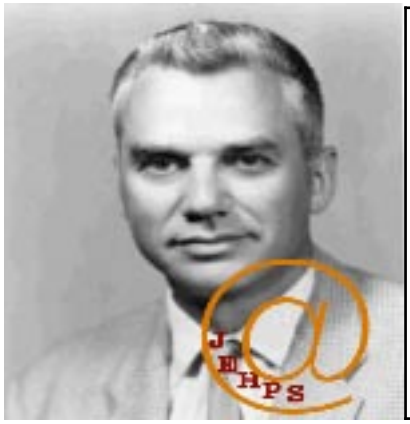

*Journ@l Electronique d'Histoire des Probabilités et de la Statistique*

*Electronic Journ@l for History of Probability and Statistics*

Vol 4, n°2; Décembre/December 2008

**www.jehps.net**

# Analyse des données et sciences humaines : comment cartographier le monde social ?

**Alain DESROSIERES**[1]

L'analyse des données à la française est elle un enfant de mai 1968 ? Diffusée dans les sciences sociales en France autour de 1970, l'analyse des correspondances telle que la promouvaient Jean-Paul Benzécri et ses collaborateurs, a eu, auprès de certains chercheurs, une image « de gauche », par opposition aux techniques économétriques, réputées « de droite ». De nos jours, cette question apparaît saugrenue : les outils statistiques ne semblent pas, par eux-mêmes, « de gauche » ou « de droite ». Comment expliquer cette image, typique de l'humeur qui a suivi 1968 ? Les arguments alors avancés par les tenants d'une analyse des données « de gauche » étaient de deux types : 1) Elle était vue comme une technique purement descriptive, sans théorie économique sous-jacente (c'est-à-dire implicitement : sans la théorie néo-classique), qui vise à dégager, sans *a priori,* des *structures fondamentales* enfouies dans un ensemble opaque de données. 2) Sa *multidimensionalité* apparaissait, dans le sillage de mai 1968, comme un gage de pluralisme, par opposition aux représentations unidimensionnelles réductrices, synonymes de monotonie et de hiérarchie[2]. Les adversaires de ce point de vue faisaient classiquement remarquer que les outils techniques n'ont pas de couleur idéologique ou politique, et que les formalismes mathématiques (diagonalisation de matrices de variance-covariance, recherche des valeurs propres et des vecteurs propres) sont les mêmes pour une analyse des correspondances et pour la résolution d'un modèle à équations simultanées[3].

Les textes sur la statistique écrits par des professionnels de celle-ci sont en général prescriptifs, sinon normatifs. Ils mettent en avant telle ou telle « bonne méthodologie », selon les cas : probabiliste ou non probabiliste (Benzécri), fréquentiste ou bayésienne, inspirée de Fisher ou de Neyman et Pearson dans le cas des tests. Parfois ils plaident pour une « complémentarité » ou une « synthèse » entre des approches vues auparavant comme antagonistes. La perspective retenue ici est plus sociologique qu'épistémologique. Il s'agit

[1] INSEE. alain.desrosieres@insee.fr

[2] En 1968, était paru, aux Editions de Minuit, un des livres cultes de cette année fameuse, celui de Herbert Marcuse, précisément intitulé *L'homme unidimensionnel*, une vigoureuse critique de la société de consommation capitaliste.

[3] Ce texte est un essai à partir de quelques débats qui ont émergé en France, dans les années 1970, lors de la diffusion des méthodes d'analyse des données dans les sciences humaines, notamment en sociologie. Mais il ne constitue en rien une étude complète de cette diffusion et de ses effets, une étude qui reste à faire.

d'identifier les réseaux d'arguments et d'usages sociaux dans lesquels sont convoquées les diverses méthodologies, et de suivre leurs genèses, leurs trajectoires et leurs retraductions.

Le débat ancien évoqué ici peut être une occasion d'examiner la façon dont les sciences humaines se sont, à partir des années 1970, emparées de l'analyse des données dite « à la française » (c'est à dire issue des travaux de Benzécri, Brigitte Cordier-Escoffier, Ludovic Lebart et quelques autres). Une première réponse à la question serait bien sûr de rappeler la « nécessaire neutralité et objectivité de la science, qui ne saurait être annexée par quelque parti que ce soit ». Mais il est aussi possible de réfléchir à la façon dont les formalismes sont intégrés dans des argumentaires *de connaissance et (ou) d'action*, et comment les structures de ces dispositifs varient selon les disciplines (ici par exemple : la psychologie, la sociologie, l'économie ou l'histoire). Nous prendrons tout d'abord pour fil conducteur la dualité de deux formalismes, celui de la *corrélation* et celui de la *régression linéaire*, bien sûr syntaxiquement jumeaux, mais sémantiquement et pragmatiquement utilisés à des fins et dans des argumentaires différents. Nous évoquerons deux cas : celui de la psychologie différentielle, entre 1900 et 1940, puis celui de de la sociologie française, dans les années 1970. Dans ces deux cas, l'analyse des données a permis de dépasser l'image classique d'un espace assimilé à une échelle unidimensionnelle, pour mettre en scène un ou plusieurs « autres axes », conduisant ainsi à des analyses plus subtiles[4].

### *L'idée de multidimensionalité apparaît d'abord en psychologie*

La corrélation et la régression ont été formalisées par Karl Pearson, dans le cadre du modèle de la loi normale à deux dimensions, pour lequel les liens (syntaxiques) entre leurs formules respectives sont particulièrement simples (Benzécri 1982). Mais, selon qu'il s'agit de relier soit les tailles des pères et des fils, soit celles des bras et des jambes d'un même individu, l'interprétation (sémantique) du lien est toute différente. Dans un cas la taille du père « explique » la taille du fils, et la régression linéaire est le formalisme (dissymétrique) qui semble convenir. Ceci suppose un modèle impliquant un *sens* de cette « explication ». En revanche, les tailles des bras et des jambes ne peuvent être reliées que par un formalisme symétrique, celui de la corrélation, puisqu'il ne peut s'agir d'une explication de l'une par l'autre[5]. Dans ce cas, il n'y pas d'hypothèse sur le sens d'une éventuelle explication. C'est là un point de départ du premier des deux arguments, les analyses en termes de corrélation ne postulent pas un sens de causalité éventuelle.

Karl Pearson lui-même, influencé par l'épistémologie empiriste du physicien autrichien Ernst Mach, récusait fermement la notion de causalité, au profit de celle de *co-relation,* c'est-à-dire de *régularités observées de co-occurences,* notamment à partir de *tableaux de contingence* (identiques à ceux de l'analyse des correspondances), additifs en lignes et en colonnes, distribuant et croisant des observations selon des critères divers. La formule de la régression linéaire n'impliquait pas, selon lui, une causalité, mais n'était qu'une façon de présenter une régularité partielle observée, et non pas un déterminisme complet. Ce n'est que plus tard qu'un glissement se produira vers l'idée « d'explication » (avec la distinction entre variables

[4] D'autres sciences sociales, à commencer par l'histoire, notamment l'histoire des sciences et même celle des statistiques, ont recouru à l'analyse des correspondances et aux classifications hiérarchiques. Ainsi, en 1991, Michel Armatte a analysé un tableau de contingence distribuant les nombres de pages consacrées par 56 traités de statistique du XIXème siècle à 21 rubriques thématiques. ( Voir aussi le JEHPS de 2006 : http://www.jehps.net/Decembre2006/Armatte.pdf ).

[5] La sentence usuelle « une corrélation n'implique pas une causalité » n'est pas seulement vraie pour les raisons classiquement avancées, mais son formalisme même ne suggère pas explicitement une idée de causalité, ce qui, en revanche, est le cas pour celui de la régression linéaire.

« explicatives » et « expliquées »), puis vers l'usage peu contrôlé de la notion de causalité, rejetée pourtant par Pearson. La différence de portée sémantique et pragmatique entre les deux outils, corrélation et régression linéaire, est donc postérieure à celui-ci.

L'idée que la corrélation n'implique pas de causalité explique le succès de celle-ci auprès des « psychométriciens » à partir de Charles Spearman (un élève de Karl Pearson), qui étudiait les relations entre les performances scolaires des élèves d'une classe dans différentes disciplines. Il lui semblait exclu d' « expliquer » les notes d'anglais par celles de mathématiques, ou l'inverse. La corrélation s'imposait donc. Elle conduisit à l'« analyse factorielle » des psychologues. Ceux-ci poursuivaient une démarche typique de la métrologie « symptomatique » des sciences de la vie (Benzécri 1982). L'histoire des controverses successives sur les outils de ces premières analyses multidimensionnelles, et sur leurs interprétations, fournit un exemple de lectures inscrites dans des argumentaires idéologiques bien différents (Gould 1983). Les protagonistes en sont les Anglais Charles Spearman (1863-1945) et Cyril Burt (1883-1971), puis l'Américain Louis-Léon Thurstone (1887-1955) [Martin 1997].

L'intelligence générale (ou " facteur *g* ") de Spearman (1904) est une variable latente, " moyenne " des résultats de *n* épreuves scolaires subies par *p* élèves. Elle est déterminée comme l'abscisse de la projection d'un point-individu sur l'axe principal d'inertie du nuage des *p* points représentant les performances des élèves dans l'espace à *n* dimensions des épreuves[6]. La théorie de Spearman est ensuite complétée par Burt, puis critiquée et démolie par Thurstone. Tout en adhérant à la notion d'intelligence générale, Burt cherche, d'une part, à analyser et interpréter la variance non expliquée par le facteur *g* (y a-t-il des facteurs secondaires reflétant des aptitudes spécifiques, indépendantes de *g* ?), et, d'autre part, à démontrer que cette intelligence générale est innée et héréditaire, comme le pensent alors la plupart des biométriciens, issus de l'école eugéniste de Galton, Pearson et Fisher (MacKenzie 1981). Son travail a été utilisé en Angleterre, de 1944 à 1965, pour tester les enfants de 11 ans et les répartir dans des filières scolaires différentes, par le test dit *eleven+*.

Mais l'Américain Thurstone va porter, en 1935, un rude coup à l'idée que l'intelligence est une grandeur unidimensionnelle révélée par le premier axe de l'analyse factorielle. Il imagine un raffinement de celle-ci, en décomposant *g*, par une habile rotation des axes, en sept aptitudes mentales primaires indépendantes les unes des autres, en s'appuyant sur le fait que les tests peuvent être regroupés en sous-ensembles bien corrélés, ce qui a l'avantage de ne pas ordonner les individus sur une échelle unique, et correspond mieux à l'idéal démocratique américain, opposé à la hiérarchie unidimensionnelle rigide de la société anglaise. Ainsi l'idée de multidimensionalité, typique de l'analyse des données, apparaît dans l'analyse factorielle des psychologues[7]. Sans ordinateurs, les psychomètres acquièrent une grande dextérité pour opérer des " rotations d'axes ", dans des espaces à beaucoup de dimensions. Surtout utilisée par les psychologues, cette technique est peu connue des sociologues, du moins en France, jusqu'aux années 1960.

---

[6] L'analyse factorielle de Spearman est formellement différente de ce qui deviendra l'analyse en composantes principales, qui résulte des travaux de Karl Pearson (1901) et Hotelling (1933). Sur ces différences, voir Pagès, Cailliez et Escoufier (1979), qui précisent que « Même si les problèmes posés respectivement par Spearman et Pearson sont différents, l'analyse en composantes principales fournit en général une excellente approximation de ce que l'on recherche en analyse factorielle au sens de Spearman ».

[7] Elle apparaît aussi, à peu près au même moment, en économétrie, notamment dans les travaux de Ragnar Frisch (1934) en termes de « confluence » (Morgan 1990).

Une expérience remarquable resta pourtant isolée et sans suites. En 1954, Jean Porte, le créateur des Catégories socio-professionnelles (CSP), effectue à l'Insee une " enquête par sondage sur l'auditoire radiophonique ", ancêtre de l'audimat. Dans une " analyse factorielle des goûts " préfigurant, vingt-cinq plus tôt, " *La distinction* " de Pierre Bourdieu (1979), il effectue une analyse factorielle d'un tableau des corrélations entre les préférences pour les divers types d'émissions. Il utilise pour cela la méthode de Thurstone dite " centroïde " : « Une telle opération ne peut guère être justifiée que par son succès, c'est-à-dire la possibilité d'interpréter les résultats » (Porte 1954). Il interprète le premier facteur comme opposant les " émissions de qualité " aux " émissions légères ", puis après une rotation d'axes (effectuée graphiquement), un second facteur oppose les " émissions musicales " aux " émissions parlées ". Mais l'analyse porte seulement sur les proximités entre émissions, et non sur leurs préférences par les diverses CSP (analysées par des méthodes plus classiques), et surtout l'analyse factorielle ne conduit pas encore à une cartographie, qui sera le propre de l'analyse en composantes principales et de l'analyse des correspondances.

### *Analyse descriptive et induction, selon Malinvaud*

Ces méthodes se diffusent rapidement dans les sciences humaines et sociales à partir de la fin des années 1960, non seulement à partir des travaux de l'équipe constituée par Benzécri, mais aussi de plusieurs autres chercheurs français[8]. Dans un numéro de 1970 des *Annales de l'INSEE* consacré à l'analyse des données, Malinvaud (le fondateur de cette revue, surtout destinée à publier des études économétriques) distingue, dans son introduction, une statistique *descriptive* et une statistique *inductive* (inférentielle), qui, selon lui, implique de poser, avant l'observation des données à analyser, un modèle s'appliquant au phénomène étudié et résultant de connaissance préalables. Il explique ensuite que :

*« Dans la pratique, les choses ne se présentent pas toujours d'une manière aussi nette. Tantôt les connaissances préalables sur le phénomène apparaissent trop imprécises pour imposer un modèle, si général et peu contraignant que celui-ci soit. Tantôt la nature des inductions auxquelles serviront les données n'est pas circonscrite a priori. Il s'agit plutôt de présenter les données sous une forme intelligible et, au moins dans une première phase exploratoire, de laisser ouverte, aussi largement que possible, la classe des interprétations auxquelles elles se prêteront. La* statistique descriptive*, qui a pour objet la présentation des données, n'est plus alors conçue seulement comme fournissant des valeurs de caractéristiques qui serviront dans certaines procédures inductives bien précisées. Elle vise plutôt à constituer la référence de base pour un examen très souple des résultats observés. Le mouvement observé en faveur de « l'analyse des données » n'est autre qu'un renouveau de faveur pour les techniques de la statistique descriptive ..., dû en grande partie à la diffusion des moyens de calcul informatique »* (Malinvaud 1970, p.8).

Ainsi Malinvaud, ouvrant les colonnes de sa revue à des auteurs favorables à l'analyse des données façon Benzécri (Anita Bensaïd, Jean-Pierre Balladur, Michel Volle), replace celle-ci dans une perspective plus générale, répondant ainsi, à sa façon toute en nuances, aux critiques, évoquées ci-dessus, de la statistique inférentielle supposant un modèle préalable[9]. Il inclut l'analyse des correspondances dans un ensemble plus vaste de techniques, comme le traitement des séries temporelles, l'analyse spectrale, l'élimination des valeurs aberrantes. Il cite

[8] Un panorama historique de ces recherches est présenté par Pagès, Cailliez et Escoufier (1979).

[9] Au même moment, certains de ses disciples économètres, moins libéraux que lui, combattent avec énergie l'éventualité d'organiser des enseignements de l'analyse des données pour les étudiants en économie.

longuement John Tukey, pour justifier la complémentarité entre l'analyse exploratoire et l'induction. En 1983, Jean-Claude Deville et Malinvaud (qui est alors Directeur général de l'Insee depuis 1974), présentent, devant la *Royal Statistical Society de Londres*, des travaux menés à l'Insee, sous le titre « Data Analysis in Official Socio-economic Statistics », en développant ces mêmes thèmes[10].

Dans le débat qui suit, Sir John Boreham, Directeur du *Central Statistical Office*, (l'équivalent britannique de l'Insee), oppose, lui, l'analyse descriptive, non pas à l'analyse *inductive*, mais à *l'analyse prescriptive : «*...qui nous concerne plus, puisque nous sommes un service statistique du gouvernement... et que notre principale tâche est d'abord la production de statistiques, et ensuite de fournir des conseils...», ce qui suggère que, pour lui, la statistique inférentielle est bien orientée vers le conseil et l'expertise en vue de l'action et de la décision. Cet échange entre les responsables des services statistiques des deux pays montre bien qu'il y une distinction entre, d'une part, analyse descriptive et exploration des données sans *a priori,* et d'autre part, « analyse prescripive », induction, modélisation et conseil au prince, même si cette distinction est moins simple qu'il n'y paraît.

Notre hypothèse de départ est que la corrélation et la régression linéaire ont été, *syntaxiquement*, des concepts jumeaux (formalismes d'origine très voisins), mais *sémantiquement* et *pragmatiquement* différents (on ne leur fait ni dire ni faire les mêmes choses). Comme dans la Bible, les jumeaux des années 1890 ont eu pour descendants des outils qui, dans les années 1970, apparaissent à certains comme rivaux. Les formalismes mathématiques sont proches, mais les usages sociaux sont distincts, même si Malinvaud souhaite en montrer la complémentarité. Les uns décrivent des groupes, les autres parlent le langage des variables. Ainsi, une différence entre ces usages est que, dans les lectures et les commentaires des analyses de données, les sujets des verbes sont en général des *groupes sociaux* ou des *branches* de l'économie (des *items* de nomenclatures sociales ou économiques) ou même des individus, tandis que dans les analyses inductives ou économétriques, ces sujets sont plutôt des *variables*, dont certaines « expliquent » d'autres, dans une perspective prescriptive de conseil et d'action, comme le dit explicitement Sir John Boreham, dans son dialogue avec Malinvaud.

### *La sociologie s'empare de l'analyse des correspondances*

Les sciences sociales se sont emparées de l'analyse des données, qui a connu alors une véritable vogue, à partir des années 1970, comme le montrent par exemple deux revues aussi différentes que les *Cahiers de l'analyse des données* (la revue de Benzécri) et *Actes de la recherche en sciences sociales* (la revue de Pierre Bourdieu). Le statisticien Benzécri et le sociologue Bourdieu occupent des positions comparables dans leurs espaces respectifs : prophètes et pionniers innovateurs, critiques, chacun dans son domaine, de l'influence américaine (*via* le quasi-monopole de la langue anglaise et de la statistique inférentielle pour Benzécri, et *via* le structuro-fonctionnalisme puis la *Rational Action Theory* pour Bourdieu). L'analyse des correspondances est apparue à Bourdieu, au début des années 1970, comme particulièrement adaptée à l'étude de la structure des « champs », telle qu'il les avait théorisés dès les années 1960. Benzécri et lui se rencontrent plusieurs fois par l'intermédiaire de Henry

[10] Jean-Claude Deville (1977) avait auparavant développé une variante originale de l'analyse des correspondances, portant sur des variables continues, sous le titre « analyse harmonique du calendrier de constitution des familles ».

Rouanet[11], notamment lors d'une journée sur l'analyse des correspondances, organisée à la Sorbonne en novembre 1978[12].

Mais, à peu près à la même période (milieu des années 1970), l'analyse des données à la façon de Benzécri a aussi été utilisée pour décrire et analyser l'espace social, dans divers contextes : recherche universitaire, Insee, entreprises publiques, instituts de sondages privés, presse grand public. Nous tenterons ici de présenter et de comparer ces divers usages, qui ont l'intérêt de coïncider dans le temps (années 1970) et par leur thème (la description du monde social). On a donc là une quasi-expérience de sociologie de la statistique, puisque trois « variables » sont fixées (méthode, date, thème d'étude).

Les sociologues, mais aussi les bureaux d'études de marché et la presse (notamment les *news magazines* hebdomadaires), ont perçu, dès le début de ces années 1970, l'intérêt de l'analyse des correspondances pour *décrire* et *représenter visuellement* un espace social multidimensionnel, par comparaison avec les austères tableaux croisés. Ces derniers leur semblent à la fois moins synthétiques et surtout moins lisibles que les schémas construits à partir des premiers axes des analyses de correspondance, bien que ceux-ci ne fassent pas apparaître directement des liaisons ou des corrélations entre variables. Une autre façon de faire parler les données se diffuse alors, différente de la démarche explicative, « corrélative » et modélisatrice induite par la statistique tabulaire. Nombre de controverses (et de malentendus) des années 1970 et 1980 résulteront de ce changement de perspective.

Le fait de présenter, dans la même figure, des proximités relatives entre des groupes sociaux (définis de façons variées), et des propriétés attachées (en probabilité) à ceux-ci, propriétés elles mêmes représentées sur la figure, rend ces schémas très suggestifs. On sait les erreurs d'interprétation possibles induites par cette apparente simplicité, due notamment au fait qu'une partie seulement de l'information et de la variance des données de base n'est bien sûr conservée par la projection sur les premiers axes. De plus, la proximité entre un point « groupe social » et un point « propriété » n'est pas directement significative : seule l'est le *système complet des positions relatives.* Mais il s'agit ici de formuler des hypothèses sur les raisons (jugées bonnes ou mauvaises par certains), du succès de cette forme de cartographie sociale dans les années 1970, indépendamment des « erreurs » ou approximations fâcheuses commises parfois. La diffusion de l'informatique et des premiers logiciels de diagonalisation de matrices joue un rôle essentiel dans cette vogue soudaine de l'analyse des correspondances. En effet, il aurait été auparavant impossible de rechercher les vecteurs propres de matrices comportant un très grand nombre de lignes et de colonnes.

L'analyse des données, formulée (ou non) à partir des travaux de Benzécri et de la thèse pionnière de Brigitte Cordier (Madame Escoffier), soutenue en 1965, est importée dans les sciences sociales à la fin des années 1960, notamment par Ludovic Lebart au Credoc (Lebart

[11] Henry Rouanet est décédé en octobre 2008. Il avait, au cours d'une carrière originale, mené des recherches dans des domaines souvent perçus comme distincts (sinon opposés) : psychologie, probabilités, théorie de l'inférence, statistique bayésienne, analyse des données. On trouve de nombreuses informations sur son parcours et son oeuvre à partir du site : http://www.math-info.univ-paris5.fr/~rouanet/

[12] Vingt ans plus tard, en octobre 1998, une nouvelle conférence sur le sujet réunira Bourdieu et Rouanet à Cologne, à l'initiative de Jorg Blasius, coauteur (avec Michael Greenacre) d'un ouvrage collectif, publié en 1994, introduisant l'analyse des correspondances dans le monde anglophone. Depuis 1991, existe un réseau anglophone, le CARME-N, consacré à « l'analyse des correspondances et les méthodes associées » et animé par Blasius et Greenacre (Blasius, Greenacre, Groenen, Velden 2008) Voir : http://www.carme-n.org/

1969), puis par Michel Volle à l'Insee (*Annales de l'INSEE*, 4, 1970). Il peut s'agir de l'*analyse en composantes principales* (Bensaïd 1970, Desrosières 1972), de *classification ascendante ou descendante* (Guibert, Laganier, Volle 1971), de l'*analyse des correspondances simples,* qui porte sur un *tableau de contingence*, distribuant une population selon deux critères additifs (par exemple la catégorie sociale et le département de résidence). Puis l'*analyse des correspondances multiples* permet d'étendre la méthode à l'exploitation d'une enquête par questionnaire, en construisant un *tableau disjonctif complet :* en lignes sont disposés tous les individus répondants, et en colonnes tous les items possibles, dûment codés, de toutes les questions. Les propriétés de l'analyse des correspondances d'un tel tableau, (qui ne comprend que des 1 et des 0, et peut ainsi être de très grande taille), ont été étudiées en détail, notamment par Lebart, Morineau et Tabard (1977). En particulier, il est possible de distinguer des *variables actives,* qui contribuent à la détermination des axes factoriels, et des *variables illustratives,* qui n'y contribuent pas, mais peuvent néanmoins être projetées et représentées sur l'espace de ces axes. Cette possibilité sera abondamment utilisée par les sociologues. Par exemple, des comportements sociaux sont traités en variables actives, et les catégories sociales en variables illustratives, ce qui permet de vérifier et de confirmer la relative généralité de la structure bidimensionnelles de l'espace social, rendue célèbre par Bourdieu (1976 et 1979), mais déjà mise en scène, indépendamment de celui-ci, dans plusieurs autres études antérieures et postérieures

Cet espace est stucturé par un « grand axe », expliquant la majeure partie de la variance des données initiales, opposant les classes populaires aux classes privilégiées. Il correspond à l'échelle sociale unidimensionnelle classique de la sociologie quantitative anglophone. Mais l'analyse des correspondances permet de faire apparaître un deuxième axe, moins classique, orthogonal au premier, qui oppose des catégories scolarisées, urbaines et proches du secteur public, à d'autres, plutôt non salariées, rurales, ou liées au monde des petites entreprises. Il est vrai que la nomenclature française des catégories socioprofessionnelles, proposée par l'Insee dans les années 1950 et utilisée largement, depuis cette époque, par les études empiriques sur la société française, facilite grandement la mise en forme de cette structure et la comparaison entre des études portant sur des thèmes très différents (Desrosières, Thévenot 2002). Avec des variantes, cette configuration de l'espace social apparaît, plus ou moins directement, dès les premières analyses de correspondances, vers 1970.

Ainsi, dans la première version, multigraphiée en 1970, de *L'analyse des données*, Benzécri rapproche deux analyses menées sur Paris, l'une de J.P. Briane sur les résultats parisiens du premier tour des élections présidentielles de 1969, l'autre de J.P. Nakache sur la structure socioprofessionnelle de Paris « vers 1955 ». Pour l'élection, le premier axe oppose, sans surprise, les 7ème, 8$^{ème}$ et 16$^{ème}$ arrondissements, qui votent Pompidou et Poher (le candidat centriste), aux, 11ème, 18$^{ème}$, 19$^{ème}$ et 20$^{ème}$ qui votent Duclos (le candidat communiste). Mais le deuxième axe révèle une opposition moins banale, qui n'est autre que celle, évoquée ci-dessus, qui sera montrée par nombre d'analyses ultérieures des années 1970 : :

*« Sur le deuxième axe nous croyons pouvoir reconnaître des distinctions souvent faites. En politique, c'est Rocard appuyé par une classe moyenne d'intellectuels habitants du 6$^{ème}$ arrondissement, contre Ducatel dont le fief est peuplé d'une autre classe moyenne, faite d'artisans et de petits commerçants qui s'activent entre les anciennes Halles et la Bastille. D'un côté, des quartiers qui sans être proprement résidentiels, ont peu d'ateliers et de négoces ; de l'autre, un pittoresque dédale dont Hausmann n'a pas eu raison, et qui survivra peut-être à Rungis… Sur la carte politique, les arrondissements 5, 6, 13, 14 et 15 sont seuls au dessous du premier axe, avec Rocard et Krivine… »* (Benzécri 1970)

A partir de ce moment, l'analyse des correspondances va être utilisée largement, par trois milieux assez différents, pour produire d'une part des *cartes*, et d'autre part des *typologies* du monde social, les unes complétant les autres. Il est intéressant de comparer les styles, les arguments et les rhétoriques d'usage de ces trois mondes. Le recours à l'analyse des correspondances dans les années 1970 en fournit l'occasion[13]. 1) Des bureaux d'étude de marketing et des intituts de sondage [Cofremca, Aesop/Agoramétrie, Centre de communication avancée(CCA)...], s'en emparent très vite, pour le compte de la presse, de partis politiques ou d'entreprises. Leur style de présentation est à la fois simplificateur et imaginatif. Peu de détails sont fournis sur les méthodes de collecte et de traitement des données. 2) La recherche universitaire y trouve un outil puissant pour mettre en oeuvre et rendre visibles des concepts théoriques, comme par exemple le champ, l'habitus et la structure du capital pour la sociologie de Bourdieu. 3) Après les travaux pionniers du Credoc sur « La morphologie sociale des communes urbaines » (Lebart, Tabard 1971), la statistique publique et des centres qui lui sont proches (Insee, Ined, Credoc) l'utilisent pour explorer les fichiers de leurs enquêtes, dans un style plus technique et austère. Dans ce cas, les sources et les méthodologies mathématiques sont plus clairement explicitées que dans les deux autres mondes.

### *L'ordinateur confirme l'Evangile*

Trois années de suite, de 1973 à 1975, l'hebdomadaire *Le Nouvel Observateur* publie, sous le titre « Le prix d'un Français », plusieurs longues études appuyées par des analyses des correspondances, portant sur une liste détaillée de professions, dont sont comparés les revenus, les handicaps et privilèges, et des trajectoires sociales fréquentes. Des cartes de grand format sont reproduites. La méthode y est ainsi présentée, en septembre 1973 :

*« Les trois cartes s'appellent, en langage de statisticien, une « analyse de données ». Elles représentent trois aspects différents d'une même image : celle d'un immense volume à 18 dimensions où se meuvent, comme autant de galaxies dans l'espace, les 150 professions étudiées. La position de chacune d'elles est un compromis complexe entre les formes plus ou moins attractives des 18 critères que nous avons choisispour dessiner le « profil » de chacun des métiers : routine, sécurité de l'emploi, temps libre, non-pénibilité physique, diplômes, intérêt du travail, perspectives d'avenir, liberté dans le travail, rémunération-salaire, avantages en nature, concurence, expérience, esprit d'entreprise, apport financier, évasion fiscale. Ainsi le point « instituteur », beaucoup plus attiré par le pôle « sécurité » que par le pôle « évasion fiscale » ou le pôle « concurrence », se situera en bas à gauche, alors que le grand notaire parisien, qui cumule les avantages financiers et les statisfactions de prestige, ira se fixer entre la considération et l'argent.*

*Ce volume immense, redessiné à plat par l'ordinateur grâce à des calculs savants mais qui conservent au mieux les disparités observées entre les professions, dessine donc une carte en forme de planisphère où des soleils entraîneraient autour d'eux des système de planètes. Pourquoi ces soleils ont-ils tous été « projetés » par l'ordinateur sur la droite de la carte ? Ce n'est pas l'effet du hasard, mais par ce que les économistes appellent « l'effet de Mathieu », l'Evangéliste qui disait : « A celui qui a tout, tout sera donné de surcroît. A celui qui a peu de chose, tout sera retiré ! » Vingt siècles après, l'ordinateur confirme l'Evangile : en situant dans*

[13] Nous ne cherchons pas ici à décrire ni la diffusion d'autres importantes méthodes d'analyse des données, comme les analyses classificatoires (analyses arborescentes), ni même les nombreux autres usages de l'analyse des correspondances. Nous nous centrons sur la progressive mise en forme de l'espace social à deux dimensions, dont la première est l'échelle sociale classique, et dont la seconde oppose les catégories plutôt diplômées à celles disposant d'un capital économique.

*la zone droite la plupart des avantages matériels et moraux, il prouve ce qu'on savait déjà : dans notre monde injuste, les avantages se cumulent et les inconvénients, hélas, s'additionnent ! »*

(Josette Alia, *Le Nouvel Observateur*, 24 septembre 1973, p.52).

Cette idée de « cumul des avantages », certes exacte et reflétée par le premier axe qui « explique » la majeure partie de la variance, peut être comparée au « facteur *g* » de Spearman en psychologie. Les analyses ultérieures, permises par des méthodologies multidimensionnelles; déploient des espaces plus complexes. De fait, en septembre1974, l'hebdomadaire reprend sa série « Le prix d'un français », et combine analyse des correspondances et classification pour présenter, d'une part, une carte faisant apparaître clairement les deux axes, et d'autre part, sur cette carte, des regroupements de catégories sociales en « cinq France », décrites dans l'article selon leurs positions relatives. Détail intéressant, la mise en page a changé. Alors que les analyses de 1973 étaient, classiquement, présentées avec un premier axe horizontal, et le second vertical, en revanche, en 1974, apparaît une version sociologiquement plus parlante. Le premier axe, correspondant à « l'échelle sociale », est *vertical*, le second, opposant, à gauche, les catégories diplômées et proches du secteur public, et, à droite, celles plutôt à capital économique, est *horizontal*. Un dessin de Wiaz résume cette façon de voir, avec deux bonshommes grimpant à un arbre, l'un à gauche du tronc, en se hissant sur des diplômes, et l'autre (coiffé d'un chapeau haut de forme) à droite, gravissant des tas d'argent. L'espace social est ainsi exprimé, en termes journalistiques, à travers le langage de cette cartographie statistique :

*« L'ordinateur a fabriqué un nuage qui, une fois projeté sur le papier, apparaît comme étiré le long d'un axe vertical. Cet axe qui est, pour parler comme les statisticiens, notre premier « axe factoriel », est « dominé » par le revenu. En bas, ceux qui gagnent le moins. En haut, ceux qui gagnent le plus. A droite, ceux qui doivent leur profession à un patrimoine mais qui n'ont pas reçu une formation très poussée : les « indépendants », la filière capitaliste. A gauche, au contraire, ceux qui occupent un emploi salarié et qui progressent grâce à leurs diplômes. En tout cinq groupes bien typés. De bas en haut : les exploités, les petits capitalistes, les classes moyennes salariées, le capitalisme traditionnel. Et, tout en haut, ce que l'on dénomme la « technostructure », le « gratin » intellectuel. ».*

(François-Henri de Virieu, *Le Nouvel Observateur*, 16 septembre 1974, p. 65.)

En 1975, l'étude est prolongée. Entre temps, la méthode d' « analyse factorielle » utilisée par l'hebdomadaire a été critiquée. Le postulat d'indépendance et de neutralité par rapport à tout modèle *a priori* est remis en cause, ne serait ce qu'à travers la nomenclature socioprofessionnelle utilisée, qui constitue un modèle parmi d'autres du monde social. Ainsi le journal précise sa position par rapport à l'outil statistique :

*« L'analyse factorielle a l'avantage de permettre le traitement simultané d'un grand nombre d'informations. Il est possible d'utiliser un faisceau de raisonnements et d'éviter l'inconvénient majeur des « modèles » mathématiques classiques qui obligent les utilisateurs à des simplifications abusives d'une réalité socio-économique complexe par définition. Mais, en dépit de ses avantages, elle reste une technique imparfaite. Ce n'est pas un outil neutre car elle est fondée, entre autres, sur trois* a priori *:*
*1) La « métrique » utilisée est celle du « Chi 2 », c'est-à-dire que les distances entre deux points résultent de comparaisons en pourcentage. Tout devient comparable mais on fait l'impasse sur la notion de taille [....].*

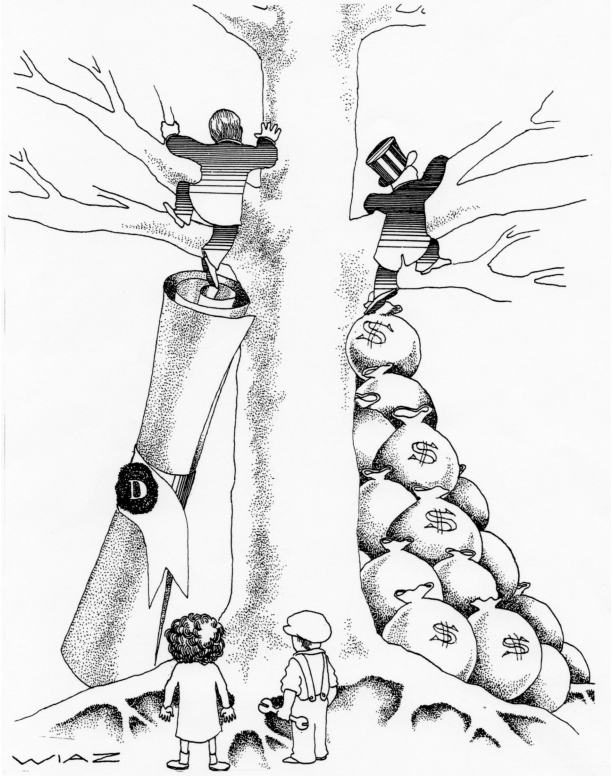


**Figure 1 : L'espace social à deux dimensions vu par le dessinateur Wiaz en 1974.**
(« Le Nouvel Observateur », *« Le prix d'un Français »,* 16 septembre 1974, p. 65)

*2) Elle ne permet pas de hiérarchiser les « variables ». Le fait d'être au chômage est une « information » qui ne pèse pas plus lourd dans l'analyse que le fait d'aller moins souvent au cinéma [...].*
*3) Elle utilise des mathématiques fondées sur le principe de non-contradiction : on est plus loin, plus plus haut, plus à gauche, plus à droite, mais on n'est jamais autre chose. Ceci évacue la contradiction, le conflit, base du jeu social.*

(Anita Duhameau[14], *Le Nouvel Observateur*, 22 septembre 1975, p. 61).

### *Penser de façon relationnelle*

Dans un tout autre monde, et sans doute indépendamment, la sociologie universitaire découvre l'intérêt de l'analyse des correspondances, vers 1975. Ainsi, Bourdieu et les chercheurs du Centre de sociologie européenne, attelés à la tâche de décrire le système complexe des relations entre les classes sociales, leurs goûts et leurs pratiques, y trouvent un puissant outil pour rendre visibles ces relations. En effet, la théorie des champs décrit les acteurs sociaux du point de vue de leurs positions relatives les uns par rapport aux autres, dans des « espaces » virtuels de relations, chaque acteur étant caractérisé par son « habitus », des propriétés acquises et incorporées, qui orientent (en probabilité) les jugements et les comportements. L'analyse des correspondances construit des espaces organisés du point de vue des relations entre des pratiques statistiquement enregistrées dans des enquêtes, et, plus précisément, des *homologies* entre espace des groupes sociaux et espaces des pratiques. « Il faut penser de façon relationnelle », tel est le mot d'ordre inspiré du Cassirer de la *Philosophie des formes symboliques*, et du Panofsky de *Architecture gothique et pensée scolastique*, dont Bourdieu a retenu le concept d'habitus, dans sa préface à l'édition française de cet ouvrage. Il exprime plusieurs fois son intérêt pour la méthode de Benzécri :

*« J'utilise beaucoup l'analyse des correspondances, parce que je pense qu'elle est essentiellement une procédure relationnelle dont la philosophie exprime pleinement ce qui, selon moi, consitue la réalité sociale. Cette procédure « pense » en relations, comme j'essaie de le faire avec le concept de champ »* (Bourdieu, préface à l'édition allemande du *Métier de sociologue*, 1991, cité par Henry Rouanet sur son site).

Ailleurs, il récuse l'opposition entre « méthode purement descriptive » et « analyse de régression », dans la mesure où, selon lui, la méthode « explique » les stratégies des agents, (en donnant ainsi au verbe « expliquer » un sens sans doute assez différent de celui des économètres, qui isolent des « variables » associées à des fins et des moyens, dans des modèles visant, directement ou non, à des actions sur le monde) :

*« Ceux qui connaissent les principes de l'analyse des correspondances multiples saisiront l'affinité entre cette méthode d'analyse mathématique et la pensée en termes de champ. Ayant pris en compte l'ensemble des agents efficients (individus et, à travers eux, institutions) et l'ensemble des propriétés - ou des atouts - qui sont au principe de leur action, on peut attendre de l'analyse des correspondances, qui,* ainsi utilisée, n'a rien de la méthode purement descriptive que veulent y voir ceux qui l'opposent à l'analyse de régression*, qu'elle porte au jour la structure des oppositions, ou, ce qui revient au même, la structure de la distribution des pouvoirs et des intérêts spécifiques qui détermine, et* explique*, les stratégies des agents »* (Bourdieu, *Les structures sociales de l'économie*, 2000, p. 128, cité par Henry Rouanet).

[14] Pseudonyme de Anita Bensaïd, alors membre de l'Insee, et une des auteurs du numéro 4 des *Annales de l'INSEE* de 1970, préfacé par Malinvaud et consacré à l'analyse des données.

De fait, l'analyse des correspondances devient, à partir de 1975, une des images de marque de la méthodologie du CSE, vingt ans avant que la question des relations entre analyse des données et méthodes de régression ne soit abordée par des chercheurs de celui-ci, notamment sous l'influence de Michel Gollac et Henry Rouanet[15]. Le schéma de l'espace social à deux dimensions est présenté en 1976 dans un numéro de *Actes de la recherche en sciences sociales* intitulé *Anatomie du goût*, puis développé en 1979, dans le gros volume *La distinction.* Les axes sont nommés : le premier (vertical) exprime le *capital total*, tandis que le second (horizontal) oppose les catégories à *capital culturel* (à gauche) à celles à *capital économique* (à droite). Des transparents sont superposés à la carte des catégories sociales. Ils représentent les divers goûts et pratiques et culturelles. Il est bien précisé que l'important n'est pas la proximité entre deux *points* de chacun des deux schémas superposés, mais plutôt l'homologie des *structures relatives* de l'ensemble des deux figures[16].

Ceci est fait, d'une part, pour l'ensemble de l'espace social (pp. 140-141 de *La distinction*), et d'autre part, séparément, pour les « variantes du goût dominant » (p. 296), et les « variantes du goût petit-bourgeois » (p. 392). Mais, paradoxalement, dans le cas de l'espace social complet, Bourdieu précise qu'il n'a pas été possible de figurer une *vraie* analyse des correspondances. Il s'en explique ainsi :

« *Bien qu'ils en aient certaines apparences et qu'on se soit aidé de différentes analyses de correspondances pour le construire, bien que nombre d'analyses de correspondances aient produit des espaces qui s'organisent selon la même structure..., les schémas présentés ici ne sont* pas des diagrammes plan d'analyse des correspondances » (Bourdieu 1979, note p.139).

En revanche, pour les deux schémas correspondants aux « goûts dominants » et aux « goûts petit-bourgeois », il s'agit bien de réelles analyses de correspondances. Dans ces deux cas, les catégories sociales ne sont pas représentées par des points, mais par des polygones de formes variées, rappelant ainsi le fait qu'à chaque catégorie correspond un *nuage de point*, c'est-à-dire une distribution statistique, dont le centre de gravité représente au mieux la position moyenne, mais non la dispersion. Cette visualisation est une des possibilités utiles offertes par les représentation cartographique de l'analyse des correspondances. Par ailleurs, comme par un jeu de fractales emboîtées, il est possible de retrouver à l'identique la structure de l'espace complet en « zoomant » une zone (les cadres : Boltanski 1982, pp. 392-393, les artisans : Zarca 1979, pp. 10-11) ou une micro-zone (les grand patrons, les évêques, les philosophes, les économistes). Un des intérêts de l'analyse des correspondances multiples est de permettre de figurer explicitement la position d'individus pour lesquels des informations sont publiques, comme c'est le cas pour les membres de certains microcosmes sociaux.

Les exemples fournis ici d'usages de l'analyse des données *par la sociologie universitaire* semblent pour l'essentiel issus du groupe de Bourdieu, et d'une seule problématique, celle de sa théorie des champs. De fait, ce fut le cas dans ce monde (mais non dans d'autres, on l'a vu), dans la période des années 1970. On peut citer un autre cas, très innovant, celui d'un membre de ce groupe (qui s'en éloignera ensuite), Luc Boltanski (1984). Dans une étude intitulée *La*

[15] Sur différents aspects de la confrontation entre ces deux familles de méthodes, voir Schiltz 1990, Nétumières 1997, Desrosières 2001, Rouanet, Lebaron, Le Hay, Ackermann et Le Roux 2002, et Rouanet et Lebaron 2006.

[16] Dans *La distinction*, les transparents sont remplacés un autre système : les intitulés des catégories sociales sont imprimés en noir, ceux des goûts en rouge. Bourdieu précise (p. 139) « qu'il faudrait introduire un troisième schéma, présentant l'espace théorique des habitus, c'est-à-dire des formules génératrices (par exemple, pour les professeurs, l'ascétisme aristocratique) qui sont au principe de chacune des classes de pratiques et de propriétés...»

*dénonciation*, il analyse un corpus de lettres de plaintes reçues par le journal *Le Monde.* Il se pose la question du *jugement de normalité* porté sur ces lettres par le journal. Pour cela, il les fait lire par un large public, à qui il demande de noter le degré de « normalité » ou d'éventuelle « folie » de ces lettres. Ce matériel est ensuite soumis à une analyse des correspondances multiples. Une notion s'en dégage, celle de *grandeur*, qui sera ensuite à l'origine d'une *sociologie des affaires*, et du courant de la *sociologie pragmatique*.

### *Budgets de famille, mariages, mensurations : d'autres lectures de l'espace social*

Le fait que, avant et après les analyses de champs sociaux menées par Bourdieu, la structure bidimensionnelle de l'espace social soit apparue dans d'autres contextes très différents, montre que son petit coup d'audace (la publication d'une *apparence* d'analyse des correspondances, tout en le reconnaissant), n'était pas complètement injustifié. Ceci peut être vu comme une hypothèse, confirmée ensuite (ou du moins non falsifiée au sens de la méthodologie poppérienne). Parmi ces validations, on peut en mentionner trois : 1) l'analyse des *budgets des ménages*, 2) celle des proximités sociales reflétées par le *choix du conjoint*, et 3) les *mensurations* du corps humain (la taille et le poids).

Ainsi l'analyse des correspondances des résultats de l'enquête « budgets de familles » de l'INSEE de 1979 reproduit de façon fidèle la structure des positions relatives des catégories sociales déjà décrite (Glaude et Moutardier 1982). L'analyse des correspondances multiples (ACM) porte sur un tableau croisant, en lignes, chaque répondant, et en colonnes, d'une part, leurs postes de dépenses traités en variables actives de l'analyse, et d'autre part, les variables socio-démographiques (catégorie socioprofessionnelle, revenu, taille de l'unité urbaine, âge de la personne de référence...), projetées en varaiables illustratives. Elle complète l'interprétation du deuxième axe de la carte désormais familière, axe bien corrélé avec la taille de l'unité urbaine de résidence, et avec l'âge. Trois catégories d'unités urbaines (plus de 100 0000 habitants, moins de 100 000 habitants, communes rurales) sont alignées le long de cet axe, de la gauche vers la droite. Les quatre catégories d'âge, des plus jeunes aux plus âgés, sont elles aussi rangées dans cet ordre, de gauche à droite.

Le choix du conjoint est un bon indicateur des proximités entre milieux sociaux (Girard 1964). Ceci pouvait être utilisé pour construire une carte de l'espace social, en un temps où la majorité des couples se mariait, et où l'Etat civil enregistrait quatre informations sur les professions : celles des deux conjoints et celles de leurs deux pères. Les professions des mariés sont peu utilisables, car peu fixés ou même inexistantes à ce moment. En revanche, celles de leurs pères sont très significatives de leurs milieux sociaux. Une analyse des correspondances simples a été faite du tableau de contingence (21X21) croisant les 21 catégories socioprofessionnelles des pères des conjoints mariés en 1972 (Desrosières 1978). Ce tableau est fortement diagonal : l'homogamie sociale est bien sûr importante. Mais l'analyse révèle les proximités entre catégories, et permet de reconstituer l'espace social à deux dimensions déjà présenté[17].

[17] L'analyse des correspondances des tableaux fortement diagonaux (comme ceux de mobilité sociale), présente une particularité. Le nuage des points projetés sur le plan factoriel des deux premiers axes a la forme d'un croissant (effet Gutmann), le deuxième axe opposant les catégories médianes aux catégories extrêmes. En revanche, le plan factoriel formé par les axes 1 et 3 reproduit fidèlement l'espace social.

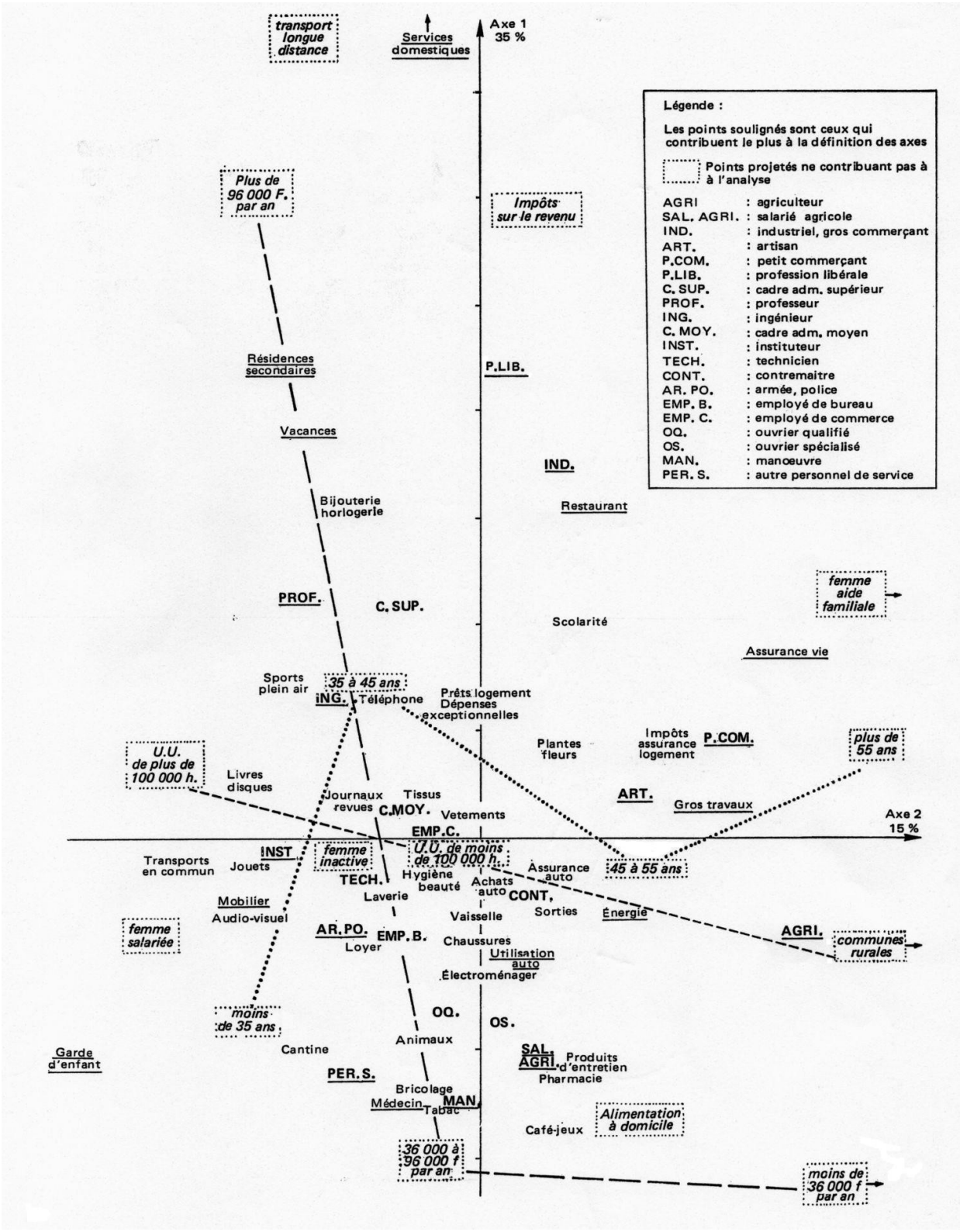


**Figure 2 : Exemple d'analyse des correspondances : « Les budgets des ménages ».**
(M. Glaude et M. Moutardier « *Economie et Statistique*, 160, p.30, Janvier 1982 »)

La taille humaine a joué de Quetelet à Galton et Pearson, on le sait, un rôle éminent dans l'histoire de la statistique Or une façon simple de construire l'espace social est d'utiliser les mensurations moyennes des catégories sociales (Charraud 1981). En portant en abscisse le poids moyen et en ordonnée la taille moyenne, le nuage de points obtenu est très semblable à ceux des analyses précédentes (et cela ne nécessite aucune analyse de correspondances). La taille d'un individu adulte varie très peu au cours de sa vie, mais, au fil des générations, la taille moyenne augmente, et celle des classes privilégiées est supérieure à celle des classes populaires. En revanche, le poids dépend fortement des comportements alimentaires et des exercices physiques, socialement très marqués. Le fait qu'une opposition nette apparaisse entre, d'une part, les catégories diplômées et urbaines, et d'autre part celles des petits patrons, montre l'intérêt sociologique de ce « deuxième axe » méconnu d'une sociologie empirique qui réduit souvent l'espace social à une échelle, elle-même synonyme du revenu.

### *Les comportements électoraux : d'une bissectrice à l'autre*

Un des schémas publiés par le *Nouvel Observateur* en 1975 faisait intervenir une autre variable, obtenue à partir de sondages : le vote aux élections législatives de 1973. En suivant de haut en bas le premier axe, on y trouve cinq partis. Les Républicains indépendants (le parti de Valéry Giscard d'Estaing) sont en haut du graphique, dans la zone des professions libérales et des cadres supérieurs. Plus bas, sur la même verticale et proche du centre de la figure, se trouve le Centre (les démocrates chrétiens, le parti de Jean Lecanuet). L'UDR (le parti gaulliste) est au même niveau que le Centre, mais plus à droite, près du pôle « non salariés » du deuxième axe. Le Parti socialiste (celui de François Mitterrand) est plus bas et nettement à gauche, du côté des cadres moyens salariés, tandis que le Parti communiste est encore plus bas, dans la zone des ouvriers.

Cette configuration est bien sûr conforme à ce que l'on sait de la sociologie électorale. Cependant, la présentation bidimensionnelle permet d'affiner l'analyse. En gros, les deux grands ensembles politiques de gauche et de droite (qui ont voté, presque pour moitié, pour Mitterrand et Giscard d'Estaing en 1974), sont partagés, sur le schéma, non pas par une ligne horizontale, mais par la « deuxième bissectrice », (N.O. - S.E.) : les électeurs de Giscard d'Estaing sont plutôt (en probabilité) les classes supérieures et les non-salariés (commerçants, artisans et paysans), ceux de Mitterrand les ouvriers, employés et les classes moyennes salariées, notamment les enseignants. Les partis d'extrême-gauche n'étaient pas représentés sur la figure, et le Front national n'existait pas encore. A partir de 1985, ce dernier, devenu électoralement significatif, compliquera ce schéma trop simple, en se trouvant sans doute socialement plutôt dans la quadrant Sud-Est de la figure, dans la zone des classes moyennes et populaires non salariées, plutôt rurales ou dans des petites villes. Mais à ce moment, la presse et les instituts de sondage n'utilisaient plus ce type de représentation.

Cette façon de représenter l'espace social aurait pu, vingt ans ou trente ans plus tard, permettre de présenter de façon éclairante les votes lors des deux référendums sur l'Europe, celui de 1992 sur le Traité de Maastricht, et celui de 2005 sur le projet de Constitution européenne. Ces votes ont dérouté les commentateurs habitués à la sociologie électorale précédente, parce que les divers groupes sociaux ne se sont pas distribués de la même façon que lors des élections opposant la gauche et la droite. Dans ce cas, c'est la « première bissectrice » (N.E. - S.O.), *perpendiculaire à la précédente*, qui a distingué (en probabilité) les partisans du *oui* et du *non*. Les catégories supérieures, les diplômé urbains, les salariés du public, ont, dans les deux cas, plus voté *oui* que les catégories populaires, les petits patrons et les ruraux. Ceci ressort des nombreux sondages qui ont été effectués en 1992 et en 2005.

### *Opinions et styles de vie*

Dès les années 1970 et 1980, les techniques d'analyse des données ont été utilisées dans un monde très différent de ceux de la recherche universitaire et de la statistique publique, pour tenter de mettre en forme des cartographies des opinions et des styles de vie. Cela a été le fait de bureaux d'études privés, travaillant pour le compte de grandes entreprises. On peut citer le groupe AESOP/Agoramétrie de Jean-Pierre Pagès, le Centre de communication avancée (CCA) de Bernard Cathelat, et la Cofremca de Alain de Vulpian[18]. En construisant des espaces d'opinions ou des typologies de styles de vie, ces organismes présentent des cartes qui ont des rapports avec les précédentes, mais ils n'utilisent pas systématiquement le critère socioprofessionnel, ou même le rejettent explicitement comme non pertinent et « dépassé » (cas du CCA). Il n'est donc pas facile de les relier à l'ensemble des travaux convergents présentés ci-dessus. Cependant, ils ont beaucoup contribué à la notoriété de l'analyse des données dans le monde des affaires et de la publicité. Nous évoquerons les deux cas de AESOP/Agoramétrie et du CCA.

Vers la fin des années 1970, l'entreprise publique Electricité de France (EDF) s'inquiète de la montée de la contestation anti-nucléaire. En ce temps, elle perçoit cette question comme une affaire de « perception » et d'opinion irrationnelle d'un public mal informé. Mais, pour analyser et comprendre cette inquiétude, il lui semble nécessaire de l'intégrer dans une étude plus large des conflits d'opinions qui parcourent la société française. Disposant déjà d'une importante tradition de recherche en sciences sociales ( Meynaud 1996), elle se tourne vers l'Association pour l'Etude des Structures de l'Opinion Publique (AESOP). Celle-ci réunit les membres de grandes administrations ou d'entreprises pour des recherches en commun sur la communication. Elle est présidée par Bernard Cazes du Commissariat au Plan[19]. Elle a effectué, entre 1977 et 1984, sept enquêtes nationales d'opinion, portant sur une cinquantaine de « questions référendums autour de conflits d'actualité », portant sur des « phénomènes de société » (comme la religion, l'avortement, la publicité, le nucléaire, l'énergie solaire, la justice, la censure, l'homosexualité, les syndicats, la violence, la famille, les grèves, la télématique...). Les interviewés doivent se situer sur une échelle à cinq paliers.

Une analyse factorielle de l'enquête de mars 1981 produit deux espaces superposés, intitulés, l'un, « le ciel », celui d'un riche plan factoriel des opinions, et l'autre, « la terre », où figurent des caractéristiques socio-démographiques (Morlat et Pagès 1984). Il est intéressant de comparer cette analyse avec les précédentes, bien que tout y soit différent. Ni l'analyse des correspondances de Benzécri, ni la théorie des classes sociales et de leurs goûts de Bourdieu ne sont mentionnées. La méthode utilisée est l'analyse factorielle de Spearman, et la théorie des mythes de Lévi-Strauss est longuement citée. La carte des opinions présentées est très riche. Celle des caractéristiques sociales ressemble à l'espace social déjà présenté, avec même une inversion de l'ordre des deux premiers axes. Le premier oppose des urbains, diplômés, jeunes, sans religion, votant à gauche, à des ruraux âgés, catholiques, agriculteurs ou commerçants, votant à droite. Le deuxième axe, apparemment moins explicatif, oppose les cadres supérieurs aux ouvriers spécialisés. Ceci confirme que, pour l'analyse des phénomènes de société, le « deuxième axe » au sens des analyses précédentes (et devenu ici le premier), est particulièrement efficace.

---

[18] La Cofremca avait participé en 1975 aux enquêtes du *Nouvel Observateur* décrites ci-dessus.

[19] AESOP sera remplacée ultérieurement par Agoramétrie, un organisme qui poursuit cette tradition d'étude de l'opinon à travers des analyses factorielles.

Le cas du CCA de Bernard Cathelat est plus déconcertant. Celui-ci parvient à populariser des analyses de « styles de vie », à partir d'analyses de données portant sur des résultats d'enquêtes. Mais il est impossible d'avoir des informations, sur les questionnaires comme sur les méhodes d'analyse produisant ces typologies, car elles relèvent du « secret des affaires ». Des catégories *ad hoc,* aux intitulés psychologisants accrocheurs, sont habilement médiatisées, et prétendent remplacer heureusement les nomenclatures de l'INSEE, jugées obsolètes. De gauche à droite sur un axe opposant l' « aventurisme » au « recentrage », on trouve : profiteurs, libertaires, dilettantes, frimeurs, militants, entreprenants, attentistes, responsables, exemplaires, vigiles, moralisateurs, utilitaristes, conservateurs. Ces études de marché soulèvent de vives controverses dans le monde de la gestion et de la publicité, mais sont ignorées par la recherche en sciences sociales et par la statistique publique, puisqu'il est impossible de les articuler avec d'autres résultats. En revanche, elles suscitent de l'intérêt dans certains milieux de l'administration, comme le Commissariat au Plan, où Cathelat est invité à parler par Bernard Cazes (Georgakakis 1997). Ainsi la logique de la concurrence entre organismes privés conduit à un émiettement des productions, et rend impossible toute tentative d'éventuel cumul des résultats de ces analyses de données.

### *Des éthiques scientifiques distinctes ?*

Un statisticien américain, sollicité pour participer à ce numéro du JEHPS sur l'histoire de l'analyse des données, eut une réaction intéressante. Assimilant l'analyse des des données à la française à l' « exploratory data analysis », il ajouta,: « L' aspect de celle-ci qui m'a intéressé est la difficulté de la concilier avec l'éthique [qui implique] de poser l'hypothèse AVANT de faire l'expérience ou l'enquête et de collecter les données ». Il est possible que cette façon de voir résulte de la tension, évoquée ci-dessus, entre, d'une part, les techniques statistiques descriptives et exploratoires, et d'autre part, la statistique inférentielle inductive, dont Malinvaud soulignait la complémentarité en 1970. Ce statisticien avait peut-être en tête les normes de recherche et de publication dans les revues scientifiques, et le contexte *publish or perish* des universités américaines. Ces normes sont fondées sur la série : hypothèse théorique, expérimentation ou enquête construites sur la base de cette hypothèse, tests pour confirmer (ou falsifier) l'hypothèse, puis, si les tests sont significatifs, publication, et enfin (point de vue plus sociologique qu'épistémologique) amélioration de la position du chercheur sur le marché des postes académiques. Dans ce cadre, les méthodes de type *data-mining*, qui consistent à rechercher mécaniquement les meilleures corrélations et les régressions les plus significatives, et à, *ensuite*, formuler des hypothèses théoriques miraculeusement confirmées par ces relations soigneusement sélectionnées, apparaissent bien sûr malhonnêtes. Ce biais, dit «de Lovell » par les économètres, minerait ainsi la production scientifique honnête (Armatte et Desrosières 2000).

Peut-être l'analyse exploratoire des données à la française est elle perçue, dans le contexte anglo-américain, à travers la grille de cette logique *theory drived,* associée à l'épistémologie dominante de la falsification. Pourtant, dans l'exemple développé ici, la progressive confirmation de la structure bidimensionnelle de l'espace social, au cours des années 1970, par cumul d'analyses des correspondances portant sur des données très variées, n'était pas, à l'origine, suggérée par une construction théorique. L'analyse exploratoire a montré sa relative stabilité. Il est vrai aussi que, à partir des années 1980, cette façon d'explorer et de décrire les données a été peu à peu supplantées, en sociologie quantitative, par les méthodes de régression logistique (modèles Logit), plus fondées sur l'idée d' « effet pur des variables » que sur celle de

co-occurence des propriétés des agents et des groupes sociaux[20]. La sociologie des *usages sociaux respectifs* des diverses techniques statistiques, exploratoires et inductives, reste à faire, au delà de leurs comparaisons proprement épistémologiques et méthodologiques, familières aux statisticiens.

De la diversité de ces usages, nous venons de voir maints exemples. Peut on en déduire que, comme certains le pensaient vers 1970, l'analyse des données serait plus « à gauche » que des méthodes hypothético-déductives telles que l'économétrie ? De façon anecdotique, on peut remarquer que Benzécri lui-même se réclame d'un catholicisme traditionaliste. Plus sérieusement, Malinvaud rappelle que, dans les années 1940 et 1950, l'économétrie avait une image « de gauche », liée à ses premiers usages pour la planification (Frisch, Tinbergen) ou pour les politiques d'intervention et la modélisation keynésienne (Lawrence Klein). De son côté, par un effet de balancier, l'économétrie nouvelle des séries temporelles (Sims) a développé des techniques fondées moins sur des modélisations a priori que sur l'exploration des autocorrélations internes à ces séries (Renault 1999). Ceci montre que les associations éventuelles d'une technique à une option politique sont largement historiques. Cependant, les débats autour des deux arguments avancés vers 1970, sur une interprétation politique de l'analyse des données (primat de l'exploration et de la description, multidimensionalité) ont été présents tout au long de ce récit très partiel de ses usages dans les années 1970 et 1980. La tension entre description et induction est au coeur de l'histoire de l'utilisation des statistiques dans les sciences sociales.

## BIBLIOGRAPHIE


A.E.S.O.P., 1985 : *Les structures de l'opinion publique en 1984,* Agoramétrie, Fontenay aux Roses.

ARMATTE M., 1991 : «Une discipline dans tous ses états : la Statistique à travers ses traités (1800-1914)», *Revue de synthèse*, IVème série, n°2, avril-juin, pp. 161-206.

ARMATTE M, 2006 : « Les images de la statistique à travers ses traités », *Journal Electronique d'Histoire des Probabilités et de la Statistique*, Vol;2, N°2, Décembre, 2006, http://www.jehps.net/Decembre2006/Armatte.pdf

ARMATTE M., DESROSIERES A., 2000 : « Méthodes mathématiques et statistiques en économie : nouvelles questions sur d'anciennes querelles », in J.P. Beaud et J.G. Prévost (éds) *L'ère du chiffre. Systèmes statistiques et traditions nationales*, Presses de l'Université du Québec, Montréal, pp; 431-481.

BENSAÏD A., 1970 : « Niveau nutritionnel des populations gabonaises. Une analyse en composantes prinicipales », *Annales de l'INSEE*, 4, pp. 11-45.

BENZECRI J.P (éd.)., 1970 : *L'analyse des données* (multigraphié). Recueil de textes divers.

BENZECRI J.P., 1982 : *Histoire et préhistoire de l'analyse des données*, Dunod, Paris.

BLASIUS J., GREENACRE M., GROENEN P, VELDEN M. van de, 2008 : « CARME-N, Correspondence Analysis and Related Methods Network », *Bulletin de Méthodologie Sociologique*, 99, pp. 73-81. Voir : http://www.carme-n.org/

BOLTANSKI L., 1982 : *Les cadres. La formation d'un groupe social*, Editions de Minuit, Paris.

BOLTANSKI L., avec Y. DARRE et M.A. SCHILTZ, 1984 : « La dénonciation », *Actes de la recherche en sciences sociales*, 51, pp. 3-40.

BOURDIEU P., avec M. de SAINT MARTIN, 1976 : « Anatomie du goût », *Actes de la recherche en sciences sociales*, 5, pp. 2-81.


[20] De fait, la cartographie sociale ainsi produite a été peu mentionnée et utilisée à partir de ces années 1980, ce qui coïncide avec la relative éclipse de la notion de classe sociale, dans une partie de la sociologie, comme dans les débats politiques.

BOURDIEU P., 1979 : *La distinction. Critique sociale du jugement* ; Editions de Minuit, Paris.
BOURDIEU P., 2000 : *Les structures sociales de l'économie*, Seuil, Paris.
CASSIRER E., 1972 : *Philosophie des formes symboliques* (3 volumes) ; Editions de Minuit, Paris.
CHARRAUD A., 1981 : La taille et le poids des Français » *Économie et Statistique ;* 132, p. 23-38.
CORDIER B. (Mme ESCOFFIER), 1965 : *Analyse factorielle des correspondances*, thèse de 3ème cycle, Université de Rennes.
DESROSIERES A., 1972 : « Un découpage de l'industrie en trois secteurs », *Economie et Statistique*, 40, pp. 25-39.
DESROSIERES A., 1978 : « Marché matrimonial et structure des classes sociales », *Actes de la recherche en sciences sociales,* 20-21, pp. 97-107.
DESROSIERES A;, 2001 : « Entre réalisme métrologique et conventions d'équivalence : les ambiguïtés de la sociologie quantitative », *Genèses*, 43, pp. 112-127.
DESROSIERES A., THEVENOT L., 2002 *Les catégories socioprofessionnelles,* La Découverte/Repères.
DEVILLE J.C., 1977 : « Analyse harmonique du calendrier de constitution des familles en France. Disparités sociales et évolution de 1920 à 1960 », *Population*, n°1, janvier-février, pp. 17-63.
DEVILLE J.C., MALINVAUD E., 1983 : « Data Analysis in Official Socio-economic Statistics », *Journal of the Royal Statistical Society*, A, Vol. 146, Part 4, pp. 335-361.
FRISCH R., 1934 : *Statistical Confluence Analysis by Means of Complete Regressions Systems*, University Institute of Economics, Oslo.
GEORGAKAKIS D.,1997 : « Une science en décalage ? Genèses et usages des « socio-styles » du Centre de communication avancée (1972-1990) », *Genèses,* 29, pp. 51-72.
GIRARD A., 1964 : *Le choix du conjoint*, PUF, Paris.
GLAUDE M., MOUTARDIER M., 1982 : « Les budgets des ménages », *Economie et statistique,* 140, pp. 15-34.
GOULD S.J., 1983 : *La mal-mesure de l'homme. L'intelligence sous la toise des savants*. Ramsay, Paris.
GREENACRE, M., BLASIUS, J. 1994 (eds.) : *Correspondence Analysis in the Social Sciences. Recent Developments and Applications.* London: Academic Press.

GUIBERT B., LAGANIER J., VOLLE M., 1971 : "Essai sur les nomenclatures industrielles", *Economie et Statistique*, 20, pp. 23636.

LEBART L., 1969 : « Introduction à l'analyse des données » , *Consommation/Revue de socio-économie,* Vol. 3, pp 57-96, et Vol. 4, 4, pp. 65-87.

LEBART L., TABARD N. 1971 : « La morphologie sociale des communes urbaines », *Consommation/Revue de Socio-Economie,* Vol. 2, pp. 97-107.

LEBART L., MORINEAU A., TABARD N., 1977 : *Techniques de la description statistique*. Dunod, Paris.

MACKENZIE D., 1981 : *Statistics in Britain, 1865-1930. The Social Construction of Scientific Knowledge*, Edinburgh University Press, Edinburgh.

MALINVAUD E., 1970 : « L'analyse des données. Statistique inductive, statistique descriptive », *Annales de l'INSEE*, 4, mai-septembre, pp. 3-8.
MARCUSE H., 1968 : *L'homme unidimensionnel*, Editions de Minuit, Paris.
MARTIN O., 1997 : *La mesure de l'esprit. Origines et développements de la psychométrie, 1900-1950*, L'Harmattan, Paris.

MEYNAUD H. (éd), 1996 : *Les sciences sociales et l'entreprise : cinquante ans de recherche à EDF*, La Découverte, Paris.
MORGAN M.S., 1990 : *The History of Econometric Ideas*, Cambridge University Press, Cambridge.
MORLAT G, PAGES J.P., 1984 : *Le ciel et la terre. Une approche structuraliste des opinions : le baromètre AESOP*. Pièce en trois actes, présentée au Colloque de l'IDATE à Montpellier en 1984.
NETUMIERES F. des, 1997 : « Méthode de régression et analyse factorielle », *Histoire et Mesure*, CNRS, Vol. XII, n° 3/4, pp. 271-298.
PAGES J.P., CAILLIEZ F., ESCOUFIER Y., 1979 : « Analyse factorielle : un peu d'histoire et de géométrie », *Revue de statistique appliquée*, Vol. XXVII, n°1, pp. 5-28.
PANOFSKY E., 1967 : *Architecture gothique et pensée scolastique*, Editions de Minuit, Paris.
PORTE J., 1954 : « Une enquête par sondage sur l'auditoire radiophonique », *Bulletin mensuel de statistique*, supplément janvier-mars, INSEE, p. 31-58.
RENAULT E., 1998 : « Le calibrage ou une controverse sur la place de la statistique dans la modélisation économique », *La Lettre du CREST*, 28, INSEE, Paris.
ROUANET H., LEBARON F., LE HAY V., ACKERMANN W., LE ROUX B., 2002 : « Régression et analyse géométrique des données : réflexions et suggestions », *Mathématiques et sciences humaines*, 160, pp. 13-45.
ROUANET H., LEBARON F., 2006 : « La preuve statistique : examen critique de la régression », communication au séminaire « Qu'est-ce que *Faire preuve* ? », CURAPP, 5 mai 2006.
SCHILTZ M.A., 1990 : « Influence du choix des traitements statistiques sur les opérations élémentaires dans un dépouillement d'enquête : hypothèses, codage et sélection des variables », communication au Congrès de l'Asociation internationale de sociologie, Madrid, juillet 1990. Etude reprise sous forme modifiée dans Greenacre et Blasius (1994), *Correspondence Analysis in the Social Sciences*.
SPEARMAN C., 1904 : « General Intelligence Objectively Determined and Measured », *American Journal of Psychology*, n° 15, pp. 201-293.
THURSTONE L.L., 1935 : *The vectors of mind,* University of Chicago Press, Chicago.
VOLLE M. *et alii,* 1970 : « L'Analyse des données et la construction des nomenclatures d'activités économiques de l'industrie », *Annales de l'INSEE,* 4, pp. 101-131.
VOLLE M., 1980 : *Analyse des données*, Economica, Paris.
ZARCA B., 1979 : « Artisanat et trajectoires sociales », *Actes de la recherche en sciences sociales*, 29, pp. 3-26.

# Annexe

**Plan factoriel de l'Analyse des Correspondances publiée par « Le Nouvel Observateur » en 1975 *(extraits)*.**