

Statistique et société

Mai 2020

Volume 8, Numéro 1

Rétrospectives

Sommaire

Statistique et société

Volume 8, Numéro 1

- 7 **Éditorial**
Emmanuel DIDIER
Rédacteur en chef de Statistique et société
-
- 9 **La montée en puissance de la statistique bayésienne, en France, depuis le milieu du vingtième siècle**
Gilles CELEUX
Institut de mathématiques d'Orsay,
Université Paris-Sud et Inria Orsay
- 11 **Entretien avec Christian Robert**
Christian P. ROBERT
Université Paris-Dauphine,
CEREMADE, Paris & University of Warwick, UK
Gilles CELEUX
Institut de mathématiques d'Orsay,
Université Paris-Sud et Inria Orsay
- 15 **Entretien avec Jean-Louis Foulley**
Jean-Louis FOULLEY
Directeur de Recherches INRA à la retraite
Gilles CELEUX
Institut de mathématiques d'Orsay,
Université Paris-Sud et Inria Orsay
- 23 **Entretien avec Jacques Bernier et Éric Parent**
Jacques BERNIER
Ingénieur EDF à la retraite
Éric PARENT
AgroParisTech/INRAe, UMR Mathématique
et Informatique Appliquées, Paris
Gilles CELEUX
Institut de mathématiques d'Orsay,
Université Paris-Sud et Inria Orsay
-

Sommaire

Statistique et société

Volume 8, Numéro 1

- 35 **Gilbert Saporta : un parcours éclectique**
Analyse, fouille, science des données
Entretien
Gilbert SAPORTA
Conservatoire national des arts et métiers, Paris
Gilles STOLTZ
CNRS – Université Paris-Sud
- 53 **2009-2019 : dix ans de publications**
sur l'enseignement de la statistique
en France
Antoine ROLLAND
ERIC EA 3083, Université de Lyon,
Université Lumière Lyon 2
-
- 73 **Appel à communication : Les chiffres au**
cœur de la crise du coronavirus

Statistique et société

Magazine quadrimestriel publié par la Société Française de Statistique. Le but de Statistique et société est de présenter, d'une manière attrayante et qui invite à la réflexion, l'utilisation pratique de la statistique dans tous les domaines de la vie. Il s'agit de montrer comment l'usage de la statistique intervient dans la société pour y jouer un rôle souvent inaperçu de transformation, et est en retour influencé par elle. Un autre dessein de Statistique et société est d'informer ses lecteurs avec un souci pédagogique à propos d'applications innovantes, de développements théoriques importants, de problèmes actuels affectant les statisticiens, et d'évolutions dans les rôles joués par les statisticiens et l'usage de statistiques dans la vie de la société.

Rédaction

Rédacteur en chef : Emmanuel Didier, CNRS, France

Rédacteurs en chef adjoints :

Thomas Amossé, CNAM, France

Jean-Jacques Droesbeke, Université libre de Bruxelles, Belgique

Chloé Friguet, Université Bretagne-Sud, France

Antoine Rolland, Université Lyon 2, France

Gilles Stoltz, Université Paris-Sud, France

Jean-Christophe Thalabard, Université Paris-Descartes, France

Catherine Vermandele, Université libre de Bruxelles, Belgique

Comité éditorial

Représentants des groupes spécialisés de la SFdS :

Ahmadou Alioum, groupe Biopharmacie et santé

Delphine Grancher, groupe Environnement et Statistique

Marthe-Aline Jutand, groupe Enseignement de la Statistique

Elisabeth Morand, groupe Enquêtes, Modèles et Applications

Alberto Pasanisi, groupe Agro-Industrie

Autres membres :

Jean-Pierre Beaud, Département de Science politique, UQAM, Canada

Corine Eyraud, Département de sociologie, Université d'Aix en Provence, France

Michael Greenacre, Department of Economics and Business, Pompeu Fabra
Université de Barcelone, Espagne

François Heinderyckx, Département des sciences de l'information, Université
libre de Bruxelles, Belgique

Dirk Jacobs, Département de sociologie, Université libre de Bruxelles, Belgique

Gaël de Peretti, Insee, France

Théodore Porter, Département d'histoire, UCLA, Etats-Unis

Carla Saglietti, Insee, France

Patrick Simon, Ined, France

Design graphique
fastboil.net

ISSN 2269-0271

Éditorial



Emmanuel DIDIER

Rédacteur en chef de Statistique et Société

Chers lectrices et lecteurs,

Dans ce numéro, bouclé en plein confinement, nous vous proposons de tourner le regard vers l'arrière pour mieux appréhender le présent et l'avenir.

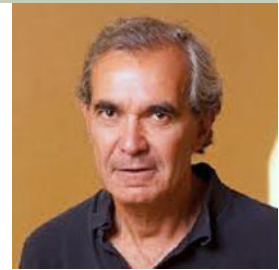
Nous revenons d'abord sur l'histoire du bayésianisme en France. Ce courant statistique avait dans notre pays une particularité : il était quasiment absent des institutions universitaires. Or il est aujourd'hui devenu majeur. Nous avons donc demandé à cinq de ses principaux représentants, Gilles Celeux, Jean-Louis Foulley, Jacques Bernier et Eric Parent, et Christian Robert de nous raconter, dans quatre entretiens séparés, comment et pourquoi ce type de raisonnement statistique s'est imposé. Nous verrons en particulier le rôle des voyages dans cette avancée scientifique, ce qui résonne fortement avec le confinement actuel : si l'on parle de changer la globalisation, espérons que celle qui adviendra n'entravera pas les progrès de la science. Puis Gilles Stoltz nous livre un long entretien avec Gilbert Saporta, que beaucoup d'entre nous (dont votre serviteur !) reconnaissent comme leur maître ès analyse des données, et qui nous explique le très intéressant tournant qu'il a effectué vers une science des données aujourd'hui omniprésente. Enfin, Antoine Rolland nous expose dans un tableau synthétique dix ans de publications concernant l'enseignement de la statistique en France. L'enseignement est un vecteur absolument majeur des rapports entre la statistique et la société, il était important que nous lui fassions une place dans la revue, que nous souhaitons de plus en plus établie.

Nous joignons à ce numéro un appel à contribution sur le rôle que joue la statistique dans la crise actuelle du Coronavirus. En effet, il n'aura échappé à aucun de nos lecteurs que la diffusion de ce virus a été accompagnée d'une prolifération indissociablement technique et politique de méthodes de quantification. Nous préparons un numéro spécial sur le thème.

Bonne lecture !

Emmanuel Didier

La montée en puissance de la statistique bayésienne, en France, depuis le milieu du vingtième siècle



Gilles CELEUX¹

Institut de mathématiques d'Orsay, Université Paris-Sud et Inria Orsay

TITLE

The rise of Bayesian statistics in France since the middle of the twentieth century

RÉSUMÉ

Cinq universitaires et ingénieurs partagent leurs points de vue sur une histoire française de la pensée bayésienne, depuis leurs débuts jusqu'à leurs fins de carrière. En France, bien avant la révolution des méthodes de Monte Carlo des années 1990, la pensée bayésienne avait pénétré dans les bureaux d'études, mais restait considérée avec quelque suspicion dans le milieu académique. Par la suite, le statisticien bayésien a trouvé sa place dans la communauté statistique bien plus tard que dans le monde anglo-saxon. Confrontée à l'avènement du Big Data et à la percée de l'intelligence artificielle dans nos sociétés, la statistique bayésienne conserve en main de beaux atouts pour l'apprentissage et la décision.

Mots-clés : *statistique, Bayes, fréquentiste.*

ABSTRACT

Five academics and engineers share their views on a French history of Bayesian thinking, from the beginning to the end of their careers. In France, long before the statistical revolution fueled by the Monte Carlo methods rendering Bayesian inference easy in the 1990s, Bayesian thought had made a breakthrough in the engineering offices, but was still a suspicious matter of concern in the academic world. Subsequently the Bayesian statistician found a place of his own within the statistical community much later than in the Anglo-Saxon world. Confronted with the advent of Big Data and the advent of artificial intelligence in our societies, Bayesian statistics still holds great potential for efficient learning and advanced decision-making.

Keywords: *statistics, Bayes, frequentist.*

« *La formule qui décrypte le monde* », c'est ainsi que le numéro de novembre 2012 de la revue Science et Vie présente la formule de Thomas Bayes. Tous s'accordent aujourd'hui pour dire qu'on la retrouve dans tous les domaines : la statistique bayésienne permet de quantifier sous la forme d'un jugement probabiliste la crédibilité d'une hypothèse, c'est-à-dire la pertinence d'un mécanisme causal (génétique, climatique, écologique, ...) à l'aune de ce qu'on observe comme conséquence (type de gènes mutés, fréquence de tornades, nombre d'extinctions, etc.). Pourtant, l'histoire de la statistique bayésienne présente une nette singularité en France. Contrairement à sa place importante en Grande-Bretagne, aux USA mais aussi dans d'autres pays comme l'Espagne, l'Italie, l'Irlande, la Belgique, le Canada, etc., elle n'occupait en France qu'une position médiocre voire négligeable dans les années soixante-dix. Pour preuve : elle était à peine mentionnée dans les enseignements universitaires de statistique ! Si l'on se focalise sur les différences entre la France et la Grande-Bretagne dans l'enseignement supérieur de la statistique, il apparaît que le caractère académique et mathématique est marqué en France et a laissé très peu de place à la statistique bayésienne, alors que l'enseignement de la statistique en Grande-Bretagne, plutôt orienté comme une science de l'ingénieur, a offert naturellement une place importante à la statistique bayésienne.

1. gilles.celeux@inria.fr

Depuis, les choses ont changé. La statistique bayésienne a accru son influence dans le monde et notre pays a suivi le mouvement. Aujourd'hui, les activités bayésiennes en applications, recherche académique et formation permanente sont nombreuses. La communauté française de statisticiens bayésiens est maintenant riche et active. C'est l'histoire de ce dynamisme, nouveau et maintenant bien établi, que ce dossier vise à analyser au travers des entretiens avec des acteurs et témoins majeurs de ce renouveau.

Avant de relater ces entretiens, indiquons que l'implantation de la statistique bayésienne en France reste fragile. Ainsi, si les statisticiens convaincus de l'intérêt de la statistique bayésienne ont fini par constituer des poches de plus en plus nombreuses dans l'enseignement supérieur, elle est loin d'avoir pénétré uniformément dans nos universités et nos écoles d'ingénieurs. Organiser un enseignement régulier de statistique bayésienne reste un acte volontariste.

J'ai interrogé quatre personnes pour constituer ce dossier : Christian Robert, Jean-Louis Foulley, Jacques Bernier et Éric Parent. Pourquoi elles? Parce que la gamme conjointe de leurs carrières couvre une large période qui va des années 1950 jusqu'aujourd'hui. Certainement aussi parce que toutes sont membres de la Société Française de Statistique et que je les connaissais. Mais surtout parce qu'ingénieurs, enseignants ou chercheurs, ce sont des acteurs importants au rôle varié de ce mouvement bayésien français.

Je n'en dis pas plus sur eux puisqu'ils vont eux-mêmes se présenter en Homo Bayesianus. Je voudrais simplement insister sur le fait que les trois plus jeunes d'entre eux ont fait des séjours aux USA ou en Grande-Bretagne décisifs pour leur orientation de chercheur.

Qu'est-ce qui différencie la statistique bayésienne de la statistique classique ? La statistique classique (ou fréquentiste) interprète la probabilité comme une limite de fréquence : elle ne probabilise donc que des quantités observables et s'appuie sur des théorèmes établis dans un contexte de répétabilité asymptotique. La statistique bayésienne, en revanche, interprète la probabilité comme un degré de crédibilité, ce qui lui permet de probabiliser aussi des inconnues non observables (concepts, hypothèses et paramètres) ou à occurrence singulière. La probabilité bayésienne est conditionnelle à un état d'information ; elle changera par mise à jour à l'aide de la loi de Bayes quand de nouvelles observations modifient nos connaissances. Pour une explication plaisante et instructive de la formule de Bayes en vidéo, suivre par exemple le lien : <https://www.youtube.com/watch?v=R13BD8qKeTg>

Entretien avec Christian Robert



Christian P. ROBERT¹

Université Paris-Dauphine, CEREMADE, Paris & University of Warwick, UK

Gilles CELEUX²

Institut de mathématiques d'Orsay, Université Paris-Sud et Inria Orsay

TITLE

Interview of Christian Robert

RÉSUMÉ

Christian Robert est un chercheur mondialement réputé, spécialiste de la statistique bayésienne et des méthodes de Monte Carlo. Il a rédigé un nombre impressionnant d'articles dans les principales revues de statistique et publié de nombreux ouvrages de référence. Ancien président de la Société internationale pour l'analyse bayésienne (ISBA), Christian Robert est sans aucun doute l'un des statisticiens bayésiens dominants de sa génération. Il faut dire que, jeune chercheur, il a fait des séjours longs et décisifs dans diverses universités américaines. Il en d'ailleurs ramené un deuxième prénom et est devenu Christian P. Robert (P pour Plessis).

Mots-clés : Bayes, algorithmes, méthodes Monte-Carlo.

ABSTRACT

Christian Robert is world wide known as an expert in Bayesian statistics and Monte-Carlo methods. He wrote numerous impressive articles in major statistical journals and he published many highly referenced textbooks. As former president of the International Society for Bayesian Analysis (ISBA), Christian Robert is without doubt one of the predominant Bayesian statistician in his generation. It must be said that as a young research fellow he had the opportunity of several long postdoctoral stays in various universities in the US. Indeed, he got there a middle name and became Christian P. Robert (P for Plessis).

Keywords: Bayes, algorithms, Monte-Carlo methods.

G. Celeux : Raconte moi tes débuts de statisticien.

Chr. Robert : Quand j'ai commencé mes études à l'ENSAE, la statistique bayésienne n'y était pas du tout enseignée, à part une très brève mention dans le cours de Christian Gouriéroux, avec la conclusion que la dépendance au choix de la loi a priori rendait cette approche inapplicable. En comparaison, le cours de théorie des jeux, enseigné par Sylvain Sorin à l'époque, comportait une section de théorie de la décision qui, naturellement, faisait la part belle aux procédures bayésiennes puisque optimales. Il n'en reste pas moins que j'étais curieusement attiré par cette approche et ai donc sauté sur l'opportunité de travailler avec Jean-Pierre Raoult à Rouen sur les estimateurs de Stein. Je pense qu'à cette époque Jean-Pierre était l'un des très rares statisticiens français à montrer un intérêt pour l'inférence bayésienne, en lien avec les économètres bayésiens du CORE comme Michel Mouchart et Jean-Marie Rolin, ainsi que Jean-Pierre Florens à Toulouse dont Jean-Pierre dirigea la thèse d'Etat. Durant ma thèse (d'Université !) il m'avait d'ailleurs fait lire une version préliminaire du livre *Elements of Bayesian Statistics* (Florens et al., 1990) de ces trois auteurs, qui m'avait alors quelque peu déconcerté du fait de son insistance sur les sigma-algèbres et le conditionnement plutôt que la théorie de la décision. Et sans lien avec les articles de Jim Berger et ses co-auteurs que j'avais étudiés durant les deux années de thèse. Une occasion ratée de rencontrer d'autres chercheurs en statistique bayésienne s'était

1. xian@ceremade.dauphine.fr

2. gilles.celeux@inria.fr

présentée avec la conférence Valencia 2, en 1987, ratée car bien que terriblement intéressé par cette rencontre, organisée par José Bernardo, je n'avais pu y assister du fait de mon statut militaire d'appelé du contingent ! J'avais cependant essayé de récupérer les articles présentés mais en ces temps ante-email, mes demandes épistolaires aux auteurs étaient restées sans réponse. (J'ai quand même fini par trouver le volume des proceedings de Valencia 2, dans un couloir de l'Université de Warwick, l'an dernier, au moment du déménagement du département.)

G. Celeux : Ton premier séjour aux USA a été décisif...

Chr. Robert : Oui... J'avais contacté quelques chercheurs américains, toujours par lettre, pour un poste de visiteur l'année suivant ma thèse (1987-1988). Et obtenu à chaque fois des réponses positives, ce qui me surprend encore car je n'avais alors que des préprints de mes quelques articles, joints à ma demande. En particulier, Jim Berger, alors en poste à Purdue, m'avait proposé de le rappeler pour discuter des conditions de ce poste et pour vérifier – je présume – la qualité de mon anglais. D'ailleurs mes premiers cours à des étudiants du MidWest se révélèrent bien difficiles, pour ces étudiants... et pour moi ! J'acceptai bien évidemment l'offre de Jim et je partis pour l'Indiana fin août 1987, quelques jours avant le début des cours. Chance incroyable, j'y découvris un groupe impressionnant de statisticiens et statisticiennes bayésien(ne)s, à la fois soudé et ouvert, avec qui collaborer était facile. Qui plus est, j'y rencontrai George Casella, en sabbatique depuis Cornell University pour y écrire son livre de statistique mathématique avec Roger Berger, lui aussi ancien étudiant de Purdue, et travailler avec les chercheurs locaux. Du fait de cet environnement fabuleux, j'appris beaucoup plus durant les premiers mois à West Lafayette que durant toute ma thèse et je pus élargir grandement mes thèmes de recherche. J'y nouai des amitiés durables qui pour certaines ont, hélas, été brisées par la disparition de Costas Goutis (1996), Bernhard Flury (1998), George Casella (2012) et Susie Bayarri (2015).

G. Celeux : Mais, il a fallu revenir au bercail...

Chr. Robert : Après une seconde année aux USA, ayant accompagné George Casella dans son retour à Cornell, je rentrai en France avec des offres de postes de plusieurs universités, ce qui est surprenant rétrospectivement du fait du thème globalement bayésien de mes recherches et aussi du style de mes publications, plus courtes et moins théoriques que la norme. En arrivant à Pierre et Marie Curie (Paris 6), je n'y trouvai pas de chercheur en statistique bayésienne, mais la possibilité d'enseigner un cours de DEA sur la théorie de la décision bayésienne et une opportunité de collaborer, avec Jean Diebolt et toi sur l'estimation bayésienne de modèles de mélanges, du fait de leur investissement dans des versions stochastiques de l'algorithme EM qui rejoignaient l'algorithme de Gibbs découvert quelques mois auparavant lors d'un exposé d'Adrian Smith (alors à Imperial College London) au Québec. Même si les aspects markoviens et stochastiques semblaient primer dans la réception positive de nos travaux, nous ne rencontrâmes aucune opposition locale quant à l'option bayésienne.

G. Celeux : Oui, je me souviens de cette époque où tu as été rapidement perçu comme le chef de file bayésien de la communauté statistique française.

Chr. Robert : En fait, il est exact qu'à une occasion, le caractère bayésien de mes travaux fut invoqué en France pour ne pas m'offrir un poste d'enseignant attaché. Déception vite effacée par les possibilités qui me furent données de prendre la responsabilité du laboratoire de statistique du CREST (1992), de publier deux livres en français sur la statistique bayésienne (Robert, 1992) et sur les méthodes de Monte-Carlo (Robert, 1996), et évidemment de devenir professeur à Rouen (1992), puis à Dauphine (2000). Le mépris qui pouvait effectivement exister dans ces années dans une partie de la communauté statistique francophone ne m'a donc jamais directement affecté. Il a d'ailleurs progressivement et insensiblement disparu. De nouvelles générations moins concernées par les querelles idéologiques s'intéressent surtout

aux opportunités pratiques offertes par les nouvelles techniques de calcul comme le MCMC. L'impact de statisticiens et statisticiennes travaillant dans des domaines appliqués comme Sylvia Richardson, Jean-Louis Foulley, Éric Parent, Bruno Lecoutre, Jean-Marie Cornuet, est notable dans cette évolution. La renaissance du minimax et l'apparition du PaC-Bayésien y jouent évidemment un rôle.

G. Celeux : Comment vois-tu l'avenir de la statistique bayésienne ?

Chr. Robert : Il reste à espérer que l'éclosion récente de techniques plus rapides d'apprentissage ne tourne pas toute la statistique, bayésienne ou non, en une approche du passé, comme avait pu le devenir la statistique bayésienne au cours du 20^e siècle. Même si les algorithmes stochastiques connaissent une reprise remarquable en France ces dernières années (Eric Moulines, Nicolas Chopin, Sylvain LeCorff, Arnak Dalalyan, Alain Durmus), ainsi que le bayésien asymptotique (Judith Rousseau, Ismael Castillo et leurs nombreux élèves).

Qu'est-ce que le MCMC ? Les méthodes de Monte-Carlo par chaînes de Markov (MCMC) sont de puissants outils computationnels d'inférence bayésienne : elles servent à échantillonner la distribution de probabilité (dite a posteriori) qui résume toute l'information disponible sur les paramètres d'intérêt d'un modèle bayésien, après le recueil des données. Ces méthodes de Monte-Carlo se fondent sur les promenades aléatoires de chaînes de Markov construites de telle sorte que leurs lois stationnaires soient les distributions à échantillonner recherchées.

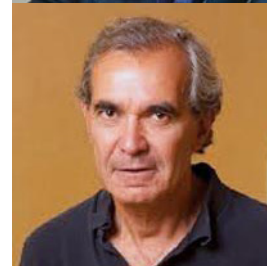
Références

Florens, J.-P., M. Mouchart, et J.-M. Rolin (1990), Elements of Bayesian Statistics, M Dekker, New York.

Robert, C. (1992), L'analyse statistique bayésienne, 393 pp., Economica, Paris.

Robert, C. (1996), Méthodes de Monte Carlo par Chaines de Markov, 340 pp., Coll. Statistique mathématique et Probabilité, Economica, Paris.

Entretien avec Jean-Louis Foulley



Jean-Louis FOULLEY¹

Directeur de Recherches INRA à la retraite

Gilles CELEUX²

Institut de mathématiques d'Orsay, Université Paris-Sud et Inria Orsay

TITLE

Interview of Jean-Louis Foulley

RÉSUMÉ

Jean-Louis Foulley est un généticien à la retraite qui a travaillé à l'Institut National de la Recherche Agronomique (INRA) dans le domaine de la sélection animale. Il a milité pour l'implantation de la statistique bayésienne dans ce domaine d'application important de la statistique. Il a bénéficié de séjours longs et décisifs aux USA pour son orientation.

Mots-clés : *statistique bayésienne, génétique, agronomie.*

ABSTRACT

Jean-Louis Foulley is a retired genetician who worked at the French National Institute for Agronomical Research (INRA) in animal selection. He was influential for the use of bayesian statistics in an important and historical domain of application of statistics. He largely benefited of longterm and crucial visiting scholarships in the US for his later scientific orientation.

Keywords: *bayesian statistics, genetics, agronomy.*

G. Celeux : Peux-tu nous indiquer le contexte de tes études supérieures?

J.-L. Foulley : J'appartiens à la génération 68, perdue pour certains voire honnie, bénie pour d'autres, mais assurément singulière. Nourri d'humanités et de géométrie dans le secondaire, j'ai complété ce régime en prépa par des doses massives de sciences et notamment de mathématique à base exclusive d'analyse. Je me suis retrouvé à l'entrée à l'Agro de Paris n'ayant jamais entendu parler de probabilité ni, *a fortiori*, de statistique, de génétique et encore moins d'informatique. Certains y verraient actuellement une lacune. J'y vois pour ma part une chance, ayant abordé ces disciplines par un enseignement rigoureux des concepts de base sous la férule de grands maîtres : Guy Lefort, Albert Jacquard, Gustave Malécot notamment. En statistique, Guy Lefort et son équipe (Paule Renaud, Camille Duby) nous concoctaient un enseignement de haut vol alliant théorie et applications et, à cette époque, dans un cadre purement classique où, rétrospectivement, je me rends compte que Fisher tout comme Benzecri n'avaient guère de place.

G. Celeux : Quand as-tu découvert la statistique bayésienne?

J.-L. Foulley : Au DEA de Génétique quantitative et appliquée, Albert Jacquard et Jean Générumont recouraient au théorème dit de Bayes en inférant la probabilité des génotypes d'individus sachant les phénotypes d'apparentés. Jacquard poussait la chansonnette plus loin en évoquant l'intérêt d'une estimation de la probabilité d'une proportion sous un modèle binomial avec

1. foulleyjl@gmail.com
2. gilles.celeux@inria.fr

un a priori beta uniforme dans le cas d'un très faible effectif. Mais, c'était beaucoup plus que succinct. En fait, la véritable confrontation avec le bayésien est venue dès mes premières années à l'INRA lors de discussions avec Guy Lefort au sujet de l'évaluation génétique des reproducteurs. Ce dernier ne comprenait pas pourquoi nous traitions les effets des taureaux comme aléatoires alors que l'échantillon des taureaux en question n'avait que peu à voir avec un choix au hasard. Lefort s'alarmait de cette incohérence des généticiens traitant un facteur, tantôt comme aléatoire, tantôt comme fixé, au gré des coutumes et techniques établies. Nous fûmes plusieurs promotions à sécher sur cette question récurrente que posait Guy Lefort sur nos travaux. À notre grande satisfaction, il résolut lui-même l'énigme en interprétant le BLUP comme un estimateur bayésien pris égal à l'espérance de la loi *a posteriori* de la valeur génétique transmise par le reproducteur, sachant que celle-ci est munie d'une loi a priori qui est « *un résumé de la connaissance plus ou moins précise qu'a le spécialiste des valeurs des paramètres* » (Lefort, 1980). Cette période (milieu des années 70) correspond d'ailleurs à l'orientation marquée de son enseignement à l'INA-PG vers la statistique bayésienne (Lefort, 1975) et à son engagement dans cette voie ; il fut le seul participant français au premier congrès de statistique bayésienne de Valencia en 1979. Son influence fut déterminante dans mon orientation personnelle. Pour nos détracteurs futurs, le ver était maintenant dans le fruit et il ne restait plus qu'à le laisser se développer sachant que les trois grands noms que j'ai mentionnés s'affirmèrent tous des bayésiens convaincus (Lefort, 1975; Jacquard, 1970; Malécot, 1947). En fait, Charles Henderson, le père du BLUP, avait établi ses équations du modèle mixte par une approche bayésienne, certes inconsciente, mais prémonitoire. Il eut recours à cette occasion, avant tout le monde, à des concepts et des outils d'avant-garde : vraisemblance des données complètes (données observées augmentées des effets aléatoires) et algorithme de type EM greffé sur les équations du modèle mixte pour estimer les composantes de la variance par maximum de vraisemblance résiduelle nommée *REML* selon son acronyme anglais (Henderson et al., 1959; Henderson, 1973).

Qu'est-ce que le modèle linéaire mixte ? Le modèle linéaire mixte généralise le modèle d'analyse de variance et de régression classiques en donnant aux résidus une structure de variance-covariance parcimonieuse induite par des effets aléatoires. Tel est le cas des modèles dits hiérarchiques à plusieurs niveaux de variabilité, par exemple 1) individus intrapopulation et 2) populations. Le modèle phare de la sélection animale qui a prévalu pendant plusieurs décennies avant la révolution génomique en constitue un très bon exemple (cf. entretien avec J.-L. Foulley). C'est un modèle qui décrit la performance d'un descendant (*k*) élevé dans un élevage (*j*) et issu d'un père (*i*). Dans ce modèle, les effets *élevage* sont traités comme fixes et les effets père comme aléatoires, si bien que les descendants d'un même père présentent une corrélation liée à cette origine commune (dite corrélation intraclasse). On peut également introduire dans ce modèle, les parentés entre les pères ainsi que les effets des grands-pères maternels qui permet une expression plus fine des relations de parenté entre individus. On aboutit ainsi logiquement au modèle dit *modèle animal* comportant les facteurs *élevage* fixe et *individu* aléatoire muni d'une matrice complète de parenté entre individus.

Pressentant qu'il se passait quelque chose qu'il ne fallait pas manquer, je mis les bouchées doubles pour me former au BLUP et plus généralement à toutes les méthodologies du modèle mixte auprès des spécialistes de Cornell, d'où un semestre sabbatique en 1979 à Guelph (Ontario) avec Larry Schaeffer et un cours d'été d'Henderson en 1981 à Landquart (Suisse) suivi par une collaboration avec Daniel Gianola lors de son séjour sabbatique à Jouy-en-Josas en 1982 et de mon année sabbatique en Illinois à Urbana-Champaign (IL) en 1985-86 puis à Madison (WI) en 1995.

Qu'est-ce que le BLUP ? Le BLUP (Best Linear Unbiased Predictor) est une méthode statistique qui généralise au modèle linéaire mixte, la méthode d'estimation des moindres carrés en autorisant les résidus à présenter des corrélations. Tel est le cas notamment pour des résidus relatifs à des descendants d'un même géniteur et plus généralement à des individus apparentés. Il fait l'objet également d'applications à la génomique sous le vocable de g-BLUP. Le BLUP possède des propriétés d'optimalité (meilleur prédicteur linéaire sans biais au sens de l'erreur quadratique moyenne). Il a été développé par un généticien de l'Université Cornell, Charles Henderson, sous la forme d'un système d'équations qui miment celles des moindres carrés auxquelles on ajoute des éléments aux blocs des effets aléatoires (par exemple une information a priori de parenté entre individus). Développées pourtant dans un contexte de statistique classique, leurs solutions peuvent très naturellement s'interpréter dans le cadre bayésien comme l'espérance de distributions *a posteriori*.

G. Celeux : Comment cela s'est-il traduit dans tes activités de recherche et d'enseignement ?

J.-L. Foulley : J'avais acquis alors une boîte à outils suffisamment fournie pour approfondir les questions de prédiction qui se posaient à nous en génétique quantitative et en sélection animale. Dans le domaine multinormal, les méthodes fréquentistes aussi bien au niveau prédiction (BLUP) qu'estimation des composantes de la variance (Henderson I, II & III, MINQUE, ML, REML) avaient fait leurs preuves et ne nécessitaient plus que des ajustements à des problématiques locales (données manquantes et effets sélectifs notamment). Pourtant, elles étaient très complexes à maîtriser et restaient toujours approximatives en dehors de ce cadre. Avec Gianola, nous décidâmes de franchir le Rubicon et de proposer, dès 1982, une méthode de prédiction des effets aléatoires d'un modèle multinomial-probit basée sur la distribution *a posteriori* des paramètres de position. À cette époque, nous n'allions pas plus loin que le MAP et l'approximation de Laplace (Gianola et Foulley, 1983). En fait, la mise en oeuvre fut difficile vue la taille gigantesque des fichiers de données de contrôle de performances récoltées en ferme. Il nous a fallu l'aide d'informaticiens avertis comme Ignacy Misztal, pour proposer un programme opérationnel d'application aux difficultés de naissance et à la fertilité des bovins. On était encore loin des MCMC même si elles se profilaient à l'horizon. Je me trouvai embarqué dans l'encadrement d'une thèse sur l'estimation REML des composantes de la variance par l'algorithme EM qui répondait très bien au statut qu'il fallait faire jouer aux effets fixes et aux effets aléatoires dans une estimation de type REML des composantes de la variance. Tu me rétorqueras que tu vois mal où se loge la statistique bayésienne dans cette affaire ? C'est exact en apparence, mais la vraisemblance marginalisée par intégration est un concept clé en statistique bayésienne pour s'affranchir des effets de certains paramètres dits parasites. En incluant les effets fixes dans le vecteur des données manquantes, on les intègre *ipso facto* dans la vraisemblance à la phase E même si leur loi *a priori* est impropre. Cela me permit de mettre en oeuvre avec succès des modèles mixtes prenant en compte l'hétérogénéité de variance et de développer des estimations type REML des composantes de variance en modèle non-linéaire via un algorithme SAEM ou en modèle linéaire généralisé par MCMC.

G. Celeux : Quel bilan tires-tu de ces travaux vis-à-vis de ton cheminement vers la statistique bayésienne ?

J.-L. Foulley : Premièrement qu'il est toujours bon de se colleter avec le réel, en l'occurrence avec des modèles qui ont donné lieu à des applications d'envergure, je pense notamment à toutes celles concernant la prise en compte de l'hétérogénéité de variance dans l'évaluation des taureaux de races laitières (travaux de Christèle Robert et Bernard Bonaiti). La statistique bayésienne est avant tout une science d'ingénieurs comme l'a bien montré McGrayne (2011)

dans son remarquable livre *The theory that would not die*. Elle a repris un second souffle à partir des années 90 grâce aux techniques de simulation MCMC et notamment à la méthode de Metropolis-Hastings et à l'échantillonneur de Gibbs. Depuis ce secteur est devenu un domaine de recherche à part entière qui mobilise beaucoup d'énergie à la frontière entre mathématique, informatique et statistique comme en attestent les ouvrages et travaux de Christian Robert. De ce point de vue, on adhérera volontiers à la position de Georges Canguilhem qui déclarait : « *La science procède de la technique et la technique doit être pensée comme création* ». Un bon exemple en est fourni par les méthodes ABC (Approximate Bayesian Computation) qui ont été introduites par des généticiens des populations (Tavaré, Beaumont, Cornuet) à la fin des années 1990 en tant qu'algorithme d'échantillonnage de *l'a posteriori* à partir des étapes constitutives d'échantillonnage dans *l'a priori* puis de simulation des données d'un modèle complexe dont on ne connaît pas la vraisemblance, celles-ci n'étant retenues que si elles s'avèrent suffisamment proches des données réelles. Ce n'est que plus tard que les théoriciens se sont attaqués aux propriétés des différents algorithmes qui avaient été proposés plus ou moins empiriquement par les généticiens, et ont pu préciser les conditions de validité de ces algorithmes. Mais il y a une contrepartie à cet essor, c'est le développement actuel d'usages opportunistes de la statistique bayésienne au nom de la disponibilité, de l'efficacité et de la simplicité conceptuelle de la boîte à outils bayésiens. Bruno Goffinet de l'INRA parlait du confort de la statistique bayésienne et Philip Dawid allait même jusqu'à la qualifier de boring. Comme le rappelait fort justement Bradley Efron (2005) : « *But using Bayes' rule does not make one a Bayesian ; always using it does, and that's where difficulties begin.* »

G. Celeux : N'y aurait-il pas un revers à la médaille bayésienne ?

J.-L. Foulley : On peut dire cela ainsi. Mais avant de parler des difficultés sinon des revers, on pourrait évoquer les points saillants de la théorie qui ont milité pour moi en faveur de la statistique bayésienne. J'en vois trois, à mon sens, fondamentaux : 1) l'inférence sur les paramètres du modèle définie conditionnellement aux données observées et non pas par rapport à des données potentiellement observables ; 2) la dualité de l'interprétation de la probabilité (épistémique et aléatoire selon la terminologie de Ian Hacking) reconnue de tous depuis très longtemps, mais enfin assumée et intégrée dans un même formalisme par la théorie bayésienne ; 3) la structure de la formule de Bayes qui permet une actualisation séquentielle de l'information, la distribution *a posteriori* d'aujourd'hui servant d'*a priori* pour demain. La combinaison de 1) et 2) permet alors de formuler des énoncés probabilistes sur les paramètres et autres variables du modèle notamment des observations à venir. Fisher avait certes intégré le point 1) dans la théorie de la vraisemblance, mais il ne franchissait jamais complètement le pas du point 2), malgré la tentative contestable de l'inférence fiduciaire et de sa version moderne des distributions de confiance.

G. Celeux : Bien, mais je reviens à ma question sur les revers de la théorie ?

J.-L. Foulley : Une difficulté fondamentale à plusieurs égards réside dans le choix des *a priori*. Elle est la source du contentieux entre fréquentistes disciples de Neyman-Pearson, fisheriens et bayésiens depuis les origines (Dale, 1999; Fienberg, 2006; Leonard, 2014; Stigler, 1986) et, en particulier, à propos des tests d'hypothèses. Ce contentieux divise même la communauté bayésienne en plusieurs courants (objectif, subjectif, empiriste). C'est aussi un facteur limitant à la diffusion plus large de l'inférence bayésienne dans des secteurs, dont le biomédical par exemple, qui se veulent à l'abri de toute critique partisane. Ce souci explique l'intérêt depuis les pères fondateurs, Bayes, Price et Laplace, pour les *a priori* dits objectifs (Berger, 2004) notamment les *a priori* non informatifs de Jeffreys (Robert et al., 2004), qui possèdent de plus la propriété d'invariance c'est-à-dire de cohérence des *a priori* quelle que soit l'échelle d'expression des paramètres. Mais poser un *a priori* non informatif relève un peu de l'homéopathie, qui ne peut pas faire de mal à défaut de faire du bien. Comme d'autres bayésiens, je considère ce type d'*a*

priori comme un point de départ « *a priori-0* » d'un processus itératif qui se nourrit de l'addition séquentielle de données passées à chaque étape au filtre de la formule de Bayes, l' *a priori* de l'étape (t) étant l'*a posteriori* de l'étape (t-1) comme je l'indiquais au point 3. C'est du moins l'approche que j'ai adoptée dans l'établissement des pronostics de football aussi bien d'une saison à l'autre qu'intra-saison d'un ensemble de journées à la journée suivante (Foulley, 2015). Je suis un peu réticent vis-à-vis de l'élicitation d' *a priori* par des experts, qui relève beaucoup du coaching d'équipes de stars en mal de leadership ! En revanche, les procédés utilisés en apprentissage avec séparation des données en échantillon d'apprentissage et échantillon test, le premier servant à construire l'*a priori*, sont à mon sens plus à même d'aboutir à des résultats pertinents. On doit se méfier de pièges tendus par la spécification quasi automatique d'*a priori* notamment en situation de lois conjuguées où l'*a priori* par défaut est beaucoup plus informatif qu'on ne le croit. Je pense notamment aux *a priori* sur les variances qui enferment, dans le cas multidimensionnel, les *a priori* dans une structure bien particulière d'échangeabilité entre composantes.

Pourquoi un problème avec l'a priori en statistique bayésienne ? La formule de Bayes donne la règle de l'apprentissage statistique : effectuer un calcul des probabilité conditionnelle qui dicte comment une loi de probabilité décrivant la connaissance d'une inconnue (loi *a priori*) doit se modifier (en une loi *a posteriori*) lorsque l'on prend en compte une nouvelle observation. Mais elle ne dit rien de l'état de connaissance initiale : comment le spécifier, en particulier quand aucune connaissance *a priori* n'est disponible ? Comme il n'existe pas de consensus chez les mathématiciens pour quantifier l'ignorance, divers critères concurrents (maximum d'entropie, invariance par reparamétrisation, etc.) peuvent être utilisés pour construire des lois *a priori* non informatives.

G. Celeux : Malgré ces difficultés techniques, pourquoi tout le monde n'est-il pas encore bayésien ?

J.-L. Foulley : Oui, c'est une bonne question que posait déjà Efron en 1986 en mettant en avant de réelles difficultés pour mettre en oeuvre l'approche bayésienne par rapport aux commodités des recettes fréquentistes. Je pourrais répondre en forme de boutade en te disant que tu fais du bayésien sans le savoir (Lecoutre, 1997) et depuis assez longtemps en réalité, depuis ta naissance si j'en crois le *bébé bayésien* de Stanislas Dehaene (2013) ! Mais je ne serai pas si péremptoire. Tout ce que je viens de raconter n'est en fait que tracas quotidien face à des questionnements beaucoup plus fondamentaux. En ce qui me concerne, c'est le paradoxe de Jeffreys-Lindley qui a servi de révélateur (Lindley, 1957). Je le rumine depuis maintenant une vingtaine d'années. Je crois à chaque fois en avoir fini avec mes doutes et pourtant j'y reviens malgré moi quelques mois ou années plus tard, comme d'autres (Robert, 2014). Il est d'autant plus perturbant que c'est le domaine où la statistique classique et la statistique bayésienne aboutissent à des résultats diamétralement opposés et à de vives polémiques (Mayo, 2018). J'avais intitulé une de mes présentations à Applibugs, fin 2013, « *Le paradoxe de Jeffreys-Lindley : pierre dans le jardin des fréquentistes ou épine dans le pied des bayésiens ?* », pour bien marquer l'embarras des deux camps face à cette divergence. Aussi déstabilisants qu'ils puissent être, les paradoxes ne sont pas que des démons maléfiques (comme celui de Maxwell), ils recèlent aussi beaucoup de vertus. C'est grâce à ce paradoxe que j'ai enfin compris pourquoi le critère BIC pénalisait la dimension d'un modèle en proportion du log de l'effectif N ce qui m'apparaissait contre-intuitif. La crise actuelle de défiance à l'égard des p-values et seuils de signification des tests d'hypothèse nulle n'a fait que raviver la plaie. Je n'en suis pas encore à désavouer l'usage des facteurs de Bayes comme certains bayésiens de premier plan l'ont fait dont Robert (2016). Leur construction pose avec encore plus d'acuité le problème du choix de l'*a priori*. J'en viens

même à penser que ce choix devrait être spécifique à la construction du facteur de Bayes.

G. Celeux : Ces questions posent donc le problème de la formation.

J.-L. Foulley : Oui, on ne peut donc faire l'économie d'un apprentissage et d'une maturation de la théorie notamment par des formations adaptées et une lecture des ouvrages des auteurs de référence tels que ceux de Jeffreys, Zellner, Box & Tiao, Berger, Robert, Robert & Casella, Marin & Robert, Gelman *et al.*, Carlin & Louis, Leonard & Hsu, Bernardo & Smith, Parent & Bernier, mais aussi Cox & Hinkley, Lehman & Casella, etc. ! Je suis redevable au groupe de lecture BaBayes, initié par mon collègue biométricien de Jouy-en-Josas, Jean-Baptiste Denis, qui m'a permis d'éplucher les ouvrages fondamentaux de C. P. Robert, *The Bayesian Choice*, puis ceux de Gelman *et al.*, *Bayesian Data Analysis*, et de Carlin & Louis, *Bayesian Methods for Data Analysis*. Dans cette perspective, Jean-Baptiste Denis et moi-même avons lancé en 2006 avec Isabelle Albert, Chantal Guihenneuc et Eric Parent, le groupe informel *Applibugs* qui se réunit deux fois par an pour des échanges sur des travaux d'application de statistique bayésienne, avec notamment de jeunes chercheurs. Je ne manquerai pas non plus la lecture quotidienne des blogs de Christian Robert et d'Andrew Gelman qui, dans des styles très différents, m'ont souvent intrigué et m'ont fait découvrir ce qui se cachait dans les coulisses de la scène bayésienne. Enfin, on ne saurait se prétendre bayésien sans mettre la main dans le cambouis. Pour ma part, j'ai essayé de diffuser les rudiments de la méthodologie bayésienne dans le cadre d'un cours annuel sur le modèle mixte donné à l'ENSAI de Rennes. À cet égard, les logiciels *Winbugs* et *Openbugs* ont été – et me sont toujours – d'une aide précieuse non seulement pour mettre en oeuvre des applications de plus ou moins grande envergure, mais aussi pour tester des modèles ou des idées. J'y ai appris beaucoup de choses non seulement sur la statistique bayésienne, mais aussi sur les modèles graphiques et en particulier sur la structure des DAG, indispensable pour savoir comment spécifier une distribution conditionnelle en présence de nombreuses variables.

G. Celeux : S'il fallait conclure d'une ou deux phrases ton parcours bayésien ?

J.-L. Foulley : Il est en accord rétrospectivement avec mes origines, ma formation et ma forme d'esprit. Quant à la statistique bayésienne, je dirais que c'est une science d'ingénieurs, simple dans sa conception, mais plus délicate dans sa réalisation et qui passionne les théoriciens au point de les diviser !

Références

Berger, J. O. (2004), «The case for objective Bayesian analysis», *Bayesian Analysis*, vol. 1, pp. 1–17.

Dale, A. I. (1999), *A History of Inverse Probability from Thomas Bayes to Karl Pearson*, Springer-Verlag.

Dehaene, S. (2013), «Les principes bayésiens de l'apprentissage : sommes-nous des scientifiques dès le berceau ?», in «<https://www.college-de-france.fr/site/stanislas-dehaene/course-2013-01-08-09h30.htm>», Collège de France.

Fienberg, S. E. (2006), «When did Bayesian inference become Bayesian», *Bayesian Analysis*, vol. 1, pp. 1–40.

Foulley, J.-L. (2015), «A simple Bayesian procedure for forecasting the outcomes of the UEFA champion league matches», *Journal de la SFdS*, vol. 156, pp. 38–50.

Gianola, D. et J.-L. Foulley (1983), «Sire evaluation for ordered categorical data with a threshold model», *Genetics Selection Evolution*, vol. 15, pp. 201–224.

Henderson, C. R. (1973), «Sire evaluation and genetic trends», in «Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr J. Lush», pp. 10–41, American Society Animal Science-American Dairy Science Association.

Henderson, C. R., O. Kempthorne, S. R. Searle, et C. von Krosigk (1959), «Estimation of environmental and genetics trends from records subject to culling», *Bometrics*, vol. 13, pp. 192–218.

Jacquard, A. (1970), *Structure génétique des populations*, Masson.

Lecoutre, B. (1997), «C'est bon à savoir ! Et si vous étiez un bayésien qui s'ignore ?», *Modulad*, vol. 18, pp. 81–87.

Lefort, G. (1975), *Cours d'Introduction à la théorie de la décision et à la statistique bayésienne*, Institut National Agronomique Paris-Grignon.

Lefort, G. (1980), «Le modèle de base de la sélection : justification et limites», *Biométrie & Génétique*, pp. 1–14.

Leonard, T. H. (2014), «A personal history of Bayesian statistics», *Wires Computational Statistics*, vol. 6, pp. 80–115.

Lindley, D. V. (1957), «A statistical paradox», *Biometrika*, vol. 44, pp. 187–192.

Malécot, G. (1947), «Les critères statistiques et la subjectivité de la connaissance scientifique», *Annales de l'Université de Lyon*, vol. 10, pp. 43–74, traduit en anglais par Gianola, D. in GSE 31, 269-298.

Mayo, D. G. (2018), *Statistical Inference as Severe Testing*, Cambridge University Press.

McGrayne, S. B. (2011), *The Theory That Would Not Die*, Yale University Press.

Robert, C. P. (2014), «On the Jeffreys-Lindley paradox», *Philosophy of Science*, vol. 5, pp. 216–232.

Robert, C. P. (2016), «The expected demise of the Bayes factor», *Journal of Mathematical Psychology*, vol. 72, pp. 33–37.

Robert, C. P., N. Chopin, et J. Rousseau (2004), «Harold Jeffreys's theory of probability revisited», *Statistical Science*, vol. 24, pp. 141–172.

Stigler, S. (1986), *The History of Statistics: The Measurement of Uncertainty before 1900*, Belknap Press of Harvard University Press, Cambridge, MA.

Entretien avec Jacques Bernier et Éric Parent



Jacques BERNIER¹

Ingénieur EDF à la retraite

Éric PARENT²

AgroParisTech/INRAe, UMR Mathématique et Informatique Appliquées, Paris

Gilles CELEUX³

Institut de mathématiques d'Orsay, Université Paris-Sud et Inria Orsay

TITLE

Interview of Jacques Bernier and Éric Parent

RÉSUMÉ

Jacques Bernier est un ancien ingénieur EDF et il fut un pionnier de l'enseignement de la décision statistique à l'ISUP. Je témoigne de la qualité de son enseignement que j'ai jadis suivi avec plaisir. Depuis son départ à la retraite en 1991, Jacques Bernier continue à être très actif scientifiquement et à publier des articles de recherche en statistique bayésienne appliquée. Il est l'un des scientifiques qui aura le plus contribué à la diffusion et à la démonstration du potentiel de l'approche bayésienne en hydrologie.

Son compère et disciple, Éric Parent, est ingénieur des Ponts, des Eaux et des Forêts. Il travaille aujourd'hui comme enseignant-chercheur en statistiques appliquées et en modélisation probabiliste pour l'ingénierie environnementale à AgroParisTech. Il a fait énormément pour la pénétration de bonnes pratiques bayésiennes en ingénierie par les nombreuses recherches appliquées qu'il orchestre.

Mots-clés : *statistiques bayésiennes appliquées, ingénierie, hydrologie, environnement.*

ABSTRACT

Jacques Bernier was an engineer at the French Electricity Company (EDF). He pioneered a course on statistical decision at ISUP. I can witness the qualities of his course which I enjoyed when following as a former student of his. Since he retired in 1991, Jacques Bernier kept his scientific activities. He continues publishing articles in applied Bayesian statistics. He counts among the statisticians who contributed the most to the diffusion and the demonstration of the potential of a Bayesian approach in hydrology.

His disciple Éric Parent belongs to the French civil corps of Bridge, Road, Water and Forest Engineers. He presently works as a Research Teacher in statistics and probabilistic modeling applied to environmental engineering at AgroParisTech. He greatly contributed to the diffusion of good Bayesian practices in engineering through the numerous applied researches he conducted.

Keywords: *applied bayesian statistics, engineering, hydrology, environment.*

Premières rencontres avec les concepts bayésiens

G. Celeux : Jacques Bernier, Éric Parent, vous avez écrit de nombreux ouvrages en français qui promeuvent la pensée de Thomas Bayes et en illustrent les applications à l'ingénierie (Bernier et al., 2000; Parent et Bernier, 2007; Boreux et al., 2010; Parent et Rivot, 2012). Quels ont été vos premiers contacts avec ce paradigme, peu répandu en France à vos débuts ?

1. jacques.bernier2@orange.fr
2. eric.parent@agroparistech.fr
3. gilles.celeux@inria.fr

É. Parent : On peut être nourri de cours de statistique classique durant les études, que ce soit Métivier et Neveu à Polytechnique en 1978, ou Deheuvels et Benzecri en DEA à Jussieu en 1983, et demeurer bien plus intéressé à la beauté des espaces hilbertiens qu'à comprendre de quel chapeau le statisticien peut bien sortir un estimateur. Ce n'est, durant ma thèse où Jacques m'a beaucoup guidé, que la lecture de son⁴ livre avec Janine Ulmo me fait comprendre qu'un estimateur n'est finalement qu'une règle de décision particulière où l'espace des décisions peut s'identifier à celui des paramètres du modèle probabiliste. Dans les premiers chapitres de leur ouvrage encore timidement bayésien, Ulmo et Bernier (1973) présentent la théorie de la décision et le théorème de la classe complète pour l'admissibilité des règles de Bayes, mais la loi *a priori* reste vue comme une simple technique de convexification de l'ensemble des règles de décision, à la mode de Wald (1950) et non comme le fondement de la théorie subjective de la probabilité justifiant l'utilisation des concepts bayésiens (Savage, 1954). Sans le formuler d'une façon explicite, ils interprètent néanmoins le test du rapport de vraisemblance de Neyman et Pearson comme un facteur de Bayes (un rapport de crédibilités *a posteriori*) avec des lois *a priori* uniformes sur chacune des hypothèses en compétition. Voilà mon premier contact avec la cohérence du paradigme bayésien, la faute originelle à Jacques Bernier ! La lecture de Tribus (1969) dans sa traduction magnifique par Pezier (1972), m'apporte également beaucoup, je comprends que :

- la théorie de la décision de Ferguson (1967) permet de conduire un raisonnement argumenté pour construire une solution au projet d'engineering, ce qui me plaît en tant que futur ingénieur,
- la probabilité fait sens comme une réponse aux principes requis pour la construction d'un ordinateur à logique inductive selon les *desiderata* de Cox et Jaynes (1963) et Jaynes (1959), ce qui me plaît en tant que mathématicien.

Je m'aperçois ainsi qu'il existe une interprétation logique des probabilités indépendante de celle fondée sur la limite de fréquences, que tous mes professeurs, tant à l'Université que dans les Grandes Écoles, avaient passée sous silence. C'est cette interprétation subjective impersonnelle qui fournit aujourd'hui une justification à l'apprentissage automatique. À l'époque, mon incursion dans le contexte bayésien me permet de mieux comprendre le paradigme classique et de le positionner par rapport au cadre bayésien :

- L'approche fréquentiste, qu'on devrait plutôt appeler fréquentielle, décrit l'incertitude qui porte sur la valeur obtenue par une procédure d'estimation si on répétait celle-ci pour un grand nombre de fois pour des répliqués générés par le modèle probabiliste dont on connaîtrait les inconnues. Il s'agit donc de la fiabilité d'une procédure de calcul algorithmique en information parfaite. L'incertitude se manifeste dans le contexte classique au travers de la variabilité de répétition des observables conditionnellement à la connaissance des inconnues.
- L'approche bayésienne évalue l'incertitude qui reste sur les inconnues, conditionnellement à l'unique jeu de données observées, sachant que l'on est parti d'une croyance initiale sur leurs variations possibles. Il s'agit donc d'un jugement probabiliste direct sous la forme d'un pari quant à la valeur de ces inconnues.

G. Celeux : Je crois savoir que Jacques de vingt ans l'ainé d'Éric, partage plutôt la conception décisionnelle de la probabilité subjective, interprétée comme un pari personnel cohérent. Ce n'est pas étonnant puisque vous avez pu assister aux conférences données à l'IHP par Léonard Savaga (1954) et Bruno de Finetti (1937) dans les années cinquante. À EDF où vous avez fait l'essentiel de votre carrière, Jacques, avez vous ressenti à l'époque, l'impulsion de ce type de recherches mettant en application les travaux de Von Neumann et al. (1953) ou d'Allais (1953) ?

4. Heureuse bénédiction pour le thésard que j'étais, la bibliographie de Jacques était alors encore succincte...

J. Bernier : De formation universitaire en mathématiques et statistique (entre 1951 et 1955 – Sorbonne puis IHP), j'ai suivi parallèlement de 1953 à 1955 les cours de Probabilités (en partie avec Georges Darmois) et de l'ISUP (au départ avec Eugène Morice puis Georges Morlat). J'y ai bien entendu prononcé le nom de Bayes. Les commentaires de Darmois sur la mise en oeuvre de la formule de Bayes étaient assez réservés. Quant à Morlat, son objectif, en deuxième année de l'ISUP, était les tests et la théorie de la Décision Statistique de Wald (parus en 1950) et bien sûr le concept de « *règle de décision bayésienne* » y était présenté sans que le paradigme bayésien y soit bien développé. Il ne faut pas oublier que Wald lui-même était un « *fréquentiste* » et le raisonnement bayésien conçu par lui était un moyen pour construire des règles admissibles (propriété éminemment fréquentiste). Connaissant Morlat (1964) comme je l'ai connu par la suite, on pourrait s'étonner de sa prudence en ce qui concerne le paradigme bayésien mais on ne saurait oublier le contexte à l'ISUP à cette époque et même après ; le cours parallèle d'Indoudjian sur l'estimation était résolument fréquentiste et fisherien. Mais en septembre 1955, j'ai été engagé par Morlat dans la Division Statistique qu'il dirigeait, au Service de Recherches Hydrauliques d'EDF et là, tout a changé pour moi. Tout d'abord, je souhaite à tout jeune débutant (j'avais 23 ans) de travailler d'emblée avec un patron tel que Morlat, avec ses grandes qualités humaines, sa façon de transmettre son expertise ainsi que les nombreuses occasions offertes à ses subordonnés de développer leur environnement. Ainsi ai-je pu rencontrer Marcel Boiteux avec lequel Morlat avait des liens privilégiés. Je suis resté quatre ans avec lui jusqu'à la dissolution de la division statistique par un nouveau directeur dont l'idée était de regrouper dans un seul service tous ceux qui utilisaient tant soit peu les mathématiques : IMA, pour Informatique et Mathématiques Appliquées⁵. Le résultat : Morlat est parti rejoindre Boiteux et moi-même, avec quelques autres compagnons d'exil, je suis parti faire un séjour de 2 ans à Gaz de France avant de revenir à EDF, rappelé par Joseph Jacquet en 1962.

La Division Statistique de Morlat à EDF a une histoire qui nous ramène de fait à Bayes. Elle est l'émanation d'un groupe de statisticiens constitué autour du normalien Etienne Halphen, dès la nationalisation de l'Électricité de 1946 et du Service Hydrotechnique regroupant les études des anciennes compagnies d'électricité. C'est Pierre Massé, futur président d'EDF, qui l'avait accueilli dans ces services dès 1940 pour le protéger des occupants allemands. Les travaux très remarquables d'Halphen dans les domaines de la statistique et de ses applications à l'économie de l'électricité se sont poursuivis et développés, après la nationalisation, avec l'arrivée en 1948 de deux jeunes universitaires Georges Morlat et Lucien Le Cam (ce dernier est parti ensuite rejoindre Neyman à l'Université de Californie à San Francisco). Halphen a poursuivi ses travaux en liaison étroite avec Morlat jusqu'à sa mort prématurée en 1954 (à 44 ans).

Engagé en septembre 1955, je n'ai donc connu Halphen que par Morlat et aussi par les papiers, notes internes et articles publiés par Halphen. Son influence sur Morlat et sur moi-même est très importante. Morlat a rapporté qu'Halphen lisait peu, sa pensée était cependant originale : il était persuadé qu'un statisticien devait avoir un rôle éthique dans ses relations avec « *son client* » en lui transmettant la cohérence des principes statistiques. Il ne connaissait pas Savage et je ne pense pas qu'il ait jamais lu de Finetti. Cependant, il avait une conception de la probabilité proche de la conception subjective et refusait le fréquentisme au moins pour les problèmes qu'il avait à traiter. « *Je ne fais pas du contrôle de fabrication* », disait-il ! Morlat avait acquis cette conception enrichie par la lecture de de Finetti (1937) et de Savage (1954). Avec lui, j'ai pu écouter Savage en 1958 lors d'une invitation à l'IHP. Je rappelle qu'outre ses activités d'enseignement à l'ISUP, Morlat était très engagé dans les activités de la RO, Recherche Opérationnelle, alors en très fort développement (époque de la création de la SOFRO et de sociétés telles la SEMA). Il m'avait poussé à donner des cours de statistique décisionnelle au CIRO (Centre Interarmées de Recherche Opérationnelle au Fort de Montrouge). De ces premiers pas avec Morlat, date mon intérêt sur les méthodes bayésiennes. En ce qui concerne ma

5. je ne peux m'empêcher de rapporter le slogan qui s'est propagé ensuite dans ce service "*l'IMA, c'est pas le PEROU !*"

propre activité de recherche à EDF, je me suis résolument tourné vers la mise en oeuvre de la philosophie et des méthodes bayésiennes surtout après mon retour aux Etudes et Recherche à Chatou en 1962. Un problème important était, à cette époque, la planification et le contrôle statistiques des expériences d'augmentation artificielle de la pluie sur les bassins alimentant les réservoirs. Dès 1962, on y a développé à la fois des techniques de test classiques et tenté la mise en oeuvre de décisions bayésiennes. Ces travaux ont fait l'objet d'une présentation au 5^e *Berkeley Symposium on Mathematical Statistics and Probability* de 1965. Je rappelle que ces rencontres étaient organisées par Neyman et Le Cam que j'ai eu l'honneur de rencontrer à ce moment. Neyman était particulièrement intéressé par la pluie artificielle et était à l'origine des séances consacrées au sujet. C'est là que le paradigme bayésien et son concept de probabilité subjective sont particulièrement bien adaptés et, dirais-je, même nécessaires dans les problèmes de décision face aux risques provoqués par les événements extrêmes. Crues de rivières, tremblements de terre, tempêtes maritimes, etc. qui peuvent menacer les ouvrages EDF; ces divers aspects ont fait l'objet d'études selon Bayes. C'est ainsi que les premières études bayésiennes de crue ont été présentées dès 1967 au symposium AIRH de Fort Collins.

Avec d'autres collègues d'EDF et d'instituts partenaires, nous avons consacré une part importante de nos activités à des développements méthodologiques mis en oeuvre dans ces problèmes ainsi que des questions environnementales diverses (propagation de la pollution des rivières et l'estimation des modèles mathématiques hydrodynamiques, etc.). Dans ces périodes, notre pierre d'achoppement était bien sûr le manque de méthodes adéquates de calcul bayésien pour les modèles d'extrêmes. Nous avons fait avec les moyens du bord : utilisation des propriétés asymptotiques et approximations (lorsque c'était possible) par les modèles gaussiens ou transformés (logarithme, etc.)

Je parle ici de cette période d'avant 1990, appelons la pré-Robert, où on ignorait en France les méthodes de calcul MCMC et autres Gibbs, Metropolis Hastings, ... et dont la succession des divers logiciels de calcul modernes, BUGS, JAGS, STAN, etc. provoque encore aujourd'hui une véritable libération bayésienne !

Statistique et statistique bayésienne dans l'enseignement en France à la fin du 20e siècle

G. Celeux : Vous nous dites que la statistique bayésienne avait commencé à se répandre en France dans les applications de recherche et développement assez tôt. Il me semble pourtant que le milieu académique français soit resté longtemps insensible à la théorie bayésienne, quoi que de nombreux développements apparaissent outre-Manche (Jeffreys, 1961) et outre-Atlantique (Raiffa et Schlaifer, 1961).

É. Parent : Je peux témoigner de mon expérience d'enseignant en Écoles d'Ingénieur vers 1990. Chargé de mettre en place un cours de statistique pour les élèves-ingénieurs de l'École des Eaux et Forêts à la fin des années quatre-vingts, j'opte néanmoins pour le paradigme classique en ce qui concerne le cours principal, mais avec des modules optionnels bayésiens pour l'analyse du risque (contrôle de la qualité et calcul des valeurs de projets en hydrologie). Pourquoi pas complètement bayésien ? Sans doute mu par une volonté de compromis pour le réaliser avec mon ami Azaïs (qui deviendra reconnu parmi les fréquentistes, voir par exemple Azaïs et Bardet (2006)) et beaucoup de conformisme pour s'inscrire plus facilement dans le programme classique : échantillons et lois empiriques/Estimation/Tests/Modèles sous l'hypothèse linéaire générale. À la même époque, Jean-Pierre Raoult⁶, me sollicite dans son équipe pour renouveler un cours de statistique, qualifié pourtant de décisionnelle à l'école des Ponts et Chaussées, mais

6. pourtant incontestablement intéressé par le caractère fructueux de l'approche bayésienne dans sa propre recherche puisqu'il a été le directeur de thèse de Christian Robert.

là encore il suit la piste fréquentiste sauf en un minuscule et unique paragraphe du nouveau manuel (que je peux rédiger). Le cas de l'Institut National Agronomique est particulièrement intéressant : dès les années soixante-dix, Guy Lefort a tenté de proposer aux étudiants un cours de statistique bayésienne (Lefort, 1975). Par un choix astucieux de la loi *a priori*, il retombe sur les formules ordinaires des tests de comparaison de moyennes, de variances, etc., ce qui permet de ne pas surcharger la mémoire des éventuels amateurs de recettes d'une boîte à outils statistiques standardisée. À sa mort prématurée, le cours de statistique donné par Duby et Daudin, retourne sous une forme plus classique qui deviendra celui enseigné aujourd'hui (Daudin, 2015).

J. Bernier : Je peux témoigner pour ma part de mon expérience à l'ISUP où j'ai commencé dès les années 1960 comme chargé de TD puis chargé du cours de décision statistique et tests en succédant à Morlat en 1968 jusqu'en 1990. Dans cette période la pensée statistique à l'ISUP était surtout influencée par le fréquentisme et Fisher. Le cours sur l'estimation de Jeanine Ulmo, parallèle au mien, était de philosophie fréquentiste en suivant les ouvrages de Lehmann. Je rappelle qu'après 1968, les jeunes enseignants universitaires de statistique et probabilités se réunissaient régulièrement pour « lire le Lehmann », réputé comme un ouvrage de statistique particulièrement « propre » sur le plan mathématique : Bourbaki avait aussi fait des émules parmi les statisticiens ! Entraîné par le courant ambiant, avouons-le, j'ai donc moi aussi fait du Lehmann dans mon cours. Toutefois, je ne laissais pas mes auditeurs ignorer que pouvaient exister certaines méthodes bayésiennes. Encore faut-il préciser que ma référence à Lehmann et ses présentations séparées des tests et de l'estimation reste fortement relativisée au départ par l'exposé de la synthèse décisionnelle de Wald.

Cependant j'ai, à la même époque, l'occasion de faire quelques cours de statistique dans d'autres enceintes (Ecole Centrale, Ponts et Chaussées, Eaux et Forêts) et, là, pour en illustrer les exposés, je n'hésite pas à présenter quelques travaux de statistique appliquée effectués à EDF où la philosophie du comportement individuel du chercheur devant l'incertitude et les méthodes bayésiennes jouent le rôle essentiel.

Je ne connais le monde enseignant statistique post-1990 que grâce à Éric Parent et je suis entièrement d'accord avec lui sur l'enseignement de la statistique en France comparé à ce qui est fait dans le monde anglo-saxon. Hors de France j'ai cependant eu le loisir de propager les idées bayésiennes et les travaux d'Halphen par une longue collaboration avec Bernard Bobée de l'Université du Québec.

Il faut reconnaître l'efficacité du système éducatif anglo-saxon. Il en résulte une production méthodologique de qualité et des outils bayésiens informatiques remarquables. On en connaît nombre de réalisations depuis les travaux précurseurs de Spiegelhalter : les logiciels *Winbugs*, *Bugs*, les utilisations des graphes représentant les modèles hiérarchiques généraux ainsi que les multiples *Bayesian packages* en R et l'aboutissement *Jags*, logiciel solveur Gibbs, particulièrement clair et pratique. On ne saurait oublier O'Hagan et ses travaux sur l'élicitation des lois *a priori* et le traitement des incertitudes et l'émulation des codes complexes (MUCM group - Sheffield University). Je peux me tromper mais il me semble que les contributions françaises à ces développements de *softwares* restent assez limitées. J'avais beaucoup parié sur le développement par l'équipe de Bordeaux du logiciel BIIPS (Todeschini et al., 2014), fondé sur des algorithmes particuliers, qui ne s'est malheureusement pas diffusé à la mesure de son mérite, me semble-t-il.

L'école américaine n'est pas en reste avec Gelman et ses contributions au développement de logiciels tels que *Stan*, le compétiteur de *Jags*, reposant sur les algorithmes Hamilton-Metropolis-Hastings, les travaux de Higdon sur les émulateurs de codes numériques, ceux de Raftery, etc.

Mais il ne faut certes pas laisser de côté les contributions nombreuses à la pédagogie bayésienne de Christian Robert et son école aussi bien dans le domaine des publications méthodologiques générales (*The Bayesian Choice 2001*, etc.) que dans les multiples applications particulières (comme les modèles de mélange). Sur le plan philosophique, il me plaît de signaler sa contribution à un article coopératif important (avec Gelman et autres) de 2019 *Abandon Statistical Significance*, paru dans le numéro 73 de *The American Statistician*.

Pour en revenir à la pédagogie statistique, il faut malgré tout reconnaître le déclin, après 2000, de la statistique décisionnelle traditionnelle à la Wald qui a dominé la seconde moitié du 20^e siècle, malgré des ouvrages tels que Parmigiani et Inoue (2009). Elle semble remplacée par la théorie de l'apprentissage à la façon de Vapnik (2013).

Il me semble que c'est la statistique à la Lehman qui reste majoritairement enseignée en France au niveau universitaire...

G. Celeux : On peut contester cela... Je dirais plutôt que c'est la statistique non paramétrique qui est maintenant privilégiée.

J. Bernier : Peut-être, en tout cas la statistique cohabite avec la nouvelle mode du *data mining*. Il y a aujourd'hui toute une panoplie d'outils bayésiens utilisés en *data mining* et *machine learning*. Il faut noter que dans tous ces outils, les fondements bayésiens invoqués appartiennent à la version de la probabilité personnelle comme extension de la logique binaire Cox-Jaynes, concurrent de la probabilité subjective décisionnelle. Cependant l'adjectif décisionnel y est souvent accolé (voir par exemple Tufféry (2012)). Il m'apparaît que les décisions impliquées ont pourtant peu à voir avec les formalisations statistiques traditionnelles. Celles-ci sont-elles obsolètes ?

Et finalement, je dois dire que je suis chagriné que l'ISUP, mon douar d'origine, ait abandonné la statistique traditionnelle, y compris sous sa forme bayésienne pour changer son nom en *Data Centre* afin d'adopter l'enseignement de ces techniques regroupant logique floue, probabilités diverses, et méthodes associées, analyses descriptives des données⁷, réseaux bayésiens, etc., inventaire dans lequel on s'attendrait à retrouver les ratons laveurs de Jacques Prévert !

Des bayésiens isolés et maudits ?

G. Celeux : La perspective bayésienne ne s'implante donc pas dans le milieu académique et de recherche français autant qu'elle vous semble présente dans le monde anglo-saxon. Les individus bayésiens se sont-ils sentis isolés ?

É. Parent : Tu as raison. Rouanet s'est alarmé par exemple que le raisonnement statistique perde sa cohérence (Rouanet et al., 1991), et je suis sûr, pour l'avoir rencontré, qu'il en souffrait presque d'un point de vue personnel. Lecoutre (1997) ne comprend pas qu'un statisticien bien éduqué ne puisse pas se rendre compte que le raisonnement bayésien se glisse diaboliquement sous les pierres de fondation de la statistique classique. En Angleterre, ce phénomène s'est produit plus tôt. Face à la prééminence de Fisher et de ses disciples, le bayésien Jeffreys (1961) a eu bien du mal à trouver sa place, mais à partir des années 60, c'est terminé : Lindley diffuse et affirme la pensée bayésienne, à la fois en théorie (Lindley, 1965) et la vulgarise efficacement (Lindley, 1991, 2006). Hors de France se fédère ensuite une communauté bayésienne anglo-saxonne forte autour des disciples de Lindley (Bernardo, Berger, Dawid, O'Hagan...) qui peut parfois donner à l'extérieur une impression quasisectaire. Ce monde anglo-saxon a exercé une

7. triomphe ultime de Benzecri !

forte attraction sur nombre de bayésiens français qui y ont fait ou y font encore de longs séjours : Christian Robert à Cornell puis à l'université de Warwick, Nicolas Chopin à l'université de Bristol, Arnault Doucet et Judith Rousseau à Oxford, moi-même à Cambridge. La statistique bayésienne ne se développe véritablement qu'au début des années 1990 – en France comme ailleurs – avec l'avènement des méthodes de simulations fondées sur les algorithmes d'importance sampling et de chaînes de Markov (Robert et Casella, 1998). Associées à la constante augmentation de la puissance de calcul de nos ordinateurs, elles ont complètement libéré, la puissance modélisatrice de l'approche bayésienne (Brooks, 2003). Spiegelhalter et son équipe développent puis offrent au domaine public le premier outil générique d'inférence bayésienne, le formidable logiciel gratuit *Bugs* (Spiegelhalter et al., 1996). En France, c'est avec l'école de Robert et al. (2004) et de Marin et Robert (2007) que se diffuse largement la perspective bayésienne dans l'université à partir des niveaux Master (Paris-Dauphine, Montpellier, etc.). Pour prendre le recul nécessaire avec cette vision trop focalisée sur les cinquante dernières années, on trouvera dans McGrayne (2011) l'intéressante épopée historique des protagonistes de la démarche bayésienne, *cette théorie qui ne voulait pas mourir*.

L'engouement actuel pour l'informatique a certainement fait basculer le débat du champ philosophique « *Quel caractère scientifique accorder à la probabilité bayésienne, revendiquée comme une quantification de la relation entre un individu et les manifestations incertaines de son objet d'études* » au cadre pratique de l'implémentation algorithmique « *Éviter de calculer une énorme intégrale afin d'évaluer une probabilité conditionnelle* ». Et, sous le poids de la nécessité technique partagée aujourd'hui avec d'autres, semblent être dissipées les aigreurs des querelles statisticiennes passées.

J. Bernier : Je ne suis pas d'accord sur ce soi-disant isolement des bayésiens dans la période pré-1990. Certes les difficultés calculatoires bayésiennes cessent à partir des années 1990, en France comme ailleurs. Mais elles n'ont pas empêché une production d'applications notables même partielles. Un ouvrage essentiel *Applied Statistical Decision Theory* de Pratt, Raiffa et Schlaifer en 1965 a marqué cette époque. On y trouve notamment la définition et une mise en oeuvre des distributions conjuguées naturelles dont l'importance s'est trouvée confirmée avec les algorithmes de Gibbs. Ces auteurs avaient publié dès 1964 dans JASA un article *The Foundations of Decision Under Uncertainty : An Elementary Exposition* exposant les bases élémentaires de l'axiomatique de Savage. Bien qu'assez tôt la théorie de la décision ait été l'objet de critiques notamment de la part de Maurice Allais concernant ceux qu'il appelait « *l'école américaine* » et l'axiome dit « *sure thing principle* » nécessaire à la mise en oeuvre de la formule de Bayes, ces discussions n'ont nui en rien aux développements des applications de la théorie, des méthodes bayésiennes et du critère de l'utilité espérée, au moins hors de France.

Sur le plan de la pratique et parmi d'autres applications, je citerais *Bayesian decision theory applied to design in hydrology* par Davis, Kisiel et Duckstein dans *Water Resources Research* en 1972, dont mes contacts après la parution m'ont permis de commencer une longue amitié avec Lucien Duckstein (Davis et al., 1979). Les bayésiens de cette époque n'étaient certes pas si isolés et maudits que cela !

Quel avenir à la statistique bayésienne ?

G. Celeux : Nous voilà donc à l'époque du *Big Data*, de l'apprentissage automatique et de l'intelligence artificielle... Que pensez vous de l'avenir de la statistique bayésienne dans ces conditions peu propices ?

J. Bernier : L'importance⁸ de l'informatique dans le développement des méthodes d'IA aujourd'hui ou celles des méthodes computationnelles bayésiennes au tournant du siècle ne doivent pas faire oublier la nécessité de la rigueur du langage, en particulier quand il s'agit d'incertitudes. Le paradigme bayésien s'impose naturellement lorsqu'un individu doit raisonner et agir face à son incertitude concernant les conséquences de ces actes. Le recueil d'informations et leurs traitements par modélisation probabiliste et algorithmes informatiques ne sont que contingents à l'objectif de base. On ne peut justifier Bayes ou n'importe quel autre paradigme uniquement sur des arguments de facilité de calcul.

L'incertitude ne décrit qu'un état psychologique d'un individu face à un environnement particulier où il doit agir. Pour cela il devra résoudre son problème de décision avec des données et au moyen de modèles traités par des algorithmes mobilisant des informations.

On ne peut donc analyser le tryptique *problème incertain/ données/ modèles & algorithmes de calcul* dans n'importe quel sens. Or il est regrettable que la tendance existe souvent dans la littérature technique où les choix dictés par les besoins de calcul l'emportent sur les choix, disons philosophiques, liés à la signification de l'incertitude et à son application au problème en main.

Dans cette optique, la mienne, on ne peut dire, par exemple, que l'incertitude se manifeste par la variabilité de répétition des observables. Cette variabilité là, dite source d'incertitude pour le statisticien, n'est qu'un état naturel des données⁹. Méfions nous des noms ou adjectifs fallacieux comme *l'intelligence* des données, autre nom d'usage du *data mining*.

É. Parent : Et même si on se revendique rigoureux en s'appuyant sur le calcul des probabilités, tout ne part pas de Kolmogorov ! Hájek (2003), comme tous les philosophes de la connaissance, distingue bien la question : *How should probability theory be formalized ?* qui est une question mathématique (à laquelle Kolmogorov apporte une réponse) et doit être séparée de : *What do statements of probability mean ?* qui est une question philosophique dont la réponse doit venir d'autre part, hors du domaine des mathématiques, c'est-à-dire de l'homme ou de son environnement naturel, et là, les réponses sont de nature subjectiviste ou fréquentiste. C'est aussi tout ce qui fait la différence entre Calcul de Probabilités et Statistique comme l'exprimait Halphen dans les années 1950 : « *Il est acquis que la statistique n'est scientifique que dans la mesure où elle se tient en liaison avec le calcul des probabilités. Mais elle ne se réduit pas à ce dernier, et ne s'y réduira jamais, pour la raison fort simple que le calcul des probabilités est une construction idéale vide de tout contenu réel : un mathématicien peut faire de la géométrie Euclidienne, Riemannienne, Hilbertienne... a priori : il faut un physicien pour savoir laquelle de ces géométries correspond à la réalité. On peut donc dire réciproquement que le calcul des probabilités ne sera fécond que s'il se tient en liaison avec la statistique : sinon il résoudra de beaux problèmes qu'il se sera donnés à lui-même pour son amusement, mais qui n'ont aucun intérêt* ». Kadane (2011) exprime la différence autrement : « *To be a good statistician requires grounding in each of the disciplines we rely on : mathematics, computing and philosophy* ».

G. Celeux : Cependant, l'inférence bayésienne bénéficie aussi des nouveaux moyens de calcul et la recherche en algorithmique bayésienne est particulièrement active.

J. Bernier : Je me réjouis de la nouvelle multiplicité des moyens de calcul à disposition du bayésien : algorithmes, logiciels (*Stan, Jags, etc.*), *packages* R et autres systèmes, actuels et en devenir. Il est clair que les statisticiens classiques en profitent aussi largement : ne peut-on pas

8. Son caractère ludique ne pourrait-il pas y être aussi pour quelque chose ?

9. État normal mais l'adjectif peut être trompeur car en météorologie par exemple, la moyenne de température a été affublée du nom de « normale » alors que la variabilité est un état tout aussi normal des processus météorologiques

considérer que l'algorithme EM n'est qu'un cas particulier de l'algorithme de Bayes variationnel (Keribin, 2010), si on étend la famille des approximations aux lois de Dirac ? Toutefois je pense qu'il est vital de ne pas laisser tomber les fondements bayésiens dans les décombres du passé, de même que les enseignements de la théorie de la décision statistique, seule façon de justifier le caractère dit décisionnel de certaines procédures à la mode. Comme le dit Éric : « *Tout ne part pas de Kolmogorov !* ». Mais l'axiomatique mathématique de la probabilité a un poids si fort aux yeux de certains qu'elle paraît incompatible avec « l'interprétation aléatoire des paramètres des modèles probabilistes ». Curieusement dans la préface de l'ouvrage d'Éric et de ses collègues (Albert et al., 2015), Jean-Michel Marin nous propose alors de discuter l'idée selon laquelle l'approche bayésienne se situerait à un autre niveau sémantique qui ne remet pas en cause cette modélisation de la réalité, en supposant que la loi *a posteriori* est un outil de résumé de l'information disponible sur les paramètres, sans remettre en cause également l'existence de ces paramètres inconnus certes mais non aléatoires. Malgré toute mon admiration pour cet ami d'Éric incontestablement bayésien, et, de surcroît, notre actuel président de la SFdS, je trouve cette position assez difficile à comprendre car ce niveau sémantique n'est jamais bien défini. Il me semble plus clair d'admettre avec Hajek et Halphen, comme le rapporte Éric Parent, que la statistique ne peut se réduire au calcul des probabilités désincarnées de Kolmogorov et que l'interprétation concrète de la probabilité demande une formulation adaptée concrètement au type d'incertitude dans laquelle est placée la personne concernée. S'agissant donc d'interprétation du concept de probabilité, deux positions cohérentes sont proposées. La première est l'interprétation logique nécessaire de Cox-Jaynes, extension de la logique binaire fondée sur des axiomes de comportement compatibles, après calculs, avec l'axiomatique de Kolmogorov. Nous n'avons pas suivi cette voie, la logique du robot raisonneur, car sur le plan de la mise en oeuvre, elle perd de sa cohérence au niveau de l'élicitation des lois *a priori* au moins pour de nombreux problèmes. Cette cohérence, comme demandée par Lindley (2006) est assurée par l'approche subjective décisionnelle de de Finetti, Savage, Raiffa, ... à laquelle Éric et moi adhérons.

É. Parent : Quel que soit le paradigme, d'accord, la modélisation a donc recours au calcul des probabilités.

Mais je pense qu'il nous faut refuser l'oecuménisme : même si la sauce mélangeant bayésien et classique a bien belle apparence, elle n'est pas digeste. Tandis que l'inférence classique requiert des techniques d'optimisation, la recherche de la loi *a posteriori* n'engendre pas une telle rupture technique : la loi *a posteriori* est elle-même un objet du calcul des probabilités et les algorithmes de simulation de Monte-Carlo font également largement recours à la mathématique des probabilités. Voilà où un même outil mathématique est à toutes les sauces et il y a là un véritable problème de compréhension et de communication, car un grand nombre d'exposés, vus sur le Net, au-delà d'une présentation mathématique axiomatique à la Kolmogorov, justifiant les concepts de variables aléatoires X , de densités de probabilités $f(x, \theta)$ et de paramètres, ne font que distinguer verbalement :

- l'approche fréquentiste : θ certain,
- et l'approche bayésienne où θ incertain est supposé appartenir à une sigma-algèbre d'évènements munie d'une densité $\pi(\theta)$.

Hélas, la démarche bayésienne y est (plus ou moins elliptiquement et très verbalement) *expliquée* par la simple distinction technique entre aléa sur X et incertitudes dites épistémiques sur θ .

J. Bernier : Pour les futurs bayésiens, la vie sera plus simple : de Finetti a prouvé l'existence conjointe de la loi *a priori* et de la vraisemblance, et il s'agira simplement de les utiliser à bon escient. L'approche subjective décisionnelle est basée sur deux piliers :

- l'élicitation des probabilités *a priori*, formalisée par Savage (1971) avec la technique des

paris et justifiée par un corps de principes de comportement devant l'incertitude dont essentiellement le principe dit « *d'évitement des pertes sûres* » et le principe de prévision dit « *sure thing principle* ».

- l'acceptation du théorème de de Finetti, dit de représentation, reposant sur l'hypothèse d'échangeabilité simple ou partielle des observations. Cette hypothèse remplace l'hypothèse d'indépendance et de distributions identiques des observations sur laquelle se fonde la statistique fréquentiste. Le théorème est fondamental pour le comportement du bayésien devant l'information et justifie, de fait, l'interprétation aléatoire des paramètres des modèles statistiques. Ce théorème devrait être un outil de modélisation fondamental pour tout bayésien. Il faut noter que des auteurs comme Robert ou Gelman y accordent peu d'importance ou ne le citent même pas. Dans son livre fondateur de 1954, Savage, après de Finetti, le présente sans lui donner la même importance qu'à son système d'axiomes. Mais cela lui a permis, l'année 1955 suivante, de généraliser, avec Hewitt, le théorème aux processus d'information les plus généraux et qui est connu maintenant sous le nom de théorème de de Finetti-Hewitt-Savage.

Personnellement, j'ai évolué dans ma perception de la théorie subjective décisionnelle de la probabilité en déplaçant sa pierre angulaire pour les applications, de l'élicitation vers le théorème de représentation en accord, d'ailleurs, avec le jugement de Bernardo (1996). Bien entendu, mon opinion personnelle ne préjuge pas des orientations que suit ou suivra le monde des statisticiens du XXI^e siècle, ma personne étant trop limitée à cet égard. Simplement mon jugement reflète une expérience assez longue de ma pratique en même temps que celle des autres, ceci grâce en soit rendue à Éric Parent qui me sollicite souvent pour me sortir de mon état de retraité.

É. Parent : La statistique, c'est ce que font les statisticiens. Si nous nous sentons heurtés aujourd'hui par la vague du *Big Data* et de l'intelligence artificielle, c'est que la dialectique qu'elle emploie nous met face à deux de nos travers.

- Nous travaillons sans doute trop seuls : c'est pour celà que le *data scientist*, comme ils disent, ressemble à un homme-orchestre, à la fois statisticien, informaticien, gestionnaire de données, analyste numérique et psychologue. Je suis surpris, par exemple, par l'irruption du bayésien dans les sciences cognitives contemporaines : nos modèles mentaux du monde verraient-ils leur plausibilité vis-à-vis des observations réévaluées selon la formule de Bayes (Seriès, 2016) ? Dehaene (2013) y voit même les fondements d'une inférence probabiliste chez le bébé!
- Nous nous faisons comprendre sans doute trop mal : c'est pour celà que la nouvelle vague veut changer la donne en changeant les termes : on ne fera plus de modélisation hiérarchique mais on construira un profond réseau de neurones ; on ne fera plus d'inférence, mais de l'apprentissage statistique voire un apprentissage automatique, algorithmique, adaptatif...

Mais sous ce grand et futile chamboulement linguistique, ne percevez-vous pas finalement la réminiscence de la formule de T. B. ? Une petite formule, jetée il y a bientôt trois siècles dans l'océan des connaissances mathématiques qui a généré des ondes qui pourraient bien s'achever en un tsunami.

Références

Albert, I., S. Ancelet, O. David, J.-B. Denis, D. Makowski, E. Parent, A. Rau, et S. Soubeyrand (2015), *Collectif BioBayes : Initiation à la statistique bayésienne - Bases théoriques et applications en alimentation, environnement, épidémiologie et génétique*, Références sciences. Edition Ellipses, URL <https://hal.archives-ouvertes.fr/hal-01194144>.

- Allais, M. (1953), «Le comportement de l'homme rationnel devant le risque : critique des postulats et axiomes de l'école américaine», *Econometrica : Journal of the Econometric Society*, pp. 503–546.
- Azaïs, J.-M. et J.-M. Bardet (2006), «Le modèle linéaire par l'exemple régression, analyse de la variance et plans d'expériences illustrations numériques avec les logiciels r, sas et splus».
- Bernier, J., E. Parent, et J.-J. Boreux (2000), *Statistique Pour L'environnement. Traitement Bayésien de L'incertitude.*, Lavoisier, TEC et DOC, Paris.
- Boreux, J.-J., E. Parent, et J. Bernier (2010), *Pratique du calcul bayésien*, Springer Science & Business Media.
- Brooks, S. P. (2003), «Bayesian computation: A statistical revolution», *Trans. Roy. Statist. Soc., series A*, vol. 15, pp. 2681–2697.
- Cox, R. T. et E. T. Jaynes (1963), «The algebra of probable inference», *American Journal of Physics*, vol. 31, no 1, pp. 66–67.
- Daudin, J.-J. (2015), *Le modèle linéaire et ses extensions : Modèle linéaire général, modèle linéaire généralisé, modèle mixte, plans d'expériences*, Technosup (Paris), Ellipses.
- Davis, D. R., L. Duckstein, et R. Krzysztofowicz (1979), «The worth of hydrologic data for nonoptimal decision making», *Water Resources Research*, vol. 15, no 6, pp. 1733–1742.
- de Finetti, B. (1937), *La Prévision : ses Lois Logiques, ses Sources Subjectives*, Institut Henri Poincaré, Paris.
- Dehaene, S. (2013), «Les principes bayésiens de l'apprentissage : sommes-nous des scientifiques dès le berceau ?», in «<https://www.college-de-france.fr/site/stanislas-dehaene/course-2013-01-08-09h30.htm>», Collège de France.
- Ferguson, T. (1967), *Mathematical Statistics, a Decision Theoretic Approach*, Academic Press.
- Hájek, A. (2003), «Interpretations of probability», in «The Stanford Encyclopedia of Philosophy (Zalta)», CiteSeer.
- Jaynes, E. T. (1959), *Probability theory in science and engineering*, 4, Socony Mobil Oil Company Field Research Laboratory.
- Jeffreys, H. (1961), *Theory of Probability*, Oxford : Clarendon Press.
- Kadane, J. B. (2011), *Principles of Uncertainty*, Texts in Statistical Science, Chapman & Hall.
- Keribin, C. (2010), «Méthodes bayésiennes variationnelles : concepts et applications en neuroimagerie», *Journal de la Société Française de Statistique*, vol. 151, no 2, pp. 107–131.
- Lecoutre, B. (1997), «C'est bon à savoir ! Et si vous étiez un bayésien qui s'ignore ?», *Modulad*, vol. 18, pp. 81–87.
- Lefort, G. (1975), *Cours d'Introduction à la théorie de la décision et à la statistique bayésienne*, Institut National Agronomique Paris-Grignon.

- Lindley, D. V. (1965), *Introduction to Probability and Statistics from a Bayesian Viewpoint*, University Press.
- Lindley, D. V. (1991), *Making Decision*, John Wiley & Sons, INC, New York.
- Lindley, D. V. (2006), *Understanding Uncertainty*, John Wiley & Sons, INC, New York.
- Marin, J.-M. et C. Robert (2007), *Bayesian core: a practical approach to computational Bayesian statistics*, Springer Science & Business Media.
- McGrayne, S. B. (2011), *The Theory That Would Not Die*, Yale University Press.
- Morlat, G. (1964), «Statistique et théorie de la décision», *Revue de Statistique Appliquée*, vol. 12, no 2, pp. 5–13, URL http://www.numdam.org/item/RSA_1964__12_2_5_0.
- Parent, E. et J. Bernier (2007), *Le raisonnement bayésien : modélisation et inférence*, Springer Science & Business Media.
- Parent, E. et E. Rivot (2012), *Introduction to hierarchical Bayesian modeling for ecological data*, Chapman and Hall/CRC.
- Parmigiani, G. et L. Inoue (2009), *Decision theory: Principles and approaches*, vol. 812, John Wiley & Sons.
- Pezier, J. (1972), *Décisions Rationnelles dans l'incertain*, Masson, traduction de "Rational Descriptions, Decisions and Designs" par M. Tribus.
- Raiffa, H. et R. O. Schlaifer (1961), *Applied statistical decision theory*, Graduate School of business administration-Harvard.
- Robert, C. P. et G. Casella (1998), *Monte Carlo Statistical Methods*, Springer-Verlag.
- Robert, C. P., N. Chopin, et J. Rousseau (2004), «Harold Jeffreys's theory of probability revisited», *Statistical Science*, vol. 24, pp. 141–172.
- Rouanet, H., M.-P. Lecoutre, M.-C. Bert, B. Lecoutre, et J.-M. Bernard (1991), «L'inférence statistique dans la démarche du chercheur», *Publications universitaires européennes. Série VI : Psychologie*.
- Savage, L. J. (1954), *The Foundations of Statistics*, Dover Publications, New York.
- Seriès, P. L. (2016), «Le cerveau est-il une machine bayésienne ?», in «Le Bayésianisme aujourd'hui», Éditions Matériologiques, Paris.
- Spiegelhalter, D. J., A. Thomas, N. G. Best, W. Gilks, et D. Lunn (1996), «Bugs: Bayesian inference using gibbs sampling», *Version 0.5,(version ii)* <http://www.mrc-bsu.cam.ac.uk/bugs>, vol. 19.
- Todeschini, A., F. Caron, M. Fuentes, P. Legrand, et P. Del Moral (2014), «Biips: software for Bayesian inference with interacting particle systems», *arXiv preprint arXiv :1412.3779*.
- Tribus, M. (1969), *Rational Descriptions, Decisions and Designs: Pergamon Unified Engineering Series*, Pergamon.

Tufféry, S. (2012), *Data mining et statistique décisionnelle : l'intelligence des données*, Editions Technip.

Ulmo, J. et J. Bernier (1973), *Éléments de décision statistique*, P.U.F.

Vapnik, V. (2013), *The nature of statistical learning theory*, Springer science & business media.

Von Neumann, J., O. Morgenstern, et al. (1953), *Theory of games and economic behavior*, Princeton University Press.

Wald, A. (1950), *Statistical Decision Functions*, Wiley.

Gilbert Saporta : un parcours éclectique

Analyse, fouille, science des données



Gilbert SAPORTA¹

Conservatoire national des arts et métiers, Paris

Gilles STOLTZ²

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 Paris³. En première année, j'étais déçu du piètre niveau des cours de mathématiques. Pour y remédier, mon camarade André Bellaïche⁴ et moi avons 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 écoles⁵, 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 pont 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 CEA⁶ 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'imaginai pas d'ailleurs concrètement). Je décidai de suivre un DEA⁷, 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'ISUP⁸, 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 avons 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 Benzécri⁹ 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'Inria¹¹) 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 avons 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 (Apur¹²), où je travaillai sur un problème d'optimisation des flux de circulation dans le 5^e 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'ISUP¹³. Ses conseils furent déterminants : elle me parla des IUT¹⁴ 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 ATER¹⁵, 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 formateur¹⁶ du futur département informatique de IUT de Paris¹⁷, 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 1969¹⁸.

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 TD¹⁹. 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 avons tous, étudiants comme enseignants, le sentiment d'être des pionniers. Comme enseignants, nous avons 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 cycle²⁰.

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 VI²¹ 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'apparis 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 α et β 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 Mines²² 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 5^e 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 LAFMA²³, 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'ENSAE²⁴ à partir de 1981 (que j'arrêtai en 2006). J'envisageai un temps de partir à

23. LAFMA : Liste d'aptitude aux fonctions de maître-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'EIASM²⁵. Alain Bensoussan, qui y intervenait, me proposa le montage suivant : me faire détacher au CNRS²⁶, 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 CEA²⁷. 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 1^{er} 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 diplôme²⁸. 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'INSEE²⁹ ; nous pûmes ainsi créer un master de statistique par unités d'enseignement capitalisables en cours du soir, le seul du genre en France³⁰.

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 DESS³¹ 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 CIFRE³²). 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).

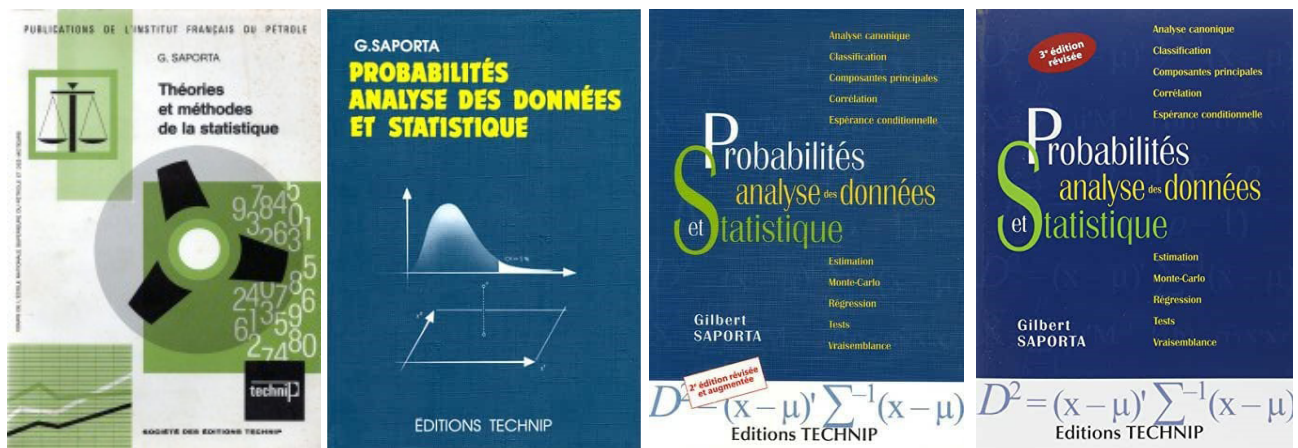


Figure 1 – Couvertures successives du manuel de Gilbert Saporta en 1978, 1980, 2006 et 2011 (de gauche à droite)

Les éditions Technip étaient alors une filiale de l'IFP³³ 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 calculatrice 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 corrélérer 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 XXI^e 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 aperçus 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 XXI^e 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 XX^e 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 bibliothèques 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'AF CET³⁷, 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*). Au près 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 SMAI³⁸, 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 Caussin. 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 Drosbeke et Tassi, 1997, ainsi que Didier, 2013). À travers ses sept associations³⁹, 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 indiscreète, 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 publier⁴⁰, 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.

2009-2019 : dix ans de publications sur l'enseignement de la statistique en France



Antoine ROLLAND¹

ERIC EA 3083, Université de Lyon, Université Lumière Lyon 2

TITLE

2009-2019: Ten years of publications on statistics education in France

RÉSUMÉ

Le journal *Statistique et Enseignement* a été publié par la Société Française de Statistique [SFdS] de 2010 à 2018. La fin du journal sous cette forme est l'occasion de faire le point de manière plus large sur les publications en didactique de la statistique, et plus largement sur les réflexions partagées concernant l'enseignement de la statistique en France. Nous avons dénombré près de 160 communications traitant de l'enseignement de la statistique dans des canaux de diffusion établis (thèses de doctorat et journaux publiés par la SFdS, les IREM et l'APMEP). La plupart de ces communications sont focalisées sur les explications mathématiques, le partage de ressources et des retours d'expériences liés aux évolutions des programmes scolaires. Il existe finalement peu d'apports théoriques conséquents dans le domaine de la didactique de la statistique. L'analyse quantitative et qualitative du corpus retenu sur cette période montre qu'une réflexion reste à mener sur les points de difficulté de l'enseignement des méthodes statistiques, en particulier dans les formations supérieures dédiées à la statistique ou à la science des données.

Mots-clés : didactique, état de l'art.

ABSTRACT

The *Statistique et Enseignement* journal was published by the Société Française de Statistique [SFdS] between 2010 and 2018. Its publication was ceased, which is an opportunity to conduct a state of the art of the publications in the field of statistics education in France during the same period. We approximately found 160 communications in this field, in various forms (PhD dissertations or articles in journals run by societies and institutions like SFdS, IREM, APMEP). Most of these communications are dedicated to theoretical explanations of concepts, share educative materials, or provide feedbacks on the evolutions of statistics curricula at the high-school level. But just a few are dedicated to theoretical considerations on statistics education itself. Both quantitative and qualitative analyses are conducted; they show that much work is needed in statistics education to identify the difficulties faced by students, especially in statistics and data science curricula at the university level.

Keywords: statistics education, state of the art.

1. Dix ans de didactique de la statistique

La Société Française de Statistique [SFdS] a pour mission de promouvoir l'utilisation de la statistique et sa compréhension, et de favoriser ses développements méthodologiques. Dans ce cadre, elle a publié de 2010 à 2018 une revue intitulée *Statistique et Enseignement*, dont la ligne éditoriale indiquait² qu'elle

visait à publier des contributions relatives à l'enseignement, l'apprentissage et la compréhension de la statistique et des probabilités, à tous les niveaux d'éducation, que ce soit dans un contexte formel (scolaire) ou informel (cadre extra-scolaire et popularisation « grand public »).

Cette revue, portée par le groupe Enseignement de la SFdS, a fusionné en 2019 avec la revue

1. antoine.rolland@univ-lyon2.fr

2. Source : éditorial du numéro 1 de la revue (2010).

Statistique et Société à la faveur d'une redéfinition de la politique éditoriale de la SFdS. L'auteur de cette communication a participé au comité de rédaction de *Statistique et Enseignement* et fait maintenant partie du comité de rédaction de *Statistique et Société*.

A l'occasion de cette fusion, un travail d'analyse a été effectué sur les articles parus dans *Statistique et Enseignement* pendant les neuf années de parution de la revue. Ce travail d'analyse a ensuite été élargi à l'ensemble des communications et articles traitant de l'enseignement de la statistique, en se limitant aux manuscrits de thèse de doctorat et aux articles publiés dans des revues françaises entre 2009 et le début de l'année 2019, soit une période d'un peu plus de dix ans. Cette communication vise à présenter les résultats de cette analyse, tant d'un point de vue quantitatif que d'un point de vue plus qualitatif, et ainsi contribuer à répondre en partie à la proposition effectuée par Régnier (2012) : « *Il y a encore nécessité de faire un état de l'art de la didactique en statistique* ». Hahn (2015) a proposé un état de l'art à partir des revues internationales mais conclut : « *En France il existe peu de travaux de recherche qui relèvent du domaine de l'éducation statistique* ». Nous espérons que cette présentation sera à même de proposer des pistes de réflexions, voire de travail pour les années à venir.

2. Méthodologie et analyse quantitative

Afin d'essayer de couvrir l'ensemble des communications effectuées en didactique de la statistique, nous avons lancé des requêtes dans les domaines Internet suivants : la revue *Statistique et Enseignement* (<http://statistique-et-enseignement.fr/>), le site des IREM³ (<http://publimath.irem.univ-mrs.fr/>), qui référence notamment les articles publiés dans le *Bulletin Vert* de l'APMEP⁴, le site de l'ARDM⁵ (<https://ardm.eu/>), ainsi que le site <http://theses.fr/>. Nous avons ensuite lu l'intégralité des articles remontés par ces requêtes, afin de déterminer si tous évoquaient effectivement le sujet de l'enseignement de la statistique. Nous avons inclus dans cette liste également les articles traitant de l'enseignement des probabilités, le lien entre les deux domaines étant très souvent évoqué dans les articles. Pour les revues comportant des rubriques, nous avons relevé celle correspondant à l'article afin de déterminer de quelle manière l'article était considéré par les éditeurs. Il est bien sûr tout à fait possible que certains articles aient échappé à notre crible. Certains choix effectués d'inclure ou non tel ou tel article dans notre liste peuvent également être discutables. Cependant, nous pensons que cela ne changerait qu'à la marge les résultats de notre étude.

2.1. Thèses

A notre connaissance, 8 thèses ont été soutenues depuis 2009 (dont 3 en 2009) sur l'enseignement de la statistique ; à titre de comparaison, plus de 100 thèses ont été soutenues en didactique des mathématiques sur la même période d'après le site theses.fr. Une thèse soutenue concerne l'école élémentaire (Coutanson, 2010), deux le lycée (Nhu, 2009; Derouet, 2016), et cinq l'enseignement supérieur (Carranza, 2009; Oriol, 2009; Kouame, 2013; Jutand, 2015; Diaz, 2016). Quatre thèses font une référence explicite dans leur titre au fait qu'elles ne concernent l'analyse didactique que d'une seule expérience (Carranza, 2009; Oriol, 2009; Kouame, 2013; Derouet, 2016). Par ailleurs, on trouve trace de deux thèses en préparation (Rodoumdje, en cours; Trunkenwald, en cours).

2.2. La revue *Statistique et Enseignement* de la SFdS

De 2010 à 2018, la revue *Statistique et Enseignement* a publié 15 numéros en 9 volumes, soit 135 articles (dont 15 éditoriaux) de 148 auteurs différents, se répartissant dans les différentes

3. IREM : Instituts de recherche sur l'enseignement des mathématiques

4. APMEP : Association des professeurs de mathématiques de l'enseignement public ; site du *Bulletin Vert* : <https://www.apmep.fr/Le-Bulletin-Vert>

5. ARDM : Association pour la recherche en didactique des mathématiques

rubriques selon le tableau 1. Les articles de la rubrique « Recherches et perspectives » proposent une réelle réflexion didactique sur l'enseignement de la statistique. Ce sont des articles de fond, qui s'attachent à proposer une approche théorisée de l'enseignement de la statistique, dépassant le simple retour d'expériences. Une analyse temporelle montre que ces articles ont été publiés très majoritairement de 2010 à 2015. Les « dossiers spéciaux », assimilables au premier abord à des « Expériences commentées » ont été publiés pendant les trois dernières années de la revue. Cela indique le passage d'une revue de soumissions à une revue de commandes et donc, peut-on supposer, une moins grande vivacité de la recherche spontanée en didactique de la statistique.

TABLE 1 – Répartition des articles de *Statistique et Enseignement* par rubrique

Rubrique	Nbr. articles
Recherches et perspectives	29
Expériences commentées	23
Notes de lecture	21
Dossier spécial	18
Editorial	15
Libres propos	14
Outils et documents	6
Chroniques	
Entretien avec un.e statisticien.ne	4
Cueillettes statistiques	2
Traverses statistiques	2
Ressources partagées	1

2.3. Les revues *Bulletin Vert* de l'APMEP et *Repères IREM*

Nous avons effectué une recherche sur le site <http://publimath.irem.univ-mrs.fr> à travers le mot-clé « statistique » et en nous limitant aux années 2009 à 2019. Cette recherche a remonté des ressources provenant essentiellement des revues *Bulletin Vert* de l'APMEP et *Repères IREM*. Nous omettons volontairement les articles des revues de vulgarisation telles que *Tangente*, également présents sur le site *Publimath* mais n'étant pas spécifiquement reliés à des questions d'enseignement.

Le *Bulletin Vert* de l'APMEP a publié 41 articles traitant de l'enseignement de la statistique ou des probabilités sur 750 articles publiés entre 2009 et 2018 (année ayant connu l'arrêt de la revue sous sa forme historique), ce qui représente 5% des articles publiés. La revue *Au fil des Maths* de l'APMEP, née en 2018 et remplaçant le *Bulletin Vert*, a publié 2 articles portant sur la statistique sur la centaine d'articles proposés jusqu'alors. Entre 2009 et 2018, la revue *Repères* publiée par l'inter-IREM national a publié 13 articles abordant la statistique ou les probabilités, sur un total de plus de 200 articles publiés, soit 6%. La plupart des articles ont été publiés en lien avec la mise en oeuvre de l'introduction des probabilités en classe de troisième (2008) et de la statistique inférentielle au lycée (2011). Par exemple, le *Bulletin Vert* a proposé trois dossiers thématiques pour accompagner cette évolution des programmes. La composition de ces revues montre un équilibre entre trois types d'articles : les expériences (plus ou moins) commentées, les articles de fond sur la didactique de la statistique, et des apports théoriques sous forme de mini-cours pour la formation continue des enseignants ; on pourra se référer aux tableaux 2 et 3.

TABLE 2 – Répartition des articles du *Bulletin Vert de l'APMEP* par rubrique

Rubrique	Nbr. articles
Dans nos classes	11
Pour chercher et approfondir	9
Dossier « Les probabilités »	6
Dossiers « Statistique » (deux dossiers)	8
Témoignage	2
Autres dossiers thématiques	5

TABLE 3 – Répartition des articles des *Repères IREM* par rubrique

Rubrique	Nbr. articles
Expériences commentées	5
Recherches et perspectives	4
Apport théorique	3
Point de vue	1

2.4. Autres revues

D'autres revues publient de temps en temps des articles en lien avec l'enseignement de la statistique :

- les séminaires de l'ADRM ont proposé deux exposés sur le thème de la statistique (Lahanier-Reuter, 2011; Dutarte, 2019) ;
- on trouve 2 articles relatifs à la statistique sur le site MathemaTICE (<http://revue.sesamath.net/>) et 9 dans la revue en ligne *Plot* de l'APMEP (qui a arrêté de paraître fin 2017 et qui est également indexée par *Publimath*) ; cette production est marginale et concerne essentiellement du partage d'expérience.

2.5. Colloques, conférences, etc.

Les différentes éditions du CFIES⁶ (2008, 2010, 2012, 2015, 2017, 2019) sont par définition des lieux où sont partagées des réflexions sur l'enseignement de la statistique. Malheureusement, le manque de maintenance et d'archivage des sites des différentes éditions fait que les archives des soumissions aux CFIES ne sont pas exploitables pour des recherches.

3. Analyse qualitative

Nous abordons ici une approche plus qualitative du contenu des publications listées précédemment.

3.1. Au collège : probabilités

Les articles de niveau collège sont très en lien avec l'apprentissage des probabilités en classe de troisième. Les publications de l'APMEP et des IREM se sont particulièrement attachées à proposer des articles théoriques sur l'importance d'un enseignement des probabilités et les points-clés d'un tel enseignement (Dacunha-Castelle, 2009; Ducl et Sausserau, 2009; Roser et Schwartz, 2009; Hennequin, 2010), et en particulier, sur la tension entre probabilités et statistique (approche fréquentiste des probabilités : Henry, 2009a; Verdier, 2009; Bonneval, 2009; Parzysz,

6. CFIES : Colloque Francophone International sur l'Enseignement de la Statistique

2009; Piednoir, 2009). On trouve aussi plusieurs articles d'expériences commentées permettant aux enseignants de mettre en oeuvre des activités au sein de leurs classes (François, 2009; Costes, 2009; Duperret, 2010; Chevalarias, 2010; Jaquet et Henry, 2014; Costes, 2014). La revue *Statistique et Enseignement* a pour sa part abordé la question de l'enseignement de la statistique au collège via plusieurs articles sur l'évolution de l'enseignement de la statistique (en 2011) ou la comparaison internationale des curriculum de statistique dans l'enseignement secondaire (en 2013); voir le paragraphe 3.4 pour davantage de détails. On trouve également une proposition de curriculum par Roser et Schwartz (2009).

On peut rajouter à ce tableau quelques articles ressources utilisables au niveau collège : sur la récolte des données par les étudiants (Girod, 2013), ou les pourcentages (Ancel, 2015), ainsi que le seul article relatant une expérience menée avec des élèves d'école élémentaire (à partir d'une planche de Galton : Drouin, 2014).

3.2. Au lycée : intervalles de fluctuations et de confiance

Les articles traitant de l'enseignement de la statistique au lycée peuvent se diviser en plusieurs groupes, inégaux en taille.

Il y a tout d'abord quelques articles qui décryptent le contenu du programme de statistique au lycée (Gasquet, 2011; Schwartz, 2012). Certains portent un regard critique, voire « polémique », qui interroge l'intérêt même d'enseigner la statistique au lycée (Perrin, 2015). Ecrits sous forme de tribune, ou « libres propos », ces points de vue sont intéressants pour comprendre la vision de l'enseignement des mathématiques au niveau lycée telle qu'elle est défendue par les uns ou les autres, ainsi que le regard que portent des mathématiciens sur la statistique, discipline en lien avec les mathématiques mais devant toujours prouver sa légitimité à rentrer dans le champ mathématique (Kahane, 2010; Maul et Vagost, 2013).

La plupart des articles proposés au niveau lycée abordent la question de la statistique inférentielle, nouveauté des programmes de 2011. Intervalles de fluctuations, intervalles de confiance, prise de décision et logique des tests d'hypothèse sont abordés à travers des articles d'apports théoriques pour la formation des enseignants (Ducel et Saussereau, 2011; Grihon, 2012b; Bonneval, 2012; Riandey, 2012; Cerclé, 2013b; Dutarte, 2013; Ducel et al., 2014; Pascal, 2014; Boulais et al., 2015). Ces articles sont généralement assez pointus et critiques, mettant en exergue, par exemple, la fausseté des énoncés proposés au programme (approximation d'une loi binomiale par une loi normale : Raymondaud, 2014) ou en exercice (l'affaire Woburn : Couallier et al., 2015; Fine, 2015). Plus rarement, certains articles proposent des mises en oeuvre pratiques avec retours d'expériences plus ou moins détaillés (Corpart et Lassalle, 2012; Grihon, 2012a; Sotura, 2013; Parnaudeau, 2013; Bülher, 2016) ou des exemples d'applications (Piednoir, 2013; Faisant et al., 2016; Zayana, 2018). Le domaine de la statistique inférentielle est reconnu comme peu maîtrisé par les enseignants en poste, qui pour la plupart n'ont jamais vu ces notions lors de leur formation initiale, et nécessite donc une formation continue.

L'autre domaine relativement présent est celui de la simulation des données, soit comme terrain d'application des enseignements d'informatique et algorithmique (Henry, 2009b; Vagost, 2010; Costes, 2011), soit comme illustration de l'approche fréquentiste des probabilités (Parnaudeau, 2009; Grihon, 2009; Henry, 2011; Kentzel, 2012). On peut noter également un article incitant à l'utilisation du logiciel **R** pour le cours de statistique au lycée (Raffinat, 2017).

Enfin, quelques articles abordent ponctuellement certaines notions ou exemples d'applications des programmes. L'étude et la construction d'indicateurs statistiques (moyenne, médiane, écart-type ou écart moyen) a fait l'objet de plusieurs articles (Dejean et Lavaux, 2009; Segouat, 2010; Cerclé, 2014; Chaput et al., 2016; Vermette, 2018), de même que la loi normale et des

lois à densité (Parzys, 2013; Cerclé, 2013a; Derouet et Alory, 2018; Derouet, 2018). D'autres notions sont plus sporadiquement représentées, telles que la notion de hasard ou d'incertitude (Kentzel, 2009; Grihon, 2017), la démographie (espérance de vie : Dittgen, 2011; Hennequin, 2012; Laurent, 2015; fécondité : Laurent, 2013), les chaînes de Markov (Bonneval, 2013), les arbres de probabilités (Larose, 2014) ou le recueil de données (Taillet, 2018).

3.3. Dans l'enseignement supérieur : un grand absent

Les articles traitant de l'enseignement de la statistique à un niveau post-bac se trouvent tous dans la revue *Statistique et Enseignement*, à l'exception de celui de Letué (2014), qui présente un panorama des formations supérieures en statistique à destination des enseignants du secondaire. Ces articles sont principalement :

- des présentations de formations en statistique (Besse et Laurent, 2016; Amini et al., 2016; Cléménçon, 2016; Niang et al., 2016; Dupuy-Chessa et al., 2016; Poggi et al., 2016; Rolland, 2017; Bazzoli et al., 2017; Delsart et Vanecloo, 2017; Besse et al., 2018; Rolland et Spennato, 2018) ou de MOOC⁷ dans le domaine (Cohen, 2012; Villa-Vialaneix, 2013; Bar-Hen, 2014; Fine et Villa-Vialaneix, 2014; Piednoir, 2014) ; ces présentations font généralement l'objet de dossiers spéciaux ;
- des présentations d'actions pédagogiques comme les challenges (Goga et Ardilly, 2017; Jollois et al., 2017; Jiao et al., 2017; Chandesris et al., 2017) ou d'utilisation de logiciels spécialisés (Grégoire et al., 2012; Jutand, 2012; Jutand et Ruiz-Gazen, 2012; Dufour, 2012; Sabourin et al., 2012) ;
- des retours d'expériences concernant l'enseignement des fondements de la statistique pour des non-spécialistes, en filière STAPS⁸ (Genolini et Driss, 2010), sciences de la vie ou épidémiologie (Palm et Allagbe, 2011; Calmant et al., 2011; Senterre et al., 2011; Dufour et al., 2017; Jutand et al., 2017), sciences humaines (Cañadas et al., 2012; Cohen et al., 2017; Zendera et al., 2017; Gélinas et al., 2018), sciences de l'éducation (Bihan-Poudec, 2012b), ou en école de management (Hahn et Stoltz, 2013; Carillo et al., 2016) ;
- quelques ressources ou situations exemples pouvant être utilisées en cours (Riandey et Widmer, 2010; Riandey, 2010; Duprez et Cros, 2010; Raymondaud, 2012).

Parmi les thèses soutenues, cinq font référence à un niveau post-bac : trois ont pour terrain des filières statistiques (Oriol, 2009; Kouame, 2013; Diaz, 2016), et deux portent sur des enseignements de statistique pour non-spécialistes (Carranza, 2009; Jutand, 2015).

Il n'y a pour ainsi dire aucune production qui s'intéresse aux difficultés ou points d'attention existants dans l'enseignement de la statistique pour spécialistes (les étudiants suivant des cours de statistique avancée et se destinant à pratiquer ou enseigner la statistique). Certains thèmes sont esquissés, par exemple la simulation de tests d'hypothèses (Dagnelie, 2010), mais de manière très partielle.

3.4. Recherches et perspectives

Les 29 articles de la rubrique « Recherches et perspectives » de la revue *Statistique et Enseignement* (plus quelques autres. . .), certains déjà cités dans cet état de l'art, permettent de dresser un panorama des questions de recherche présentes ces dix dernières années en didactique de la statistique de manière transversale. Un tiers des articles parus dans cette rubrique sont

7. MOOC : *massive open online course*, que l'on peut traduire par « cours en ligne ouvert et massif »

8. STAPS : Sciences et techniques des activités physiques et sportives

consacrés à l'analyse des curriculum de statistique d'un point de vue général (Fine, 2013; Burrill et Biehler, 2013), historique (Henry, 2010; Dutarte, 2011) ou centré sur la comparaison entre pays (Raoult, 2013a; Vermandele, 2013; Batanero et al., 2013; Sanchez et Hoyos, 2013; Gattuso et Vermette, 2013; Rouan et El Idrissi, 2014). Plusieurs autres articles insistent sur la nécessaire formation des enseignants à la statistique (Batanero et Diaz, 2010; Fine, 2010; Gauvrit, 2013; Serrado Bayes et al., 2014; Gonzalez, 2014; Arteaga et al., 2015; Droesbeke, 2017).

Un deuxième tiers des articles présentent des réflexions plus larges sur la didactique de la statistique. Deux articles sont consacrés à des états de l'art de la didactique de la statistique, d'un point de vue français (Régnier, 2012) ou international (Hahn, 2015). Les autres proposent des réflexions sur la manière d'aborder la statistique comme discipline scolaire, seule (Gattuso, 2011; Bihan-Poudec, 2012a; Lahanier-Reuter, 2012), en interdisciplinarité (Raoult, 2011; Fine, 2012b; Raoult, 2013b) ou dans son historicité (Desrosières, 2010; Armatte, 2010).

Les derniers articles abordent des sujets précis de méthodes statistiques, sans être rattachés à un niveau scolaire ou universitaire particulier : promenades aléatoires (Soto-Andrade, 2015), notion de moyenne (Vandeschrick, 2017; Droesbeke et Vermandele, 2018), lois à densité (Derouet, 2017), « datavisualisation » (El Methni, 2018), concepts de proportions (Fine, 2012a).

4. Conclusion et pistes de travail

Ce bref panorama de la recherche sur l'enseignement de la statistique dans la période 2009–2019 en France montre qu'il reste encore du travail à faire dans le domaine de la didactique de la statistique, au-delà de quelques cas particuliers. La statistique ne correspond qu'à 5% de la production d'articles ou thèses du domaine de la didactique des mathématiques. Certes, les enseignants non formés à la statistique (lycée) ou faisant face à un public d'étudiants non-spécialistes (enseignement supérieur) se posent des questions pratiques sur la façon d'enseigner les fondements de la statistique (récolte des données, statistique descriptive, statistique inférentielle élémentaire). Les articles publiés ces dix dernières années répondent à ces besoins. Mais encore trop rare est la recherche didactique en France sur l'approche de l'enseignement de la statistique à des futurs statisticiens : pré-supposés, méthodes, difficultés des apprenants, etc. Quelques articles fondamentaux ont préparé le terrain (Gattuso, 2011; Régnier, 2012; Fine, 2013; Burrill et Biehler, 2013; Hahn et Stoltz, 2013; Hahn, 2015; Perrin, 2015; Besse et Laurent, 2016). Mais au moment où la science des données se développe et où les formations supérieures dans le domaine statistique se multiplient, il y a là une piste prometteuse de recherche.

Références

Amini, M.-R., J.-B. Durand, O. Gaudoin, E. Gaussier, et A. Iouditski (2016), «Data science : une formation internationale de niveau master en science des données», *Statistique et enseignement*, vol. 7, no 1, pp. 95–102.

Ancel, L. (2015), «Autour de la pomme», *Bulletin Vert APMEP*, vol. 512, pp. 46–53.

Armatte, M. (2010), «Le rôle de l'histoire dans l'enseignement de la statistique», *Statistique et enseignement*, vol. 1, no 2, pp. 23–47.

Arteaga, P., C. Batanero, J. M. Contreras, et G. Canadas (2015), «Statistical graphs complexity and reading levels : a study with prospective teachers», *Statistique et enseignement*, vol. 6, no 1, pp. 3–23.

- Bar-Hen, A. (2014), «MOOC qui peut (MOOC qui veut)», *Statistique et enseignement*, vol. 5, no 1, pp. 91–99.
- Batanero, C. et C. Diaz (2010), «Training teachers to teach statistics: what can we learn from research?», *Statistique et enseignement*, vol. 1, no 1, pp. 5–20.
- Batanero, C., J.J. Ortiz, R. Roa, et L. Serrano (2013), «la statistique dans le curriculum en Espagne», *Statistique et enseignement*, vol. 4, no 1, pp. 89–106.
- Bazzoli, C., M. Villanova, et F. Letué (2017), «L’alternance et les licences professionnelles du département STID de Grenoble», *Statistique et enseignement*, vol. 8, no 1, pp. 73–87.
- Besse, P., B. Guillouet, et B. Laurent (2018), «Wikistat 2.0 : ressources pédagogiques pour l’intelligence artificielle», *Statistique et enseignement*, vol. 9, no 1, pp. 43–61.
- Besse, P. et B. Laurent (2016), «De statisticien à data scientist, développements pédagogiques à l’INSA de toulouse», *Statistique et enseignement*, vol. 7, no 1, pp. 75–93.
- Bihan-Poudec, A. (2012a), «Statistiques ou statistique ? Que peut apporter la théorie des représentations sociales à la compréhension de son enseignement?», *Statistique et enseignement*, vol. 3, no 1, pp. 37–62.
- Bihan-Poudec, A. (2012b), «Un contrôle surprise pour l’enseignant ! L’évaluation comme révélateur des préconceptions de la statistique chez les étudiants», *Statistique et enseignement*, vol. 3, no 1, pp. 63–72.
- Bonneval, L.-M. (2009), «Aux urnes, collégiens», *Bulletin Vert APMEP*, vol. 484, pp. 627–632.
- Bonneval, L.-M. (2012), «Fluctuation et confiance au lycée», *Bulletin Vert APMEP*, vol. 500, pp. 494–504.
- Bonneval, L.-M. (2013), «Chaînes de Markov au lycée», *Bulletin Vert APMEP*, vol. 503, pp. 155–162.
- Boulais, P., M. Diumenge, M. Vergnac, et C. Vergne (2015), «Une initiation à la statistique en classe de seconde», *Repères IREM*, vol. 98, pp. 21–51.
- Burrill, G. et R. Biehler (2013), «Les idées statistiques fondamentales dans le curriculum scolaire», *Statistique et enseignement*, vol. 4, no 1, pp. 5–24.
- Bülher, M. (2016), «Probabilités conditionnelles et tests de dépistage en TS», *Plot, APMEP*, vol. 56, pp. 20–25.
- Calmant, P., M. Ducarme, et M. Schneider (2011), «Obstacles a priori à l’apprentissage de l’analyse statistique inférentielle», *Statistique et enseignement*, vol. 2, no 1, pp. 43–59.
- Carillo, K., N. Galy, C. Guthrie, et A. Vanhems (2016), «“J’aime pas les stats !” : mesure et analyse de l’attitude à l’égard du cours de statistique dans une école de management», *Statistique et enseignement*, vol. 7, no 1, pp. 3–31.
- Carranza, P. F. (2009), *La dualité de la probabilité dans l’enseignement de la statistique : une expérience en classe de BTS*, Thèse de doctorat, Université Paris 7, sous la direction de Alain Kuzniak.

- Cañadas, G., C. Batanero, C. Diaz, et R. Roa (2012), «Psychology students' understanding of the chi-squared tests», *Statistique et enseignement*, vol. 3, no 1, pp. 3–18.
- Cerclé, V. (2013a), «Bon sang mais c'est bien sûr ! Une approximation des coefficients binomiaux à partir de la fonction de densité de la loi normale», *Bulletin Vert APMEP*, vol. 502, pp. 41–47.
- Cerclé, V. (2013b), «Quelques interrogations du professeur de lycée autour des intervalles de fluctuation», *Repères IREM*, vol. 91, pp. 51–69.
- Cerclé, V. (2014), «Et si on s'intéressait à la moyenne des écarts ?», *Bulletin Vert APMEP*, vol. 510, pp. 398–404.
- Chandesris, M., P. Gandubert, S. Pelloux, et P. Bouche (2017), «SNCF Data Challenge : une première à l'ENSAI», *Statistique et enseignement*, vol. 8, no 2, pp. 137–147.
- Chaput, B., C. Ducamp, et H. Hadidou (2016), «Compte rendu d'un travail interdisciplinaire mathématiques – sciences physiques», *Repères IREM*, vol. 102, pp. 615–626.
- Chevalarias, T. (2010), «Le chapitre probabilités en troisième», *Repères IREM*, vol. 78, pp. 59–69.
- Cléménçon, S. (2016), «Le mastere spécialisé Big Data de Télécom Paris Tech», *Statistique et enseignement*, vol. 7, no 1, pp. 103–110.
- Cohen, A. (2012), «Utilisation de la chaîne éditoriale SCENARI pour un cours de statistique», *Statistique et enseignement*, vol. 3, no 1, pp. 73–101.
- Cohen, A., A. Lamme, C. Vermandele, et M. Sraieb (2017), «Les enjeux des supports de cours : cas d'un enseignement de statistique en sciences humaines», *Statistique et enseignement*, vol. 8, no 1, pp. 21–44.
- Corpart, A. et N. Lassalle (2012), «Un biberon comme outil de simulation au lycée», *Bulletin Vert APMEP*, vol. 500, pp. 477–493.
- Costes, R. (2009), «Les probabilités entrent au collège : quels enjeux, quelles activités», *Plot, APMEP*, vol. 26, pp. 2–11.
- Costes, R. (2011), «Méthodes et pratiques scientifiques : un espace pour un enseignement scientifique interdisciplinaire en seconde», *Bulletin Vert APMEP*, vol. 496, pp. 517–528.
- Costes, R. (2014), «Vive les arbres ! Ou l'art de représenter graphiquement», *Plot, APMEP*, vol. 45, pp. 18–21.
- Couallier, V., L. Gerville-Réache, et G. Stoltz (2015), «L'enseignement de l'affaire Woburn, suite», *Statistique et enseignement*, vol. 6, no 2, pp. 41–55.
- Coutanson, B. (2010), *La question de l'éducation statistique et de la formation de l'esprit statistique à l'école primaire en France : étude exploratoire de quelques caractéristiques de situations inductrices d'un enseignement de la statistique au cycle III*, Thèse de doctorat, Université Lumière Lyon 2, sous la direction de Jean-Claude Régnier.
- Dacunha-Castelle, D. (2009), «Le difficile chemin de l'aléatoire dans les programmes», *Bulletin Vert APMEP*, vol. 484, pp. 595–597.

Dagnelie, P. (2010), «Dans l'enseignement de la statistique, la simulation a des limites», *Statistique et enseignement*, vol. 1, no 2, pp. 67–68.

Dejean, T. et F. Lavaux (2009), «Les statistiques, un thème d'interdisciplinarité en seconde», *Plot, APMEP*, vol. 28, pp. 15–23.

Delsart, V. et N. Vaneecloo (2017), «L'alternance dans un master de statistique et d'informatique décisionnelle. Le cas du parcours data sciences du master SIAD de Lille 1», *Statistique et enseignement*, vol. 8, no 1, pp. 89–96.

Derouet, C. (2016), *La fonction de densité au carrefour entre probabilités et analyse en terminale S : Etude de la conception et de la mise en oeuvre de tâches d'introduction articulant lois à densité et calcul intégral*, Thèse de doctorat, Université Paris 7, sous la direction de Alain Kuzniak et Fabrice Vandebrouck.

Derouet, C. (2017), «Emergence historique des lois à densité. Des pistes pour l'enseignement», *Statistique et enseignement*, vol. 8, no 2, pp. 3–24.

Derouet, C. (2018), «L'histogramme sous une autre facette», *Au Fil des Maths, APMEP*, vol. 528, pp. 1–5.

Derouet, C. et S. Alory (2018), «Une séquence d'enseignement articulant les lois de probabilité à densité et le calcul intégral en terminale S», *Repères IREM*, vol. 113, pp. 45–80.

Desrosières, A. (2010), «Un enseignement sur l'histoire de la statistique et de l'économétrie pour les élèves de l'ENSAE», *Statistique et enseignement*, vol. 1, no 1, pp. 21–33.

Diaz, D. (2016), *Les facteurs influençant la réussite des activités collaboratives médiées par les TICE dans une situation de formation universitaire à la statistique*, Thèse de doctorat, Université Lumière Lyon 2, sous la direction de Jean-Claude Régnier.

Dittgen, A. (2011), «La pierre philosophale ?», *Bulletin Vert APMEP*, vol. 494, pp. 334–336.

Droesbeke, J.-J. (2017), «La formation à la statistique des enseignants-chercheurs. L'expérience des journées d'étude en statistique», *Statistique et enseignement*, vol. 8, no 2, pp. 25–41.

Droesbeke, J.-J. et C. Vermandele (2018), «Faciliter l'enseignement de la moyenne arithmétique en se servant de son histoire», *Statistique et enseignement*, vol. 9, no 1, pp. 99–120.

Drouin, F. (2014), «Des p'tits clous, des p'tits clous, toujours des p'tits clous», *Plot, APMEP*, vol. 45, pp. 28–31.

Ducel, Y., D. Fourny, M. Fourny, et B. Saussereau (2014), «Calcul de risques de première et de seconde espèces à travers un exemple», *Repères IREM*, vol. 94, pp. 46–70.

Ducel, Y. et B. Saussereau (2009), «Quelle problématique pour un enseignement des probabilités en troisième ?», *Repères IREM*, vol. 77, pp. 53–65.

Ducel, Y. et B. Saussereau (2011), «La prise de décision de la seconde à la première», *Repères IREM*, vol. 85, pp. 31–49.

Dufour, A.-B., J. Lobry, et I. Amat (2017), «Enseigner le recueil des données : explorer la variabilité biologique, au chaud, dans une salle de cours», *Statistique et enseignement*, vol. 8, no 2, pp. 79–85.

Dufour, B. (2012), «La part du logiciel R dans l'enseignement de la statistique en biologie : le site web de Lyon 1», *Statistique et enseignement*, vol. 2, no 2, pp. 41–47.

Duperret, J.-C. (2010), «De la modélisation du monde au monde des modèles (2) ; des statistiques aux probabilités», *Bulletin Vert APMEP*, vol. 486, pp. 67–89.

Duprez, J.-M. et M. Cros (2010), «Accompagner étudiants et chercheurs dans l'exploitation des sources statistiques : l'expérience de la plateforme universitaire de données de Lille (PUDL)», *Statistique et enseignement*, vol. 1, no 1, pp. 65–73.

Dupuy-Chessa, S., S. Lambert-Lacroix, et G. Blanco-Laine (2016), «un parcours big data en alternance dans une licence professionnelle», *Statistique et enseignement*, vol. 7, no 1, pp. 121–126.

Dutarte, P. (2011), «Evolutions de la pratique statistique dans l'enseignement du second degré en France», *Statistique et enseignement*, vol. 2, no 1, pp. 31–42.

Dutarte, P. (2013), «Du bon usage d'un intervalle de fluctuation», *Bulletin Vert APMEP*, vol. 505, pp. 454–466.

Dutarte, P. (2019), «Probabilités, statistique et citoyenneté : inscrire le développement du jugement critique du futur citoyen dans le cadre des programmes de mathématiques de l'enseignement secondaire», in Barrier, T. et C. Chambris, éditeurs, «Actes du séminaire national de didactique des mathématiques 2017», pp. 110–127, IREM de Paris – Université Paris Diderot, URL [hal-02001693](https://hal.archives-ouvertes.fr/hal-02001693).

El Methni, J. (2018), «Data visualisation et enseignement de la statistique au travers d'exemples historiques en R», *Statistique et enseignement*, vol. 9, no 1.

Faisant, J., D. Lanier, J. Lejeune, R. Morello, et D. Trotoux (2016), «La statistique du chi-deux : son usage dans l'étude britannique de 1950 concernant l'association entre cancer du poumon et tabac.», *Bulletin Vert APMEP*, vol. 520, pp. 467–479.

Fine, J. (2010), «Probabilités et statistique inférentielle, approche sondage versus approche modèle», *Statistique et enseignement*, vol. 1, no 2, pp. 5–21.

Fine, J. (2012a), «Quelle est votre définition de "pourcentage" ? Proposition pour l'enseignement», *Statistique et enseignement*, vol. 3, no 2, pp. 87–97.

Fine, J. (2012b), «Statistique, informatique, mathématiques et interdisciplinarité», *Statistique et enseignement*, vol. 3, no 2, pp. 33–59.

Fine, J. (2013), «Le rapport GAISE (US) cadre d'un curriculum statistique de la maternelle à la terminale», *Statistique et enseignement*, vol. 4, no 1, pp. 25–54.

Fine, J. (2015), «A propos de l'enseignement des affaires Woburn et Castaneda», *Statistique et enseignement*, vol. 6, no 1, pp. 45–49.

Fine, J. et N. Villa-Vialaneix (2014), «Interview de Sylvie Méléard», *Statistique et enseignement*, vol. 5, no 1, pp. 83–89.

François, G. (2009), «L'aléatoire pour introduire les fréquences en classe de cinquième», *Repères IREM*, vol. 77, pp. 83–101.

- Gasquet, S. (2011), «Programme scolaire : de la conception. . . à la pratique en classe», *Bulletin Vert APMEP*, vol. 490, pp. 573–574.
- Gattuso, L. (2011), «L'enseignement de la statistique : où, quand, comment, pourquoi pas ?», *Statistique et enseignement*, vol. 2, no 1, pp. 5–30.
- Gattuso, L. et S. Vermette (2013), «L'enseignement de statistique et probabilités au Canada et en Italie», *Statistique et enseignement*, vol. 4, no 1, pp. 107–129.
- Gauvrit, N. (2013), «Conception du hasard et biais probabilistes chez des enseignants du second degré : effet d'une formation courte», *Statistique et enseignement*, vol. 4, no 2, pp. 53–66.
- Genolini, C. et T. Driss (2010), «Eveiller l'intérêt pour la statistique par l'exemple», *Statistique et enseignement*, vol. 1, no 2, pp. 49–57.
- Girod, F. (2013), «Une activité sur le thème de la météorologie en troisième et première ES : un problème "ouvert" sur le monde», *Bulletin Vert APMEP*, vol. 504, pp. 278–286.
- Goga, C. et P. Ardilly (2017), «Présentation du challenge "graines de sondeur"», *Statistique et enseignement*, vol. 8, no 2, pp. 103–111.
- Gonzalez, O. (2014), «Secondary mathematics teacher's professional competencies for effective teaching of variability-related ideas: a Japanese case study», *Statistique et enseignement*, vol. 5, no 1, pp. 31–51.
- Grihon, D. (2012a), «Fluctuation d'échantillonnage», *Plot, APMEP*, vol. 38, pp. 21–23.
- Grihon, P. (2009), «Les urnes de Polya», *Bulletin Vert APMEP*, vol. 485, pp. 737–746.
- Grihon, P. (2012b), «Introduction à la statistique inférentielle», *Bulletin Vert APMEP*, vol. 500, pp. 468–476.
- Grihon, P. (2017), «De l'aléatoire dans les mesures», *Bulletin Vert APMEP*, vol. 522, pp. 80–91.
- Grégoire, G., F.-X. Jollois, J.-F. Petiot, A. Qannari, S. Sabourin, P. Swertwaegher, J.-C. Turlot, V. Vandewalle, et S. Viguier-Pla (2012), «Les logiciels et l'enseignement de la statistique dans les départements STID des IUT», *Statistique et enseignement*, vol. 2, no 2, pp. 5–24.
- Gélinas, S., E. R. Berger, M. Balbinotti, D. Lalande, et M. Cantinotti (2018), «L'apprentissage par la pratique : vécus d'étudiants en psychologie impliqués dans la création de tutoriels informatisés en méthodes quantitatives», *Statistique et enseignement*, vol. 9, no 1, pp. 23–41.
- Hahn, C. (2015), «La recherche internationale en éducation statistique : état des lieux et questions vives», *Statistique et enseignement*, vol. 6, no 2, pp. 25–39.
- Hahn, C. et G. Stoltz (2013), «Savoir académique, savoirs pratiques : tensions et recherche d'équilibre», *Statistique et enseignement*, vol. 4, no 2, pp. 19–52.
- Hennequin, P.-L. (2010), «Le rôle de l'APMEP pour l'introduction des probabilités et de la statistique dans le secondaire», *Bulletin Vert APMEP*, vol. 490, pp. 541–543.
- Hennequin, P.-L. (2012), «Quelques remarques sur l'espérance de la vie humaine», *Bulletin Vert APMEP*, vol. 497, pp. 70–80.

Henry, M. (2009a), «Emergence de la probabilité et enseignement : définition classique, approche fréquentiste et modélisation», *Repères IREM*, vol. 74, pp. 67–89.

Henry, M. (2009b), «À propos du programme de statistique en seconde : remarques sur la simulation informatique», *Bulletin Vert APMEP*, vol. 484, pp. 598–602.

Henry, M. (2010), «Evolution de l'enseignement secondaire français en statistique et probabilité», *Statistique et enseignement*, vol. 1, no 1, pp. 35–45.

Henry, M. (2011), «Simulations d'expériences aléatoires en classe : un enjeu didactique pour comprendre la notion de modèle probabiliste, un outil de résolution de problèmes», *Bulletin Vert APMEP*, vol. 496, pp. 536–550.

Jaquet, F. et M. Henry (2014), «Approche de la notion de probabilité chez des enfants de 10-15 ans», *Repères IREM*, vol. 94, pp. 5–20.

Jiao, Y., B. Goutorbe, C. Grauer, M. Cornec, et J. Jakubowicz (2017), «The categorization challenge organized by cdiscount on datascience.net in 2015 : analysis of the released dataset and winning contributions», *Statistique et enseignement*, vol. 8, no 2, pp. 125–135.

Jollois, F.-X., F. Muri, E. Ottenwaelter, A. Rolland, et S. Viguier (2017), «Concours DATAVIZ : retour d'expérience en première année de DUT STID», *Statistique et enseignement*, vol. 8, no 2, pp. 113–124.

Jutand, M.-A. (2012), «Tutoriels pour logiciels et cours interactifs de statistique», *Statistique et enseignement*, vol. 2, no 2, pp. 75–79.

Jutand, M.-A. (2015), *Études des phénomènes de transposition didactique de la statistique dans le champ universitaire et ses environnements : une contribution à la pédagogie universitaire*, Thèse de doctorat, Université Bordeaux, sous la direction de Bernard Sarrazy.

Jutand, M.-A., K. Leffondré, M. Savès, et V. Kiewsky (2017), «Enseigner le recueil des données : étude de cas en épidémiologie», *Statistique et enseignement*, vol. 8, no 2, pp. 87–101.

Jutand, M.-A. et A. Ruiz-Gazen (2012), «l'enseignement à distance en statistique et l'usage de logiciels au travers de deux exemples», *Statistique et enseignement*, vol. 2, no 2, pp. 25–39.

Kahane, J.-P. (2010), «La statistique vue par un mathématicien», *Statistique et enseignement*, vol. 1, no 2, pp. 59–65.

Kentzel, J.-F. (2009), «Différencier une suite aléatoire d'une autre qui ne l'est pas», *Bulletin Vert APMEP*, vol. 481, pp. 181–187.

Kentzel, J.-F. (2012), «Série simulée ou inventée : une activité instructive et ludique au lycée», *Statistique et enseignement*, vol. 2, no 2, pp. 63–73.

Kouame, K. (2013), *Conception et expérimentation d'un outil d'aide au transfert de connaissances enseignées : cas d'une formation en alternance de Techniciens supérieurs en statistique et traitement informatiques de données*, Thèse de doctorat, Université Lumière Lyon 2, sous la direction de Laurent Veillard et Andrée Tiberghien.

- Lahanier-Reuter, D. (2011), «Études des enseignements-apprentissages en statistique : questions de disciplines», in Abboud-Blanchard, M. et A. Flückiger, éditeurs, «Actes du séminaire national de didactique des mathématiques 2011», pp. 41–56, IREM de Paris – Université Paris Diderot, URL [hal-02001693](https://hal.archives-ouvertes.fr/hal-02001693).
- Lahanier-Reuter, D. (2012), «La statistique est-elle une discipline scolaire ?», *Statistique et enseignement*, vol. 3, no 2, pp. 23–32.
- Larose, V. (2014), «Dessine-moi un arbre, et plus encore», *Plot, APMEP*, vol. 45, pp. 22–25.
- Laurent, E. (2013), «Les mesures de la fécondité comme support pédagogique à l'enseignement de la statistique», *Statistique et enseignement*, vol. 4, no 2, pp. 67–85.
- Laurent, E. (2015), «L'espérance de vie : une ressource pour l'enseignement», *Statistique et enseignement*, vol. 6, no 1, pp. 25–43.
- Letué, F. (2014), «Allier statistique et informatique : quelles filières après le bac ?», *Bulletin Vert APMEP*, vol. 509, pp. 331–339.
- Maul, A. et D. Vagost (2013), «La statistique : un nom galvaudé pour une discipline universelle», *Plot, APMEP*, vol. 43, pp. 2–12.
- Nhu, T. H. V. (2009), *Une étude didactique sur l'introduction dans l'enseignement mathématique vietnamien de notions statistiques dans leurs liens avec les probabilités*, Thèse de doctorat, Université Joseph Fourier (Grenoble), sous la direction de Nicolas Balacheff.
- Niang, N., G. Saporta, M. Cruvianu, et P. Rigaux (2016), «Le certificat "analyste de données massives" du CNAM», *Statistique et enseignement*, vol. 7, no 1, pp. 111–119.
- Oriol, J.-C. (2009), *Formation à la statistique par la pratique d'enquêtes par questionnaires et la simulation : étude didactique d'une expérience d'enseignement dans un département d'IUT*, Thèse de doctorat, Université Lumière Lyon 2, sous la direction de Jean-Claude Régnier.
- Palm, R. et G. Allagbe (2011), «Simuler pour comprendre : un didacticiel pour l'apprentissage de notions de base en statistique inférentielle», *Statistique et enseignement*, vol. 2, no 1, pp. 77–84.
- Parnaudeau, J.-M. (2009), «Bandes dessinées et statistiques», *Plot, APMEP*, vol. 28, pp. 20–23.
- Parnaudeau, J.-M. (2013), «Statistiques inférentielles : un débat scientifique en classe de seconde», *Bulletin Vert APMEP*, vol. 505, pp. 397–404.
- Parzysz, B. (2009), «De l'expérience aléatoire au modèle, via la simulation», *Repères IREM*, vol. 74, pp. 91–103.
- Parzysz, B. (2013), «La longue genèse de la loi normale», *Bulletin Vert APMEP*, vol. 502, pp. 29–40.
- Pascal, G. (2014), «Le risque des statistiques», *Repères IREM*, vol. 95, pp. 56–62.
- Perrin, D. (2015), «Remarques sur l'enseignement des probabilités et de la statistique au lycée», *Statistique et enseignement*, vol. 6, no 2.

- Piednoir, J.-L. (2009), «Du hasard : des statistiques aux probabilités, des probabilités à la statistique», *Bulletin Vert APMEP*, vol. 484, pp. 603–614.
- Piednoir, J.-L. (2013), «Le statisticien et le radariste», *Bulletin Vert APMEP*, vol. 502, pp. 55–57.
- Piednoir, J.-L. (2014), «“Fondamentaux en statistique”, un MOOC de France Université Numérique (FUN)», *Statistique et enseignement*, vol. 5, no 1, pp. 101–104.
- Poggi, J.-M., C. Bouveyron, G. Hebrail, et F.-X. Jollois (2016), «Un DU d’analyste Big Data en formation courte, au niveau L3», *Statistique et enseignement*, vol. 7, no 1, pp. 127–134.
- Raffinat, P. (2017), «R au lycée : plus qu’un logiciel de statistiques !», *site Mathématique*, vol. 56.
- Raoult, J.-P. (2011), «De l’erreur et du rôle de la philosophie dans l’enseignement de la statistique», *Statistique et enseignement*, vol. 2, no 1, pp. 85–87.
- Raoult, J.-P. (2013a), «La statistique dans l’enseignement secondaire en France», *Statistique et enseignement*, vol. 4, no 1, pp. 55–69.
- Raoult, J.-P. (2013b), «L’enseignement des probabilités et de la statistique face à celui des sciences de l’ingénieur en France», *Statistique et enseignement*, vol. 3, no 2, pp. 61–75.
- Raymondaud, H. (2012), «Le logiciel R comme outil d’initiation à la statistique descriptive : enquêtes sur les dépenses des ménages», *Statistique et enseignement*, vol. 2, no 2, pp. 49–62.
- Raymondaud, H. (2014), «Quelques remarques et propositions sur les sujets de bac S 2013 probabilités, statistiques et algorithmique», *Bulletin Vert APMEP*, vol. 509, pp. 253–264.
- Riandey, B. (2010), «Déontologie, législation et secret statistique : un enseignement pour tous», *Statistique et enseignement*, vol. 1, no 1, pp. 75–82.
- Riandey, B. (2012), «Inférence statistique et sondages électoraux», *Bulletin Vert APMEP*, vol. 500, pp. 505–512.
- Riandey, B. et I. Widmer (2010), «L’enseignement des sondages à l’usage du plus grand nombre : quelques réflexions tirées de l’expérience», *Statistique et enseignement*, vol. 1, no 1, pp. 47–63.
- Rodoumdje, D. (en cours), *Les initiatives des élèves par rapport à l’apprentissage de la statistique : cas de la classe de terminale S en Centrafrique.*, Thèse de doctorat, Brest, sous la direction de Ghislaine Gueudet et Caroline Poisard.
- Rolland, A. (2017), «L’alternance en DUT STID : exemple de l’IUT Lumière Lyon II», *Statistique et enseignement*, vol. 8, no 1, pp. 63–71.
- Rolland, A. et M. Spennato (2018), «L’attractivité des départements STID. Étude à partir des données APB», *Statistique et enseignement*, vol. 9, no 1, pp. 63–76.
- Roser, E. et C. Schwartz (2009), «L’esprit des probabilités, de l’école au lycée», *Bulletin Vert APMEP*, vol. 484, pp. 633–644.
- Rouan, O. et A. El Idrissi (2014), «Enseignement de la statistique au secondaire marocain», *Statistique et enseignement*, vol. 5, no 1, pp. 53–82.

- Régnier, J.-C. (2012), «Enseignement et apprentissage de la statistique : entre un art pédagogique et une didactique scientifique», *Statistique et enseignement*, vol. 3, no 1, pp. 19–36.
- Sabourin, S., A. Qannari, et L. Reboul (2012), «Un outil pour l'enseignement de l'analyse statistique textuelle et les systèmes de gestion de bases de données», *Statistique et enseignement*, vol. 3, no 1, pp. 103–112.
- Sanchez, E. et V. Hoyos (2013), «Analyse du curriculum statistique du collège au Mexique d'un point de vue de la formation à la statistique», *Statistique et enseignement*, vol. 4, no 1, pp. 131–142.
- Schwartz, C. (2012), «À propos du document de ressources pour la classe de première "statistiques et probabilités"», *Bulletin Vert APMEP*, vol. 498, pp. 136–144.
- Segouat, G. (2010), «Impliquer les élèves dans les statistiques descriptives», *Plot, APMEP*, vol. 29, pp. 19–23.
- Senterre, C., Y. Coppieters, A. Levêque, et M. Dramaix (2011), «Présentation et analyse d'un dispositif d'apprentissage en analyse multivariable appliquée à l'épidémiologie», *Statistique et enseignement*, vol. 2, no 1, pp. 61–75.
- Serrado Bayes, A., M. Meletiou-Mavrotheris, et E. Papanastasiou (2014), «Early statistics : a course for developing teacher's statistics technological and pedagogical content», *Statistique et enseignement*, vol. 5, no 1, pp. 5–29.
- Soto-Andrade, J. (2015), «Une voie royale vers la pensée stochastique : les promenades aléatoires comme "pousse d'apprentissage"», *Statistique et enseignement*, vol. 6, no 2, pp. 3–24.
- Sotura, B. (2013), «Une activité pour initier à la statistique inférentielle en classe de seconde», *Bulletin Vert APMEP*, vol. 502, pp. 48–54.
- Taillet, J. (2018), «Faire des mathématiques au moyen d'une station météo», *site Mathématique*, vol. 61.
- Trunkenwald, J. (en cours), *Approche fréquentiste et approche combinatoire dans l'enseignement des probabilités. Nature du travail mathématique mis en jeu*, Thèse de doctorat, Université Paris 7, sous la direction de Alain Kuzniak.
- Vagost, D. (2010), «L'algorithmique en seconde : un exemple de mise en oeuvre dans la classe», *Bulletin Vert APMEP*, vol. 486, pp. 42–46.
- Vandeschrick, C. (2017), «La moyenne : l'approche de Chisini revisitée. exemples et enseignements», *Statistique et enseignement*, vol. 8, no 1, pp. 3–20.
- Verdier, J. (2009), «Le hasard au collège», *Bulletin Vert APMEP*, vol. 484, pp. 615–626.
- Vermandele, C. (2013), «Le curriculum statistique en Belgique francophone», *Statistique et enseignement*, vol. 4, no 1, pp. 71–88.
- Vermette, S. (2018), «Les connaissances d'enseignants du secondaire sur les concepts d'écart moyen et d'écart type», *Statistique et enseignement*, vol. 9, no 1, pp. 3–21.

Villa-Vialaneix, N. (2013), «J'ai testé pour vous un MOOC», *Statistique et enseignement*, vol. 4, no 2, pp. 3–17.

Zayana, K. (2018), «Pas de proba, pas de chocolat», *Au Fil des Maths, APMEP*, vol. 527, pp. 10–13.

Zendrera, N., V. Dubreil-Frémont, J.-M. Marion, et A. Bihan-Poudec (2017), «Les données et leur production : réflexion sur une lacune paradoxale en éducation statistique», *Statistique et enseignement*, vol. 8, no 2, pp. 59–78.

Appel à communication

Les chiffres au cœur de la crise du coronavirus

2 mai 2020

La revue Statistique et société <http://statistique-et-societe.fr/> prépare un numéro spécial sur « la crise du Coronavirus confirmée par les chiffres » et lance par la présente un appel à communication.

L'épidémie de Coronavirus que nous traversons est très éprouvante. Restrictions des libertés publiques, maladie, décès de proches, nous sommes tous affectés, plus ou moins violemment, par cette crise. Mais cette violence n'empêche pas, sans doute bien au contraire, que l'événement soit aussi largement confirmé par des chiffres. Les statistiques comme le nombre de morts, de personnes infectées, de personnes soignées, à l'hôpital, dans les établissements d'hébergement pour les personnes âgées, en population générale qu'on nous communique chaque soir ; les courbes issues des modèles statistiques qu'utilisent les épidémiologistes, les cartes chiffrées reflétant la diffusion de la maladie ; les projections économiques des conséquences de l'épidémie, les besoins en masques, tests et médicaments pour lutter contre elle ; toutes les mesures politiques, tous nos discours, toutes nos réflexions sont orientés et conformés, à un degré rarement atteint, par des outils quantitatifs. Les chiffres et les affects vont de pair, ils sont au cœur de la crise.

Lorsque l'urgence sera derrière nous, nous souhaiterions proposer une réflexion sur la construction, la circulation et le rôle des statistiques dans la série d'événements que nous aurons traversée. Nous appelons ceux qui le souhaitent à envisager une communication sur ce thème. Nous étudierons toutes les intentions et soumissions d'article reçues dans les mois qui viennent et préparerons alors un numéro spécial sur la crise du coronavirus confirmée par les chiffres. Vous pouvez dès que vous le souhaitez écrire à contribution-statsociete@framalistes.org.

D'ici là, prenez soin de vous.

Emmanuel Didier
Rédacteur en chef de *Statistique et société*.