Prendre les données au sérieux et rêver d'une sociologie autre

Philippe Cibois

Laboratoire Printemps, Université de Versailles-St-Quentin, France

BMS *Bulletin de Méthodologie Sociologique*
2018, Vol. 137–138 70–93

Résumé

J'ai voulu dans ce texte montrer comment une formation en sociologie et des débuts de carrière en informatique pour les sciences humaines m'ont conduit à mettre l'accent sur l'analyse des données et sur la pratique du dépouillement d'enquête (Cibois, 2015a) mais aussi à me poser des questions sur le lien entre la sociologie et d'autres disciplines. J'ai essayé de montrer que la perspective de la linguistique pouvait avoir du sens en sociologie en distinguant des jeux d'oppositions qui définissent des groupes sociaux. J'envisage de poursuivre cette interrogation en travaillant sur la notion d'endogamie et ses liens avec la systématique en biologie.

Mots clés

informatique pour les sciences humaines, sociologie, analyse de données, analyse des correspondances, PEM (pourcentage de l'écart maximum), métiers de la recherche.

***

> « La problématique de mon œuvre » interpelle des scientifiques à la carrière plus avancée et leur demande de revenir sur les grandes étapes de leur travail afin d'interroger le rôle joué par les questions de méthode dans les inflexions/revirements/progressions de leur pensée et de leur lecture de la société ».

Comment répondre à un tel cahier des charges quand on n'a pas la prétention d'avoir fait une œuvre en sociologie mais d'avoir plus prosaïquement aidé les sociologues à traiter leurs données d'enquêtes ? Je vais plutôt essayer de suivre une logique d'exposition chronologique pour montrer comment se sont constituées mes interrogations et les tentatives pour les prendre en compte.

Pourquoi faire de la sociologie ?

Il existait dans les années soixante dans l'Eglise catholique en France une institution qui s'intitulait les « missionnaires diocésains ». A l'origine il s'agissait de prêtres chargés de prêcher dans les diocèses français des sessions intensives afin de réchauffer la foi des paroissiens et qui, pour faire ce travail, pour bien comprendre la population qu'ils allaient rencontrer, avaient adopté des méthodes de la sociologie et utilisaient ce qu'on appelait alors une « sociologie pastorale ».

Dans la paroisse parisienne où j'étais alors jeune prêtre, la monographie sociographique qu'ils avaient élaborée sur le secteur où j'étais m'a paru très éclairante et leur discours tout à fait digne d'intérêt et même tellement passionnant que je demandais des explications supplémentaires. Après la lecture d'un ouvrage d'initiation à la sociologie qui je crois était les *Eléments de sociologie générale* d'Henri Mendras qui dataient de 1963 et que j'avais appréciés, on[1] me fit lire immédiatement ensuite *Les règles de la méthode sociologique* puis *Le suicide*. Comme après peu de temps, je demandais la suite, on m'expliqua que pour devenir sociologue, il fallait suivre la filière universitaire de la première année du Deug d'alors à la thèse de 3eme cycle. L'année universitaire suivante (1968–1969) je commençais donc à suivre les cours de sociologie à Censier.

La sociologie générale était enseignée par un althussérien de strict observance et je me rendis rapidement compte qu'il employait une méthode que je connaissais bien puisqu'il s'agissait de l'exégèse des textes sacrés, ceux de Marx ou d'Althusser évidemment. De fait mon impression ne fut pas très favorable mais il y avait beaucoup d'autres choses à apprendre dans ce Deug qui était voulu alors comme une formation large : l'ethnologie où Jean Guiart nous parla de la Nouvelle Calédonie ; la linguistique où André Martinet nous expliqua la phonologie ; les mathématiques ou était présenté le travail de Marc Barbut sur leur utilisation en sciences humaines.

J'avais plutôt une vision positive des évènements de 68 mais il fallait traiter les faits sociaux comme des choses et non être pilotés par eux. Echaudé par des présentations de la sociologie qui me semblaient essentiellement politiques, je trouvais dans la statistique un enseignement que je pensais à l'époque comme plus scientifique et, lors de ma deuxième année de Deug, je pris tous les enseignements possibles en mathématiques et en statistiques. J'avais rompu définitivement avec l'Église et les perspectives professionnelles me semblaient davantage assurées par ce type de formation. Toujours dans l'objectif de l'approfondissement d'une méthode plus scientifique à mes yeux, je choisis de faire en licence un certificat d'analyse des données, nouvellement créé par des collaborateurs de Jean-Paul Benzécri qui venait de mettre au point les techniques d'analyse factorielle des correspondances. Un stage était prévu dans ce certificat et je fis le miens au Musée des Arts et Traditions Populaires (ATP).

Expériences d'analyse des données

En 1971 le musée des Arts et Traditions Populaires brillait de tous les feux de la modernité : un bâtiment neuf dans un site prestigieux, dans lequel on s'apprêtait à ouvrir une exposition permanente et une galerie d'études selon des principes muséographiques mis au point par Georges–Henri Rivière et qui associaient un objet traditionnel et une méthode scientifique. En un mot le paradis sur terre pour un étudiant qui en franchissait les portes avec émotion pour venir y commencer un stage dans le service informatique. Car ce jeune musée disposait de ce symbole de la scientificité qu'était alors l'ordinateur qui était considéré comme

[1] Il s'agissait de Joseph Debès, spécialiste de l'action catholique ouvrière.

nécessaire pour se livrer à des études formelles et structurales de collections d'objets domestiques et de proverbes par le biais des méthodes nouvelles de la classification automatique et de l'analyse des correspondances. Car j'y venais pour faire un stage d'informatique, pour appliquer sur des vraies données les techniques que j'étais en train d'apprendre dans le certificat d'analyse des données.

Cet enseignement m'avait également permis d'avoir mon premier contact avec l'informatique lors d'un stage d'initiation dans le service de calcul de la Maison des Sciences de l'Homme (MSH). La MSH, comme les ATP, disposait d'un service informatique équipé d'ailleurs du même ordinateur, un 1130 IBM doté de 8K de mémoire à la MSH et de 2K aux ATP. Les raisons en étaient les mêmes : dans des bâtiments neufs, dans des structures neuves, il fallait montrer que la Science de l'Homme avait partie liée avec la recherche la plus scientifique qui soit, celle qui utilise les techniques les plus avancées.

En descendant donc au deuxième sous–sol de la MSH pour y apprendre le Fortran, j'eus le sentiment très vif de participer à une expérience à la fois nouvelle, la découverte scientifique de la réalité sociale, et ancienne, ce regard scientifique sur le monde qui du mythe pythagorique aux expériences scientifiques d'Archimède, s'ancrait dans une antiquité aussi respectable que celle de la tradition judéo-chrétienne et qui semblait l'assumer à mes yeux.

Le Fortran par sa rudesse, se révéla une école de rigueur : qu'il faille, pour afficher un résultat numérique, prévoir le nombre de blancs à sauter, la largeur du libellé à inscrire et tenir compte de la forme numérique du résultat me fit prendre conscience immédiatement que l'ordinateur n'était pas un être mythique doté de tous les pouvoirs mais une bête machine à laquelle il fallait mâcher le travail. Je me mis à la programmation qui emplit rapidement mes jours, et aussi mes nuits comme pour tout programmeur débutant et inexpérimenté. En effet pour les besoins du stage aux ATP je devais procéder à la programmation d'outils de dépouillement d'enquête : tri à plat, tri croisé, recodage, mise en codage disjonctif. Je mis là au point des morceaux de programmation dont certains me servent encore. J'éprouvais du plaisir à la chose car j'y trouvais une synthèse de la science et de l'étude des phénomènes humains.

Cette étude se situait à un niveau modeste : il s'agissait de ré-exploiter une enquête sur le patrimoine qui n'avait été exploitée que sous forme de tris croisés. Inutile de dire que pour le milieu de l'analyse des données où se faisait l'enseignement et le stage, il s'agissait là d'une sous–exploitation grave qu'il fallait pallier par des analyses factorielles. Ce milieu de l'analyse des données dans lequel se trouvaient des anciens élèves de J.-P. Benzécri (Brigitte Le Roux, Georges Oppenheim et Claude Deniau pour l'enseignement, Michel de Virville pour le stage), vivait dans l'enthousiasme de la découverte que l'analyse factorielle des correspondances, récemment mise au point, permettait de prendre en compte en même temps plusieurs questions d'une enquête afin d'en avoir une image globale.

Appliquée à l'enquête sur le patrimoine, je vis apparaitre que l'idée que l'on se fait du rôle du patrimoine dépend de sa possession. Que les cadres supérieurs, qui en ont un, y croyait alors que les cadres moyens qui n'en n'ont pas n'y voyait qu'un principe moral, des valeurs qu'il faut transmettre. Cela me paraissait fort logique et je n'eus pas le sentiment

d'avoir fait avancer de beaucoup la connaissance scientifique des réalités humaines : ce ne fut pas l'avis de M. de Virville qui, insistant sur l'aspect méthodologique de l'utilisation de l'analyse factorielle, me proposa immédiatement d'exposer mes résultats à un colloque scientifique.

Ce colloque, qui eut lieu au couvent de l'Arbresle[2]2 et qui dans l'esprit d'un certain nombre de ses participants devait permettre la confrontation des techniques formelles avec les sciences de l'homme (c'est ce qu'espérait par exemple le créateur de l'analyse de similitude Claude Flament), se révéla entre les mains de son organisateur, J.-P. Benzécri, un instrument à la gloire de l'analyse des correspondances et de la classification automatique. Je présentais mon plan factoriel (mal, j'avais oublié de représenter les axes), Benzécri me posa quelques questions techniques et s'extasia sur la puissance de la méthode. Ce fut ma première rencontre avec lui, personnage déjà mythique à la longue barbe, au béret et aux bottes de caoutchouc, drapé, quelle que soit la température dans un vaste vêtement à la couleur indéfinissable mais qui d'une voix chaude et claire savait poser les questions qui montraient son intérêt profond pour les données qui étaient traitées devant lui.

A la fin de l'année, j'étais bien formé au dépouillement d'enquête en utilisant l'analyse factorielle mais je ne voyais pas trop ce que cela représentait car pour ce qui est de la méthode elle–même, je savais la faire à la main sur de petites données en utilisant règle à calcul et machine mécanique à additionner (bonnes aujourd'hui pour les antiquaires), mais je fus noyé dans l'algèbre linéaire qui ne m'apporta pas beaucoup de lumières sur la technique. J'avais beau ne pas être bloqué face aux mathématiques qui me semblaient le support de toute science, la barre avait été mise un peu haute. Il n'empêche, notre groupe d'étudiants essuyait les plâtres d'un nouvel enseignement et il le faisait avec des enseignants à l'enthousiasme communicatif : des solidarités sont nées qui subsistent encore.

Sociologie quantitative

Le cours de Raymond Boudon l'année suivante fut le premier cours de sociologie qui me parut répondre au programme durkheimien que j'avais adopté à ma première rencontre avec la sociologie. Avec la découverte de la sociologie empirique introduite en France par les ouvrages de R. Boudon et Paul Lazarsfeld (1965, 1966), il me semblait être au cœur du sujet : j'avais sous les yeux une sociologie qui prenait les données au sérieux mais qui me posait quand même quelques problèmes. Certes R. Boudon n'utilisait pas l'analyse des données mais il me semblait normal qu'il l'ignorât : il suffisait de lui en montrer les mérites et c'est pourquoi je voulu en faire une démonstration dans mon mémoire de maitrise passé avec lui.

Mais plus profondément, ce qui me posait encore problème c'était l'usage constant qui était fait de concepts qui me semblaient typiquement être des prénotions durkheimiennes : quand on se posait le problème de savoir si la mobilité sociale avait cru ou décru, il me semblait que les indices les plus sophistiqués qui tenaient compte de la mobilité structurale n'étaient quand même que l'opérationnalisation de l'idéologie nord– américaine de la

[2] Colloque de l'Arbresle sur l'analyse des données : 21–22 avril 1971 (Charraud et al., 1971).

société ouverte. De même quand R. Boudon nous présentait ce qui allait devenir *L'égalité des chances* (Boudon, 1973), je ne pouvais m'empêcher de penser que savoir si l'éducation s'est démocratisée ou non est foncièrement un problème politique et social, non un problème sociologique ni *a fortiori* scientifique, quelle que soit la sophistication des méthodes employées.

J'avais eu auparavant ma dose de « social » et bien sûr de « religieux » dans les institutions ecclésiastiques et je n'avais nulle envie de me replonger dans des problèmes de cet ordre, même avec le regard armé du sociologue. C'est pourquoi après un court passage auprès d'Henri Desroche, qui m'appris ce qu'est le travail intellectuel (Desroche, 1971), je renonçais à me spécialiser en sociologie de la religion où évidemment j'avais un terrain tout trouvé, des introductions, une problématique, une compétence. Je refusais même une spécialisation sur un terrain quelconque et décidais de continuer dans ce qui m'avait séduit, c'est–à–dire le terrain des méthodes d'analyses de données, pour voir en quoi elles pouvaient être utiles au sociologue.

Etude d'une cohorte d'é lèves

Mon mémoire de maitrise fut un premier galop d'essai : je récupérai des données non traitées sur une cohorte d'élèves saisis en fin de troisième et suivis pendant trois ans. Je sortis des services statistiques du Ministère à Vanves avec un paquet de cartes perforées sous le bras et une totale liberté. Comme d'autre part R. Boudon était plutôt un directeur de maitrise libéral, je n'eus pas beaucoup de contraintes et pu faire fonctionner à ma guise les techniques d'analyse de données : je fréquentais assidûment le Centre inter-régional de calcul électronique (CIRCE) du CNRS à Orsay où la carte perforée était alors reine. L'IBM 360 était alors utilisé et j'appris une syntaxe de cartes de contrôle qui allait me servir pendant environ 15 ans jusqu'à l'arrivée de la micro-informatique.

Le problème de la cohorte n'était pas trop compliqué : il y avait des variables de statut social (sexe, origine sociale, etc.) et des variables de comportement comme la filière suivie ou l'année scolaire. Il suffisait d'appliquer la technique alors récente des variables supplémentaires en analyse factorielle. Mais je commençais à me poser un problème qui allait m'occuper de nombreuses fois dans la suite : quel était le statut du plan factoriel obtenu ? On y voyait certes que les modalités relatives aux filières nobles étaient proches des modalités élevées de statut social. De ce point de vue tout était compréhensible, mais quelle était la réalité de cette proximité ? Pour y voir plus clair je commençais par simplifier au maximum l'analyse pour ne plus avoir comme variables actives que les situations successives dans le temps. En projetant en même temps les numéros des individus, je vis alors que les configurations de trois situations (en seconde, en première, en terminale) des diverses filières du système correspondait à un nombre limité de cas de figure. En faisant l'inventaire des individus qui correspondait à ces cas de figure je m'aperçus que j'étais en train de réinventer le tri de profondeur trois. Pour chacune des modalités des trois situations, il s'agissait de faire un inventaire lexicographique de tous les cas observés et de compter ensuite les individus. Pour revenir aux données à partir de l'analyse factorielle, il

suffisait donc, pour chacune des situations répertoriées, de croiser la sous–population obtenue avec les variables de statut. On obtenait ainsi les relations entre variables.

Le retour aux données

En faisant cette recherche, je mis à jour une démarche dont je m'aperçois aujourd'hui qu'elle continue à me guider dans ma manière d'utiliser l'analyse factorielle. En effet, en traitant une enquête par l'analyse des correspondances, on voit qu'un plan factoriel montre des proximités entre des modalités de réponses issues de plusieurs questions. Quelle est la nature de ces proximités, quel est leur caractère de généralité ? Quand on voit une proximité entre filière défavorisée et situation sociale inférieure, doit-on dire que l'analyse des correspondances montre que les défavorisés sont exclus ? Ou plus encore le prouve ? A cette époque je résolus empiriquement le problème par une utilisation des techniques graphiques de Jacques Bertin.

« Sémiologie graphique »

Ce livre était récent (Bertin, 1967). Il rassemblait les résultats obtenus au Laboratoire de graphique de la 6e section de l'Ecole pratique des hautes études (EPHE) de l'époque. J. Bertin y montrait brillamment que la perception visuelle a un pouvoir synthétique d'une puissance telle que des résultats numériques doivent être impérativement traduits en mode graphique si on veut dépasser l'analyse locale. Comme je venais de découvrir l'analyse des correspondances et son mode d'exposition sous forme de plan factoriel, je ne pouvais qu'être séduit par cette prise de position qui était d'ailleurs rendue inattaquable par une multitude d'exemples issus de recherches réelles.

Dans la priorité que j'ai donnée depuis aux expressions graphiques, il y avait plus qu'un hommage mérité à J. Bertin, il y avait un choix d'une méthode qui a été validée également par John Tukey dans son livre *Exploratory Data Analysis* de 1977. On y insiste également sur la représentation graphique des phénomènes et la « boite à moustaches » inventée par cet auteur qui représente une distribution, sa moyenne, sa médiane, son écart-type et éventuellement ses individus déviants est particulièrement célèbre et utile.

R. Boudon me proposa alors d'être son collaborateur technique dans son équipe CNRS du GEMAS[3]. Je commençais par vérifier les chiffres de *L'inégalité des chances* : je mis au point un programme qui permettait de faire varier les paramètres des équations. Les autres travaux concernant la sociologie de l'éducation, ce ne furent plus des travaux de débutant. En même temps, M. Barbut me proposa de faire des travaux de programmation pour des historiens et des archéologues.

[3] Groupe d'Etude des Méthodes de l'Analyse Sociologique.

Premiers travaux informatiques

Il s'agissait de répondre à la demande précise d'un historien, Roland Mousnier qui pensait avoir un problème qui ne pouvait trouver une solution que dans l'informatique. Si cette question n'est pas directement pertinente pour la sociologie, elle l'est pour l'organisation de la recherche utilisant l'informatique et l'expérience acquise sur ce terrain fut utile pour l'application de l'analyse des données à la sociologie.

Le problème semblait au chercheur gigantesque : il avait une nomenclature d'un millier de titres et il cherchait à voir qui se mariait avec qui. Il avait donc besoin d'un tableau de 1000 sur 1000 pour entrer ses données : il alla voir les gens d'IBM et leur posa le problème. Un ingénieur lui répondit qu'un tel tableau ne pouvait tenir que dans les plus gros ordinateurs de l'époque. Il fit un devis que le chercheur transmis à son président d'université qui déclara forfait immédiatement. Il alla pleurer misère chez l'EPHE sa voisine : et après quelques péripéties M. Barbut me confia le dossier.

Spontanément je proposai de réduire les catégories de 1000 à 100 en arguant de l'exemple du code des catégories socio–professionnelles : ce ne fut pas considéré comme une offense au vu de ma bonne volonté évidente mais on m'expliqua qu'une société où la hiérarchie est basée sur le métier peut bien avoir un code réduit de professions, tandis qu'une société d'ancien régime, qui était une société d'ordre, ne pouvait admettre pareille réduction.

L'examen des données traitées ramena le problème à ses justes proportions : il y avait peut–être un million de cas possibles mais il y avait moins d'un millier de mariages à étudier. Il suffisait donc de mettre en ordre lexicographique les dossiers en triant d'abord sur le titre du mari, puis à égalité de ce point de vue sur le titre du père de la mariée. Le problème fut vite réglé à la grande satisfaction du chercheur.

La première conclusion méthodologique que j'en tirai fut que pour pouvoir utiliser correctement l'informatique et les méthodes quantitatives, il faut connaitre à la fois la technique et le problème à traiter. Il faut donc accepter de passer beaucoup de temps pour entrer dans la problématique du chercheur : le chercheur lui, ne fait pas la démarche inverse puisqu'il a sous la main quelqu'un de compétent. On a là la limite du travail « à façon » en informatique et en statistique : pour travailler correctement, il faut que le spécialiste qui dispose de la technicité passe beaucoup de temps à comprendre la problématique du chercheur qui lui n'a pas d'acquisition à faire. Evidemment, les spécialistes traités de cette façon, devant l'asymétrie des investissements ont le sentiment de se faire exploiter et se refusent à « perdre du temps » à écouter le problème du chercheur. La situation redevient symétrique : plus personne ne fait d'effort pour comprendre l'autre et le résultat peut être prévu à l'avance, il sera insatisfaisant pour les deux parties. Le spécialiste aura le sentiment que des méthodes géniales sont mal utilisées par le chercheur et le chercheur pensera que le spécialiste n'a à lui proposer que des méthodes polyvalentes et passe–partout qui ne s'appliquent évidemment pas à son cas, si ce n'est qu'au prix de réductions qu'il juge intolérables.

Un deuxième travail me fut proposé en collaboration avec Jean–Pierre Bardet qui exploitait les rôles fiscaux de Rouen aux XVIIe et XVIIIe siècles. Comme j'avais tiré les leçons de l'expérience précédente, je commençais par me faire expliquer en profondeur le problème à traiter (faire des statistiques et les cartographier), la nature des données et des codages déjà effectués puisque les données avaient déjà été encodées. Ce fut ce dernier point qui me posa d'ailleurs le plus de problèmes : il fallut pratiquement faire un programme de reconnaissance des formes pour identifier tous les cas plus ou moins complexes qui avaient été prévus et créer une base de données exploitable dans la suite par des programmes créés pour la circonstance (si les systèmes de gestion de bases de données existaient déjà, ils n'étaient pas encore disponibles sur le CIRCE à Orsay).

Grâce à une bonne volonté réciproque, le traitement arriva à ses fins (Bardet, 1983) mais du fait de la taille de l'investissement en temps qui nous fut nécessaire, J.–P. Bardet décida ensuite d'apprendre la programmation afin de retrouver la liberté d'exécution qu'un chercheur juge normale. Pour ma part, j'en tirai ma deuxième conclusion méthodologique : plutôt que de forcer un chercheur à passer du temps (à fonds perdu) pour expliquer son problème à un spécialiste, il valait mieux lui faire gagner du temps en lui apprenant la programmation. L'apprentissage est rentable pour le chercheur puisqu'il devient autonome et peut multiplier ses demandes. Il peut même infléchir sa recherche en fonction des premiers résultats, ce qui semble la démarche normale du chercheur. Au contraire quand le travail à façon est fait dans sa manière traditionnelle, le spécialiste commence par demander l'intégralité des travaux à effectuer avant de faire son estimation du coût nécessaire (en temps, en général), qui ne peut plus être modifiée ensuite au grand désespoir du chercheur.

Informatique et sciences humaines

Je commençais à bien réfléchir sur ces problèmes de méthodologie de l'informatique en sciences humaines et de ce fait M. Barbut m'offrit de participer au comité de rédaction de la revue *Informatique et Sciences Humaines* et il me proposa de faire une enquête pour le compte de la Délégation générale à la recherche scientifique et technique (DGRST ; Cibois, 1973) sur l'utilisation de l'informatique dans les sciences de l'homme.

Je circulai dans une cinquantaine d'équipes de recherches utilisant l'informatique en sciences humaines et j'interviewai les chercheurs sur les problèmes rencontrés : j'y vérifiai ce que j'avais déjà vécu pour mon compte, c'est-à-dire que les résultats sont bons quand il y a une interface de connaissances entre le chercheur et l'informaticien. *A contrario* je produisis de nombreux exemples de mauvais résultats quand cette interface n'existait pas. Au comité de rédaction d'*Informatique et Sciences Humaines* ce rapport fut apprécié, en particulier par Henry Rouanet avec qui ce fut mon premier contact.

Pour les besoins de la revue, j'entrai en relation avec divers milieux utilisant l'informatique et en particulier avec un laboratoire qui investissait beaucoup dans le

domaine de la formalisation, le CADA[4] à Marseille. Je fus ainsi amené à assister au colloque de 1972 sur les bases de données archéologiques consacré aux amphores.

Le problème de la formalisation des données relatives aux amphores est intéressant du point de vue de la formalisation en sciences de l'homme en général. A l'origine les amphores sont de simples récipients, la boite de conserve de l'antiquité, mais ce sont pour nous de beaux objets, recherchés comme tels. Cependant la vision des archéologues étant passée au fil des siècles de la recherche des beaux restes de l'antiquité à l'anthropologie et à ses méthodes, les restes d'amphores servent aujourd'hui à repérer les flux économiques.

Pour les besoins du CIL, *Corpus des inscriptions latines*, vaste compilation d'inscriptions commencée au XIXe siècle, un chercheur, Heinrich Dressel, mit au point une typologie des formes d'amphores afin de ne pas avoir à décrire le détail de chaque récipient sur lequel se trouvait une inscription. On peut ainsi parler des "Dressel 2–4", amphores classiques, hautes, ou des « Dressel 20 », amphores ventrues. Cette typologie purement visuelle était considérée comme insatisfaisante et des archéologues collaborèrent avec l'équipe de Jean–Claude Gardin qui avait introduit le souci de la formalisation documentaire, puis de la formalisation des objets en archéologie[5].

Les profils d'amphores furent d'abord saisis d'une manière exhaustive en notant la coordonnée de chaque point de la courbe. Chaque amphore était décrite avec une extrême précision mais il apparut vite que cette masse de chiffres, si elle autorisait un stockage et une reproduction parfaite, ne permettait ni de retrouver les types, ni de les améliorer. Des algorithmes performants ne sortaient rien de ces données, noyés qu'ils étaient sous la masse des chiffres non pertinents.

Pour sortir quelque chose d'intéressant de ces données, il fallait trouver les traits pertinents, sortir de la perspective « étic » (de l'anglais *phonetic*) pour passer à la perspective « émic » (de *phonemic*) et donc passer, analogiquement, de la phonétique à la phonologie ; de la description physique, complète, gigantesque, à la recherche des traits pertinents, ceux qui manifestent les oppositions sur lesquelles viennent se greffer l'effet de sens. Cette manière de faire, inspirée du structuralisme linguistique, donna de bons résultats sur les amphores : il suffisait en effet de considérer quelques rapports de forme comme celui du périmètre maximum sur la hauteur. On voyait que certains types correspondaient à certaines plages de ce rapport et que combiné avec la présence ou l'absence de certains traits on retrouvait et précisait la classification de Dressel.

A travers cet exemple je fus convaincu de la nécessité de propager à d'autres secteurs des sciences humaines, et en particulier en sociologie, ce paradigme linguistique, et l'analyse factorielle me sembla un bon moyen pour y arriver dans la mesure où la perspective « (phon)étique » conduisait à une description précise mais exagérée, redondante, dont une méthode descriptive ample d'analyse des données pourrait venir à bout et suggérer au

[4] Centre d'Analyse Documentaire pour l'Archéologie.

[5] « Bibliographie » (Gardin, 1991).

chercheur une perspective « (phon)émique » c'est-à-dire analogue à la phonologie qui lui permettrait de découvrir les oppositions pertinentes.

J'avais la méthode, j'avais le paradigme, la quête du Graal pouvait commencer et je n'étais pas près d'oublier la leçon de J.-Cl. Gardin. L'article que j'en tirai pour *Informatique et Sciences Humaines* (Cibois, 1975) fut remarqué par Mario Borillo, un collaborateur de J.-Cl. Gardin, qui me proposa de participer à la réorganisation du Service de Calcul de la MSH : ce fut l'expérience du LISH, le Laboratoire d'Informatique pour les Sciences de l'Homme.

L'approximation

A cette époque (1975), j'assistai à une session de formation du CNRS « Mathématiques et sciences humaines », organisée par le CMS, *Centre de Mathématiques Sociales* de l'EHESS, et animée par Georges Guilbaud.

G. Guilbaud, appelé par Claude Lévy–Strauss et Fernand Braudel à créer au sein de la 6e section de l'EPHE un pôle de mathématiques qui reprendrait la mathématique sociale de Condorcet, avait créé le *Centre de Mathématiques Sociales*. Avec la participation de M. Barbut et de Jacques Maitre, il fut dans les années soixante et au–delà un centre actif de diffusion des techniques mathématiques, en particulier en sociologie au sein du Centre d'études sociologiques de la rue Cardinet.

Mathématicien que Lacan consultait (Roudinesco 1986 : II, 564–567), créateur de films mathématiques, G. Guilbaud m'a fait comprendre ce qu'étaient les mathématiques et également ce que doit être un cours, ce mélange de séduction, de mise en scène, d'attention aux manières de voir des auditeurs, cette simplification extrême du problème quand on le domine, ce passage par le calcul qui évite le formalisme mathématique. Il revisita devant nous l'analyse factorielle en la replaçant dans la logique qui est la sienne, celle de la gestion correcte de l'approximation, de l'à–peu– près (Guilbaud, 1985).

En effet, il suffisait de considérer le tableau dont on voulait faire l'analyse factorielle comme un tableau à décomposer en une somme de tableaux simples, connus par leurs marges (de rang un) mais dont le premier soit la meilleure approximation possible (Cibois, 1983a). En analyse des correspondances, c'est le cas avec le tableau des effectifs théoriques sous l'hypothèse d'indépendance entre lignes et colonnes. On peut dire que le tableau se décompose en indépendance (de rang un) et écarts à l'indépendance. On recommence la décomposition avec les écarts à l'indépendance dont on trouve une approximation de rang un grâce à un algorithme des puissances itérées, tout à fait faisable à la main. On extrait ainsi un résumé ligne et colonne qui, par multiplication, donne la première approximation des écarts, et ainsi de suite jusqu'à épuisement du rang de la matrice de départ.

Ce sont ensuite ces résumés, ces facteurs, qui permettent à la fois la reconstruction d'une matrice d'approximation et une représentation factorielle. Si un point ligne et un point colonne sont en proximité angulaire, cela manifeste que dans les premières approximations il y a un écart positif à l'indépendance et donc une liaison. Si c'est une quadrature, il y a indépendance, si c'est une opposition, un déficit.

Avec quelques concepts simples, bien connus déjà des utilisateurs de sciences humaines comme la matrice de rang un analogue de l'indépendance ou la notion de facteur comme résumé d'un tableau de rang un, G. Guilbaud venait de reléguer le formalisme géométrique de l'algèbre linéaire au rang des outils complexes, indispensables au mathématicien mais inutilisables pour le commun des praticiens. Il lui substituait un corps de concepts accessibles à tous ceux qui voudraient faire l'effort de se l'approprier. Non seulement j'avais une bonne méthode mais je disposais maintenant d'un mode d'exposition génial pour la faire comprendre.

Le séminaire de G. Guilbaud devint pour moi un endroit de rêve ou j'appris comment on fait un calendrier, et en particulier pourquoi il y a des hésitations dans le calendrier arabe, comment on compte les vis et les boulons en se servant des logarithmes, comment est faite la gamme, en quoi consistent les rapports de Pythagore et de l'harmonie, en quel sens Archimède connaissait les logarithmes ou au moins le travail sur les exposants : ce que l'on découvre en lisant l'Arénaire, ce que je fis avec attention et où je découvris l'astronomie antique et comment on faisait des expériences de physique dans l'antiquité[6].

L'enseignement supérieur court

Mon premier travail de sociologue professionnel se fit au GEMAS sous la direction de Janina Lagneau : il s'agissait d'un contrat de l'Organisation de coopération et de développement économiques (OCDE) qui consistait à comparer les enseignements supérieurs courts de trois pays de l'OCDE : France (IUT), Grande–Bretagne (Polytechnics) et Yougoslavie (Visa Skola). Il me permit de comparer les résultats que J. Lagneau avait produits pour son travail sur les IUT (Lagneau et al., 1973), avec des données analogues élaborées en Grande–Bretagne et en Yougoslavie. Les difficultés furent nombreuses : le questionnaire anglais était différent, mais les données étaient encodées sur une bande standard. Si pour la Yougoslavie le questionnaire était le même que pour les IUT, le codage informatique était assez exotique. Je fis ainsi l'apprentissage que le plus difficile en analyse secondaire n'est pas seulement de se procurer les données, car les institutions sont plutôt fières qu'on veuille utiliser les données qu'elles ont produites, le plus dur, c'est de les lire car, du fait de la division du travail entre sociologue et informaticien, les descriptions ne correspondent pas toujours à la réalité physique de ce qui est enregistré. Il faut se livrer à un travail de détective, recouper les résultats obtenus avec les résultats déjà publiés, en un mot faire des hypothèses de codage avant de faire des hypothèses de recherche.

Voulant absolument travailler sur les données d'origine et non sur des tableaux croisés faits auparavant, j'arrivai à mes fins et pus ainsi mener une analyse comparative sur des bases bien assurées. Je pus ainsi montrer que contrairement à la stratégie française de l'époque en matière d'enseignement supérieur court, il était assez contradictoire de vouloir augmenter le prestige des IUT pour favoriser leur développement. Les expériences anglaises des Polytechnics et yougoslaves des Ecoles supérieures montraient que l'enseignement

[6] Tome II des *Œuvres d'Archimède* de l'association Guillaume Budé, Paris, les Belles Lettres, 1971.

supérieur court se développe quand il n'est pas choisi pour lui–même mais faute d'avoir pu entrer dans l'enseignement supérieur long sélectif. C'est donc au contraire le prestige augmenté des filières nobles qui permet l'augmentation quantitative des filières inférieures.

L'OCDE apprécia le travail et édita le rapport dans un ouvrage qui fut également traduit en anglais (Cibois et al., 1976).

La représentation en surface des tableaux croisés

Du point de vue méthodologique et comme les commanditaires de l'enquête ne souhaitaient pas voir utiliser d'analyse des correspondances mais en rester aux tableaux croisés, je mis au point un système de visualisation des tableaux croisés que j'appelle la représentation en surface. Ce procédé que j'ai exposé dans *L'analyse des données en sociologie* (1984 : 22–43), est une modification de ce que J. Bertin appelle les « matrices pondérées » (Bertin, 1967). Cela consistait pour lui à représenter en surface l'effectif de toutes les cases d'un tableau : mon apport a été de représenter seulement ce qui est pertinent dans un tableau c'est–à–dire ce qui s'écarte de la valeur théorique sous l'hypothèse d'indépendance. De ce fait, on voit immédiatement sur la représentation graphique si l'on a affaire à une attraction (dans ce cas il y a une surface au–dessus de l'indépendance), à une répulsion, un déficit (surface au–dessous de l'indépendance) ou si l'on est dans la situation d'indépendance (pas de surface représentée). La seule information perdue est ce qui correspond à la moyenne d'une des distributions marginales, mais cette information peut être donnée par un graphique en forme de « camembert »[7].

J'ai dans la suite amélioré ce type de graphique d'abord en le formalisant (Cibois, 1983b) puis en permutant lignes et colonnes du tableau afin de le « diagonaliser »[8]. Ce concept de diagonalisation étant à prendre dans le sens de J. Bertin (1967) bien que finalement cette opération corresponde au même résultat que la diagonalisation d'une matrice de variance/covariance pour en extraire les valeurs propres. En effet la diagonalisation mathématique a pour résultat de ranger l'information pertinente (les valeurs propres) sur la diagonale d'une matrice et pour J. Bertin, l'opération consiste à mettre sur la diagonale du tableau les forts écarts à l'indépendance, simplement en permutant lignes et colonnes du tableau. J'ai montré dans ma thèse (Cibois, 1980 : 153–176) que les deux opérations étaient tout à fait comparables dans leurs résultats et que ce que faisait J. Bertin de manière empirique, l'analyse factorielle le faisait d'une manière formalisée[9].

[7] Les Anglo–saxons, dont certains trouvent nos fromages trop agressifs, parlent de graphique en part de tarte (*pie-chart*).

[8] Technique rendue opérationnelle dans le logiciel Trideux que j'ai mis au point à l'époque et amélioré depuis et qui permet de dépouiller des enquêtes avec les techniques d'analyse factorielle et de régression logistique. Ce programme est accessible librement sur le site personnel de l'auteur.

[9] J.-P. Benzécri a d'ailleurs montré quelles en étaient les raisons (1973, t. I, 261–287).

En utilisant ce résultat, pour traiter un tableau, on commence par faire une analyse factorielle afin de dégager l'ordre du premier facteur qui va servir à réordonner les lignes et les colonnes du tableau d'origine. Ensuite on peut faire une représentation en surface et les grandes masses d'écarts à l'indépendance se trouveront sur une diagonale du tableau[10].

Sur cette question mon apport théorique a été de mettre en avant dans un tableau croisé l'écart à l'indépendance dont j'ai découvert ensuite qu'il jouait un rôle clé :

- dans l'interprétation des tableaux croisés car son signe permet de repérer attractions et rejets ;
- dans l'interprétation des plans factoriels où les correspondances en question se retrouvent sous la forme de conjonction ou d'opposition (et de quadrature pour l'indépendance) ;
- dans l'interprétation des similitudes entre lignes (ou respectivement entre colonnes) qui ont des profils d'écarts similaires dans un tableau et qui se trouvent proches dans un plan factoriel.

A cette époque, j'avais bien vu le rôle clé de l'écart à l'indépendance, mais j'étais resté dans une perspective « morphologique » : je m'intéressais à la forme des écarts. La prise en compte de l'intensité de l'écart sera un apport ultérieur (le PEM, Pourcentage de l'écart maximum). Je formalisais ainsi la pratique habituelle des sociologues quand ils utilisent des tableaux croisés et se servent des pourcentages en ligne et les comparent avec le pourcentage en ligne de la marge.

Une réexploitation des données des élèves du Panel DEP (Direction de l'Evaluation et de la Prospective) du Ministère de l'éducation nationale (situation en 1991) pour le compte d'un travail avec François de Singly (1993), m'a permis de mettre au point une nouvelle technique dite des « profils ». En effet la pratique du sociologue est de regarder dans plusieurs tableaux croisés le comportement d'une modalité qui l'intéresse et ainsi d'arriver à en rendre compte, à l'expliquer. Cette pratique conduisait en général à ce genre littéraire hautement soporifique qu'était le commentaire de tableaux croisés où une modalité était ainsi comparée de tableaux en tableaux.

Pour formaliser cette pratique, plusieurs décisions doivent être prises :

- définir l'univers de référence : quelles sont les autres modalités de l'enquête que l'on souhaite voir associées avec la modalité dont on veut rendre compte ? En étant large, on peut fixer à environ 200 le nombre de modalités soit de comportement, soit de statut, que l'on désire croiser avec une modalité ;
- définir l'indicateur de liaison : prendra-t-on l'écart à l'indépendance comme dans la première version de mon logiciel Tri-deux ? Ou un indicateur qui tienne compte des marges. C'est cette dernière option qui est retenue à travers le PEM (Cibois, 1993) ;

[10] On peut en option laisser l'ordre d'origine quand il est pertinent.

- définir la fiabilité de la liaison observée : une liaison peut être due à des effectifs non significatifs. Pour avoir une idée de cette fiabilité, on calcule un Khi–deux (muni de la correction de Yates) et l'on indique les seuils standards de significativité (10%, 5%, 1%).

Pour juger de l'intensité de la liaison, je mis au point un indice, le PEM qui s'interprète comme son nom l'indique comme le rapport entre l'écart à l'indépendance observé et l'écart à l'indépendance *qu'il y aurait si la liaison était maximum*. Cet indice a le mérite d'être intelligible pour le sociologue averti : il a le défaut de devoir nécessiter quelques calculs mais comme il est programmé, ce défaut n'en est plus un.

Je pense aujourd'hui que j'ai commis une erreur de communication scientifique en prenant comme indice un « pourcentage » : cela aurait été beaucoup plus noble de prendre la même réalité, mais en proportion, variant de 0 à 1, plutôt que de 0 à 100, ce qui fait trivial (et de lui donner comme nom une lettre grecque plutôt qu'un acronyme). L'avantage cependant c'est qu'un pourcentage parle plus aux profanes qu'une proportion.

Si mon parcours de sociologie de l'éducation s'était révélé fructueux au plan méthodologique car il m'a permis de formaliser la pratique de la lecture des tableaux croisés, de l'automatiser et de l'améliorer ; sur le plan des résultats, je le trouvais assez décevant car toutes ces enquêtes avaient été faites à partir d'interrogations sociales précises et donc limitées.

Cependant, les instruments méthodologiques mis au point permettaient aux chercheurs qui utilisaient des enquêtes de répondre rapidement à des questions précises par le biais du tableau croisé simple ou par une association de tableaux croisés sous forme de profils. Il manquait en revanche encore une perspective apportée par l'analyse des données, celle qui permet d'avoir une vision d'ensemble sur les répondants d'une enquête, d'en faire une typologie. C'est cette perspective qui va être envisagée maintenant.

L'expérience du LISH

Ayant été engagé comme secrétaire scientifique lors de la création du *Laboratoire d'informatique pour les sciences de l'homme* (LISH–CNRS) en 1975, mon rôle fut d'aider les chercheurs fréquentant le service de calcul de la Maison des Sciences de l'Homme à utiliser l'informatique. Je tirai les leçons de mes expériences passées : l'informatique n'est correctement utilisée que s'il y a de la part de ses utilisateurs un investissement tant sur l'informatique elle–même que sur les techniques statistiques utilisées.

Pour ce qui est de l'utilisation de l'informatique, je soutins, en collaboration étroite avec la responsable du service, Monique Renaud, que le centre de calcul ne pouvait correctement fonctionner qu'en libre-service, c'est-à-dire avec un mode de fonctionnement où les utilisateurs créaient eux-mêmes leurs cartes perforées et sortaient eux-mêmes les listings qui en résultaient. Cela pouvait sembler peu de chose mais cette organisation supprimait l'ancien système du travail « à façon » où un ingénieur prend (mal) en charge des travaux qui sont ensuite examinés (sans investissement suffisant) par un chercheur.

La nouvelle organisation permettait au chercheur de faire lui-même les travaux informatiques à la condition qu'il y soit formé : des stages nombreux permirent à des utilisateurs de s'initier aux joies du système d'exploitation IBM 360 du CIRCE et le LISH, grâce à cette nouvelle organisation, devint rapidement un centre de calcul performant et très fréquenté. Cependant, il ne suffisait pas de savoir se servir matériellement de la machine, il fallait encore maitriser les techniques statistiques utilisées. Comme bien souvent les utilisateurs venaient avec des enquêtes qu'ils voulaient dépouiller, j'éprouvais le besoin de mettre au point une méthode d'enseignement et des stages de formation adaptés aux chercheurs afin de leur faire comprendre l'analyse des correspondances qui était la technique la plus à même de voir l'ensemble des données dont ils disposaient.

Je repris à cette fin le mode d'exposition mis au point par G. Guilbaud et avec un jeu d'exemples compréhensibles par des chercheurs de toutes disciplines, je commençais à faire des stages de quatre jours, à partir de 1979 (et jusqu'en 1989) à raison de 20 à 30 chercheurs par stage quatre à cinq fois dans l'année. En effet en plus des stages trimestriels du LISH, je fus amené à faire ce stage pour l'EHESS, pour la formation permanente du CNRS, pour des écoles d'été du CNRS à Grenoble et à Lille, à l'Institut national de recherches pédagogiques, à l'ENS Fontenay, à l'Institut national de recherches agronomiques de Paris et pour diverses autres institutions ainsi qu'à l'étranger.

Mes nombreux contacts avec des chercheurs en situation de recherche me firent voir aussi les limites de l'analyse des correspondances appliquée au dépouillement d'enquête et ce fut pour moi l'occasion de recherches méthodologiques que je proposais dans ma thèse de doctorat de troisième cycle soutenue en juillet 1980 sous la direction de R. Boudon (les autres membres du jury étant MM. M. Barbut et H. Rouanet), sous le titre de *La représentation factorielle des tableaux croisés et des données d'enquête : étude de méthodologie sociologique* (Cibois, 1980).

Thèse de troisième cycle

L'apport de cette thèse s'est fait dans trois directions :

1) **couplage des techniques de J. Bertin avec l'analyse des correspondances.** J'ai tout d'abord formalisé la technique de la représentation en surface des écarts à l'indépendance (étude des limites des indices). J'ai montré que la représentation en surface pouvait être utilisée pour la représentation de la contribution au Khi– deux d'une case et qu'il devenait ainsi possible de représenter un facteur ; que la technique Bertin des matrices pondérées était en quelque sorte une analyse factorielle manuelle et qu'il était tout à fait intéressant de coupler les deux méthodes pour éclairer l'une par l'autre, en particulier, après une analyse factorielle, pour retrouver d'une autre manière dans les données ce qui avait été découvert ;
2) **que ce retour aux données était indispensable pour que le sociologue,** qui avait peu de chance d'être un spécialiste de l'analyse factorielle, puisse l'utiliser sans risque en

prouvant par des techniques ensemblistes, ce qu'il avait trouvé dans un plan factoriel. Je montrais qu'il était possible de construire des « variables idéal-typiques », dont la logique est de compter le nombre d'individus qui possèdent un nombre plus ou moins important de modalités spécifiques d'une partie d'un plan factoriel. Le nom choisi faisait évidemment référence à Weber et à sa notion d'ensemble intellectuellement cohérent même s'il est peu attesté statistiquement.

Dans le même ordre d'idée, c'est là que pour la première fois j'ai envisagé (en utilisant la recherche précédente sur les maximums des indices de représentations en surfaces des écarts), la procédure qui consiste à rechercher quelle pourrait être, pour un tableau à marges données, la situation de dépendance maximum et à comparer par exemple le Khi-deux observé avec le Khi–deux maximum. En utilisant un algorithme fourni par G. Guilbaud et qui a une solution unique quand on dispose d'un ordre sur les lignes et sur les colonnes, j'ai montré qu'on pouvait ainsi construire un tel indice qui permet de voir le degré de liaison mieux que ne le fait le V de Cramér dont c'était un perfectionnement (ce qui allait devenir un peu plus tard le PEM) ;

3) enfin dans une dernière partie j'ai tenté **une analyse sociologique du développement de l'analyse des données**. J'ai montré que les motivations du créateur de l'analyse des correspondances, J.-P. Benzécri étaient sans aucun doute de nature religieuse et que cela expliquait un certain nombre de choses et en particulier pourquoi il insistait tant sur l'examen des facteurs et non pas sur les données elles–mêmes. C'est qu'en effet pour Benzécri, les données sont invalidées par les a priori des chercheurs tandis que les facteurs, fondés sur de vastes ensembles, permettent de retrouver l'essence des choses telle qu'elle a été créée par Dieu. Le rôle du chercheur est de découvrir une réalité préexistante, non d'en construire une par des hypothèses hasardeuses.

Comme cette insistance sur les facteurs était prise pour un résultat mathématique et non comme une position philosophique, élèves et utilisateurs de Benzécri transmettaient cette doctrine comme quelque chose d'indiscutable. L'intérêt de mon analyse fut de montrer qu'on pouvait prendre ses distances avec cette pratique et que le retour aux données était une attitude légitime. On pouvait ainsi mettre l'accent sur des techniques diverses comme la représentation d'un tableau en écart à l'indépendance ou la construction de variables idéal–typiques qui permettent de vérifier ce qu'on a découvert par une analyse des correspondances.

Je montrais également qu'une partie du succès de l'analyse des correspondances venait de ce que j'appelais l'effet « d'homothétie » qui fait que c'est la *structure* des écarts à l'indépendance et non leur *force* qui est mise en valeur par une représentation factorielle. Ainsi de faibles écarts étaient magnifiés par l'analyse alors qu'ils n'auraient peut–être pas été pris en compte si on les avait simplement vus dans un tableau croisé. La conclusion n'était pas que les chercheurs travaillaient sur des écarts insignifiants mais qu'au contraire, il n'était pas nécessaire d'avoir des écarts gigantesques pour dire des choses intéressantes.

Je découpais ensuite cette thèse en plusieurs publications qui mirent l'accent sur la présentation de l'analyse factorielle (Cibois, 1983a), sur la méthodologie de son utilisation en sociologie (Cibois, 1984) et sur les problèmes de l'interprétation de la philosophie de Benzécri (Cibois, 1981).

L'article de la Revue du MAUSS

« Pour une science sociale synchronique » est paru dans la *Revue du MAUSS* du deuxième trimestre 1989. Il m'a permis de faire le point d'une manière précise et de poser les jalons de l'évolution ultérieure. C'est un article « radical » en ce sens qu'il pousse des idées jusqu'à leurs conséquences pour en voir la solidité. L'articulation en est la suivante : elle repose sur l'opposition entre science et savoir.

Un savoir est vraisemblable, il est fait de connaissances empiriques, de concepts qui nous viennent de la société par le biais des idéologies, des croyances ou de savoirs antérieurs : il n'y a pas de solution de continuité entre ce « savoir » du sociologue et le travail de l'expert, ni entre celui–ci et celui du journaliste, puis avec celui de tout un chacun. C'est le fruit du travail du sociologue quand il travaille sur son terrain, avec ses concepts qui sont l'opérationnalisation de son point de vue.

Une science du social au contraire n'est pas vraisemblable mais sûre : son modèle est la phonologie, c'est-à-dire qu'elle cherche à trouver la règle du jeu des oppositions significatives, des traits pertinents à un moment donné. Elle ne peut se constituer que par observation de la manière dont des traits sociaux de situation, de comportement, d'opinion, classent les individus dans des groupes.

Mon projet durkheimien arrive ici à son acmé : il se radicalise. La scientificité du structuralisme linguistique en est le ressort : la compréhension wébérienne est rejetée dans le vraisemblable du savoir, utile certes pour la vie courante, pour la politique, pour l'utilité sociale, mais la barre de la scientificité est placée plus haut. Projet durkheimien mais très critique de Durkheim dont plusieurs ambigüités sont relevées : en particulier est bien montrée la liaison entre son souci de lutter contre les prénotions et son idéologie sociale très datée qui voyait dans la corporation le creuset social qui allait remplacer une religion respectée mais éliminée comme source du lien social.

Un point qui se révèlera essentiel est ici bien mis en lumière : il faut soigneusement distinguer la scientificité qui ne peut être que synchronique, être une règle de grammaire, une règle du jeu, une prise en compte des lois d'un système à un instant donné, en un mot un jeu d'opposition ; il faut distinguer cette scientificité du savoir social qui porte sur les évolutions de la société, qui s'enracine dans l'histoire, la tradition, les idéologies, les croyances. Il n'y pas de science de l'évolution sociale, mais il n'y a pas non plus de « compréhension » des systèmes d'oppositions. On ne peut pas mettre d'idéologie sur une grammaire ni de rationalité sur l'histoire. Il y a un croisement de deux directions qui sont irréductibles l'une à l'autre.

Comme cet article ne posait que des principes, il était indispensable de tenter de les mettre en œuvre à travers une réalisation concrète dont je pourrais tirer des leçons. L'appel d'offre pour une ré-exploitation de l'enquête sur les pratiques culturelles des français de 1989 m'en fournit l'occasion.

Les pratiques culturelles des Français

Il s'agissait de mettre en œuvre un principe théorique sur des données concrètes qui n'avaient pas été conçues à cette fin mais pour tester l'avancée ou le recul de pratiques culturelles. Cependant, comme on avait voulu éviter les effets d'une bonne volonté culturelle, le questionnaire était présenté aux 5000 interviewés comme une enquête sur les loisirs, ce qu'elle était en partie d'ailleurs.

Mon premier travail consista à critiquer l'analyse factorielle publiée dans le rapport du ministère (Donnat, 1990) et qui prétendait montrer que la principale opposition était faite par des activités « branchées » très minoritaires. Je découvris à cette occasion l'existence d'un « effet de distinction » en analyse des correspondances qui vient du fait que s'il existe dans une population étudiée un sous–groupe qui recherche des pratiques distinctives, ce sont ces pratiques qui créeront obligatoirement l'opposition du premier facteur du fait de l'utilisation de la distance du Khi–deux qui met en relief les attractions liées à de faibles effectifs (Cibois, 1992a).

Je mis donc au point une analyse non pondérée qui, sur le même objet pertinent que l'analyse des correspondances, les écarts à l'indépendance d'un tableau de Burt, utilise une distance euclidienne ordinaire. Cela revient à faire de l'analyse en composantes principales sur un tableau centré mais non réduit. Je pus ainsi faire une analyse d'ensemble qui montra que les oppositions majeures étaient le fait du niveau socio-culturel (repéré par le niveau d'étude BAC ou plus, le fait d'être ouvrier ou le fait de ne pas avoir de diplôme ou le CEP – certificat d'études primaires), de la situation dans le cycle de vie (célibataire, marié, retraité) et du sexe. Je croisai l'ensemble des pratiques de l'enquête avec ces huit caractéristiques fondamentales : je regardai pour chacune l'intensité de la liaison entre modalités de pratique et modalités de statut.

Styles de vie

En croisant les huit modalités pertinentes de statut avec les pratiques culturelles ou de loisir, on obtient pour chaque cas une « signature », c'est-à-dire un ensemble de huit valeurs du PEM. On peut de ce fait repérer des similitudes entre signatures, par exemple entre deux émissions de télévisions que sont *Des chiffres et des lettres* et *Les enquêtes du commissaire Maigret* qui ne sont pas du tout de même genre mais qui, on le voit sur cet exemple, sont suivies par des gens aux caractéristiques proches :

Tableau 1. Pourcentage de l'écart maximum entre le statut et les pratiques culturelles ou de loisir

	Niveau social			Position		Sexe		
	BAC	OUV	CEP	CEL	MAR	RTR	MAS	FEM
Maigret	74	91	118	47	95	117	98	102
Chiffres et lettres	54	64	134	57	70	133	98	102

Note de lecture. Les PEM sont ici centrés sur 100 : un PEM de + 17 est égal à 100 + 17 = 117, un PEM de – 26 est égal à 100 - 26 = 74, 100 correspond donc à l'indépendance.
BAC = Bac ou plus, OUV = ouvrier, CEP = pas de diplôme ou Certificat d'études
CEL = célibataire, MAR = marié, RTR = retraité
MAS = Masculin, FEM = Féminin.

On constate que les deux signatures sont très proches : même rejet des diplômés et des plus jeunes, même choix des retraités peu diplômés. Quant au sexe, même petite différence pour les femmes.

Pour traiter l'ensemble et à défaut d'une classification automatique qui à l'époque n'était pas stabilisée sur le critère de Ward comme elle l'est aujourd'hui, j'ai simplifié les profils. Si on regarde par exemple les deux émissions citées plus haut, on s'aperçoit qu'elles sont socialement marquées par l'âge et le faible niveau de diplôme. En revanche l'émission Apostrophes est surtout marquée par les Bac et plus (comme *Océaniques*), tandis que *Ushuaia* et *Musicales* le sont par les jeunes. De même on pourra mettre ensemble *Maguy*, *Avis de recherche, La roue de la fortune* et *Barbara* qui sont spécifiques d'un faible niveau de diplôme et du sexe féminin.

Pour simplifier l'information, on n'en retiendra que la présence ou l'absence de ces marques sociales, le fait qu'il y a un niveau d'attraction suffisant (en général un PEM de þ10%) entre une émission et une caractéristique de statut. Avec cette méthode on peut ainsi classer d'une manière lexicographique les 200 profils dans la mesure où seront semblables les comportements qui sont en attraction positive suffisante avec des modalités de statut. Soit par exemple les deux premiers profils qui correspondent aux jeunes (CEL) diplômés (BAC) soit de sexe masculin soit de sexe féminin. On s'aperçoit que les garçons jouent au tennis tandis que les filles ont des activités artistiques ; que les garçons lisent des quotidiens ou des livres de sciences et techniques, des magazines, des BD ou de la science–fiction tandis que les filles lisent de la littérature classique ; que les garçons vont au concert de jazz ou de rock tandis que les filles vont au spectacle de danse ou à des expositions de peinture ; que les garçons vont au cinéma (sans autre indication), tandis que les filles vont voir des films d'auteurs.

Stéréotypes que toutes ces oppositions ? Il se peut en effet que par le biais d'une enquête on ne fasse que retrouver des rôles intériorisés, mais ce que l'on oublie c'est qu'il s'agit des stéréotypes *en milieu jeune et diplômé*. Si l'on prend la même opposition dans d'autres contextes de statut, on rencontre beaucoup d'autres cas analogues qui permettent de préciser les oppositions.

Cette vision simplifiée est simpliste et peu intéressante. Ce qui fait l'intérêt de la formalisation effectuée, ce passage du (phon)étique au (phon)émique (ou phonologique), c'est :

1) la possibilité offerte de caractériser un ensemble de modalités de statut par un ensemble de comportements, de donner ainsi un statut formel au concept si galvaudé de « style de vie » (et critiqué justement, cf. Herpin, 1986) : dire que le style de vie de la femme d'un certain âge (mariée ou retraitée) sans diplôme c'est de recevoir à diner ses parents, de faire du jardinage, de tricoter et de regarder *Au théâtre ce soir*, tandis que celui de l'ouvrière mariée est de rechercher de nouvelles recettes de cuisine dans des livres et de lire des romans Harlequin ; c'est donner ainsi une précision aux styles de vie qui me semble évidemment plus intéressante que de parler de recentrés, de décalés ou de vigiles.
2) la possibilité de retrouver des oppositions stéréotypées (de sexe, d'âge ou de niveau d'études), mais en voyant le contexte social, parfaitement délimitées, où elles se situent.

A cela on peut répondre qu'il existe des marquages qui dépassent les contextes : que le livre, la cuisine, la couture sont marqués au féminin, que la voiture, la technique, la force physique sont marquées au masculin. C'est quelquefois faux donc sans aucune utilité : ce sont les garçons diplômés qui évoquent les surgelés : ce sont les hommes diplômés qui parlent d'autres lectures que de romans ; ce sont les femmes diplômées qui parlent de gymnastique. On pourra trouver des explications pour chacun des cas cités : on peut toujours rationaliser, pêcher à la ligne des bonnes raisons : ce dont on a besoin c'est d'un cadre général qui permette de situer avec précision chaque observation.

Aller plus loin ?

On peut souhaiter aller plus loin en approfondissant la notion de système général d'opposition dont on a présenté ici une esquisse. Si je ne suis pas arrivé à ce système général, c'est certainement par manque d'imagination mais aussi peut–être parce que les données n'ont pas été constituées dans ce but. Il faudrait refaire une vaste enquête où les différentes pratiques soient bien différenciées : où par exemple on ne mélange pas le bridge, le tarot et la belote sous la seule rubrique « jeu de cartes », où de même on distingue entre les divers « sports collectifs » car *foot, hand, rugby* ou *basket* n'ont pas la même connotation. Il faudrait de plus trouver un répertoire d'activités et de comportements qui soit aussi détaillé selon les milieux sociaux, les âges ou les sexes alors que l'enquête sur les Pratiques culturelles des Français était implicitement centrée sur la « vraie » culture, celle qui est plutôt pratiquée par les milieux situés en haut de l'échelle sociale.

A la suite de cette expérience, j'en suis arrivé à la conclusion que pour que la sociologie puisse devenir une science selon le paradigme de la linguistique, il fallait qu'elle puisse disposer d'instruments d'observation autrement plus puissants que ceux qui nous sont fournis par les instruments actuels, enquêtes qui ne visent qu'un aspect de la réalité sociale : socio–économique, pratiques culturelles, éducation, démographie, consommation. Un outil

d'observation qui envisagerait tous ces éléments à la fois semble hors de portée aujourd'hui car il ne serait qu'un outil de recherche théorique. La société n'a pas cette attitude vis–à–vis de la physique théorique car on sait que ces recherches théoriques ont des retombées pratiques. On peut penser qu'il en serait de même en sociologie fondamentale s'il est permis d'intituler ainsi une telle attitude de recherche.

A défaut de pouvoir réaliser un tel observatoire, un premier pas d'intégration de la sociologie dans le domaine d'une science serait de profiter de l'institution du mariage pour accrocher l'étude des groupes sociaux à la biologie. En effet, la définition d'une espèce en biologie se base sur le fait que mâles et femelles d'un groupe sont interféconds. Cette définition si on s'en tient à elle définit seule l'espèce humaine dans son ensemble et n'est d'aucun secours pour la sociologie. Mais la définition actuelle de l'espèce (Mayr, 1942) a aussi un autre critère que l'interfécondabilité dont la définition est : « groupe de populations naturelles effectivement ou potentiellement interfécondes, isolé par rapport aux groupes similaires au plan de la reproduction ». La notion d'isolement au plan de la reproduction est tout à fait pertinente en sociologie pour définir des groupes sociaux.

L'exemple classique pour faire comprendre de quoi il s'agit est celui–ci : si une population de papillons est coupée en deux par un glacier, les deux populations séparées, bien que théoriquement interfécondes deviennent progressivement des espèces séparées qui ont leur évolution propre. Mon idée est que ce concept serait utile en sociologie en ce sens qu'il conduirait à fonder les groupes sociaux sur un critère sûr, celui du mariage.

Le fait de se marier avec des semblables est vécu comme une exigence dans certains milieux par exemple juifs (Mathieu, 2013) ou de la haute bourgeoisie (Pinçon et Pinçon-Charlot, 1989, 2007) mais elle semble correspondre à une pratique générale comme l'attestent toutes les études démographiques depuis *Le choix du conjoint* (1964) d'Alain Girard[11].

Une étude menée à partir de l'enquête permanente des conditions de vie (EPCV) de 2003 présentée avec Germain Barré au 7e congrès de l'AFS à Amiens a montré que l'on pouvait isoler quatre groupes sociaux,

- groupe 1 : classe populaire issue de l'immigration ;
- groupe 2 : classe ouvrière ;
- groupe 3 : classe moyenne ;
- groupe 4 : classe supérieure ;

Si l'endogamie est forte au niveau des classes extrêmes, elle est moins claire entre la classe ouvrière et la classe moyenne et là encore des données plus précises seules permettraient d'affiner la définition des groupes sociaux en termes d'homogamie. Dans cette nouvelle perspective, on bute à nouveau sur l'absence d'un observatoire sociographique plus complet que ce qui se fait actuellement, bien que le dispositif Elipss (Etude Longitudinale par Internet Pour les Sciences Sociales) puisse être considéré comme

[11] Bozon et Héran, 2012 ; Vanderschelden, 2006 ; Bouchet-Valat, 2014, 2015.

une réalisation de cet observatoire sociographique qui permettra des études plus fines (il faudrait aussi disposer de données socio-linguistiques sur les pratiques langagières des groupes sociaux).

***

A ce stade de l'aventure, je reste à la fois persuadé qu'on doit pouvoir traiter les données sociales comme un langage dont les oppositions sont pertinentes mais en même temps je suis bien conscient qu'il s'agit peut-être d'un rêve né aux beaux jours du structuralisme. Peu importe si ce rêve était réalisable ou non, il a été pour moi l'occasion d'investir dans les domaines de l'analyse des données, de rendre accessibles à tous par des écrits, des enseignements et par un logiciel les méthodes factorielles, de visualiser les résultats d'enquêtes par des techniques diverses de représentation des tableaux croisés, de graphiques triangulaires et de mettre au point un indice de liaison entre modalités adaptable à une case ou à l'ensemble du tableau (PEM simple ou généralisé) : heureuse faute comme disaient les anciens, sérendipité comme disent les modernes.

Aujourd'hui, du fait de mes contacts avec la biologie pour écrire un livre consacré à l'héritage de Linné (Cibois, 2015b), je fais l'hypothèse que les principes de la systématique qui ont fait leurs preuves en biologie animale peuvent nous aider à mettre au cœur de l'identification des groupes sociaux la pratique de l'endogamie. Non pas pour des raisons pratiques, mais pour ancrer sur ce point la sociologie dans une pratique scientifique ancienne ; car celle-ci repose sur la classification ascendante des espèces et c'est la dimension temporelle qui sert d'échelle à la classification. On retombe dans des problèmes d'analyses de données : un vaste programme pour la suite . . .

Références

Bardet JP (1983) *Rouen aux XVIIe et XVIIIe siècles*. Paris : SEDES.

Benzécri JP (1973) *L'analyse des données*. Paris : Dunod.

Bertin J (1967) *Sémiologie graphique*. Paris : La Haye, Mouton.

Bouchet-Valat M (2014), Les évolutions de l'homogamie de diplôme, de classe et d'origine sociales en France (1969–2011) : ouverture d'ensemble, repli des élites. *Revue française de sociologie*, 55(3) : 459-505.

Bouchet-Valat M (2015) *Les rouages de l'amour et du hasard*. Thèse de sociologie soutenue à Sciences-Po Paris sous la direction de Louis-André Vallet le 8 décembre 2015.

Boudon R (1973) *L'inégalité des chances*. Paris : Armand Colin.

Boudon R et Lazarsfeld P (1965) *Le vocabulaire des sciences sociales*. Paris : La Haye, Mouton.

Boudon R et Lazarsfeld P (1966) *L'analyse empirique de la causalité*. Paris : La Haye, Mouton.

Bozon M et Héran F (2012) *La formation du couple*. Paris : La Découverte.

Charraud A, Cibois Ph, Gossiaux JF et de Viville M (1971) *Communications présentées au colloque de l'Arbresle sur l'Analyse des données*. Ronéotype, Centre d'Ethnologie Française.

Cibois Ph (1973) L'utilisation de l'informatique dans les recherches économiques et sociales. *Informatique et sciences humaines* 17 : 29-68.

Cibois Ph (1975) Le colloque de Rome sur les typologies d'amphores : d'une vision synthétique à un discours analytique. *Informatique et Sciences Humaines* 25 : 11-30.

Cibois Ph (1980) *La représentation factorielle des tableaux croisés et des données d'enquête : étude de méthodologie sociologique*. Thèse de doctorat de 3e cycle, Université Paris V.

Cibois Ph (1981) Analyse des données et sociologie. *L'Année sociologique* 31 : 333-348.

Cibois Ph (1983a) *L'analyse factorielle*. Paris : Presses Universitaires de France, coll. "Que sais-je ?", 2095, 5e édition 2000.

Cibois Ph (1983b) Méthodes post-factorielles pour le dépouillement d'enquête. *Bulletin de méthodologie sociologique* 1 : 41-78.

Cibois Ph (1984) *L'analyse des données en sociologie*. Paris : Presses Universitaires de France.

Cibois Ph (1989) Pour une science sociale synchronique. *La revue du MAUSS* 4 : 70-84.

Cibois Ph (1992) *Soixante styles de vie et de pratiques culturelles*. Rapport CERSOF : Université Paris V.

Cibois Ph (1993) Le PEM, pourcentage de l'écart maximum : un indice de liaison entre modalités d'un tableau de contingence. *Bulletin de méthodologie sociologique* 40 :43-63.

Cibois Ph (2015a) *Les méthodes d'analyse d'enquêtes*. Lyon : ENS Editions.

Cibois Ph (2015b) *Parler latin pour classer la nature. L'héritage de Linné*. Saint–Nazaire : Editions Petit Génie.

Cibois Ph et Markiewicz–Lagneau J (1976) *Les étudiants dans l'enseignement supérieur court*. Paris : OCDE.

Desroche H (1971) *Apprentissage en sciences sociales et éducation permanente*. Paris : Editions ouvrières.

Donnat O et Cogneau D (1990) *Les pratiques culturelles des Français*. 1973–1989. Paris : La Découverte, La Documentation française.

Gardin JC (1991) *Le calcul et la raison*. Paris : EHESS.

Girard P (1964) *Une enquête psycho-sociologique sur le choix du conjoint dans la France contemporaine*. Paris : Brodart et Taupin.

Guilbaud GT (1985) *Leçons d'à-peu-près*. Paris : C. Bourgois.

Herpin N (1986) Socio-style. *Revue française de sociologie*. VII : 265-272.

Lagneau J, NetterMet Lorieux J (1973) *Les étudiants des Instituts universitaires de technologie en France*. Rapport OCDE.

Mathieu S (2013) Couple mixte. *In* : Leselbaum J et Spire A (dir.), *Dictionnaire du judaïsme français depuis 1944*. Paris : A. Colin; Lormont : Le bord de l'eau.

Mayr E (1942) *Systematics and the Origin of Species*. New York : Columbia University Press.

Pinçon M et Pinçon-Charlot M (1989) *Dans les beaux quartiers*. Paris : Seuil.

Pinçon M et Pinçon-Charlot M (2007) *Les ghettos du Gotha*. Paris : Seuil.

Roudinesco E (1986) *Histoire de la psychanalyse en France*. TI, 1885–1939, TII, 1925–1985. Paris : Seuil.

Singly F de (1993) Les jeunes et la lecture. *Les dossiers Education et Formations* 24, Ministère de l'Education et de la Culture.

Tukey J (1977) *Exploratory Data Analysis*. Reading, MA : Addison-Wesley.

Vanderschelden M (2006) Homogamie socioprofessionnelle et ressemblance en termes de niveau d'études : constat et évolution au fil des cohortes d'unions. *Economie et Statistique* 398-399