# Gilbert Saporta : un parcours éclectique
## Analyse, fouille, science des données

Gilbert SAPORTA[1]
Conservatoire national des arts et métiers, Paris
Gilles STOLTZ[2]
CNRS — Université Paris-Sud

**TITLE**
Gilbert Saporta : An eclectic career – Data analysis, data mining, data science

**RÉSUMÉ**
Gilbert Saporta est professeur émérite du Conservatoire national des arts et métiers (CNAM), spécialiste de l'analyse des données. Dans cet entretien, il nous dresse d'abord un portrait du monde de la statistique appliquée en France, au tournant des années 1970. L'évocation de son manuel *Probabilités, analyse des données et statistique* (Éditions Technip, 1980), un vrai succès de librairie encore de nos jours, est l'occasion ensuite de discuter de l'évolution sémantique des termes d'analyse, de fouille et de science des données. Enfin, nous revenons avec lui sur l'évolution des sociétés savantes de statistique en France au tournant des années 2000… et constatons qu'il a réalisé un grand chelem de présidence de ces sociétés !

***Mots-clés :*** *souvenirs, évolution sémantique, sociétés savantes.*

**ABSTRACT**
Gilbert Saporta is an emeritus professor at CNAM (Conservatoire National des Arts et Métiers). His main research themes were around data analysis. In this interview, he first recalls the French statistical community in the late 1960s and early 1970s. His best-selling textbook "*Probabilités, analyse des données et statistique*" (Probability theory, data analysis and statistics, published in 1980) is then an opportunity to study the semantic evolution of the concept of data analysis to data science via data mining. Finally, we discuss the learned societies of statistics in France in the last 1990s and early 2000s… and we realize that he achieved the grand slam of being the president of each of them!

***Keywords:*** *memories, semantic evolution, learned societies.*

L'entretien a été mené par échange de courriels entre le 23 octobre 2019 et le 2 mars 2020.

## 1. Parcours académique et professionnel

**GSt :** Où en étais-tu de ta vie vers 20 ans, comment se déroulaient tes premières années d'études supérieures, quels étaient tes projets d'avenir à l'époque ? En particulier (mais pas uniquement), quelle profession et dans quel univers te voyais-tu exercer ?

**GSa :** De 19 à 22 ans, c'est-à-dire de 1965 à 1968, j'étais élève à l'École Centrale de Paris[3]. En première année, j'étais déçu du piètre niveau des cours de mathématiques. Pour y remédier, mon camarade André Bellaïche[4] et moi avions organisé avec un certain succès une inscription simultanée de dizaines d'élèves aux cours de mathématiques de la Faculté des sciences de

1. gilbert.saporta@cnam.fr / désigné par les initiales GSa dans cet entretien.
2. gilles.stoltz@math.u-psud.fr / désigné par les initiales GSt dans cet entretien.
3. Ou Centrale (tout court) dans la suite de ce texte : surnoms de l'actuelle École centrale des arts et manufactures, désormais CentraleSupélec après fusion avec Supélec (surnom de l'École supérieure d'électricité).
4. André Bellaïche quitta l'école en fin de première année pour intégrer l'École normale supérieure. Il est maintenant le gérant des éditions Cassini.

Paris. Je passais ensuite pas mal de temps à militer à l'Union des grandes écoles[5], dont je devins secrétaire général. Une de nos revendications était « pour un savoir professionnellement utile et épistémologiquement fondé », *sic* ! Dans cette période de plein emploi, on ne se préoccupait pas du chômage. Le syndicalisme étudiant fut aussi une belle école de formation, très utile par la suite.

Mon classement s'en était ressenti : à la fin de la deuxième année, j'étais parmi les derniers de ma promotion et je n'avais pas pu obtenir les deux choix d'option de troisième année qui m'intéressaient le plus, à savoir, mathématiques appliquées ou économie. On m'avait alors proposé, ou plutôt imposé, l'option chimie, car elle était bien peu demandée. Je ne l'ai pas regretté car le professeur Henry Brusset qui dirigeait l'option était un visionnaire : pour lui, l'avenir du génie chimique passait par les ordinateurs ; comme il avait compris que je n'étais guère motivé pour la chimie classique, il m'orienta donc vers l'informatique encore balbutiante.

Henry Brusset était un ponte qui cumulait allègrement son poste à Centrale, une chaire à la Faculté des sciences de Paris et un poste au laboratoire de chimie de l'École normale supérieure. Son carnet d'adresses était impressionnant et il put m'obtenir un stage avec deux autres camarades au Département de calcul électronique du CEA[6] à Saclay. C'était là que l'on pouvait utiliser les machines les plus puissantes de l'époque, des IBM 360, ce qui était sans comparaison avec les calculateurs Olivetti à ruban perforé de l'École Centrale.

Notre stage consistait à programmer, en Fortran, l'optimisation de la fabrication d'ammoniac avec une adaptation discrète du maximum de Pontryagin. Le travail était théorique, je ne vis jamais de réacteur chimique. Mais c'est sans doute de cette époque que j'ai gardé une grande sympathie pour ce que l'on n'appelait pas encore la chimiométrie. Le stage fut un peu écourté en mai et juin 1968… mais l'approche nous avait tellement plu que nous avons convaincu Dunod de traduire le livre de Liang Tsen Fan et Chiu Sen Wang sous le nom de *Décisions économiques séquentielles optimales par le principe de Pontryagin*, publié en 1971 et hélas mis au pilon quelques années plus tard.

Je n'étais pas spécialement attiré par une carrière industrielle et comme j'avais eu la chance d'être réformé, je considérais que je pouvais faire une année d'études de plus avant de rentrer dans la vie active (vie active que je n'imaginais pas d'ailleurs concrètement). Je décidais de suivre un DEA[7], et j'hésitais entre l'informatique naissante, mon premier choix, et des mathématiques appliquées – mais lesquelles ? Comme j'avais bien réussi l'examen de probabilités et statistique sans trop assister aux cours, je me disais que ce serait une orientation facile et que je pourrais aisément faire le tour de cette discipline… quelle naïveté !

Je réussis à décrocher des rendez-vous avec des professeurs de Jussieu en plein mois de juin 1968. L'accueil que je reçus en informatique me refroidit : visiblement mon interlocuteur n'aimait pas les élèves des grandes écoles, tandis que Daniel Dugué, directeur de l'ISUP[8], terré dans son bureau, m'accueillit à bras ouverts tout en me prenant à témoin de la chienlit qui régnait sur le campus. Lâchement, je me gardai bien de le contredire. Et c'est ainsi que je m'inscrivis simultanément au DEA de statistique et au cycle supérieur de l'ISUP et me préparais à devenir statisticien.

L'été 68, je partis avec trois amis entreprendre un voyage en URSS qui tenait un peu de

5. Acronyme UGE ; c'était une organisation étudiante créée indépendamment de l'UNEF (Union nationale des étudiants de France) en 1947 et qui a co-existé avec elle jusqu'à la fin des années soixante, où l'UGE a été dissoute dans l'UNEF. Il se trouve qu'en mai 2019, l'UNEF a recréé l'UGE en son sein. [*Note de GSt*]
6. CEA : Commissariat à l'énergie atomique, désormais Commissariat à l'énergie atomique et aux énergies alternatives.
7. DEA : Diplôme d'études approfondies, correspondant de nos jours à un master 2 orienté recherche.
8. ISUP : Institut de statistique de l'Université de Paris ; c'est la plus ancienne formation de statistique en France, fondée en 1922 par le mathématicien Émile Borel.

l'aventure : nous avions acheté une Peugeot 203 d'occasion et notre circuit commençant par la Scandinavie nous mena de Leningrad à Moscou, Kiev, et Odessa, après quoi le retour eut lieu par la Roumanie et la Yougoslavie. Nous nous trouvions à Bucarest en août quand Brejnev décida de mettre un terme au Printemps de Prague par l'envoi des chars soviétiques. Nous assistâmes en direct aux discours de Ceausescu, qui nous parut alors un héros.

**GSt :** Merci beaucoup, Gilbert, pour ce récit haut en couleurs de tes années comme étudiant généraliste ! Ton récit s'achève à un moment-clé : celui où, comme je l'imagine, tu vas découvrir la statistique, en tomber amoureux et y consacrer ta vie professionnelle. À quoi ressemblait l'enseignement de la statistique alors, quels étaient les enjeux de la recherche dans le domaine, et qu'es-tu devenu après cette année de DEA ? J'imagine que ta famille te poussait sans doute à trouver une occupation rémunérée et je suppose que cette dernière a été une bourse pour préparer une thèse…

**GSa :** J'ai en effet découvert la statistique au cours de cette année de DEA. La querelle des anciens et des modernes battait son plein. Le DEA était dirigé par Jean-Paul Benzécri[9] qui prenait son auditoire à témoin des principes qu'il publiera en 1973 : le premier était que « statistique n'est pas probabilité »[10] et le deuxième, que « le modèle doit suivre les données et non l'inverse ». L'enseignement de l'ISUP semblait plus classique. Étant inscrit aux deux, je pouvais panacher les cours de l'ISUP et du DEA, et ma mémoire a du mal à les distinguer. Je découvrais en même temps l'analyse des données, l'estimation et les tests d'hypothèses ; ces derniers me donnèrent du fil à retordre jusqu'à ce que je lise le lumineux chapitre 22 du monumental traité de Kendall et Stuart (1961).

L'enseignement des probabilités à l'ISUP était assez vieux jeu, et sur les conseils d'une camarade (Anne Schroeder, qui devint à la fin des années 80 directrice du centre de recherche de Rocquencourt de l'Inria[11]) j'allais suivre au fond de l'amphi les cours de Jacques Neveu, qui était un pédagogue exceptionnel. Les cours d'informatique se déroulaient sans ordinateur ou presque, et soumettre des paquets de cartes perforées était une épreuve d'endurance.

L'esprit de 68 était bien vivant, en voici quelques exemples. Des expérimentations pédagogiques avaient lieu, comme les travaux dirigés à deux enseignants qui ne nous laissaient pas souffler. Quand la faculté était fermée à cause de divers mouvements, notre groupe de travail se repliait avec son enseignante Janine Ulmo dans l'appartement d'un de nos camarades près des arènes de Lutèce. Avec une petite délégation d'étudiants et chercheurs, nous avions eu le culot d'aller au Ministère demander le remplacement du directeur de l'ISUP, ce qui n'aboutit pas.

Cette première année d'études de la statistique fut passionnante et je rencontrais des personnalités qui allaient m'influencer. Je décerne une mention toute spéciale à Georges Morlat, qui me mit le pied à l'étrier sur de nombreux sujets. Conseiller scientifique à EDF, disciple de Leonard J. Savage, auteur de l'article « Statistique » de l'*Encyclopædia Universalis*, il transmettait son enthousiasme pour la recherche et les applications.

Mon stage de DEA se déroula à l'Atelier parisien d'urbanisme (Apur[12]), où je travaillai sur un problème d'optimisation des flux de circulation dans le 5e arrondissement. Les données de comptage de véhicules étaient fournies par des organismes différents (le service de la voirie

9. Qui vient de décéder le 24 novembre 2019.
10. Voir J.P. Benzécri et coll. (1973).
11. Inria : Institut national de recherche en informatique et en automatique ; le centre de recherche de Rocquencourt a été transféré à Paris en janvier 2016. [*note de GSt*]
12. « Association à but non lucratif dont le but est l'étude et le dialogue avec les grands acteurs de la Métropole [de Paris] sur les sujets des évolutions urbaines et sociétales participant à la définition des politiques publiques d'aménagement et de développement » (source : Wikipedia, consulté le 19 novembre 2019) ; l'association a été créée le 3 juillet 1967 par le Conseil de Paris et regroupe actuellement 27 acteurs institutionnels. [*note de GSt*]

de la ville de Paris pour les rues et les « Ponts et Chaussées » pour les ponts), à des moments différents, et n'étaient pas cohérentes. Le nombre de véhicules entrant à un carrefour ne correspondait pas à celui qui en sortait ! Nous dûmes faire des comptages nous-mêmes en mobilisant une petite équipe d'étudiants. Le stage se poursuivit par une étude par sondage de la qualité du fichier du cadastre préalable à sa numérisation. Là encore, il fallut aller sur le terrain pour comparer *de visu* la fiche cadastrale papier avec ce qui existait réellement sur une parcelle.

Après l'année de DEA, il me restait encore à effectuer la deuxième année de l'ISUP pour en avoir le diplôme et je n'étais pas mûr pour une thèse. Je voulais aussi être indépendant, même si mes parents étaient prêts à financer une nouvelle année (mon père avait une bonne situation dans une compagnie d'assurances et j'étais fils unique). J'étais très tenté de devenir enseignant.

Je m'ouvris de ce projet à Françoise Laurant, l'une de mes professeurs de l'ISUP[13]. Ses conseils furent déterminants : elle me parla des IUT[14] qui étaient en plein essor et allaient recruter de nombreux assistants. Un DEA suffisait pour candidater et il n'était pas nécessaire de se soumettre à toutes les épreuves actuelles (qualification, sélection, audition, etc.). Les assistants, qui ont été remplacés par les ATER[15], formaient un corps de fonctionnaires. J'écrivis à quelques responsables d'IUT dans la région parisienne. Deux d'entre eux me répondirent : celui de l'IUT de Cachan, et le formateur[16] du futur département informatique de IUT de Paris[17], situé avenue de Versailles. J'enfourchai ma Mobylette pour leur rendre visite. Les deux me proposèrent un poste. Comme Cachan était un peu loin en deux-roues du nord-ouest de Paris où j'habitais, je choisis l'IUT de Paris et j'y fus ainsi nommé assistant non agrégé au 1er octobre 1969[18].

**GSt :** Ainsi, pendant un an, tu as étudié tout en enseignant un peu ? Quelles matières t'étaient-elles confiées à l'IUT, les étudiants (et les étudiantes ?) avaient-ils un goût pour elles, quelles étaient leur attitude et leurs attentes, eux qui venaient décrocher un diplôme nouvellement créé de l'enseignement supérieur ? Et enfin… qu'est-il advenu au bout de cette dernière année de scolarité à l'ISUP ?

**GSa :** La situation était un peu étrange : enseignant d'un côté, étudiant de l'autre.

A l'IUT, on m'avait confié le cours de mathématiques de première année en amphithéâtre, et je devais recruter des chargés de TD[19]. Je n'enseignai la statistique, les probabilités et un peu de recherche opérationnelle qu'en 1970-71, quand la deuxième année fut ouverte. L'ambiance était excellente : nous avions tous, étudiants comme enseignants, le sentiment d'être des pionniers. Comme enseignants, nous avions des responsabilités étendues et peu de hiérarchie. Très rapidement, des échanges d'expériences se mirent en place avec les collègues chargés des mathématiques dans les départements d'informatique des IUT en régions. Les étudiants savaient que les débouchés en informatique étaient prometteurs et la quasi-totalité se faisaient recruter comme analystes-programmeurs à l'issue des deux ans. Il faut noter qu'à cette époque des débuts des IUT, les filles étaient presque aussi nombreuses que les garçons, ce qui changea plusieurs années après. Les étudiants étaient très bien encadrés et la formule IUT connut un succès extraordinaire : nous avons eu certaines années plusieurs milliers de candidats pour 140 places environ. Je garde un excellent souvenir des quinze ans que j'ai passés à l'IUT de Paris.

13. Elle partit ensuite rejoindre le département de statistique de l'IUT de Grenoble.
14. IUT : Institut universitaire de technologie ; quatre IUT ont été créés à titre expérimental en 1965 et onze autres ont été institués en janvier 1966, au sein d'universités ; ils préparent à des fonctions d'encadrement technique et professionnel dans l'industrie et les services, et dispensent une formation supérieure en deux ans. [*note de GSt*]
15. ATER : Attaché temporaire d'enseignement et de recherche ; ce sont des postes à durée déterminée, typiquement des contrats d'un an, renouvelables seulement un petit nombre de fois. [*note de GSt*]
16. Terme administratif en vigueur à l'époque ; le formateur exerçait des fonctions d'administrateur provisoire.
17. Rattaché à l'Université Paris V, désormais dénommée Université Paris-Descartes.
18. Mon salaire mensuel net était de 1700 F ce qui correspond à 1868 € selon le convertisseur de l'INSEE, plus que celui d'un ATER actuel. Avec quelques heures complémentaires, on vivait très correctement.
19. TD : Travaux dirigés.

A l'ISUP, des normaliens (Jacques Chevalier, Paul Deheuvels) qui avaient suivi le DEA en même temps que moi, nous faisaient désormais cours : ils avaient sur nous l'avantage de connaître une semaine à l'avance les exercices qu'ils nous posaient ! J'avais un peu de mal à me comporter comme un étudiant et j'en avais assez de passer des examens. Mais les cours restaient passionnants et c'est Jean-Pierre Pagès, chercheur au CEA, bien plus que Jean-Paul Benzécri, qui me convertit à l'analyse des données. Il devint plus tard mon véritable directeur de thèse de troisième cycle[20].

À l'issue de ma scolarité à l'ISUP, je me consacrai à l'enseignement, sans vraiment mener de travaux de recherche, mais en assistant à divers séminaires : celui du BURO (Bureau Universitaire de Recherche Opérationnelle) à l'Université Paris VI[21] et le « petit séminaire », groupe peu formel dans l'esprit de 68 qui réunissait des chercheurs comme Jean-Pierre Raoult, Guy Romier, Christian Pozzo et bien d'autres que je ne peux tous citer.

Deux nouvelles expériences d'enseignement se présentèrent l'année 1970-71. Un ami me proposa d'animer des cours-TD pour des étudiants de licence de psychologie à l'Université Paris X de Nanterre. En ce temps-là, l'Université Paris X était desservie par un train de banlieue et la station se nommait « La Folie – Complexe universitaire ». Sur le plan pédagogique, c'est peut-être là que j'appris le plus. Les étudiants (surtout des étudiantes) étaient venus suivre des études de psychologie sur un malentendu qui persiste d'ailleurs aujourd'hui : ils croyaient que la psychologie allait les aider à mieux se connaître et découvraient à la place une discipline scientifique exigeante. Faire passer les notions d'intervalle de confiance, de risques $\alpha$ et $\beta$ à des étudiants qui pensaient ne plus jamais faire de mathématiques et avaient en horreur les équations, n'était pas facile. Mais ce fut gratifiant.

Au cours de l'année 1970, je me trouvais dans le bureau de Françoise Denizot à l'ISUP, quand elle reçut un appel téléphonique urgent de la direction des Mines[22] de Paris : les élèves avaient boycotté l'examen de statistique car un des exercices portait sur un thème que le professeur avait explicitement exclu du programme, sans apparemment en avertir le chargé de TD qui avait conçu le sujet. Il fallait d'urgence trouver quelqu'un pour préparer les élèves à un nouvel examen. Puisque j'étais là, Françoise Denizot me demanda si j'étais intéressé et j'acceptai au culot. J'assurai donc quelques séances de rappels de cours et de résolution d'exercices sous l'œil attentif de Lucien Vielledent, le directeur des études, qui y assistait, ce qui était un peu intimidant. L'examen se déroula bien, mais les élèves persistèrent à demander le renvoi du professeur de statistique, et à l'époque, on cédait volontiers aux pressions des étudiants. Le directeur des études me proposa donc de reprendre l'année suivante les cours de Jean Mothes. À 24 ans, j'allais donc remplacer une personnalité du monde industriel (il deviendra directeur général des sources Perrier), spécialiste du contrôle de qualité et auteur d'un ouvrage de plus de 600 pages (Mothes, 1968) ! J'ai occupé cette charge de cours pendant 13 ans, j'ai pu créer un cours supplémentaire optionnel d'analyse des données. Outre le fait de pouvoir enseigner à un niveau supérieur à des élèves doués, l'ambiance intellectuelle des Mines était très stimulante. Les cours de statistique étaient rattachés au département d'économie et au CGS (Centre de Gestion Scientifique) dirigé de main de maître par Claude Riveline. J'étais invité aux réunions, dont une partie était à chaque fois consacrée à l'étude d'un des cours relevant du département. Chaque enseignant passait sur la sellette et devait défendre ses choix épistémologiques et pédagogiques. Les débats étaient passionnés et passionnants. Lorsque vint mon tour, j'étais un peu inquiet mais il n'y eut guère de discussions : la statistique ne soulevait pas les mêmes ardeurs que l'économie ou la sociologie. Je n'ai jamais retrouvé une telle atmosphère dans les

20. À l'époque, le DEA (première année du troisième cycle universitaire) se poursuivait par un travail d'initiation à la recherche, mené pendant un ou deux ans, et conduisant à la rédaction et à la soutenance d'une thèse. [*note de GSt*]
21. Située à Jussieu dans le 5e arrondissement de Paris et constituant une des universités ayant donné naissance à l'actuelle Sorbonne Université. [*note de GSt*]
22. École nationale supérieure des mines de Paris, désormais Mines ParisTech.

nombreux conseils que j'ai fréquentés, où les aspects administratifs l'emportent souvent sur les problèmes de fond.

**GSt :** Si j'ai bien suivi le récit ci-dessus, ta carrière a dû prendre un tournant au milieu des années 1980 : sans doute le moment où tu es passé professeur au CNAM. Comment et avec qui es-tu revenu plus intensément à la recherche, passé tes premières années d'IUT ? Et comment es-tu devenu professeur au CNAM ?

**GSa :** C'est grâce à Georges Morlat et Jean-Pierre Pagès que je me suis mis à la recherche. Georges Morlat m'a fait fréquenter l'ASU (Association des statisticiens universitaires, ancêtre de la Société Française de Statistique). J'ai participé à mes premières Journées de l'ASU en 1972 à Clermont-Ferrand. Avec Jean-Marie Bouroche et Michel Tenenhaus, j'ai découvert les travaux de statisticiens américains comme J. Douglas Carroll (1939-2011), alors chercheur aux Bell Labs, qui avait proposé en 1968 une généralisation de l'analyse des corrélations canoniques à plus de deux ensembles de variables. Je m'intéressais également aux méthodes de codage numérique des modalités de variables qualitatives (*optimal scaling*) dans l'optique de développer une analyse discriminante sur variables qualitatives un peu différente de celle que Michel Masson avait mis en œuvre. Jean-Pierre Pagès m'orienta vers les travaux d'Yves Escoufier et les opérateurs qui portent désormais son nom. Ce furent les thèmes de ma thèse de troisième cycle soutenue en mai 1975. Internet n'existait pas et la diffusion de la recherche en France se faisait souvent par les thèses que l'on imprimait en un grand nombre d'exemplaires. Inscrit sur la LAFMA[23], je devins maître-assistant, toujours à l'IUT. Je collaborai jusqu'à la fin des années 70 à la COREF, société de conseil créée par Jean-Marie Bouroche et Patrice Bertier où nous développâmes la méthode de *scoring* connue sous le nom de Disqual. C'est à cette époque que Jean-Marie Bouroche et moi-même (1978) écrivîmes le *Que Sais-je* sur l'analyse des données ; il a connu neuf éditions jusqu'en 2005, mais n'a plus été réédité ensuite. En 1977, Jean-Pierre Pagès me fit entrer comme enseignant à l'École nationale supérieure du pétrole et des moteurs (désormais dénommée IFP School).

Pour devenir professeur des universités, il me fallait soutenir une thèse de doctorat d'État. Après plusieurs années de recherche en analyse des données en dimension finie, je décidai d'étudier le cas de la dimension infinie, ce que l'on appellera plus tard l'analyse de données fonctionnelles. Il y avait une grande effervescence sur ces thèmes avec en particulier l'école toulousaine (Philippe Besse, Rachid Boumaza, Jacques Dauxois, Alain Pousse entre autres) autour de Henri Caussinus. Je me rapprochai d'une part de Jean-Claude Deville, dont j'avais fait la connaissance aux Journées de l'ASU de Montpellier en 1975 et qui avait publié un article fondamental sur l'analyse en composantes principales de processus à temps continu, et d'autre part, de Paul Krée, professeur à l'Université Paris VI, qui avait été membre de mon jury de thèse de troisième cycle. Paul Krée accepta d'être mon patron mais me demanda ce que je connaissais en matière d'opérateurs de Hilbert-Schmidt. Devant ma réponse évasive, il me dit « tu reviendras quand tu auras lu les deux premiers tomes de Dunford et Schwartz (1958, 1963) », ce qui me prit quelques mois. Je découvris un monde mathématique étendant de manière très élégante l'algèbre linéaire. Avec Jean-Claude Deville, nous publiâmes en 1979 un article généralisant l'analyse des correspondances multiples à des données à temps continu. Je finis par soutenir ma thèse de doctorat d'État en juin 1981.

Qualifié ensuite aux fonctions de professeur, il me restait à trouver un poste. Mais au début des années 80, il y en avait bien peu de publiés et quelques années passèrent. Je pris en charge des cours à l'ENSAE[24] à partir de 1981 (que j'arrêtai en 2006). J'envisageai un temps de partir à

23. LAFMA : Liste d'aptitude aux fonctions de maitre-assistant.
24. ENSAE : École nationale de la statistique et de l'administration économique, alors située à Malakoff, dans les Hauts-de-Seine.

Bruxelles à l'EIASM[25]. Alain Bensoussan, qui y intervenait, me proposa le montage suivant : me faire détacher au CNRS[26], qui me mettrait à la disposition de l'Inria, qui m'enverrait à Bruxelles… mais la manœuvre échoua !

Puis je reçus courant 1983 un appel téléphonique de Patrick Lascaux, directeur du département de mathématiques et informatique du CNAM et ancien chercheur (et chef du service de mathématiques appliquées) du CEA[27]. Nous ne nous connaissions ni l'un ni l'autre mais Jean-Pierre Pagès lui avait donné mon nom (la filière nucléaire !), car le CNAM avait décidé de publier un poste de professeur des universités avec pour profil « analyse des données ». Je ne connaissais du CNAM que son activité de cours du soir, mais cela me convenait. Tout alla vite et je fus classé premier, puis nommé au 1er février 1984 par un décret de mai 1984 (!). J'y resterai pour le reste de ma carrière.

**GSt :** Pour cette dernière question relative à ton parcours académique et professionnel, j'ai envie de proposer un défi au statisticien disert que tu es : résumer en une page environ les trente années passées au CNAM, de ta prise de poste en 1984 jusqu'à ton éméritat en 2014…

**GSa :** Je pris mes fonctions en février 1984. Un de mes souvenirs marquants fut l'accueil chaleureux de Claude Kaiser, professeur d'informatique, qui, pour faire ma connaissance, m'invita à déjeuner dans un restaurant chinois de la rue de Turbigo, disparu depuis, à l'enseigne du « Mandarin des Arts et Métiers » : cela ne s'invente pas !

La statistique était alors représentée au CNAM par Paul Jaffard, titulaire de la chaire de calcul des probabilités et statistique mathématique, collègue d'une grande courtoisie, auteur d'un bon manuel mais dont le centre d'intérêt était l'algèbre, et par Jacqueline Fourastié, la fille du célèbre économiste Jean Fourastié, enseignante au département économie-gestion. Mais les deux départements de mathématiques et informatique et d'économie-gestion ne collaboraient guère.

Une des particularités du CNAM est son corps propre de professeurs titulaires de chaires qui a échappé avec ceux du Collège de France et du Museum national d'histoire naturelle à la réforme Edgar Faure de novembre 1968. Les professeurs du CNAM étaient recrutés selon une procédure spécifique datant de 1920, requérant l'avis de l'Institut de France, parmi des universitaires et des experts du monde économique et industriel presque sans conditions de diplôme[28]. Lorsque Paul Jaffard partit à la retraite, je fus élu en 1993 sur sa chaire, renommée « statistique appliquée ».

Compte tenu de la demande croissante de formation et des projets de recherche, je n'eus de cesse de demander la création d'une deuxième chaire de statistique, malgré l'incompréhension de collègues qui pensaient qu'il était bien plus confortable d'être seul mandarin. Cela prit quand même six ans : une nouvelle chaire de « modélisation statistique » fut créée en 1999 avec pour titulaire Alain Monfort, un des économètres français les plus réputés, venant de l'INSEE[29] ; nous pûmes ainsi créer un master de statistique par unités d'enseignement capitalisables en cours du soir, le seul du genre en France[30].

En 1997, Sylvie Thiria, professeur d'informatique à l'Université de Versailles Saint-Quentin-en-Yvelines et ex-maître de conférences au CNAM, nous mit en contact avec son université pour

25. European Institute for Advanced Studies in Management : c'est le nom à la fois d'un institut, d'un réseau et d'une société savante dont la mission est de contribuer à l'amélioration de la recherche et des études doctorales en sciences de gestion. [*note de GSt*]
26. CNRS : Centre national de la recherche scientifique.
27. Il retournera au CEA en 1987. [*note de GSt*]
28. La procédure de recrutement des professeurs du CNAM vient d'être modifiée par le décret 2019-1122 du 31 octobre 2019.
29. INSEE : Institut national de la statistique et des études économiques ; c'est l'organisme de statistique publique français.
30. Michel Béra succéda à Alain Monfort en 2010.

monter un DESS[31] devenu ensuite master « Ingénierie de la statistique ». Cette formation initiale, puis par apprentissage, fut une expérience très intéressante, mais un peu hors des missions du CNAM. La collaboration cessa en 2012.

Au début des années 2000, je fus associé de très près par Jean de Kervasdoué, professeur d'économie et de gestion des services de santé, à la création de l'École Pasteur-CNAM de santé publique pour la partie « Méthodes quantitatives ». Un mastère spécialisé finit par ouvrir en 2007 et est un bel exemple de collaboration transdisciplinaire.

À mon départ à la retraite, on comptait une équipe de 10 statisticiens : 3 professeurs et 7 maîtres de conférences, 19 unités d'enseignements (au lieu de 2 à mon arrivée) et 4 diplômes et certifications.

J'avais noté, au hasard de rencontres à la cantine, que des collègues d'autres disciplines avaient une grande culture statistique : Claude Genty (méthodes physico-chimiques d'analyse), André Allisy (métrologie) et Alain Delacroix (chimie industrielle). Avec Alain Delacroix, je participai à partir de 1985 à l'organisation de colloques au CNAM sur les plans d'expérience et l'optimisation. Ces conférences furent parmi les premières du GFC, Groupe français de chimiométrie, qui rejoindra la SFdS en 2008. Je suis toujours membre de son comité scientifique dont j'apprécie l'esprit pragmatique.

Lorsque je fus nommé au CNAM, je choisis délibérément d'y mener toute ma recherche, mais il n'y avait alors aucune activité de recherche en mathématiques ; je me rapprochai rapidement des informaticiens pour rejoindre ce qui allait devenir le CÉDRIC (Centre d'études et de recherches en informatique et communication), un des principaux laboratoires du CNAM avec actuellement près de 170 membres. J'y créai une équipe d'analyse des données avec les premiers maîtres de conférences de la chaire de statistique appliquée et quelques doctorants. En 2004, cette équipe fusionna avec celle de « réseaux de neurones » pour donner naissance à l'équipe MSDMA (Méthodes statistiques de *data mining* et apprentissage), que j'ai animée jusqu'en 2014.

Au cours de mes 30 années d'activité au CNAM, j'ai dirigé 29 thèses, souvent en collaboration avec des entreprises (en particulier, 9 de ces thèses ont eu lieu en mode CIFRE[32]). J'ai en effet privilégié ces thèses développées sur des problèmes industriels, qui non seulement apportent des sujets d'actualité qui ont des retentissements sur la formation, mais contribuent également au financement de l'équipe de manière plus souple que les projets sur appel de candidature.

## 2. Manuel *Probabilités, analyse des données et statistique* et évolution sémantique

**GSt :** Les plus jeunes du comité de rédaction de Statistique et Société, dont je fais partie, t'ont connu d'abord au contact de ton manuel publié chez Technip (*Probabilités, analyse des données et statistique*), une vraie bible à mes yeux de statisticien mathématicien. Je l'ai beaucoup utilisé pour présenter des exemples de tests d'hypothèses à des étudiants mathématiciens, à qui parfois on ne présente que la théorie des tests. Ton ouvrage est concret et va droit à l'essentiel, il permet d'avoir une vue d'ensemble, quitte à approfondir des points ou récupérer les démonstrations dans d'autres ouvrages. (D'ailleurs, tu ne t'en souviens sans doute pas, mais nos premiers échanges ont eu lieu vers 2006 ou 2007, quand je t'ai posé une question à propos

31. DESS : diplôme d'études supérieures spécialisées, qui correspond dans le système actuel à la seconde année d'un master à finalité professionnelle hors recherche ou enseignement.
32. CIFRE : Conventions Industrielles de Formation par la Recherche ; cela signifie que le doctorant est recruté et payé par l'entreprise au bénéfice de qui les travaux de recherche sont menés, et elle obtient pour cela des subventions publiques ; un accord de gestion de la propriété intellectuelle des résultats est signé entre l'entreprise et le laboratoire d'accueil du doctorant, et l'entreprise verse une rémunération au laboratoire, comme y fait allusion Gilbert Saporta dans la suite de sa réponse. [*note de GSt*]

d'une affirmation présente dans cet ouvrage, et détaillée par Kendall et Stuart, 1961…)

Pourrais-tu commencer par nous retracer l'histoire de la rédaction de ce manuel que l'on appelle « le livre de Gilbert Saporta chez Technip » (et même juste « le Saporta ») et qui a connu plusieurs éditions ?

**GSa :** Je ne résiste pas au plaisir de montrer les couvertures des versions successives de l'ouvrage, pour commencer ma réponse (voir Figure 1).

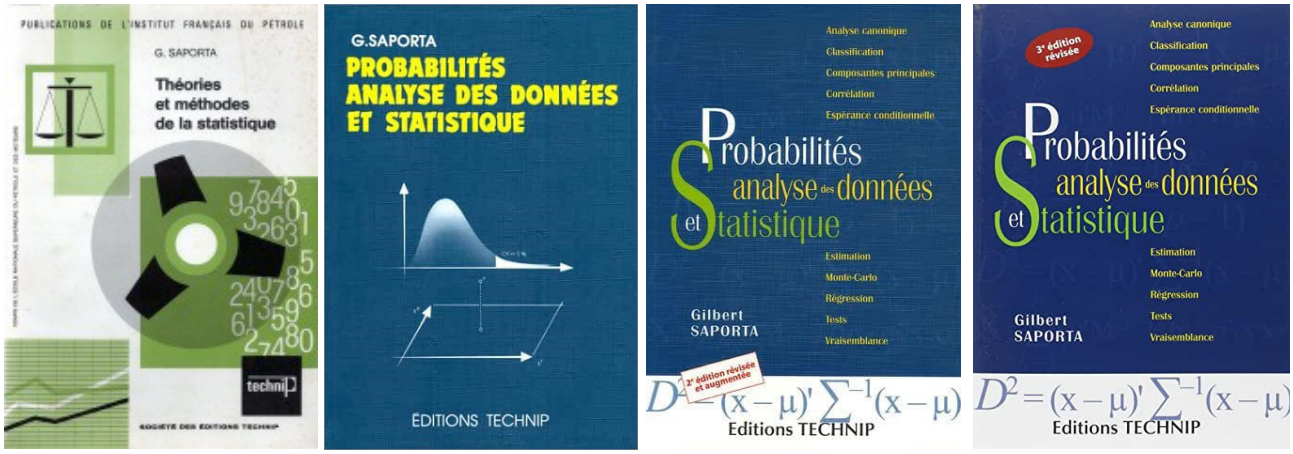


**Figure 1** – *Couvertures successives du manuel de Gilbert Saporta en 1978, 1980, 2006 et 2011 (de gauche à droite)*

Les éditions Technip étaient alors une filiale de l'IFP[33] et à ce titre avaient entre autres missions celle d'éditer les cours de l'École nationale supérieure du pétrole et des moteurs, l'actuelle IFP School, dont nous avons déjà parlé dans cet entretien. C'est ainsi que je fus sollicité pour écrire un manuel issu de mes cours qui fut publié en 1978 sous le titre de *Théories et méthodes de la statistique*. Sans le savoir, j'avais presque plagié le titre du traité *Théorie et méthodes statistiques* de Pierre Dagnelie (1973, 1975).

Mon manuel ayant connu un certain succès, la décision fut prise de le refondre en une édition avec une typographie professionnelle et une couverture solide. C'est ainsi que parut en 1980 la première édition de *Probabilités, analyse des données et statistique* qui avait pour ambition de fournir un cours complet de statistique pour des utilisateurs ayant un niveau licence ou école d'ingénieurs. Le titre fit l'objet de débats avec des collègues : fallait-il séparer ainsi l'analyse des données de la statistique, puisqu'analyser des données est l'objet même de la statistique ? De mon point de vue, cela avait l'avantage de la clarté et de faire comprendre que le lecteur y trouverait bien à la fois des probabilités, de la statistique inférentielle et des méthodes exploratoires multidimensionnelles.

Les retours de lecteurs conduisaient à la publication d'erratas et des corrections lors des retirages. La deuxième édition en 2006 se caractérisa à la fois par un changement d'aspect et des ajouts importants sur le recueil des données (sondages et plans d'expériences), la régression logistique et une introduction aux méthodes d'apprentissage – 120 pages de plus. La couverture sert depuis de modèle pour les ouvrages de statistique publiés par les Éditions Technip. La troisième édition en 2011 est une mise à jour.

Je compte bien aboutir dans un futur proche à une quatrième édition complétée selon les suggestions de collègues et amis qui aimeraient bien trouver des chapitres sur les méthodes

33. IFP : Institut français du pétrole, désormais IFP Énergies nouvelles.

multi-blocs, les modèles linéaires généralisés, etc. Mais chut… En tous cas il n'y aura rien sur les séries temporelles et les processus, car ce n'est pas ma spécialité et il existe de bons ouvrages. Je n'ai qu'un regret, celui de n'avoir pu convaincre mes collègues au CNAM ou ailleurs d'écrire un livre compagnon avec exercices et études de cas.

**GSt :** Sans être indiscret, peux-tu nous donner une idée du volume des ventes des différentes éditions du manuel ? Et as-tu une idée du lectorat ? Le projet initial, tel que tu le mentionnes, était d'écrire un « cours complet de statistique pour des utilisateurs », que tu imaginais dotés d'une licence ou en école d'ingénieurs…

**GSa :** En moyenne, il s'en vend environ 300 exemplaires par an. Ce n'est pas mal pour un ouvrage qui remonte à 40 ans. Cela peut s'expliquer sans doute par un certain manque de concurrence qui lui-même s'explique par le fait qu'écrire des manuels n'est pas valorisé dans la carrière académique en France. Qui sont les lecteurs ? J'avoue ne pas bien savoir quelle est la part d'élèves ingénieurs, d'étudiants en sciences, en informatique ou en économie, mais le niveau semble celui que j'avais visé. Tous ne le conservent pas car on peut en acheter d'occasion chez Gibert !

**GSt :** Ce manuel, et ses évolutions, sont en réalité une occasion pour moi de t'interroger sur l'évolution (et la permanence ?) de notre discipline. Dans les années 70, on parlait d'analyse des données, termes que tu reprends dans le titre du manuel ; plus tard, est venue la fouille de données (« *data mining* ») ; désormais, on a inventé la science des données, après une étape intermédiaire de « *big data* ». D'aucuns disent que tout n'est que statistique. Toi qui as toujours mêlé méthodes statistiques et considérations informatiques, de quel œil peut-être goguenard vois-tu cette évolution ? À vrai dire, tu pourrais commencer par nous préciser ta définition de la statistique et de son champ…

**GSa :** Je n'ai pas la prétention d'ajouter ma définition de la statistique aux innombrables existant déjà. Celle de l'*Encyclopædia Universalis*[34] me va très bien :

> « *Le mot "statistique" désigne à la fois un ensemble de données d'observation et l'activité qui consiste dans leur recueil, leur traitement et leur interprétation.* »

Même si le terme données d'observation demanderait à être étendu aux expérimentations et aux données recueillies automatiquement, cette définition a le mérite d'insister sur le fait que les données sont premières et que la statistique est aussi un métier. Pour moi, la statistique n'est pas une simple branche des mathématiques appliquées, elle utilise les mathématiques, surtout les probabilités, tout autant que l'informatique ; cf. les principes de Jean-Paul Benzécri déjà évoqués, et cette citation de Jerome Friedman (2001) : « *We may have to moderate our romance with mathematics*[35] ». Je n'ai pas la prétention de faire des mathématiques mais plutôt de contribuer à résoudre des problèmes. J'aime bien également rappeler la distinction, certes franco-française, entre *la* statistique, en tant que discipline ou activité, et *les* statistiques, au pluriel, qui désignent des recueils de données. Ce distinguo constitue d'ailleurs l'introduction du rapport de l'Académie des Sciences (2000) consacré à la statistique.

J'ai eu la chance d'assister à la double révolution qu'a connue la statistique : celle des moyens de calcul et celle des données. Les changements ont été inouïs. On est passé tout d'abord de la calculette à des logiciels généraux comme BMDP dans les années 70, puis SAS, et maintenant à des logiciels libres avec R et Python. Tout est devenu plus simple ; des méthodes connues mais

34. Dont j'ai déjà parlé plus haut : Georges Morlat, « Statistique », *Encyclopædia Universalis* [en ligne], consultée le 28 janvier 2020. URL : http://www.universalis.fr/encyclopedie/statistique/.
35. Il faut peut-être réfréner notre romance avec les mathématiques. [*traduction de GSt*]

inapplicables sont devenues facilement utilisables, d'autres sont nées avec le développement de l'informatique. Il n'est plus concevable maintenant de proposer une nouvelle méthode sans fournir le code correspondant. L'explosion de la taille des données s'est produite dans les deux directions : observations et variables. Souvenons-nous qu'encore dans les années 70, les « grands » échantillons étaient ceux au-delà de 30 observations. Mais c'est surtout l'augmentation du nombre de variables, ce que l'on appelle la grande dimension, qui bouleverse les perspectives avec la possibilité de corréler tout avec tout, ce qui n'est pas sans risque. Certains ont prophétisé que les données massives (« *big data* ») rendaient les théories obsolètes et que des corrélations suffisaient : « *correlation is enough* » selon Chris Anderson (2008). On sait maintenant que c'est illusoire et on ne peut que se féliciter des tendances actuelles vers plus d'interprétabilité et du retour de l'inférence causale dans le contexte de données massives.

J'observe avec intérêt l'engouement actuel pour la science des données, après celui pour la fouille de données. Le « *data mining* » insistait sur la découverte de relations et de structures dans des grandes bases de données. La science des données se fixe des objectifs plus vastes incluant la prévision, la visualisation, l'inférence. Le nombre d'offres d'emploi de « *data scientists* » explose et ce métier a été même qualifié comme le plus sexy du XXIe siècle (Davenport and Patil, 2012) ! Pour moi la statistique, c'est la science des données, même si ce point de vue est contesté pour des motifs souvent opportunistes par des chercheurs et praticiens venant d'autres disciplines ; voir l'excellent article de Donoho (2012). Au-delà de développements algorithmiques originaux et puissants, on redécouvre par exemple les vertus de cette bonne vieille analyse en composantes principales.

Il faut prendre garde au risque de ringardisation de la statistique qui serait vue comme une discipline rendue obsolète par le développement de l'apprentissage machine (« *machine learning* ») et de l'intelligence artificielle. Effacer son nom, comme on le voit dans les masters de science des données qui fleurissent, est dangereux et il ne suffit pas de se répéter la phrase attribuée à Bradley Efron : « *Those who ignore statistics are condemned to reinvent it*[36] ». Pour pouvoir traiter les données omniprésentes, il faut avoir le sens de la donnée, et la statistique permet d'acquérir ce savoir-faire.

**GSt :** Comment définirais-tu le « sens de la donnée » et surtout, comment l'enseigner ? Est-ce uniquement un savoir d'expérience ? Plus généralement, n'hésite pas à nous donner des exemples d'écueils que des « *data scientists* » pourraient rencontrer, du fait d'un manque d'un tel « sens de la donnée ».

**GSa :** Les données ne sont pas seulement des enregistrements sur lesquels faire des calculs : elles ont une histoire, un contexte. Elles peuvent être erronées, imprécises, biaisées ou même manquantes. On parle de données brutes qu'il faut prétraiter comme l'on polit des pierres. Visualiser les données est indispensable pour les appréhender et les modéliser. Avoir le « sens de la donnée », c'est aborder tous ces points de vue pour savoir tirer des conclusions valides. Bien sûr, l'expérience est indispensable, mais il est trop facile de la renvoyer aux stages et à l'apprentissage sur le tas. Le « sens de la donnée » peut et doit s'enseigner : il faut faire travailler les étudiants sur des données réelles, éventuellement piégées. Les compétitions et autres hackathons sont très utiles et font maintenant partie des bonnes formations de *data scientists*.

J'ai assisté récemment aux résultats d'un « *challenge* » où, comme il est d'usage, on récompense les meilleures prévisions sur des données de test où la variable cible est cachée. Les données de test suivaient une distribution un peu décalée par rapport aux données d'apprentissage. Les concurrents qui n'avaient pas visualisé les données, avec par exemple une analyse en composantes principales, ne s'en sont pas aperçu et ont fait de mauvaises prévisions.

36. Ceux qui ignorent la statistique sont condamnés à la réinventer [*traduction de GSa*].

J'ajouterais volontiers que la formation au « sens de la donnée » doit aussi inclure la responsabilisation du statisticien quant au recueil et à l'usage des données, en particulier des données personnelles. La déontologie ne concerne pas que les statisticiens publics !

**GSt :** Plus généralement, pourrais-tu nous parler des relations entre statistique et informatique, statisticiens et informaticiens, dans leurs communautés académiques plus vastes, d'une part, et au sein du CNAM, d'autre part ? Sans vouloir trop influencer ta réponse, j'aurais tendance à dire que si, comme tu le soulignes, les méthodes statistiques récentes doivent s'accompagner d'un code (souvent un *package* R ou Python), cela permet à des utilisateurs avancés de les mettre en œuvre, mais ne permet pas réellement que l'industrie et des univers tiers s'emparent de ces méthodes, pour diverses raisons. L'architecture de collecte et traitement des données sont souvent à penser, notamment en termes de bases de données, et les statisticiens ne se penchent pas naturellement sur ces questions.

**GSa :** Vaste sujet… Au XXIe siècle, la statistique ne peut se pratiquer sans des compétences en informatique : codage, bases de données évidemment, calcul parallèle et distribué également. Si les statisticiens du XXe siècle ne s'en préoccupaient pas trop en effet, on insiste maintenant sur le profil équilibré du *data scientist* qui doit être performant en statistique, en informatique, avoir des connaissances métier et savoir communiquer ! Je crois davantage aux vertus du travail d'équipe avec des personnes maîtrisant chacune deux ou trois de ces quatre compétences, qu'à la quête de celle ou celui qui les réunit toutes. Mais quoi qu'il en soit, il est nécessaire que les statisticiens soient formés à l'informatique.

Ta question sur les relations entre les disciplines statistique et informatique renvoie plutôt selon moi à la partie de l'informatique qui se préoccupe d'apprentissage : le « *machine learning* », que l'on peut rattacher à l'intelligence artificielle. Mais je n'utiliserai pas ici ce terme maintenant trop galvaudé. Les informaticiens du *machine learning* développent des outils très efficaces, purement algorithmiques et bien éloignés de la modélisation traditionnelle. Le but n'est pas nécessairement de comprendre, mais de bien prévoir comme le font les réseaux de neurones convolutifs du « *deep learning* » qui remportent des succès spectaculaires dans la reconnaissance faciale, la traduction automatique et bien d'autres domaines. Si certains informaticiens ont pu être tentés de renvoyer les statisticiens à leur gloire passée avec leurs petits modèles, la paix me semble revenue d'autant que des statisticiens et non des moindres se mettent à l'apprentissage et que la frontière entre les deux disciplines dans ce champ est maintenant ténue, tant pour les théoriciens que pour les praticiens. Je note que même en économétrie, on recommande de s'intéresser aux méthodes d'apprentissage (Varian, 2014).

À titre personnel, je me suis toujours senti bien accueilli par les informaticiens. Si le département d'enseignement « mathématiques et informatique » de mes débuts au CNAM s'est rapidement scindé en deux du fait de l'explosion des effectifs en informatique, je suis toujours resté membre du laboratoire d'informatique dans une équipe réunissant statisticiens et neuro-miméticiens et j'espère que cela va continuer ainsi, d'autant que les autres équipes sont de plus en plus confrontées à des problématiques de traitement de données.

Je ne suis pas complètement d'accord avec ta réflexion sur la difficulté pour l'industrie de s'emparer des nouveaux développements avec des *packages* qui seraient trop académiques. Tout d'abord l'industrie change : la numérisation est partout, les données sont une ressource abondante et précieuse, les ingénieurs sont mieux formés aux métiers de la donnée. Utiliser des méthodes avancées peut procurer un avantage compétitif et améliorer les performances économiques (je pense à la maintenance préventive), sans parler du fait que ces programmes sont gratuits. Bien sûr, certains *packages* R ou librairies Python ne sont pas utilisables directement et doivent avoir été validés ; je note que les grands systèmes comme SAS et IBM SPSS développent des interfaces qui permettent de lire et d'exécuter du code R ou Python, ce

qui va faciliter leur diffusion dans l'industrie et fournir du travail pour des professionnels de l'intégration.

**GSt :** Oui, tu as raison, au moins en ce qui concerne les nombreux secteurs d'activités où existe une culture bien ancrée de logiciel statistique (par exemple, dans le secteur pharmaceutique). Pour les autres, les difficultés s'entremêlent : « découverte » (en un sens) de la démarche statistique, et volonté d'utiliser des méthodes récentes dans l'air du temps, la première difficulté surpassant sans doute la seconde.

Nous venons d'évoquer presque cinq décennies de statistique française, avec ses mutations, à travers ton manuel. Je te propose de les évoquer désormais à travers notre société, la Société française de statistique (SFdS).

## 3. Rôle fondateur dans la Société Française de Statistique et rôle dans d'autres sociétés savantes

**GSt :** Je crois que tu as joué un rôle fondateur pour la Société Française de Statistique (SFdS), qui (je viens de l'apprendre pour poser cette question), est récente : elle a été fondée en 1997.

Mais avant que nous ne parlions de cela, je voudrais mieux cerner le contexte et tes motivations. Peux-tu commencer par nous dire quelles sociétés savantes tu fréquentais dans les décennies 1970, 1980 et 1990, et en quoi tu les trouvais utiles à la communauté ? Tu as déjà mentionné l'ASU (Association des statisticiens universitaires) et ses journées annuelles.

**GSa :** Ma première participation aux Journées de statistique remonte à 1972, donc un an après le dépôt des statuts de l'ASU. Ces journées se déroulèrent à Clermont-Ferrand et nous étions si peu nombreux pour l'excursion dans le parc des volcans qu'il a suffi de mobiliser quelques voitures des organisateurs. Depuis, à deux exceptions près, j'ai participé à toutes les Journées de statistique ! J'ai animé pendant 10 ans, à partir de 1978, le groupe « Analyse des données » de l'AFCET[37], où l'on pouvait rencontrer universitaires et praticiens de la recherche opérationnelle. Sur le plan international, je fus élu en 1983 membre de l'Institut international de statistique (en anglais, ISI : *International Statistical Institute*). Auprès de Jean-Louis Bodin et sous la direction d'Edmond Malinvaud, j'ai pris une part active à l'organisation du congrès mondial de 1989 à Paris, qui rassembla un millier de participants. Les congrès de l'ISI sont souvent considérés comme trop généralistes par certains collègues qui préfèrent des rencontres plus ciblées et à faible effectif sur leurs thèmes de recherche. Pour moi, ils étaient des occasions exceptionnelles de rencontrer des statisticiens dans toute leur diversité : où pouvait-on croiser en même temps des mythes vivants comme C.R. Rao, David Cox, des directeurs d'Instituts nationaux, et des spécialistes de tous bords ?

Au sein de l'Institut international de statistique, je suivais les activités de l'IASC (*International Association for Statistical Computing*), à l'origine des congrès européens Compstat (*International Conferences on Computational Statistics*). Dans les années 80, ces congrès permettaient de rencontrer les éditeurs de logiciels et de s'informer des derniers développements informatiques. Ces congrès sont devenus maintenant des conférences classiques de statistique appliquée.

L'utilité des sociétés savantes était évidente dans les décennies citées dans la question : elles étaient des lieux d'échanges et d'information. Maintenant que l'on trouve tout sur Internet, le rôle de source d'information a évolué, mais celui de lieu de rencontre et d'échanges reste fondamental : les jeunes peuvent y rencontrer leurs aînés, nouer des collaborations, confronter

37. Association française pour la cybernétique économique et technique, créée en 1968, disparue en 1998 (faillite).

oralement leurs travaux. On voit d'ailleurs bien le succès de tous les colloques, écoles, séminaires, groupes de travail à côté des journées annuelles.

Si la SFdS a été fondée en 1997, elle est l'héritière de la SSP (Société de statistique de Paris) créée en 1860 et de l'ASU. L'article de Jean-Jacques Droesbeke (2006) retrace l'essentiel de son histoire, mais je vais le compléter par quelques souvenirs personnels. Dans les années 80, l'ASU était essentiellement une association d'universitaires, comme son nom l'indiquait, et l'analyse des données « à la française » occupait une bonne part de ses publications et journées. Quand je fus élu président en 1986, Andreas Zipfel vint me voir pour me demander si l'ASU pouvait accueillir un groupe de statisticiens de l'industrie pharmaceutique, en leur laissant une certaine autonomie. Il avait fait la même demande à la SFB (Société française de biométrie), qui l'avait éconduit. Je vis immédiatement une occasion de développement de l'ASU à ne pas rater. Le conseil était d'accord, mais nous achoppions sur deux points : le U pour « universitaires », qui ne convenait pas aux industriels, et la nécessité de ne pas avoir qu'un seul groupe. Ces deux points furent résolus lors de l'assemblée générale des Journées de statistique de Lausanne, en 1987 : l'ASU devint l'« Association pour la statistique et ses utilisations », selon la suggestion habile de Claude Langrand, et aux côtés du groupe Biopharmacie naquit le groupe Enseignement.

La fin de la décennie 1990 vit la fusion de la SSP, de l'ASU et de la SSF (Société de Statistique de France), qui servait de fédération à diverses associations proches de la statistique. Le paysage était morcelé, et pour nos partenaires internationaux la situation était un peu confuse. Les probabilistes et statisticiens mathématiciens se retrouvaient au sein du groupe MAS de la SMAI[38], créé en 1991, et les statisticiens publics fréquentaient plutôt la SSP dont l'audience déclinait au fur et à mesure que l'âge des participants augmentait. Sous l'impulsion de Félix Rosenfeld, Georges Le Calvé, Ludovic Lebart, Jean-Louis Bodin, Henri Caussinus et moi-même, la fusion fut enclenchée par une assemblée générale commune en octobre 1996. Les autres parties prenantes de la SSF (groupe MAS, la SFC [Société francophone de classification] et la SFB) souhaitèrent garder leur autonomie. Sur la suggestion de Félix Rosenfeld, nous fîmes appel à son notaire pour rédiger un « traité de fusion », tel un contrat de mariage ! Ce traité fut signé et approuvé par les trois associations : la SFdS était née.

Il restait à régler la question de la reconnaissance d'utilité publique : seule la SSP était reconnue d'utilité publique. La nouvelle société pouvait-elle hériter de cette reconnaissance ou fallait-il faire une nouvelle demande qui risquait de prendre plusieurs années pour aboutir ? Avec François Sermier, alors secrétaire général de la SFdS, je me rendis en mars 1997 au Ministère de l'intérieur, qui assure la tutelle des associations et fondations d'utilité publique. L'administrateur civil qui nous reçut fut d'une efficacité rare et nous assura qu'il ferait le nécessaire pour que le caractère d'utilité publique soit transmis de plein droit à la SFdS, puisque la nouvelle société reprenait l'objet social de la SSP. Nous étions comblés et la reconnaissance d'utilité publique intervint en décembre 1998.

**GSt :** Merci beaucoup pour le récit personnel de cette période charnière dans le monde des sociétés savantes de statistique ! Je n'ai pu m'empêcher de consulter la note de Jean-Jacques Droesbeke (2006) et j'en ai déduit qu'en plus de l'ASU, tu avais également présidé, sur des périodes différentes, la SSP, la SSF et la SFdS. Tu as donc, en un sens, réalisé le grand chelem des sociétés savantes de statistique !

**GSa :** Je m'en souvenais bien pour la SSP dont j'ai été le dernier président (années 1995-97) et la SFdS (années universitaires 2000-01 et 2001-02). J'ai consulté la note historique de Félix Rosenfeld (2007) pour vérifier que j'ai bien été président de la SSF en 1995 et 1996, juste avant

38. MAS : Modélisation aléatoire et statistique ; SMAI : Société de mathématiques appliquées et industrielles

Henri Caussinus. Je constate donc que j'ai en effet réalisé ce grand chelem !

**GSt :** Venons-en maintenant aux années 2000. Leur début est marqué par la disparition de Lucien Le Cam, en avril 2000. La SFdS a très rapidement mis en place un hommage, en la « conférence Le Cam », prononcée par un orateur de renom lors des Journées de statistique annuelles. Tu as été président de la SFdS à ce moment : peux-tu nous raconter comment la communauté statistique française a vécu cette disparition et comment cette idée d'hommage a été proposée et décidée ?

**GSa :** Lucien Le Cam était bien sûr connu des statisticiens théoriciens. Beaucoup de jeunes statisticiens connaissaient son nom mais peu avaient lu ses travaux réputés difficiles. Nous savions tous que Lucien Le Cam était une vedette internationale, un français établi et reconnu aux USA à l'instar de Gérard Debreu. L'idée de la conférence en son honneur revient à Marc Hallin de l'Université libre de Bruxelles. Après la première édition de cette conférence Le Cam, où Lucien Birgé avait été distingué lors des Journées de statistique de Nantes, en 2001, et avec la perspective d'une deuxième conférence aux Journées de 2002 à Bruxelles, Marc Hallin mena une intense campagne pour pérenniser cet hommage et le rendre annuel. Ayant été convaincu de l'intérêt de cette conférence qui pouvait faire venir nos collègues théoriciens à la SFdS, je présentai le projet de règlement au conseil de la SFdS fin 2001, qui l'approuva. La conférence Le Cam est maintenant bien établie et la liste des récipiendaires est éloquente.

**GSt :** Toujours pour ces années 2000, je lis dans ta biographie que de 2005 à 2007, tu as également été vice-président de l'ISI (*International Statistical Institute*, Institut international de statistique), dont nous avons déjà parlé. D'ailleurs, tu as écrit plus haut que tu en avais été élu membre, ce qui me laisse perplexe : il ne suffit donc pas de payer une cotisation ? Peux-tu nous parler un peu de cette association internationale, de ses missions, et peut-être également nous confier quelques anecdotes sur l'époque où tu en as été le vice-président ?

**GSa :** L'ISI est une des plus anciennes sociétés savantes internationales ; elle a été fondée en 1885. Au départ, c'était plutôt un club où se retrouvaient les dirigeants des instituts nationaux de statistique. L'ISI s'est ouvert à tous les débats concernant la statistique publique et est même allée au-delà : je pense aux débats ayant eu lieu au tournant entre les XIX$^{e}$ et XX$^{e}$ siècles quant à l'utilisation de ce que l'on appellerait aujourd'hui des échantillons représentatifs pour éviter d'effectuer des dénombrements exhaustifs (pour plus de détails, voir Droesbeke et Tassi, 1997, ainsi que Didier, 2013). À travers ses sept associations[39], l'ISI couvre presque tous les champs de la statistique. C'est une organisation élitiste, il faut en effet être parrainé par trois de ses membres pour prétendre être élu. Mais depuis quelques années, l'ISI a créé une nouvelle catégorie de « *regular members* », qui n'ont qu'à payer une cotisation. Malgré ses efforts, l'ISI ne représente encore qu'imparfaitement la statistique mondiale et reste plutôt occidentale : l'Asie et l'Afrique sont sous-représentées parmi les 4500 membres de l'ISI et de ses associations. Reconnue par l'ONU, l'ISI joue un rôle important de rencontres, de coordination et de promotion de la statistique, en particulier pour les pays en développement. Ses actions en matière de déontologie et d'indépendance de la statistique sont également très importantes.

Pendant plusieurs années le Groupe des membres français de l'Institut international de statistique (GMFIIS) s'est préoccupé de susciter des candidatures de statisticiens français à l'ISI. Le GMFIIS a disparu, mais je recommande aux jeunes collègues d'adhérer à l'ISI : c'est une expérience très enrichissante, même si le fonctionnement de l'ISI peut sembler opaque.

39. En l'occurrence [*note de GSt*] : Bernoulli Society, International Association for Official Statistics (IAOS), International Association for Statistical Computing (IASC, dont nous avons déjà parlé plus haut), International Association for Statistical Education (IASE), International Association of Survey Statisticians (IASS), International Society for Business and Industrial Statistics (ISBIS), The International Environmetrics Society (TIES).

Enfin, je n'ai pas d'anecdotes particulières sur ma période comme vice-président : beaucoup de réunions administratives mais dans un bon esprit de travail. J'ai quand même réussi à monter la première enquête de satisfaction sur un congrès (celui de Lisbonne en 2007) et présenté au Conseil de l'ISI une étude de l'évolution des thèmes de sessions sur huit ans à l'aide d'une analyse textuelle.

**GSt :** Est-ce que je peux évoquer la décennie 2010 en quelques mots ? Je crois que pendant ces années, tu as présidé (et présides encore) la fondation « La science statistique ». Peux-tu brièvement nous parler de cette fondation et de ses actions ?

**GSa :** Cette discrète fondation d'utilité publique a été créée en 1927 pour soutenir l'ISUP qui était alors la seule formation de statisticiens en France. Elle a pour objet social de promouvoir la statistique : elle effectue des dons, accorde des bourses et subventions, et parraine des prix, dans la mesure de ses ressources qui proviennent des revenus de son capital. Elle travaille en étroite collaboration avec la SFdS, dont elle soutient certaines actions de communication. La Fondation gère le legs du docteur Norbert Marx, qui finance le prix du même nom décerné tous les deux ans par la SFdS. Ses moyens sont hélas limités car les dons n'affluent pas, malgré une fiscalité favorable pour les donateurs : à bon entendeur, salut !

## 4. Les mots de la fin

**GSt :** Le temps de conclure cet entretien est venu. Pour cela, je voudrais te demander, de manière peut-être indiscrète, de raconter ton présent : voilà déjà quelques années que tu es professeur émérite. Quelles activités as-tu conservées, de quelles occupations (et soucis) t'es-tu débarrassé ? Ne vivrais-tu pas, soudainement, dans la plus grande liberté scientifique ?

**GSa :** La position de professeur émérite est en effet plutôt confortable : je continue à être membre de mon laboratoire, à publier[40], à assister à des congrès tant que des collègues veulent bien m'inviter. Je peux être membre de jurys de doctorat, même si je ne peux plus diriger de thèses. Au CNAM, j'ai encore la responsabilité de la liaison avec le centre du Liban, mais plus de tâches administratives : finis les conseils et autres comités de recrutement ! Je regarde avec sérénité sans intervenir. À la SFdS, j'effectue mon dernier mandat de co-organisateur des Journées d'étude en statistique (JES), après 26 ans de service, et je préside le jury du prix de thèse décerné en l'honneur de Marie-Jeanne Laurent-Duhamel. Je reprends quelques activités de consultant. Je ne fréquente plus que des gens fréquentables. Bref : que du plaisir.

**GSt :** Merci beaucoup, Gilbert, de t'être prêté au jeu de cet entretien, et d'avoir répondu à mes questions avec tant de verve !

## Références

Académie des sciences (2000), *La statistique*, Rapport sur la science et la technologie n°8, Tec & Doc.

Anderson C. (2008), « The end of theory: the data deluge makes the scientific method obsolete », *Wired*, https://www.wired.com/2008/06/pb-theory, page consultée le 26 janvier 2020.

Benzécri J.-P. & collaborateurs (1973), chapitre « Les principes de l'analyse des données », *in L'analyse des données, tome 2 : L'analyse des correspondances*, pp. 3-17, Dunod.

40. On notera que bizarrement, les publications des émérites ne sont pas comptabilisées pour l'attribution de crédits aux laboratoires !

Bouroche J.-M. et G. Saporta (1978), *L'analyse des données*, Que Sais-Je n°1854, Presses Universitaires de France.

Dagnelie P. (1973), *Théorie et méthodes statistiques : applications agronomiques, tome 1 : La statistique descriptive et les fondements de l'inférence statistique*, Les presses agronomiques de Gembloux.

Dagnelie P. (1975), *Théorie et méthodes statistiques : applications agronomiques, tome 2 : Les méthodes de l'inférence statistique*, Les presses agronomiques de Gembloux.

Davenport T. H. and D. J. Patil (2012), « Data scientist: the sexiest job of the 21st century », *Harvard Business Review*, vol. 90, n°10, pp. 70-76.

Deville J.-C. (1974), « Méthodes statistiques et numériques de l'analyse harmonique », *Annales de l'INSEE*, 15, pp. 5-101.

Didier E. (2013), « Histoire de la représentativité statistique : quand le politique refait toujours surface », in Marion Selz (éd.), *La représentativité en statistique*, INED éditions, pp. 15-30.

Donoho D. (2017), « 50 years of data science », *Journal of Computational and Graphical Statistics*, vol. 26, n° 4, pp. 745-766.

Droesbeke J.-J. (2006), « Les Racines de la SFdS », https://www.sfds.asso.fr/sdoc-1651-0c05b7f0fe23ac31fd355c5496c29e3a-les_racines_de_la_sfds_23_02_06.pdf, note consultée le 11 février 2020.

Droesbeke J.-J. et P. Tassi (1997), *Histoire de la statistique*, Que sais-je n°2527, deuxième édition, Presses Universitaires de France.

Dunford N. and J. T. Schwartz (1958), *Linear Operators, Part I: General Theory*, Wiley & Sons.

Dunford N. et J. T. Schwartz (1963), *Linear Operators, Part II: Spectral Theory, Self Adjoint Operators in Hilbert Space*, Wiley & Sons.

Friedman J. H. (2001), «The role of statistics in the data revolution? », *International Statistical Review*, vol. 69, n° 1, pp. 5-10.

Kendall M. and A. Stuart (1961), *The Advanced Theory of Statistics, volume 2: Inference and relationship*, Griffin.

Mothes J. (1968), *Prévisions et décisions statistiques dans l'entreprise*, deuxième édition, Dunod.

F. Rosenfeld (1997), « Histoire des sociétés de statistique en France », *Journal de la société française de statistique*[41], vol. 138, n° 3, pp. 3-18.

H. Varian (2014), « Big Data: new tricks for econometrics », *Journal of Economic Perspectives*, vol. 28, n° 2, pp. 3-28.

41. En réalité, à l'époque, ce journal était encore le Journal de la Société de Statistique de Paris.