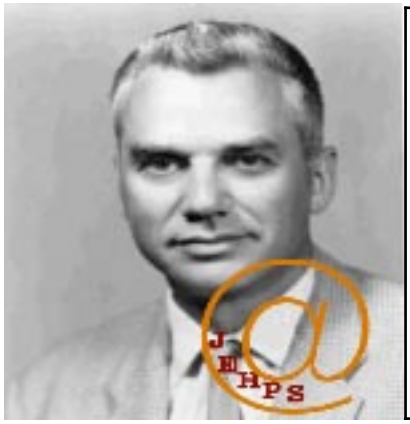

Journ@l Electronique d'Histoire des Probabilités et de la Statistique

Electronic Journ@l for History of Probability and Statistics

Vol 4, n°2; Décembre/December 2008

www.jehps.net

Histoire et Préhistoire de l'Analyse des données par J.P. Benzecri : un cas de généalogie rétrospective[1]

Michel ARMATTE[2]

Dans les années 1970, L'école d'analyse des données "à la française" – AD dans ce qui suit - est en plein développement. Elle fait même l'objet "*d'un véritable phénomène de mode chez les statisticiens, caractérisé à la fois par l'engouement et le rejet*" selon M. Volle (1980). Nous en rappelons d'abord les grands moments.

Elle a été impulsée par le professeur Jean-Paul Benzécri, normalien et élève de Cartan[3], Maître de conférence à l'Université de Rennes depuis 1960 . Son travail à Rennes centré sur une approche inductive de la linguistique par l'analyse de tables de distributions de mots (dans la lignée des travaux de Z.S. Harris, opposée à celle de Chomsky), forme dès 1962 le cadre de l'invention de l'analyse des correspondances comme méthode d'analyse factorielle de tableaux de contingence munis d'une métrique du χ^2 imposée par le principe d'équivalence distributionnelle. Un exposé au Collège de France début 1963 est l'occasion une première synthèse de ses résultats, et la thèse (de 3^e cycle) de Brigitte Cordier présentée le 6 mai 1965 conforte les principales démonstrations mathématiques de cette approche, établit le principe de dualité qui permet la représentation simultanée des deux ensembles en correspondance, et surtout, traite pour la première fois un exemple complet sur un IBM 1620.

---

[1] Je remercie Ludovic Lebart et Pierre Cazes pour leur relecture attentive de ce texte. Je dédie ce texte à la mémoire de Henry Rouanet, décédé le mois dernier, qui fut une référence de la Statistique mathématique et de l'Analyse des Données et un grand admirateur critique de l'œuvre de Benzécri. Il a plus que largement contribué à la diffusion de ce qu'il appelait l'approche géométrico formelle de l'analyse des données multidimensionnelles, en France (voir Rouanet et Le Roux 1993) et aux Etats-Unis (voir Rouanet 2004) avec la préface de P. Suppes.

[2] Université Paris Dauphine et Centre A. Koyré. michel.armatte@dauphine.fr

[3] avec qui il a fait une thèse de topologie (1955)

Benzécri, nommé Professeur à la Faculté des Sciences de Paris (Sorbonne) est alors accueilli par le professeur Dugué, Directeur de l'ISUP et le doyen Zamanski dans le nouveau laboratoire de statistique mathématique de la Faculté des Sciences de Paris. Il y fait d'abord un cours d'option avant d'en prendre la Direction en 1967. La chaire et le laboratoire associé ont été le cadre essentiel de l'élaboration et de la diffusion de cette analyse des données à la française. Ses éléments ont été établis à l'occasion des cours de DEA de 1965-66, et 1967-68, enrichis par la thèse de Maurice Roux (1968) sur la classification de données écologiques. En janvier 1968, Ch. Masson lit au colloque d'Honolulu consacré à la reconnaissance des formes un rapport de Benzécri, publié l'année suivante, qui peut être considéré comme la première synthèse en anglais de ses travaux.

Cette même année 1968 est bien évidemment marquée par les événements de mai. L'innovation principale est que les stages de DEA, introduits dès 1966 vont se substituer partiellement aux enseignements suspendus (ainsi que toute l'activité du laboratoire) et devenir obligatoires. Benzécri présente cela comme une reprise de sa proposition et comme une réponse à la critique d'une Université sclérosée, mais je me souviens aussi qu'ils furent la solution d'un contrôle des connaissances interrompu par les événements du Quartier Latin. L'examen fut repoussé en septembre et complété par une note de rapport de stage[4]. En tout cas ces stages dans les laboratoires du CNRS, de l'Université, ou des grandes entreprises nationalisées, ont été une source ininterrompue de données et de problèmes posés au Laboratoire, stimulant la recherche qui y était faite. Mais ils ont été aussi un puissant vecteur de sa diffusion dans les domaines les plus variés, débouchant souvent sur des emplois et en tout cas sur des réseaux de collaboration animés par des anciens élèves se vivant comme des missionnaires de l'AD.

En 67-68, l'effectif des étudiants de DEA est déjà de 94 inscrits, il est quelques années plus tard, dans les années 1972-76 de l'ordre de 180 à 200. La méthode a acquis une stabilité mathématique et commence à diffuser dans les différents milieux de la recherche, toutes disciplines de l'homme et de la nature confondues, par le jeu conjugué de l'ordinateur et des bataillons de jeunes statisticiens envoyés en stage de DEA dans leurs laboratoires[5]. Pierre Cazes[6] qui

---

[4] Les négociations entre les étudiants et Benzécri ont été menées par Michel de Virville, délégué des étudiants au conseil de gestion du laboratoire, et L'AG du 10 juin 1968 a décidé que l'admission se ferait sur la base d'un compte-rendu de stage ou de tout autre travail statistique. Mais le rapport de force ayant un peu changé dans l'été, un examen de septembre eu lieu en complément.

[5] C'est une stratégie explicite de Benzécri : "Or pour analyser il fallait des données. ..je résolus d'envoyer les étudiants du DEA cueillir des gerbes de ces précieuses fleurs. Dès mon cours j'annonçai qu'ils devaient faire des stages pratiques . Mais cet appel répété de semaine en semaine ne trouvait pas d'écho (…) Le joli mois de mai allait tout changer!" (Benzécri 2008). Benzécri lui-même (HPAD et Réponse à Rouanet et Lépine 1976) reconnaît que ces élèves mal aguerris lancés dans la nature avec les premiers programmes d'AD sur ordinateurs ont menés une guerre de conquête qui n'a pas été sans dégâts collatéraux : "A partir de 1965, des

assure alors en tandem avec Benzécri la gestion de ces stages. confirme que cette gestion était facilitée par le fait que les offres de stages et les demandes de traitement de données affluaient au laboratoire, et que Benzécri en traitait personnellement un grand nombre. Dans le prolongement de ces stages, les thèses se font de plus en plus nombreuses. Cazes (entretien) chiffre à 40 thèses par an le flux de thèses annuelles faites sous la direction de Benzécri dans les années 1970. Quant au rôle de l'ordinateur, toujours mis en avant par Benzécri, il faut le resituer dans le contexte des équipements de calcul de la Recherche française des années soixante[7]. Le Laboratoire mobilise 60% des ressources du Centre de Ressources Informatiques de Paris-6, et une grande partie des calculs, programmés en FORTRAN par de jeunes statisticiens formés le plus souvent non pas dans le cadre de leurs études antérieures, mais dans les TP du DEA, se fait également soit à l'Institut Blaise Pascal[8] (1946-69), soit au CIRCE[9] à Orsay, soit enfin sur les ordinateurs des partenaires comme EDF, CEA…

La publication des deux volumes (Corespondance et Taxinomie) de *l'Analyse des données,* chez Dunod en 1973, réédités en 1976, consacre ce point de non retour de l'Ecole d'Analyse des Données. Le tome II "Correspondances" rassemble à la fois les "principes et formules de l'Analyse", les leçons sur l'Analyse Factorielle rédigées par Benzécri et certains de ses proches élèves, des applications dans des champs très variés (linguistique, anthropologie, sociologie, économie, marketing, hydrologie…) et un programme complet en Fortran. Ces deux volumes, bible de l'Ecole d'Analyse des Données, ne seront jamais traduits en anglais, et de ce fait ces travaux n'auront qu'une diffusion réduite outre Atlantique. Dès 1976, cette publication se voit complétée, à l'initiative principalement de Michel Jambu, par *Les Cahiers de l'Analyse des Données* qui paraîtront de 1976 à 1997[10].

***

---

paquets de cartes ont servis sans que nous sachions ni à qui ni à quoi…L'Analyse des correspondances est une méthode; elle est aussi un outil. A la philosophie de la méthode l'outil doit son efficacité; mais marteau sans maître, celui-ci frappe désormais librement"

[6] Pierre Cazes, ancien élève de l'ENST et étudiant de Benzécri en 1967-68 sort major du DEA et devient l'assistant de Benzécri. Il est chargé à la fois d'une étude sur contrat avec Elf sur données géologiques, des Travaux Dirigés à l'ISUP et dans le DEA, de la gestion des stages, et bientôt des publications, en particulier les *Cahiers de l'Analyse des Données* à partir de 1976, quand le Laboratoire devient une ERA du CNRS, et se voit subventionné pour cette publication.

[7] En 1955, il y a selon Pierre Mounier Kuhn exactement 6 ordinateurs en France (SEA, IBM, et Bull)

[8] Créé en 1946, en réunissant le Laboratoire de calcul numérique de Couffignal à l'ONERA, et le centre de calcul de Borel puis Fréchet et Darmois à l'IHP, l'Institut Blaise Pascal, est dirigé par de Possel, un bourbakiste dissident, et ses adjoints Jean Porte puis André Lentin. Il s'installe ensuite dans les locaux de la rue du Maroc, et est démantelé en 1969. (Voir Mounier-Kuhn 1991 et Louis Nollin 1998)

[9] Centre interrégional de calcul électronique (?)

[10] Entre 1980 et 1987 paraîtrons également les 5 tomes de *Pratique de l'Analyse des Données* dédiés à des exposés élémentaires co-rédigés par Jean-Paul et Françoise Benzécri, son épouse (t.1 : 1979) et (t.2 : 1980) puis à des études sectorielles : Lexicologie (t.3 : 1981), Médecine (t.4 : 1996), Economie (t.5 : 1986).

C'est à ce moment-là – printemps 1975 - que J.P. Benzécri commence des recherches historiques sur la statistique mathématique qui lui fourniront les matériaux de cinq articles publiés dans les *Cahiers* entre mars 1976 et mars 1977, et réunis ultérieurement (1982) dans un petit ouvrage de 160 pages, publié chez Dunod, et intitulé "*Histoire et Préhistoire de l'Analyse des Donnnées*" (HPAD dans la suite[11]). Si le dernier chapitre de cet ouvrage est un développement de cette histoire courte dont Benzécri fut l'acteur principal et que nous venons de résumer, l'ouvrage dépasse le cadre classique de l'introduction historique à une œuvre, en plongeant résolument et complètement dans une histoire plus longue des "sciences du hasard" : c'est ainsi que Benzécri nomme la discipline protéiforme qui regroupe le calcul des probabilités, la statistique et l'analyse des données pour s'affranchir – comme nous le faisons aussi – des avatars qui ont présidé aux rapports de ces disciplines depuis trois cents ans.

L'exercice que nous proposons dans ce dossier consacré à J.P. Benzécri est de comprendre d'abord (I) le rôle que cette investigation historique a pu jouer dans la pensée du "maître", et plus largement dans le corpus des publication de l'Ecole d'Analyse des données qu'il a fondée; ensuite (II) de nous intéresser au contenu de cet ouvrage et à la place qu'il a prise dans l'historiographie de la Statistique, celle qui existe dans les années 1970, et enfin (III) de le faire dans celle qui s'est développée depuis trente ans. L'objectif plus large étant alors de pointer les traits spécifiques d'une telle histoire, par rapport à ce qui apparaît comme une généalogie, écrite par celui-là même qui lui a donné ce qu'il considère comme un aboutissement.

Mais pour ce faire, répondons à la question posée à tout locuteur dans les amphithéâtres des années 1970 : *mais d'où parles-tu camarade?*

La question qui peut paraître saugrenue dans un univers scientifique bien fermé sur son paradigme et dans lequel n'entre aucun intrus, n'est pas sans importance dès lors qu'il s'agit d'histoire. Qui produit cette histoire? Dans le cas des mathématiques, cette question est d'autant plus importante que son histoire a parfois été écrite par des mathématiciens qui tiraient leur légitimité de la connaissance intime qu'ils avaient des formalismes et théories produite par leurs collègues, et parfois elle l'a été par des historiens des sciences "n'ayant jamais produit le moindre théorème" donc pas mathématiciens selon la définition de Dieudonné, mais qui se réclamaient d'une autre discipline – l'histoire - et d'une pratique – celle des différents corpus, sources et archives mobilisés par l'historien – qui constituent à leurs yeux une légitimité au moins aussi forte pour parler d'une œuvre ancienne, fusse-t-elle mathématique.

[11] A notre connaissance cet ouvrage n'a pas reçu d'autre revue critique que la courte note de Bresson (1978) dans *MSH.*

En ce qui me concerne j'aurai l'audace de revendiquer ces deux légitimités: je parle d'abord du point de vue de l'élève de Benzécri que je fus en 1967-68 et qui a longtemps enseigné et pratiqué l'Analyse des données, cherchant par l'AFC et la CAH à faire surgir des "données" recueillies par des sociologues, des anthropologues ou des économistes, certaines structures Guttmaniennes ou hierarchiques qui faisaient sens[12]. Mais je parlerai aussi, et surtout, du point de vue de l'historien des sciences que je suis devenu plus tard après un DEA "Sciences Technique et Société" et une thèse sur *l'histoire du modèle linéaire* en économétrie. La première situation me donnera donc la familiarité nécessaire avec les objets dont il est question en AD - et le prétexte à quelques anecdotes vécues – tandis que la seconde m'autorisera à une réflexion à la fois plus distanciée et plus approfondie sur ce que pourrait être une histoire de l'Analyse des Données.

I. Le rôle de HPAD dans le développement de l'AD.

Pourquoi Benzécri a-t-il éprouvé le besoin de faire une plongée dans les oeuvres de ses prédécesseurs? Le parcours antérieur de ce normalien né à Oran dans les années 1930, nourri de mathématiques bourbachiques ne l'avait pas spécialement conduit à une approche historique. Le peu d'éléments que l'on a sur la biographie de Benzécri laisse supposer que ses premiers centres d'intérêt, ses premiers travaux sur la linguistique l'ont très vite entraîné à se détourner de l'analyse classique (calcul différentiel et calcul fonctionnel, topologie) et plus généralement de la mathématique pure de ses maîtres Cartan et Choquet qui l'a nourri[13].

Il semble, à écouter plusieurs témoignages de ses élèves, qu'il ait été quelque peu déstabilisé par l'accueil assez froid de son paradigme dans le monde anglo-saxon. Son séjour des années 1950 à *l'Institute for Advanced Studies* l'avait familiarisé à la fois avec la langue, les publications et les outils – y compris les computer (encore souvent analogiques) - de la recherche mathématique américaine. Le désamour a semble-t-il commencé lorsque Benzécri fait un nouveau voyage aux Etats-Unis dans l'été 1965 pour présenter l'

[12] Mes travaux au sein du Musée des Arts et Traditions Populaires avec Jean Cuisenier et Michel de Virville m'ont familiarisés avec les problématiques de description et classification des objets de musée aussi bien qu'avec les questionnaires des sociologues du Centre d'Ethnologie Française abrité dans les mêmes lieux, l'AFC offrant une alternative au moins économique et écologique (si ce n'est épistémologique) aux listings de centaines de tableaux croisés que nous produisions avec ferveur à leur requête.

[13] Sa thèse avec Cartan en 1955, la rédaction d'un exposé de sur la théorie des capacités d'après G. Choquet pour le Séminaire Bourbaki (1954-56), et la publication pour la SMF (1960) d'un papier "sur les variétés localement affines" sont les dernières traces d'un début de carrière de mathématicien "pur". Son séjour à Princeton en 1954-55 a pu jouer un rôle dans ce virage, à moins que cela ne soit d'avoir servi dans un Groupe de Recherche Opérationnelle de la Marine nationale de 1959 à 1960 (Cf. Benzécri 2006)

Analyse des Correspondances aux *Bell Labs*. Ecoutons Henry Rouanet raconter comment l'échec de la rencontre fut patent :

> "*J'attendais avec curiosité le résultat de cette visite. A son retour Benzécri s'est montré évasif. Que s'est il passé lors de cette visite? Dans les publications du groupe MDS postérieures à 1965, plus de mention aucune de Benzécri! Lors de ma visite à la Bell Telephone en 1968 on m'a davantage évoqué les extravagances vestimentaires de Benzécri que ses innovations statistiques. Personnellement je date de 1965 l'attitude de grand renfermement de Benzécri vis-à-vis du monde anglo-saxon…*"[14]

Mais d'autres "timides et maladroits apôtres" de J.P. Benzécri – comme se définit rétrospectivement Ludovic Lebart en 2008 – soulignent que Benzécri a été relativement vexé par l'article de Mark Hill, publié en 1974 sous le titre prometteur de "*Correspondence Analysis : A neglected Multivariate Method*". Cet article loin de reconnaître l'œuvre de Benzécri déjà bien établie dans ses cours publiés de 1967-68, et publiée l'année précédente en deux volumes (Benzécri 1973), ne signale que l'article sur la reconnaissance des formes, lu par Ch. Masson au colloque d'Honolulu en janvier 1968, et publié l'année suivante (Benzécri 1969). Plus grave encore, Hill présente l'analyse des correspondances dans une filiation bien particulière : "*Guttman (1941), Torgerson (1958) and Hill(1973) introduce correspondence analysis as a method of scaling rather than of contingency table analysis*" écrit-il.

Ce n'est pas la dernière fois où l'analyse des données selon Benzécri est la victime d'un certain ostracisme de la part des écoles de la *multivariate analysis*. En 2005, l'ouvrage de l'irlandais Fionn Murtagh, disciple de Benzécri est l'objet de vives critiques de Jan de Leeuw . On en retiendra la principale: "*this book is charmingly interesting and maddeningly insular (...) It is no longer true that the technique is the exclusive property of a somewhat esoteric group of French statisticians and that it is associated with an equally esoteric philosophy of science and data analysis*". Comme le premier chapitre de Murtagh est une reprise de HPAD, Jan de Leeuw peut ajouter que ce livre repose "on Benzécri's *very limited* history of data analysis"[15].

A la suite de l'article précité de Hill, Benzécri s'enferme plusieurs semaines à la Bibliothèque Nationale et consacre son printemps 1975 à une investigation en profondeur de la statistique, majoritairement anglo-saxonne du siècle précédent. S'agissait-il de vérifier les références placées par Hill sur le chemin de l'analyse des correspondances comme Hotelling (1933), Hirschfeld

---

[14] "entretien de H. Rouanet avec Philippe Bonnet sur "Benzécri et l'Analyse des données", février 2007, Site de Henry Rouanet : www.math-info.univ-paris5.fr/~ROUANETBEN.html

[15] La réponse de Fionn Murtagh à cette critique fort musclée se trouve sur son site.

(1935), ou Fischer (1940) ? Ou bien d'en trouver d'autres qu'il pourrait revendiquer afin de se réinsérer dans cette historiographie qui ne lui laissait pas sa place? Ou bien encore de trouver des justifications historiques à son point de vue radicalement empirique et anti-probabiliste? Un peu tout cela à la fois sans doute. Mais la réponse est certainement, au moins entre les lignes, dans le texte même de *Histoire et Préhistoire de l'Analyse des Données*.

II. HPAD et l'historiographie de la Statistique.

Le récit historique de Benzécri se veut "soucieux de rendre à chacun ce qui lui est dû" et s'efforce " de retrouver par quelles voies la statistique est parvenue au point ou nous la voyons aujourd'hui"; au point où *il* la voit aujourd'hui faudrait il dire plus exactement. Suivons donc le fil de cette reconstruction en repérant d'où proviennent ses matériaux.

a) Les probabilités

Le premier chapitre , intitulé "la préhistoire" est principalement l'occasion pour Benzécri de régler sa dette (et son compte) au calcul des probabilités, dont "on cherchera l'origine dans les méditations d'Aristote sur le hasard et la fortune" L'auteur arrive directement aux pères du calcul des probabilités que sont Pascal, Fermat, Huyghens, T. Bayes, Jacques Bernoulli, Laplace et Gauss. Sa référence principale et quasi unique pour en traiter est le livre de Todhunter (1865) dont il reprend la démarche d'une histoire des idées philosophiques et mathématiques sur le hasard. Le résumé qu'en fait Benzécri pourrait constituer la base de ce qui deviendra une doxa de l'histoire du calcul des probabilités en France dans les années 1970, avec ses légendes fondatrices (la correspondance Pascal-Fermat, la querelle de priorité entre Legendre et Gauss sur les moindres carrés…), ses propos hagiographiques nombreux et son style légèrement pompeux[16], le souci de redonner les formules des uns et des autres dans les notations modernes qui sont celles de ses cours, et le pointage discret des difficultés et apories de ce calcul qu'il s'agisse de la définition circulaire de la probabilité chez Laplace ou du postulat des probabilités a priori égales chez les Bayesiens qui ne lui semble pas "réaliste". De l'Arithmétique politique il n'est pas question globalement bien que soient évoquées les études démographiques et théologiques de Süssmilch (qu'il cite d'après Gnedenko 1950, sa seconde source après Todhunter) et les controverses autour de l'inoculation de la variole.

La théorie des erreurs semble d'abord se résumer un peu simplement à la découverte des moindres carrés par Legendre. Heureusement Benzécri s'attarde

---

[16] "Nous ne tenterons pas d'imiter ici l'illustre géomètre qui dans *l'Essai philosophique* donne de son analyse un exposé dépourvu de toute formule mathématique, mais devions au passage saluer un monument"

"A la fois grand analyste, grand mécanicien, grand probabiliste et ce qui - singulièrement dans la langue de ce temps – dit plus que tout : grand géomètre, Laplace était tout désigné pour de tels travaux"

un peu sur l'hypothèse de normalité qui est mobilisée par Laplace et Gauss pour justifier cette méthode. Et le ton du commentaire sur l'hypothèse de normalité des erreurs se fait un peu railleur, dans la veine de la boutade de l'astronome Lippman, rapportée par Poincaré[17], ou encore du traité de Joseph Bertrand (1889). Mais pour qui connaît un peu cette littérature touffue de la théorie des erreurs[18], il est remarquable que Benzécri y repère assez bien les avancées successives de Gauss et Laplace, avec un petit penchant pour Laplace. Benzécri a bien lu la section correspondante de la *théorie analytique* et son empathie très forte pour Karl Pearson ne l'empêche pas de le contredire sur ce point : la loi normale multidimensionnelle est bien là, déjà, et la justification des moindres carrés n'est point dans la normalité des observations mais dans celle des estimateurs auxquels il applique son théorème central limite. L'auteur de HPAD ne fait guère de cas de la seconde théorie de Gauss, celle de 1833 et 1826 qui le conduit à privilégier, pour justifier les moindres carrés, non pas l'approche du maximum de vraisemblance qui avait été la sienne en 1809, non pas l'approche asymptotique de Laplace, mais celle du minimum de variance des estimateurs linéaires. Or c'est cette approche qui sera principalement reprise dans la tradition économétrique du 20^{e} siècle.

Pour terminer ce chapitre Benzécri s'engage dans un bilan de la théorie cinétique des gaz de Maxwell et Boltzmann, mais ce bilan tourne court pour sauter directement à Einstein. Est-ce faute de matériaux puisque ici Todhunter n'est plus d'aucune utilité? Est-ce parce que Benzécri n'a pas une connaissance directe ou indirecet des avancées de la mécanique statistique[19] ? Ou bien encore parce que cette discipline ne rendre pas dans le champ des sciences d'observation auxquelles s'adresse l'Analyse des données?

Que conclut Benzécri de ce rapide tour d'horizon sur l'histoire de la probabilité? D'abord que si certains principes de symétrie s'imposent a priori dans la probabilité des jeux et des phénomènes physiques, "tout autres sont les probabilités du statisticien" : la répartition uniforme du risque variolique n'est plus une invariance de principe mais une hypothèse hasardeuse dans le mauvais sens du terme. Du coup les probabilités fréquentistes, fussent elles révisées par un processus bayésien, lui apparaissent encore moins fondées que les probabilités subjectives de de Finetti à qui il emprunte la boutade suivante : *ce sol n'est pas assez ferme : c'est du sable; enlevons le sable (les probabilités) et fondons la bâtisse sur le vide (la formule de Bayes)*". Rien ne tient, tout s'effondre dans ce calcul, hors la physique quantique et les rares phénomènes

[17] Lippman : "Les expérimentateurs s'imaginent que c'est un théorème de mathématique et les mathématiciens que c'est un fait expérimental".

[18] Armatte 1995 et 2004

[19] Par exemple avec les articles de synthèse de Borel et de Paul et Tania Ehrenfest dans l'Encyclopedie des mathématiques pures et appliquées de Jules Molk.

ergodiques que l'on peut identifier. Devant tant de déboires, Benzécri se raccroche à la dernière branche, celle de l'axiomatique de Kolmogorov qui vise à bien "séparer la détermination des probabilités des règles de leur calcul", mais avec le regret de voir que "les jeunes générations formées à l'austère théorie de la mesure ne reçoivent plus dans les paradoxes à la Joseph Bertrand la tradition de l'expérience séculaire". Cette branche axiomatique n'est donc pas la panacée.

La dernière solution qui s'offre à lui c'est de renoncer aux méthodes probabilistes, en étudiant des échantillons suffisamment homogènes et complets pour être signifiants à défaut d'être significatifs. De quelle population le seraient ils d'ailleurs s'ils sont pris "dans la nature" sans aucun processus d'échantillonnage, sans aucune "randomisation"? Dés lors, le slogan "*Statistique n'est pas Probabilité*" termine ce chapitre de façon assez brutale et radicale, comme il avait ouvert avec force le tome "Correspondances" de son *Analyse des Données* un ou deux ans plus tôt, lui fournissant explicitement son premier principe : "*sous le nom de statistique mathématique, des auteurs (qui je vous le dis en français n'écrivent guère notre langue…) ont édifié une pompeuse discipline, riche en hypothèses qui ne sont jamais satisfaites dans la pratique. Ce n'est pas de ces auteurs qu'il faut attendre la solution de nos problèmes typologiques*"

Benzécri abandonne donc la probabilité en rase campagne, au moment même où elle obtient ses lettres de noblesse, et nous ne saurons rien de la suite de son histoire, pas même de l'Ecole Française de Borel, Fréchet et Lévy qui a dominé les années suivantes.

Notons pour finir que Benzécri n'est pas le premier à faire ce choix : nous avons montré[20] que Lucien March avait dès 1908 une position très tranchée sur l'usage de la probabilité en Statistique : *Une différence essentielle sépare la fréquence observée en statistique et la probabilité mathématique. En statistique nous ignorons les circonstances initiales des faits observés; notre connaissance peut tout au plus s'étendre à quelques parties de l'enchaînement intermédiaire entre les conditions originelles et le résultat (...) Dans le schéma des probabilités les combinaisons et répétitions qui contiennent en puissance le résultat final sont complètement connues; seul le jeu du déclenchement qui fait apparaître certaines combinaisons demeure imperceptible en raison de l'exiguïté de son action. L'accord d'une distribution de fréquences, observée en statistique, avec une distribution de probabilités n'implique donc qu'une analogie apparente entre l'enchaînement des faits statistiques et la formation des probabilités (...) Aussi serait-il opportun, en statistique, de renoncer à l'emploi du mot probabilité pour exprimer l'attente que fait naître la constatation d'une fréquence; car si dans la théorie des probabilités, la*

[20] Armatte, jehps, 2005

convention sur laquelle repose cette attente inspire une parfaite confiance, en statistique le degré de confiance que mérite cette attente est souvent modifié par l'étude des liaisons des faits, par les enseignements des sciences sociales.

Benzécri n'a eu que peu de chances de lire ces lignes d'un statisticien davantage classé comme ingénieur et administrateur que comme mathématicien. Mais ce discours a été repris ou réinventé par de nombreux statisticiens économistes comme Divisia, Pareto, et surtout par Karl Pearson, le héros du chapitre 2 et le mentor qu'il s'est choisi. Chez Pearson la position par rapport à la probabilité est plus complexe : fondée d'abord sur une philosophie – celle de la *Grammaire de la Sciences* – qui privilégie les phénomènes et ne voit dans les modèles a priori (probabilistes ou non) qu'un résumé sténographique de leur perception[21], cette position autorise à décrire a posteriori les phénomènes par les schèmes probabilistes mais nécessite du coup qu'on en multiplie la variété – c'est le fameux travail de Pearson sur le système des courbes de distribution qu'il invente à la fin des années 1890 pour ajuster des observations. De ce fait les contemporains comme Lucien March, lui aussi admirateur de K. Pearson, peuvent se situer dans le débat pour ou contre la probabilité dans une situation médiane, proche également des thèses de Lexis, et visant à abandonner le dogme laplacien au profit d'une pluralité des modèles : "*Deux genres de théories s'opposent actuellement, l'une rattachant toutes les distributions à la loi des erreurs de Gauss, l'autre, avec Pearson, cherchant des lois où interviennent des probabilités variables. Il ne me semble pas qu'on doive condamner aucune de ces tendances*". C'est cette voie que suit Benzécri[22].

La Statistique mathématique entre 1880 et 1960, qui occupera le reste de l'ouvrage HPAD n'est donc pas celle qui prolonge le projet Laplacien : ni Bienaymé, ni Quetelet ni Edgeworth, ni Bortkiewicz, ni Cheybychev, ni Borel pour finir ne seront à l'honneur dans le tableau d'honneur dressé par Benzécri. Le travail de sape de Joseph Bertrand aidé par son humour décapant a fait le reste. Dès la première page de l'ouvrage, Benzécri affirme que "*l'abus du calcul des probabilités a nui à la statistique*" et en tire argument pour que "l'histoire de ces deux disciplines soit dans ses grandes lignes décrite comme l'histoire d'une seule science : celle du hasard".

---

[21] Benzécri écrit (p.44) : "A la différence de maints statisticiens ses cadets il reconnaissait le primat des données sur les modèles".

[22] Au pointage par Rouanet et Lépine (1976) d'une "confusion entre le langage probabiliste et le langage proprement statistique" chez Benzécri, celui-ci répond : "une fois admis que le formalisme probabiliste va au-delà de la nature, où les probabilités stables sont rares, on reste libre d'utiliser le langage probabiliste dont la richesse analogique est très grande.".

b) La biométrie

L'école biométrique anglaise, celle qui résulte des travaux de Francis Galton, Karl Pearson et Ronald Fischer, sera l'unique objet du chapitre II. Ce choix tient en grande partie au fait que la source historique la plus sollicitée, en dehors des textes mêmes de ces auteurs, est principalement l'ensemble de ceux que Maurice Kendall et Egon Pearson ont rassemblés dans le premier tome des S*tudies in the History of Statistics and Probability* (1970). Benzécri n'a hélas pas pu profiter du second tome publié en 1977 par Plackett et Kendall. Il s'appuie également sur les travaux historiques de Pearson lui-même[23], car, comme il le dit lui-même toujours aussi élégamment : "*les auteurs illustres qui nous ont laissé leur témoignage sur le progrès de la statistique ont plus souvent remémoré les éclairs de leur propre pensée que les reflets qui leur étaient venus d'ailleurs*". Benzécri reconnaît d'entrée de jeu, avec ce petit brin d'habitus normalien redoublé de suffisance bourbachique, que ces auteurs, "à la différence d'un Laplace ou d'un Gauss ne sont pas de très grands mathématiciens". Et pourtant on va le voir fasciné par leur traitement des tables de contingences. Le premier réflexe de Benzécri dans l'étude des travaux de Galton[24] est de trier le bon grain de l'ivraie, à savoir d'en extraire ce que la statistique moderne peut revendiquer en héritage, et de refouler dans l'ombre d'une part ce qui relève de la recherche biologique - "*Il ne nous appartient pas d'apprécier ce que la biologie a acquis par ce labeur*" – d'autre part ce qui ferait tache aujourd'hui, à savoir le cadre "héréditariste" et plus encore eugéniste des travaux de cette école: on perd donc toute idée des motivations et du programme de recherche de Francis Galton qui est inséparablement cognitif et politique. Par contre le résumé formel des découvertes par Galton de la réversion, puis de la régression et enfin de la corrélation est tout à fait canonique. Benzécri ne tombe pas totalement dans les filets de Pearson qui voudrait faire croire que la découverte de Galton est originale et bien différente de celle des géomètres de la théorie des erreurs : "*Qui a été, comme nous, élevé sous le sceptre de N. Bourbaki, s'étonnera plutôt que les mathématiciens n'aient pas découvert par eux-mêmes toutes les propriétés de la loi normale multidimensionnelle; que Galton n'ait pas été devancé*". Et Benzécri de rappeler les résultats de Laplace et de reprendre ceux de Bravais dans lesquels il repère même la première occurrence du mot "corrélation". Mais il reconnaît comme valide l'argument de Pearson que lui seul postule des observations premières corrélées. Un argument qu'il faut rattacher aussi à sa philosophie phénoménaliste, et dont Pearson (1920) fait une véritable coupure

---

[23] Pearson 1920 et 1978.

[24] Il n'en est pas de même pour Weldon: Benzecri fait un long développement sur ses travaux, avec cependant la même retenue de principe : "il ne nous revient pas de décrire les travaux de Weldon, mais il importe à l'analyse des données d'en citer des exemples pour mieux en méditer les leçons". Quant à la controverse entre Pearson et les mendéliens elle est également assez longuement traitée et, suivant Pearson fils et père, se conclut par la complémentarité des deux approches.

épistémologique : r coefficient de régression (sur données centrée et réduites) devient l'outil de révélation de la corrélation de structure, que Benzécri juge "fondamentale pour l'analyse des données".

Tantôt Benzécri se laisse embarquer dans le débat de la discipline qui mobilise des statistiques, tantôt il quitte sa logique propre pour s'émerveiller de ce qui préfigure certains concepts de l'AD : L'espace cognitif de la génétique a forgé le terme de facteur, et celui de l'anthropométrie celui de matrice de corrélation : le premier volume de *Biometrika* contiendrait "la première matrice complète de corrélation jamais imprimée". Pearson décompose ses variables normales pour avoir un modèle des corrélations héritées. "*On voit donc que tous les éléments étaient réunis pour fonder l'analyse factorielle*"(p.43). Allant encore plus loin dans ce sens, Benzécri se saisit du problème de la corrélation partielle – dont il ne dit pas que c'est Yule qui l'a principalement développé – pour le critiquer au nom des résultats de l'analyse des données de 1975 : avec juste raison il rappelle que le modèle multi-normal ne peut conduire qu'à "des sections de la loi normale multidimensionnelle toutes égales entre elles à une translation près" alors que le nuage de l'AFC peut présenter une torsion telle que d'une section à l'autre les corrélations (partielles) s'inversent". Cet argument très souvent repris au titre de la supériorité de l'AD sur le modèle linéaire me semble pour la première fois exprimé et illustré géométriquement ici.

S'émerveillant de la mise au point par Pearson du test du χ^2 avec la distance du même nom qui suit une loi du même nom – mais dont le paramètre est erroné chez Pearson, Benzécri ne manque pas de reprendre l'interprétation de ce χ^2 comme trace des valeurs propres d'une AFC. Mais il ne discute pas l'idée de Pearson d'en déduire un "coefficient de corrélation généralisé pour des données catégorielles – par la transformation $\varphi^2 = \chi^2 / (1+\chi^2)$, ni surtout de la controverse entre Pearson et Yule qui en résulte.

Impossible de quitter Pearson pour Fisher sans citer la phrase de conclusion de Benzécri : "*On le voit pour l'analyse des données de 1975, K. Pearson est un précurseur. Intrépide dans la collecte des données, fécond mais imparfaitement exact dans les constructions mathématiques, plus aventuré encore dans la philosophie, intraitable dans bien des querelles d'Ecole, Karl Pearson a vu son rôle minimisé par la génération de statisticiens élevés dans la doctrine de R.A. Fisher; lequel corrigea maintes inexactitudes de son devancier et donna à la méthode statistique une forme plus cohérente mais dirons nous moins accueillante aux flots de la nature. On nous permettra de choisir ici le patronage de K. Pearson".*

Tout est donc déjà dit sur Fisher qui constituera pourtant l'objet d'une longue Partie III intitulée *era piscatoria* (l'ère du pêcheur) : "*En presque tout, autant qu'il le put, Fisher s'opposa à Pearson ; mais maintenant que le silence s'est fait sur le champ de bataille, on peut affirmer sereinement que celui-là fut*

le continuateur de celui-ci". Contrairement à l'Ecole historique de Bath, Benzécri préfère prolonger ce silence plutôt que réveiller les controverses avec Yule sur la mesure de la contingence, avec Fisher sur la méthode des moments[25], avec les mendéliens sur la Loi ancestrale de l'hérédité. Pour faire vite, disons que Benzécri se voit obligé de faire le tri des innovations de Fisher : peu intéressé par l'échafaudage des lois dérivées de la loi multinormale que Student et Fisher utilisent pour de petits échantillons, assez sceptique au sujet des plans d'expériences sur données randomisées, et plus encore au sujet des tests d'hypothèses probabilistes, Benzécri est un grand admirateur de la représentation géométrique dans R^n des variables statistiques et de leurs moyennes, variances et corrélations, ainsi que de la décomposition de ces espaces vectoriels en sous espaces orthogonaux qui sont mobilisés en analyse de variance et qui joueront un rôle clé en analyse factorielle. Fisher, pour une raison que l'on a parfois attribué à sa quasi cécité, use facilement de ces représentations pour ses démonstrations sans toujours expliciter ses visions, ne serait-ce que parce que l'algèbre linéaire n'est pas encore usuelle à cette époque. L'ouvrage historique de Benzécri rend explicite, et parfois prolonge lui-même dans des notations modernes, ces visions géométriques de Fisher, les complétant avec des schématisation graphiques qui aident beaucoup le lecteur. La théorie de l'estimation de Fisher et la justification du maximum de vraisemblance est ainsi réinterprétée par Benzécri comme une "projection orthogonale, avec la métrique du χ^2 , de la loi de fréquence de l'échantillon sur l'espace des lois des paramètres à estimer (p.57). Mais Benzécri ne suit pas Fisher quand il attribue comme objectif à l'inférence l'estimation de paramètres de lois hypothétiques, et quand il fait de cette théorie de la vraisemblance, comme il le fit plus tard de la probabilité fiduciaire, une arme contre l'approche bayésienne[26]. La statistique des années 1960 et 1970 lui semble au contraire s'intéresser de plus en plus au non paramétrique d'une part, et aux méthodes neobayésiennes d'autre part.

Intéressé par l'Analyse de variance développée par Fisher pour mesurer l'importance de diverses causes parce que les notions de variances intra-classe et interclasses, sont utiles en analyse de données, et en analyse discriminante, il ne goûte guère la théorie des plans d'expérience. Conçue par Fisher dans le cadre de recherches agronomiques à la station de Rothamsted, la méthode des plans d'expérience qui suppose à la fois un modèle causal a priori défini, des expériences contrôlées "toutes choses égales par ailleurs" ou une randomization c'est-à-dire la considération d'échantillons aléatoires de populations infinies normales, n'est pas exportable en dehors d'un contexte expérimental. Elle n'est d'aucune utilité pour l'analyse de données d'observation.

---

[25] "Une méthode raisonnable mais rien ne permet de démontrer qu'elle soit optimale (…) Ce problème mit entre Pearson et Fisher une opposition que le temps ne put réduire (…) Nous ne tenterons pas ici d'analyser puis d'arbitrer cette controverse : il suffira de rappeler quelques principes" .

[26] Voir par exemple Armatte (1988)

"*L'expérimentation à la Fisher comme celle à la Stuart Mill (...) n'est pas un outil suffisant de découverte d'une loi y = f(x); elle précise et confirme seulement la découverte préalable qu'il fallait centrer l'étude sur y et x. Selon nous cette découverte a son origine dans l'observation. Ce qui l'engendre, s'appelle d'un mot simple et grand : le génie : libre démarche de l'esprit que ne règle aucun algorithme. Mais l'examen ordonné d'un ensemble d'observations suggestives, peut aider la statistique : en écologie, en psychologie, en sociologie, partout où les formules y = f(x) ne sont qu'une façon de parler, non une loi déjà découverte, nous préférons généralement l'analyse d'observations judicieusement recueillies sur une base naturelle, à l'expérimentation suivant un plan combinatoire (d'ailleurs souvent inapplicable).*" (p.70-71).

Benzécri ne s'attarde pas davantage sur la théorie des tests d'hypothèse de Fisher. Bien sûr il la mentionne et il mentionne la controverse entre l'approche inductive de Fisher et l'approche décisionnelle de Neyman et Pearson, endossant le point de vue de Kendall selon lequel cette opposition recouvre celle du point de vue anglais contre l'américain, soit "les pays où ce que fait un homme compte plus que ce qu'il pense contre ceux où c'est l'inverse". Mais une fois de plus il conclura par la formule habituelle : "*il ne nous appartient pas d'arbitrer ce différend*". Et nous n'aurons point d'analyse de ce qui résulte aujourd'hui de ce différend, à savoir une théorie hybride dont le mésusage et l'abus dans des champs comme la psychologie expérimentale ou l'économétrie ont été maintes fois dénoncés[27].

c) la psychométrie

L'avant dernier chapitre de HPAD, est consacré à la psychométrie, domaine quasi unique (hormis l'écologie) ayant servi de matrice à l'analyse factorielle. Ses sources ici sont principalement la littérature anglo-saxonne des années cinquante : *Multiple factor Analysis* de Thurstone en 1947, *A course in multivariate analysis* de Kendall (1957), mais aussi les actes du colloque CNRS de 1955 sur l'analyse factorielle et ses applications. Nous sommes donc désormais à faible distance historique de ces objets. Suivant les travaux de Spearman (1904) et Thurstone (1931-32) sur la décomposition de l'intelligence en facteurs spécifiques et facteurs communs – multiples chez le second, mais unique chez le premier (c'est le facteur général) - Benzécri pointe les questions de doctrine aussi bien que l'avancée des formalismes. Par exemple le fait que l'on peut partir uniquement des observations ou d'un modèle a priori, linéaire, distingue au début analyse en composantes et analyse factorielle, mais il y a finalement des équivalences formelles entre ces deux approches.

[27] Voir par exemple le dossier du Journal de la Société Française de Statistique, 145-4, 2004.

Les outils de calcul restent déterminants pour les progrès du domaine : les déterminations successives des facteurs par tâtonnement (méthode du centroïde), par régression (chez Thurstone) et par itération et orthogonalisation (Hotelling) sont marquées par les conditions du calcul d'avant guerre et l'absence d'ordinateurs.

Les épreuves de validité – sous hypothèse de normalité – introduites par Hotelling et plus prudemment par Thurstone n'ont guère la faveur de Benzécri : "Thurstone lui-même s'efforçait d'échapper aux griffes de la statistique mathématique" et de citer la préface de son ouvrage : *"...mieux vaut une mesure significative sans distribution d'échantillonnage, qu'une mesure triviale ou irrelevante (sic) choisie parce que la distribution en est connue*"

Quant aux interprétations des facteurs, elles renvoient à certaines visions structurelles. Benzécri pense ces interprétations non pas en s'aidant des cas évoqués par ces auteurs, mais à partir des exemples mobilisés dans ses propres recherches (l'analyse des données du concours d'admission à l'Ecole polytechnique)

Il se penche pour finir sur l'analyse des données non métriques - on dira aussi catégorielle ou nominales – qui est la caractéristique de la future Analyse Factorielle des Correspondance (AFC). Ses sources sont alors principalement Torgerson (1958) : *Theory and methods of scaling,* qui formalise les réponses comparées à des stimulus dans des espaces métriques lui permettant d'établir des matrices de distances; Guttman 1941: *The quantification of a class of attributes* qui cherche à rendre compte de structures particulières (scalogrammes) des pattern de réponses à un questionnaire[28]; Lazarsfeld dont les travaux sur les structures latentes[29] ont fortement influencé les sociologues français après la traduction qu'en a donné Raymond Boudon ; C.H. Coombs (1964) : *A theory of Data* qui produit une somme synthétique sur la mesure, le scaling, et la classification de sujets soumis à des stimuli à partir de données qui sont de quatre types : choix ou préférences, réponse à un stimulus simple, comparaisons de stimulus, données de similarité. Ce dernier auteur fournit à Benzécri l'occasion de mettre en avant les opérations de codage, essentielle en AD.

d) l'analyse des données

Dans le dernier chapitre (5) sur l'analyse des Correspondances, Il n'est plus tant question des errances de l'histoire que de son aboutissement avec les

[28] Les études de Guttman portent sur la confiance du soldat américain en l'armée qu'il sert. Les tableaux sont traités par des procédés mécaniques de permutation des lignes et colonnes.. Les calculs faits ensuite révèlent des relations polynomiales entre facteurs

[29] On fait l'hypothèse que les individus i répondent à des questions de façon indépendante et aléatoire, en fonction d'une variable cachée continue, découpée en classes "latentes".

travaux de Benzécri. Cette partie n'est "*qu'une mise au point chronologique susceptible de réduire à leur juste proportion les controverses de priorité*". L'ensemble de l'ouvrage converge donc vers cette problématique des priorités, plutôt à courte vue, bien qu'on devine qu'elle soit comme toujours à forte charge affective : "*l'analyse des correspondances remonte à l'automne de 1962, et le premier exposé fut donné par J.P. Benzécri au Collège de France dans une leçon du cours Pécot de l'hiver 1963*". L'auteur revendique à juste titre, non pas les premières formules de cette analyse mais les différents ingrédients qui font systèmes et qui transforment l'invention en innovation : des formalismes nouveaux, des règles de codage, des interprétations géométriques, des algorithmes implantés sur différents systèmes d'exploitation, des aides à l'interprétation, des investigations effectives déployées sur de nombreux champs de recherche. La méthode a permis "*de réduire à l'unité le traitement des problèmes posés par les données les plus diverses*."

Benzécri retrace alors les étapes de cette innovation, plus ou moins celles qui font doxa aujourd'hui, et que nous avons reprises au début de cet article pour situer l'Analyse des données.

Il revient à cette occasion sur l'épisode de sa visite aux Bell Laboratories, dont il rend compte avec humour, en particulier quant à ses échanges avec Caroll. Il y reconnaît une troisième voie d'arriver aux formules de B. Cordier permettant la reconstitution des données à partir des facteurs. Une quatrième interprétation lui est fournie par les écrits de Guttman qu'il découvre à la bibliothèque de la Bell, et ceux de Hayashi. "Quant à remonter dans le temps avant Guttman (1941) comme Hill (1974) y invite, nous serons plus réservé". Benzécri rend compte néanmoins des travaux de l'Ecole anglaise : Hirschfeld (1935), Fisher (1940) et Maung (1941) dont Hill revendique l'héritage "Après l'article de Hill, nous avons lu ces références…" Benzécri fait donc allégeance.

C'est aussi dans cette partie qu'il justifie les notations tensorielles (avec indices haut et bas) qu'il a introduites systématiquement à partir de 1965 : Pour Benzécri le calcul des formules de transitions résulte du calcul des transitions probabilistes qui est une variante du calcul tensoriel, apprise à l'école de Bourbaki, chez Lichnérowicz plus particulièrement. Il permet de rendre compte de la relation de dualité entre espace vectoriel et espace dual, entre espace des fonctions et espace des mesures, et de transporter par une fonction qui généralise les applications ensemblistes à la fois les éléments des ensembles et leurs poids.

Benzécri termine par la description sommaire des perfectionnements apportés à la méthode : où l'on voit qu'elles sont de quatre sortes : 1°) l'extension de l'analyse à d'autres types de données que des tableaux de contingence, en particulier les tableaux de données quantitatives additives et les tableaux logiques mis sous forme disjonctive complète. 2°) Les améliorations des algorithmes de diagonalisation. 3°) les enrichissements des programmes par les aides à l'interprétation et la cartographie (contributions, projection d'éléments

supplémentaires, les jugements de stabilité. 4°) l'articulation avec la classification.

III. Histoire et généalogie rétrospective.

Nous avons jusque là collé au texte, cherchant uniquement à en rendre l'esprit et la richesse, ou à en expliciter quelques éléments laissés dans l'ombre. Prenons maintenant un peu de recul, un recul que nous pouvons avoir de deux manières . D'une part parce que nous sommes 30 ans plus tard et que ce temps écoulé donne de la distance qui permet de réévaluer un moment particulier à l'aune de ses développements ultérieurs. D'autre part parce que pendant ces trente ans, l'histoire de la Statistique que Benzecri découvre a été revisitée par des historiens avec des questionnements et des grilles d'analyse différents. Et posons nous la question suivante : de quelle histoire s'agit-il donc dans ce petit traité?

Une évaluation à l'aune des développements ultérieurs? C'est bien souvent ce que font les scientifiques. Mais il s'agit d'évaluer non pas l'analyse des données mais l'histoire qu'en produit Benzécri dans *HPAD*. En 1975, sa performance est tout à fait exceptionnelle. Il n'existe tout simplement aucun ouvrage en langue française qui donne un tel panorama de la statistique mathématique sur 2 siècles. Alain Desrosières a plusieurs fois raconté que lorsque l'INSEE a organisé les séminaires qui devaient aboutir aux deux tomes de *Pour une histoire de la Statistique* (1977 et 1987), toutes les problématiques et les branches de la statistique administrative y sont passées, mais aucun de ces polytechniciens n'a pensé à l'histoire de la statistique mathématique et des méthodes de traitement de données. L'ouvrage de Benzécri a constitué pour beaucoup d'étudiants de ma génération une entrée privilégiée sur l'histoire de la discipline qu'ils enseignaient et pratiquaient. Pour moi elle a même constitué un déclic de reconversion possible que j'ai pu valider dès 1983-84 en entamant des recherches plus approfondies. Comme Benzécri je n'ai pas trouvé mieux que l'ouvrage de Todhunter et surtout les recueils de Kendall pour commencer à travailler. Mais bien vite les choses ont évolué et dans les quelques années suivantes ont paru des ouvrages qui devaient révolutionner l'historiographie de la statistique mathématique. Je pense en particulier aux deux tomes issus du séminaire de Bielefeld (Kruger et al. 1977 et 1987), au bel ouvrage longtemps indépassable de Stigler (1986), à *The emergence of probability* de Hacking (1975), et aux thèses séminales de Mackenzie (1981) sur *Statistics in Britain.* Ces ouvrages ont reconfiguré le paysage et ouvert la voie aux publications de Brian (1990) Desrosières (1993), Armatte (1995) et bien d'autres ensuite.

Or cette production soudaine de matériaux nombreux mais surtout de réflexions réarrangées autour de problématiques nouvelles – l'Etat et les

Statistiques, La Révolution probabiliste, Homme moyen et variabilité…- ont construit une nouvelle historiographie dont les questions aussi bien que les réponses ont mis à mal les approches du petit ouvrage de Benzécri.

Que pouvions nous reprocher à ce petit livre après l'avoir bien utilisé ? Principalement de ne pas être une histoire de la statistique, ni une préhistoire de l'Analyse des Données mais plutôt une généalogie rétrospective de celle-ci. En prenant comme point focal ses propres travaux, Benzécri se condamnait à ne prélever dans le flot des productions statistique que de ce qu'il jugeait important pour la branche qu'il cultivait. Comme si la science était un arbre (généalogique ou botanique) dans lequel on pouvait se contenter du seul chemin qui mène à cette branche terminale. Or tout le mode sait aujourd'hui, et Benzécri le savait déjà en 1975, que le graphe de l'histoire s'il existe n'est pas un arbre, pas plus qu'en généalogie d'ailleurs : c'est un réseau aux mille et un nœuds que l'historien ou le généalogiste peut parcourir librement le long du chemin qu'il se choisit et qui n'est qu'un possible parmi d'autres. Avec deux cas particuliers qui limitent considérablement ces possibles : partir d'une situation unique d'un lointain passé et développer une généalogie descendante (dont on ne sera qu'un des millions de points terminaux) ou partir de soi et développer une généalogie ascendante. Benzécri lui-même parle d'une "introduction historique à l'AD" (p.58) et ne réfuterait pas le terme plus exact de généalogie rétrospective.

L'approche généalogique n'est pas sans intérêt et elle produit du sens en particulier pour les contemporains. Dissertant sur les façons d'écriture l'histoire de la Statistique, Alain Desrosières (2000) écrivait à propos de l'histoire de sa mathématisation:

"*L'ambition du livre de Jean-Paul Benzécri, Histoire et Préhistoire de l'analyse des données (1982) n'est pas seulement celle d'une historiographie érudite. Pour Pearson comme pour Benzécri, écrire une histoire qui est en même temps la leur et celle de la façon dont ils pensent leurs prédécesseurs et précurseurs (le second parle longuement du premier de façon admirative) est une manière de prendre du recul et de s'inscrire dans une longue histoire que l'on assume tout en la transformant profondément. Ce n'est pas une activité de retraité nostalgique, ni une adresse présidentielle pompeuse dans quelque société savante. Plus tard, cette culture historique et philosophique, et ce souci de se situer dans une histoire longue repensée en profondeur, disparaîtront en partie chez les spécialistes pointus, pour qui, comme dans une chaîne de Markov, tout le passé est supposé résumé dans l'avant dernier état de la science et dans le pas franchi qui leur a permis de passer au dernier état*".

La plus value pédagogique de cette approche dans un objectif de formation des praticiens de l'analyse des données, en leur permettant eux aussi de se situer dans une lignée, est également incontestable.

Mais histoire et généalogie me semblent devoir être distinguées. Qu'est-ce que l'histoire direz vous? Bien des spécialistes de cette discipline ont essayé d'y répondre sans qu'un consensus se dégage vraiment. C'est que cette définition a fortement dépendu des écoles et des périodes. Paul Veyne (1971) répond (sur plus de quatre cent pages) que l'histoire c'est juste un récit qui est vrai. "Puisque tout est historique, l'histoire sera ce que nous choisirons" écrit il. Voilà qui pourrait justifier qu'elle se ramène à une généalogie du point de vue d'un acteur. Mais il se reprend et écrit plus loin que l'histoire est la description de ce qui est ni universel ni singulier, mais *spécifique*, c'est à dire compréhensible, dans les événements humains". L'accent est donc mis sur la radicale altérité et spécificité du passé qu'il convient d'expliquer plus que sur la continuité généalogique. Jacques Revel (2001) essayant de répondre à cette même question penche pour la définition de Raymond Aron – "l'histoire en tant que connaissance est reconstruction et reconstitution de ce qui a été à partir de ce qui est". Mais ce que l'on met en général derrière l'expression "ce qui est" c'est principalement les traces – archives, témoignages – de ce passé qui ont été conservées et ce n'est pas le positionnement d'un auteur-acteur dans le présent. Le même Revel reprend de l'historien allemand Reinart Kosellek l'idée d'une coupure post-révolutionnaire entre deux régimes d'historicité. Dans le nouveau régime d'historicité, "*le passé ne nous parle plus au présent en nous offrant le miroir de situations exemplaires, de conduites modèles, un réservoir de précédents prêts à servir. Il est définitivement passé et il n'est lié à notre présent que par ce qu'il a d'irréductiblement distant et différent. Faire de l'histoire c'est désormais reconstruire la chaîne des raisons qui permettent de nous relier de façon intelligible à ce qui nous est désormais étranger*" [30]. Revel ajoute : "*La pensée du progrès qui domine alors la production historiographique et qui va en nourrir l'imagination de façon privilégiée pendant près de deux siècles, offre certes un moyen de référer l'après à l'avant; mais elle pose précisément qu'ils sont distants et disjoints*". Or il est évident que l'histoire des sciences telle que la pratique Benzécri n'a pas cette propriété. Elle cultive non pas l'étrangeté et la distance mais la similarité des démarches des statisticiens du siècle passé vus comme des précurseurs de l'analyse des données.

Et sans cette altérité du passé, sans la reconnaissance qu'il est différent du nôtre, la reconstruction peut s'éloigner beaucoup du réel vécu par les anciens. En effet, le choix de l'optique généalogique ascendante - se prendre pour point focal - n'est pas innocent. Il produit évidemment des déformations et des biais que nous avons bien identifiés dans le compte rendu de HPAD : risques de téléologie, risques de rétro-projection des problématiques et concepts d'aujourd'hui sur les événements d'hier ; risques d'anachronismes, de sur-interprétation, de filtrage des événements et des textes, d'incompréhension des

[30] J. Revel, Les sciences historiques, in J.M. Berthelot, *Epistémologie des sciences sociales*, PUF, 2001.

catégories et motivations "indigènes"…. Celui qui sait déjà n'est pas historien, or le pratiquant d'une généalogie sait déjà qu'elle mène à lui.

Ajoutons que la pratique historique a elle-même évolué depuis cette coupure révolutionnaire dans un sens qui n'est pas favorable à l'histoire d'un champ par ses acteurs d'aujourd'hui. La célèbre querelle de 1903 entre l'historien Seignobos et le sociologue Simiand débouchait sur une mise en cause de certaines idoles des historiens et sur une nouvelle ère incarnée par l'Ecole des Annales. Ses caractéristiques? Le passé qui ne suffit pas à constituer l'objet de l'histoire, l'individu exceptionnel et l'événement qui sont moins intéressants que les structures qui dominent la longue durée, la reconstruction du passé qui ne suffit pas si elle n'est pas accompagnée d'une élaboration raisonnée et d'une explication (Le fait historique ne se donne pas, il est construit par l'historien), et l'enrôlement des sciences sociales dans ces tentatives d'explication. Les Annales se sont peu occupées d'histoire des mathématiques. Mais les historiens des mathématiques ont été largement influencés par ces thèses.

Tout autant certainement que par les thèses de Kuhn réduisant le progrès à une suite de paradigmes et de révolutions lui ôtant tout caractère de continuité. Un paradigme est inséparablement un certain agencement conceptuel et une certaine organisation sociale de la science. L'école de Bourdieu va dans le même sens en définissant le champ scientifique comme "*système de relations objectives entre les positions acquises, et lieu d'une lutte de concurrence qui a pour enjeu scientifique le monopole de l'autorité scientifique inséparablement définie comme capacité technique et comme pouvoir social*". Comment dès lors être juge et partie dans ce jeu concurrentiel qu'est le champ scientifique?

La nouvelle école des "sciences studies" fondée sur une approche anthropologique va encore plus loin dans cette réunification de l'histoire, de la sociologie et de la philosophie. Elle ne sépare plus le texte du contexte, le contenu cognitif produit des conditions sociales de sa production. Elle restitue à l'homme de science toute sa complexité stratégique (choix de principes, d'objectifs, de méthodes, d'alliés..) qui mêle éléments naturels et culturels, humains et non humains. Adoptant dans son programme fort quelques principes comme l'impartialité vis-à-vis de la vérité ou de l'efficacité des dispositifs, la symétrie du traitement entre les perdants et les gagnants d'une controverse, et la réflexivité de ses analyses, la *socio-logique* de Bloor, Woolgar, Shapin, MacKenzie, Latour et Callon initiée dans les années 1980 décrit des controverses, raconte les micro-stratégies de la vie de laboratoire, et décrit sans état d'âme "la science telle qu'elle se fait" dans une approche plus constructiviste que conventionnelle pour ce qui est des valeurs de vérité et d'efficacité.

Dès lors, pour revenir au cas de l'Analyse des données, faut-il penser qu'un acteur présent du domaine est le mieux placé pour en faire l'histoire parce qu'il possède les compétences et les codes de ce champ et qu'il peut plus facilement que tout autre se "mettre dans la peau" de tel ou tel acteur? Sans

doute oui si l'on visitait le passé comme le font les héros du film *Les Visiteurs*, en allant directement y prendre part. Sans doute non si l'on veut que cet homme puisse prendre la place de *tous* les acteurs à tour de rôle. Un acteur essentiel de la science statistique comme Benzécri n'est pas le mieux placé pour étudier froidement et méthodiquement le champ de bataille et les stratégies des acteurs passés de la statistique. Il trahit régulièrement sa volonté de se mettre dans le paysage, de ce situer dans la lignée de savants reconnus glorifiés, de choisir ses pères et ses pairs, de porter des jugements de valeur qui sont les siens au lieu de reconnaître que plusieurs valeurs s'affrontent en telle ou telle occasion, sa volonté de dresser un cordon sanitaire entre programmes politiques et programmes pour mieux développer une vision irénique de la science comme lieu d'émergence du vrai, et sa vision des "données" comme productions d'une nature qui se donne à voir, et pas comme des constructions sociales. Enfin il s'efforce de protéger le pur joyau du savoir des enjeux sociaux et des controverses, qu'il refuse d'affronter - "il ne nous appartient pas de commenter…d'arbitrer ce différend" – sans doute parce que, plus acteur qu'observateur distant, il aurait à entrer dans des conflits d'intérêt et que les conflits font souffrir.

Conclusion.

L'histoire des sciences est une discipline bien trop importante pour la laisser aux scientifiques…que l'on veut étudier!

L'historien d'hier, et plus encore celui d'aujourd'hui, soucieux de substituer une histoire multiple et constructiviste à une généalogie forcément hagiographique, ne peut pas être lui-même partie prenante de la construction scientifique qu'il décrit, dès lors que cette construction n'est pas seulement une activité cognitive de mise en relation de concepts, de formes, de preuves, mais qu'elle possède de nombreuses composantes sociales et stratégiques.

Qu'on ne s'y méprenne pas toutefois. Notre propos ne vise pas à décrédibiliser l'analyse des données, et encore moins de nuire à notre maître J.P. Benzécri, mais seulement de dire que cet ouvrage HPAD reflète autant les orientations de son auteur que les caractéristiques propres des ancêtres qu'il s'est donné.

Car s'il s'agissait maintenant, avec ce même regard d'historien d'aujourd'hui, d'évaluer non pas son essai historique, mais la totalité de son œuvre, l'admiration l'emporterait largement au vu de la véritable *success story* de l'AD: il ne fait aucun doute que l'analyse des données à la française est loin d'être le non événement que certains ont dénoncé. Si on la juge non pas avec les seuls critères du jeu des citations, mais avec ceux de la sociologie des sciences, dans la façon dont elle a déplacé des lignes et imposé son paradigme, c'est une innovation qui a réussi et dont les fructifications continuent à se produire

aujourd'hui. En plein boum du structuralisme mathématique, Benzécri a pris le contrepied de ses collègues Bourbakistes, mais aussi des linguistes anthropologues et sociologues marqués par l'œuvre de Levi-Strauss, et recherchant comme lui des structures fondées sur des invariances simples se donnant comme des formes a priori de l'entendement. Levi-Strauss avait construit son structuralisme avec un certain esprit de système, au sens ancien de théorie complète constituant une doctrine - pensons au *Traité des systèmes* de Condillac - et avec l'aide dit on des mathématiciens, mais contre les statisticiens : il fustigeait les fondements empiriques et statistiques des (bio-, socio-, éco-, psycho-)"*métries*",: "*les mathématiques humaines veulent résolument échapper au désespoir des grands nombres, ce radeau ou agonisaient les sciences sociales perdues dans un océan de chiffres*" écrivait il dans les années 1950. L'école de Benzécri aura réussi à conjuguer la statistique avec la recherche de structures, pensées dans un sens plus moderne comme ensembles de positions et relations entre des caractérisations. Son point de vue épistémologique empiriste aura su redonner toute sa noblesse à des mots du XIXe siècle, au mot "observation", au mot "donnée"[31]. Allié objectif de Tukey et de sa statistique exploratoire il aura réaffirmé le message sceptique de Bertrand et de quelques autres sur la vanité du modèle probabiliste normal, et de tous les modèles quand on pratique une science d'observation pré-Newtonienne dans laquelle le mot loi n'a pas sa place. Comme un ingénieur, il aura redonné sa légitimité au problème sans nier la théorie : "le problème pratique provoque ou aiguillonne le développement des idées théoriques; et le perfectionnement des outils rend paresseuse la spéculation théorique" (2006). Il aura trouvé les mathématiques – ensemblistes, tensorielles, géométriques – qui produisent des représentations de ces structures factorielles, ordinales ou hiérarchiques pouvant être facilement communiquées et discutées. Il aura enfin su saisir toutes les opportunités qu'offraient ces nouvelles machines – les ordinateurs – pour traiter et représenter les données. Le devenir de l'AD qui a été entre autre de s'allier au savoir faire des informaticiens - Bases de données – et des managers - Knowledge Management - pour intégrer ces dimensions en une nouvelle discipline - le Data Mining – montre bien le potentiel de cette approche.

On peut donc penser que Benzécri est le prototype même du savant dont la stratégie globale, appuyée sur un parti pris affirmé, des outils cognitifs, logiciels et matériels bien choisis, a donné à son paradigme toutes les chances de s'imposer. Il a su redéfinir le champ scientifique, *problématiser* la recherche de structures comme un traitement des données d'observation, *intéresser* d'autres acteurs à son projet dans toutes les disciplines et *enrôler* un grand nombre d'étudiants et de chercheurs devenus ses élèves, *mobiliser* des alliés dans

[31] Avec un risque de fétichisme non contrôlé : il n'a pas assez répété comme Bourdieu que les données ne sont pas données…

l'univers du calcul électronique ou de la linguistique et des sciences humaines, pour finalement opérer une *traduction* au sens de Michel Callon (1986), c'est-à-dire un déplacement et détournement des objectifs, des forces et des intérêts de nombreux hommes de sciences auxquels il a dit : faites un détour par ma méthode et ce sera un raccourci pour vous!

Bibliographie

Armatte M., 1988, "La construction des notions d'estimation et de vraisemblance chez Ronald A.Fisher*", Journal de la Société Statistique de Paris* , tome 129, N°1-2, 1988

Armatte M., 1994, "Invention et intervention statistiques : une conférence exemplaire de K. Pearson (1912)", Politix, N°25, p.21-45.

Armatte M., 1995, *Histoire du Modèle linéaire. Formes et usages en Statistique et en Econométrie jusqu'en 1945*, Thèse EHESS, sous la dir. de J. Mairesse.

Armatte M., 2004, « La théorie des erreurs (1750-1820), enjeux, problématiques, résultats, in *Histoires de Probabilités et de Statistiques,* E. Barbin et J.P. Lamarche (ed), IREM, Ellipses.

Armatte M., 2005, « Lucien March: statistiques sans probabilité », *Journal électronique d'histoire des probabilités et de la statistique*, Vol.2/N°1, mars 2005. www.jehps.net.

Benzécri J.P., 1965-66, Leçons sur l'analyse factorielle et la reconnaissance des formes, Rennes

Benzécri J.P., 1969, Statistical Analysis as a tool to make patterns emerge from data, in Wanatabe, Proceeding of Honolulu Conference on methodologies of pattern recognition.

Benzécri J.P., 1973, *L'Analyse des Données*, t.1 : *Taxinomie* ; t.II : *L'Analyse des Correspondances*, Bordas, Paris (1re édition 1973, 2e édition 1976, 3e édition 1980, 4e édition 1982)

Benzécri J.P., 1982, *Histoire et Préhistoire de l'Analyse des Données*, Bordas. (Première parution dans les *Cahiers de l'Analyse des données,* N° 1-2-3-4, 1976 et N°1, 1977)

Benzécri J.P., 1983, L'avenir de l'analyse des donées, *Behaviormetrika,* N°14, p.1-11

Benzécri J.P., 2006, l'analyse des données : histoire, bilan, projets,…, perspective, In Memoriam : Pierre Bourdieu, *Revue Modulad,* N°35, p.1-5 .

Benzécri J.P., 2008, Сижавзз…Si j'avais un laboratoire, Revue Modulad, N° 38, p. 1-12.

Bertrand J., 1889, *Calcul des probabilités*, 1ère ed. 1889; 2ème ed. 1907, 3ème édition, New York, Chelsea, 1971, XLIX + 317 p.;

Bresson D., 1977, Note bibliographique sur Benzécri J.P., "Histoire et Préhistoire de l'Analyse des Données"*, Les Cahiers de l'Analyse des donnée,* dans *Mathématiques et Sciences humaines,* N°60, 1977, p.63-65.

Callon M., 1986, "Eléménts pour une sociologie de la traduction. La domestication des coquilles Saint-Jacques et des marins pêcheurs dans la baie de Saint-Brieuc", *L'Année Sociologique*, 36, 169-208.

Cibois Ph., 1984, *L'analyse des données en sociologie*, PUF, Le Sociologue.

Cordier Brigitte, 1965, *L'Analyse des Correspondances*, Thèse de Doctorat (troisième cycle), Faculté des sciences de l'Université de Rennes, p.1-65 (jury Boclé, Dugué, Benzécri)

De Leeuw J., 2005, Review of Correspondence Analysis and data Coding with Java and R, *Journal of Statistical Software,* sept 2005, vol. 14, Book review 5.

Desrosières A., 2000, Histoire de la statistique : styles d'écriture et usages sociaux, in J.P Beaud et J.G. Prévost, *L'ère du chiffre,* Presses Universitaires du Quebec, p. 37-57.

Fisher R.A., 1940, The precision of discriminant function, *Annals of Eugenics,* 10, p. 422-429.

Gnedenko, 1950, *Cours de théorie des probabilités,* avec complément historique, Moscou et Leningrad..

Hill M.O., 1974, Correspondence Analysis : A neglected Multivariate Method, *Applied Statistics,* 23-3, p.340-354.

Hirschfeld H.D., 1935, A connection between Correlation and Contingency, *Proc. Camb. Phil. Soc.,* 31, p.520-524.

Hotelling H., 1933, Analysis of a complex of statistical variables into principal components, *J. Educ.Psy.,* vol. 24, p.417-441, p. 498-520.

Kendall M. G., Pearson E. S. , 1970, *Studies in the history of Statistics and probability*, vol 1, Griffin, London.

Kendall M. G., Plackett R. , 1977, *Studies in the history of Statistics and probability*, vol 2, Griffin, London

Lebart L., Morineau A., Fénelon J.P., 1979, *Traitement des données statistiques, méthodes et programmes,* Paris, Dunod.

Murtagh F., 2005, *Correspondence Analysis and Data Coding with Java and R,* Préface J.P. Benzécri, Chapman et Hall.

Murtagh F., 2005, Reply to Review by Jan de Leeuw of Correspondence Analysis and Data Coding with Java and R, site Web de l'auteur.

Pagès J.-P., Caillez F., Escoufier Y., 1979, Analyse factorielle : un peu d'histoire et de géométrie, *Revue de Statistique appliquée*, tome 27, N°1, p. 5-28/ Numdam

Pearson K., 1901, On lines and planes of closest fit to a system of points in space, Phil. Mag., p. 559.

Pearson K., 1920, "Notes on the History of Correlation", *Biometrika*, 13, 25-45; repris dans *Studies in the History of Statistics and Probability*, ed.Kendall et Pearson, 1970 , p 185-206

Pearson K., 1978, *The History of Statistics in 17th and 18th century*, Lectures given at Univ.College of London 1921-1933, E.S. Pearson eds., London, Griffin.

Revel J., 2001, Les sciences historiques, in J.M. Berthelot, *Epistémologie des sciences sociales,* PUF.

Rouanet H. et Le Roux B., 1993, *Analyse des données multidimensionnelles*, Paris, Dunod.

Rouanet H. et Lépine D., 1976, A propos de l'analyse des données selon Benzécri, l'Année Psychologique, 76, p. 133-144.

Todhunter I., 1865, *History of the Mathematical Theory of Probability,* réimpr. Chelsea New York, 1949

Van Metter K., Schiltz M-A., Cibois P., Mounier L., 1994, *Correspondence Analysis : A history and French sociological perspective,* in Greenacre M. et Blasius J., *Correspondence analysis in the social sciences,* London, Academic Press, p. 128-137.

Veyne P., 1971, *Comment on écrit l'histoire,* suivi de *Foucault révolutionne l'histoire,* Le Seuil; Collection Points-Histoire.

Volle M., 1980, *Analyse des données,* Paris, Economica.