la perception de Ph. Kaminski, *Propos iconoclastes sur l'analyse des correspondances*, Besançon, INSEE, 1977: : « Lancés à travers le pays, les Bencécri's boys ont évangélisé à tour de bras. Tout le monde a suivi, dans le cohue des grandes transhumances, abandonnant sur place paquetages et provisions » (*in* Épilogue).

8. In *Encyclopedia Universalis*, vol. 17, Organum, Paris 1973, pp. 11-24.

9. In *La Pensée catholique*, n° 118. Paris, 1969, pp. 58-76.

10.Cf. le troisième chapitre de notre thèse citée plus haut.

11.La place de l'a *priori, op. cit.*, p. 21.

12.*Ibid.*, p. 24.

13.*Ibid.*, p. 18.

La philosophie sous-jacente est explicitée par ailleurs : « Nous croyons, quant à nous, que dans les phénomènes, l'esprit humain retrouve sans cesse des principes ontologiques immuables, objet de la *philosophia perennis* »[14].

La *philosophia perennis* désigne la philosophie thomiste, philosophie officielle de l'Eglise catholique.

Cette référence au thomisme n'est pas sans importance statistique : il existe un lien entre cette philosophie « réaliste » au sens technique du mot et l'interprétation des facteurs en analyse des correspondances. En effet, selon Benzécri, si on fait de telles analyses, « c'est dans l'espoir de découvrir les axes propres à un équilibre existant réellement dans la nature (...), on aspire à découvrir les propriétés cachées qui, situées plus haut dans la hiérarchie naturelle des causes que celles qui tombent sous le sens, régissent celles-ci »[15]. Ou en d'autres mots : comme les réalités de ce monde sont des choses créées par Dieu, le travail du statisticien est de remonter des faits à l'essence des choses, forme que le Créateur leur a donnée ; il ne peut accomplir cette démarche que s'il dispose d'une ample moisson de faits qu'il devra « distiller » jusqu'à en tirer des idées.

Résumons-nous: Benzécri polémique avec les sciences humaines trop marquées à son gré par des « meneurs à la mode » : il le fait au nom de la philosophie thomiste dont il nous dit qu'il l'a découverte en approfondissant sa foi catholique.

Il semble légitime de rattacher la pensée de Benzécri à un courant du catholicisme qui prône activement la philosophie thomiste, qui critique l'idéalisme des sciences humaines et qui publie dans *La Pensée catholique,* c'est-à-dire au courant intégriste. Ce rattachement est éclairant dans la mesure où il permet de comprendre que la lutte politique contre l'idéalisme philosophique se fait sur un terrain bien précis, celui des sciences humaines. En effet, l'intégrisme voit une filiation entre l'idéalisme cartésien, la philosophie de Kant et celle de Hegel, puis celle des « maîtres du soupçon » selon l'expression de P. Ricoeur, c'est-à-dire Marx, Freud et Nietzsche dont est issu l'athéisme moderne. Ces philosophies n'étant pas sans lien avec certains développements des sciences humaines.

Benzécri veut faire oeuvre sociale, donc, selon la pensée intégriste, œuvre religieuse du même coup, en s'attaquant aux sciences humaines non pas en tant que telles mais en leur proposant une solution technique de rechange : l'analyse des correspondances. En utilisant cette méthode, les chercheurs découvriront dans

14. Ibid., p. 15.
15. *L'analyse des données,* t. 2*, op. cit.,* p. 48.

les facteurs la véritable nature des choses et abandonneront ainsi « sans douleur » leurs *a priori* condamnables.

On comprend mieux ainsi les raisons de l'apostolat qui est fait de l'analyse des correspondances : il s'agit d'un apostolat à motivation religieuse, mais dont les moyens sont d'ordre temporel, liaison faisant partie de l'idéologie intégriste[t6].

Cette remise en cause des sciences humaines explique que la méthode de Benzécri se soit diffusée en premier lieu dans des institutions semi-publiques du genre CREDOC : ces institutions en effet sont pratiquement tenues de ne pas avoir sur la réalité sociale un regard critique. L'analyse des correspondances leur a permis de tenir un discours neutre, ne faisant que refléter les *a priori* implicites contenus dans les instruments de recueil des données, c'est-à-dire les questionnaires.

On s'explique moins par contre le succès rencontré par l'analyse des correspondances dans les milieux de la recherche, et en particulier chez des sociologues qui sont loin de partager la philosophie de Benzécri. En fait, pour rendre compte de cette diffusion, il faut replacer le phénomène en question dans le cadre du développement des techniques d'analyse de données, développement lié à l'emploi des ordinateurs. Pour mieux voir cet effet, nous comparerons la situation française avec son équivalent dans les pays anglo-saxons.

Les techniques anglo-saxonnes d'analyse de données

Si nous regardons les pratiques et non les intitulés, c'est aux techniques de *multidimensional scaling* qu'il faut comparer l'analyse des correspondances et les techniques françaises d'analyse de données, non à l'analyse factorielle traditionnelle..

Examinons les points communs et les différences : dans le cas de l'analyse multidimensionnelle comme dans le cas de l'analyse des correspondances, on cherche à avoir une perception visuelle des données en les représentant dans un espace à deux dimensions, éventuellement trois, alors que dans l'analyse factorielle traditionnelle les nombreux facteurs sont représentés sous forme de matrices de coordonnées, si bien que les structures sous-jacentes ne sont guère plus apparentes que dans les données brutes.

En analyse multidimensionnelle comme en analyse des correspondances on ne vise pas, au contraire de l'analyse

18. Thème du Christ-Roi que l'on retrouve d'ailleurs sous une forme iconographique dans *Les Cahiers de l'Analyse des données* (dessins de "vœux" des derniers numéros de 1977 et 1978).

factorielle traditionnelle, à fournir un modèle des données : la perspective est de montrer au chercheur ses données sous le meilleur jour possible, de façon qu'il puisse y découvrir les structures sous-jacentes. Ces méthodes ont une fonction heuristique ; les structures d'opposition sont en général repérées par les directions de la représentation graphique : plus on va dans une direction, plus les objets représentés ont en commun un certain nombre de propriétés.

Les deux méthodes correspondent à des pratiques communes : elles sont liées au développement de l'informatique dans la recherche, et dans les deux cas, elles sont utilisées comme instrument pratique d'observation des données. Comme il s'agit d'un point de départ, le chercheur peut être assez ignorant de la manière dont sont obtenus les résultats ; l'analyste de données français, comme son homologue anglo-saxon, veut une description de ses données et non un modèle. Pour arriver au même but, des techniques différentes ont été envisagées, mais l'expérience montre que sur des mêmes données les résultats sont tout à fait comparables.

On peut donc dire que les techniques d'analyse de données centrées autour de l'analyse des correspondances et de la classification automatique, sont l'avatar français d'un phénomène international, fruit du mariage d'un désir, celui de voir l'ensemble de ses données, avec une technique, l'informatique.

On pourrait discuter longuement pour savoir quelle est la méthode la meilleure et mettre en avant l'exigence française de pureté mathématique de la méthode de Benzécri, opposée à l'empirisme anglo-saxon des algorithmes non métriques développés par Kruskal et autres. Il s'agit là d'un débat qui n'intéresse pas directement le sociologue. Par contre on comprend mieux pourquoi les techniques de *multidimensional scaling* sont peu utilisées en France et les techniques de Benzécri peu employées dans le monde anglo-saxon. On ne doit pas interpréter ce fait comme un provincialisme d'un côté et un ostracisme de l'autre : cela vient tout simplement du fait que chacun dispose de techniques analogues répondant aux mêmes besoins de chercheurs, et n'éprouve pas de ce fait le besoin d'en acclimater d'autres.

L'usage de l'analyse des données en sociologie

Reprenons maintenant les réactions spontanées que le sociologue peut avoir devant la littérature d'analyse des données ; l'intérêt pour lui peut être grand s'il en voit bien la finalité. Il s'agit de méthodes heuristiques ayant une finalité descriptive. Devant une masse importante de données, il est possible maintenant d'en avoir rapidement une vision synthétique. La valeur du

résultat est fonction de la méthode mais aussi de celui qui regarde: un profane mis devant un microscope ne voit rien, il ne sait pas ce qui est pertinent. Il en est de même en analyse factorielle des correspondances et on s'explique mieux la pauvreté de certains commentaires présentés par des analystes de données qui ne sont pas spécialistes d'un domaine. Devant des données même intéressantes, ils ne voient que les oppositions qu'un profane peut voir, oppositions qui peuvent sembler évidentes et sans intérêt pour le spécialiste.

Il est impossible d'isoler l'opération de vision des données d'un processus complet de recherche : l'observation des données n'est qu'un moment de la recherche. Si certains analystes de données ont encore la naïveté de l'adepte de l'empirisme qui croit pouvoir par sa méthode observer « la réalité » directement, le sociologue sait bien que son observation est pour une part construite et qu'il retrouvera donc dans une vision d'ensemble de ses données quelque chose des concepts qui ont été mis en place pour le recueil de l'observation. Toute la démarche scientifique réside dans le fait que malgré tout, une observation des données peut conduire à des surprises, à des agencements inattendus.

Regarder l'ensemble de ses données pour y découvrir des agencements inattendus n'est pas le tout de la démarche scientifique : on a souvent des hypothèses à vérifier, des modèles à tenter. Il n'empêche qu'à un moment ou à un autre le chercheur a envie d'avoir une vision d'ensemble parce qu'il se trouve dans un domaine nouveau.

C'est typiquement le cas dans la pratique du dépouillement d'enquête pour lequel l'analyse factorielle des correspondances est un instrument commode. En effet, si l'on décide de faire une enquête, c'est que l'on suppose qu'un certain nombre de phénomènes ont entre eux des liaisons, mais que l'on ne sait ni lesquels exactement ni comment. Dans ce cas, l'analyse des correspondances permet d'avoir une vue globale de ces liaisons si elles existent.

Analyse des correspondances et dépouillement d'enquête

L'usage de l'analyse des correspondances dans le cas du dépouillement d'enquête est certainement le cas où la méthode apporte le plus, mais où aussi beaucoup de difficultés d'interprétation subsistent.

Rappelons-en le principe : dans le cas standard, l'analyse des correspondances traite un tableau de contingence où à l'intersection d'une ligne et d'une colonne se trouve l'effectif des individus qui ont en même temps la caractéristique de la ligne et celle de la

colonne. Par contre, dans le cas du dépouillement d'enquête, le tableau utilisé est un tableau d'occurrence. A chaque individu de l'enquête correspond une ligne du tableau et à chaque modalité de réponse aux différentes questions correspond une colonne. L'intersection d'une ligne et d'une colonne compte le nombre de fois où un individu a ou non cette caractéristique : par définition, il ne peut l'avoir que une fois ou zéro fois. Ce codage « logique » (en zéro/un) donne des résultats très différents de ceux d'un simple tableau de contingence et en particulier possède cette propriété très particulière que le Khi-deux du tableau (l'inertie dans le langage des analystes des données) *nest en aucun cas fonction des données* mais dépend uniquement du nombre de questions de l'enquête et du nombre de modalités de réponses que ces questions ont engendrées.

De ce fait les règles d'interprétation des contributions des facteurs sont entièrement modifiées et les valeurs observées ne doivent en aucun cas être comparées avec les valeurs analogues observées dans le cas d'un tableau de contingence.

Le deuxième point de divergence avec le cas standard porte sur l'interprétation des plans factoriels : alors que dans le cas du traitement d'un tableau de contingence, quand une ligne et une colonne sont en conjonction angulaire par rapport au centre, cela signifie qu'on a de bonnes chances d'observer un écart positif à l'indépendance dans le tableau d'origine à l'intersection de la ligne et de la colonne. D'une manière différente, dans le dépouillement d'enquête, on ne représente pas les lignes (les individus sont en général trop nombreux pour être observés individuellement), mais on observe des conjonctions de modalités de réponse issues de plusieurs questions.

Quel est le statut de ces configurations ? L'expérience a montré que l'on pouvait les assimiler à des types idéaux wébériens. En effet, si l'on repère une configuration de quatre modalités issues de quatre questions différentes en conjonction étroite, on peut facilement vérifier par comptage que le nombre d'individus qui possèdent en même temps les quatre modalités est assez faible[17], par contre si l'on est moins exigeant, si l'on compte ceux qui en possèdent trois ou même simplement deux, on voit les nombres croître assez rapidement. Nous sommes bien dans le cas d'un type idéal : configuration assez rare statistiquement si l'on recherche le type pur mais par contre bien attestée numériquement si l'on recherche des approximations du type idéal.

Ceci nous explique la perplexité de beaucoup de chercheurs devant leurs résultats : ils sont à la recherche de liaisons que l'on

17. Mais supérieur à l'effectif théorique correspondant à l'Indépendance.

pourrait qualifier de « durkheimiennes »[18] entre des variables ; avec une variable explicative, une variable à expliquer et éventuellement des variables autres qui masquent cet effet et ils découvrent présentées par l'analyse des correspondances des configurations « wébériennes » de variables faites de types idéaux quelquefois très peu marqués (c'est-à-dire portant seulement sur une sous-population restreinte).

En sociologie il y a des méthodes « durkheimiennes » et des méthodes « wébériennes », les premières liées au coefficient de corrélation et à toutes les méthodes qui l'utilisent *(path analysis* et régression), les autres liées aux méthodes graphiques qui présentent des configurations de modalités de réponse *(rnultidimensional scaling* et analyse des correspondances). Les premières sont développées en France sous forme de modèles avec rétroaction dans les cas les plus élaborés, les secondes sont souvent utilisées par les chercheurs qui se réclament d'une inspiration wébérienne[19].

Comme on le voit, nous avons laissé entièrement de côté l'idéologie benzécrienne du réalisme des facteurs. Nous n'avons pu le faire que parce que nous avons repéré l'origine extra-mathématique de cette manière de voir inspirée par des considérations philosophiques et religieuses tout à fait honorables, mais qui relèvent du choix individuel. Il est évident que beaucoup de chercheurs ont du mal à s'en débarrasser parce que Benzécri est mathématicien et que c'est en profondeur que son discours mathématique est sous-tendu par son discours philosophique. Lui-même en est d'ailleurs tout à fait conscient puisqu'il dit que « l'analyse des correspondances est une méthode ; elle est aussi un outil. A la philosophie de la méthode l'outil doit son efficacité... »[20]

Ce que nous pouvons souhaiter c'est qu'ayant fait cette découverte, les chercheurs utilisent l'analyse des correspondances dans une perspective heuristique, pour la recherche de types idéaux wébériens, et ce d'une manière consciente. Dans le dépouillement d'enquête, l'analyse des correspondances est un point de départ, une méthode rapide pour découvrir des configurations intéressantes qu'en tout état de cause, il faudra ensuite comptabiliser avec précision et vérifier. Méthode synthétique, l'analyse des correspondances ne peut que proposer des approximations des données : il appartiendra toujours au chercheur de revenir à ses données

18. Pour reprendre une remarque de R. Establet.

19. On notera par exemple que 1'. Bourdieu utilise l'analyse des correspondances non pas tellement à des fins heuristiques qu'à des fins de présentation synthétique des résultats. Pour lui une configuration de modalités exprime bien le type Idéal qu'il veut manifester.

20. J.-P. Benzécri, Histoire et préhistoire de l'analyse des données. Les *Cahiers* de *l'analyse des données (1), 1977;* p. *39.*

après l'analyse des correspondances à des fins de vérification et d'affinement des résultats[31].

Présentation de la méthode

A des degrés divers tous les ouvrages étudiés ici ont tous le même mode d'exposition : il s'agit toujours de la juxtaposition d'un formalisme mathématique avec des études de résultats faits sur des cas issus de diverses disciplines de sciences humaines ou biologiques.

Comme nous l'avons vu, il est difficile de reprocher aux auteurs le peu d'intérêt des résultats auxquels ils arrivent : en effet, des résultats intéressants ne peuvent être trouvés qu'au niveau d'une recherche spécifique et par un chercheur spécialisé dans un champ de recherche. Les exemples présentés n'ont donc qu'une vertu pédagogique : ils permettent de voir comment se présente un résultat.

Par contre sur la question du formalisme mathématique qui est utilisé, nous devons être très critiques car tous les auteurs ont identifié un formalisme mathématique qui permet de *valider* la méthode avec un mode d'exposition qui en donne la compréhension nécessaire à sa bonne utilisation.

Chez Benzécri il y a une liaison intime entre la philosophie et la pratique du mathématicien : il n'est pas facile de faire des distinctions qui sont pourtant nécessaires. Nous pouvons poser cependant quelques jalons de ce type d'analyse.

Le point central est la philosophie du réalisme des facteurs liée au formalisme mathématique de « l'espace » : que l'on relise la toute dernière présentation de la méthode (Benzécri, 1980), on y voit que

1) Ce sont les lignes et les colonnes du tableau *dans leur ensemble* qui constituent les objets d'étude : chaque ligne ou chaque colonne constitue un objet appelé *profil.* Ce n'est pas, notons-le bien, l'intersection de la ligne et de la colonne qui est prise en compte, c'est tout l'ensemble des nombres de la ligne ou de la colonne. II y a là une option qui engage la méthode vers un certain type de présentation : nous verrons qu'une autre option est possible.

2) Ces *profils,* c'est-à-dire ces lignes et ces colonnes auxquelles on a donné le statut d'objet, forment une « réalité multidimensionnelle »

21. Le souhait évoqué plus haut est déjà en partie réalisé puisque dans le même passage cité Benzécri note que son outil tel un marteau sans maître frappe désormais librement, et que cette situation lui inspire des regrets, car des gens ont pris l'outil sans la philosophie sous-jacente.

dont il faut donner une vue sensible « grâce à une réduction optima de la dimension » (Benzécri, 1980, p. 11).

3) Ici la philosophie se manifeste clairement : ce sont les facteurs qui sont la réalité profonde, car si l'on prend suffisamment de données, « c'est dans l'espoir de découvrir les axes propres à un équilibre existant réellement dans la nature (...), on aspire à découvrir des propriétés cachées qui, situées plus haut dans la hiérarchie naturelle des causes que celles qui tombent sous le sens, régissent celles-ci »[22].

En un mot, la pensée créatrice de Dieu telle que l'homme peut la découvrir dans la nature est trop vaste pour son esprit limité. Seule une réduction à quelques dimensions peut lui permettre de l'appréhender.

Comme la multidimensionnalité du Créateur est première dans l'analyse de Benzécri, on comprend le formalisme mathématique mettant l'accent sur la notion d'espace et plus particulièrement d'espace vectoriel, ce qui a pour conséquence de faire de la ligne ou de la colonne l'objet de départ.

Si l'on ne partage pas la philosophie de Benzécri, il est possible de donner de la méthode une présentation toute différente, qui met l'accent sur d'autres formalismes mathématiques, et qui répond mieux aux exigences intellectuelles des sociologues.

En effet, et nous suivons ici une présentation de G.-Th. Guilbaud, on peut légitimement centrer son attention non sur la ligne et la colonne comme objet, mais sur les intersections des lignes et des colonnes, c'est-à-dire sur le contenu du tableau de données dans son ensemble. Le but de l'analyse des correspondances devient alors *descriptif :* il s'agit de trouver la meilleure *approximation* du tableau en se servant du minimum de facteurs. Les facteurs ne sont plus alors des substances que l'on découvre mais des résumés approximatifs mais pratiques puisque par multiplication du vecteur ligne par le vecteur colonne, on obtient une approximation facile à lire des écarts à l'indépendance..

Dans cette présentation le terme de facteur est à prendre dans le sens de « mise en facteur » d'une expression. Par une opération purement formelle, on présente à nouveau les mêmes réalités d'une façon plus agréable pour l'usager. Plutôt que d'analyse factorielle nous préférons parler de *représentation* factorielle, de présentation nouvelle des mêmes données (ou de leur approximation) sous un jour plus agréable.

Avec cette manière d'exposer la méthode, l'expérience nous a montré qu'en quelques heures, des chercheurs pouvaient comprendre la nature des opérations effectuées

22. *L'analyse des données, t.* II, p. 48.

dans une analyse des correspondances. I1 s'agit de *comprendre,* c'est-à-dire de démonter le mécanisme, ses articulations, de pouvoir être capable d'en reproduire les procédures à la main dans des cas simples ; il ne s'agit pas d'être capable d'en faire la théorie.

Tous les ouvrages présentés ici veulent au moins ébaucher sinon faire la théorie des méthodes qu'ils présentent. Si un tel investissement peut dans certains cas être légitime, il ne l'est pas pour l'utilisateur moyen. Autant il est dangereux d'utiliser une méthode sans en avoir une expérience théorique et pratique, autant c'est un piège de laisser croire au chercheur qu'il doit être capable de valider la méthode.

En premier lieu le sociologue a tout à fait le droit d'utiliser des méthodes statistiques validées par d'autres que lui, mais c'est un piège car l'impossibilité pratique où il se trouve d'effectuer cette validation permet de faire l'impasse sur toute tentative d'explicitation des procédures suivies. Les procédures mathématiques servent ici d'écran à la philosophie de Benzécri avec un raisonnement de ce style : « Chers amis utilisateurs, je vais vous expliquer l'analyse des correspondances : il s'agit de trouver les axes principaux d'inertie dans l'espace des profils et, pour ce faire, on est amené à diagonaliser une matrice carrée symétrique... Ah ! Vous ne suivez plus, eh bien ce n'est pas grave, passons tout de suite à une étude de cas ». Il s'agit là d'une caricature car l'enseignement mathématique peut être correctement fait, mais tant que le facteur sera la réalité sous-jacente, les considérations géométriques seront fondamentales. Si l'on envisage un point de vue descriptif, on peut présenter l'analyse factorielle comme la décomposition d'un tableau de données en une somme de tableaux de rang un, présentation tout à fait abordable pour le sociologue pour lequel le tableau croisé constitue un instrument familier-.

Analyse des données et pratique du sociologue

Au terme de ce compte rendu, nous voudrions souligner que la perspective de l'analyse des données peut être d'une grande utilité au sociologue à deux conditions :

- Qu'il se trouve dans une situation exploratoire en ce qui concerne l'articulation de plusieurs variables et qu'il se serve de ces techniques à des fins descriptives, et non modélisantes. En un mot, que sa perspective soit plus e wébérienne » que « dukheimienne », mais il est possible de passer de l'une à l'autre dans le déroulement d'une recherche.

23. Pour une telle présentation, nous renvoyons au premier chapitre de notre thèse citée plus haut.

L'analyse des données doit se situer *au début* d'une recherche, elle sert à faire avancer une recherche, elle n'en est jamais la phase finale. Outil de départ elle peut même être absente de la publication de résultats trouvés par son intermédiaire, mais vérifiés et précisés par la suite par des méthodes directes.

- Qu'il acquière une *expérience* théorique et pratique de la méthode. Expérience pratique en faisant des essais sur des données bien connues de lui pour en quelque sorte étalonner pour son propre compte la méthode. Expérience théorique en refusant de se contenter d'une vague image spatiale mais aussi en refusant l'obligation d'un apprentissage des formalismes mathématiques (l'algèbre linéaire) qui fondent la méthode. Entre ces deux extrêmes on peut, par un examen précis des procédures utilisées, se faire une idée suffisante de la méthode pour voir ses avantages et ses dangers.

II nous semble que pour admettre ces deux conditions, il importait de voir que les pratiques contraires diffusées par la littérature française d'analyse de données sont sous-tendues par une philosophie qui n'est consciente que chez Benzécri, philosophie que l'on est libre de ne pas partager.