ENTRETIEN AVEC UN-E STATISTICIEN-NE

Dans cette chronique « Entretien avec un-e statisticien-ne », nous partons à la rencontre de celles et ceux qui font la statistique, tant des praticiens qui l'utilisent comme un outil essentiel dans le cadre de leur activité professionnelle que des universitaires qui la développent dans le cadre de leurs travaux de recherche et d'enseignement. Une ou plusieurs série(s) de questions mettent en particulier en lumière le rapport à l'enseignement. Les praticiens sont appelés à exprimer les besoins de formation qu'ils jugent prioritaires, les compétences qu'ils apprécient chez ceux qu'ils recrutent. Les universitaires sont interrogés sur leur vision personnelle de la statistique et de son importance, sur leur art de la transmission du savoir statistique, entre théorie et applications, idéalisation professorale et réalités étudiantes.

NADINE MANDRAN : COMMENT ENSEIGNER LA PRATIQUE MÉTIER ?

(Il faut apprendre à donner du sens aux données.)

Nadine MANDRAN[1] et Gilles STOLTZ[2]

Nadine Mandran est actuellement ingénieure d'études (IE) au CNRS, en gestion et analyse de (bases de) données. Son laboratoire d'affectation est le laboratoire d'informatique de Grenoble (LIG). Le CNRS lui a décerné sa « médaille de Cristal » en 2016, pour ses contributions, « aux côtés des chercheurs, à l'avancée des savoirs et à l'excellence de la recherche française ». Elle finit actuellement de préparer son manuscrit de thèse de doctorat, qui porte sur la recherche en informatique centrée sur l'humain, les processus expérimentaux à suivre et les indicateurs de traçabilité à retenir à cet effet. Nadine Mandran a donné plusieurs cours d'analyse de données, à des niveaux variant entre le L3 et le M2.

© Marie-Hélène Le Ny / « Infinités pluri*elles* »

L'entretien a été mené par des échanges de courriels entre le 29 août et le 16 novembre 2016.

Parcours académique et professionnel

GS : *Nadine, merci beaucoup de te prêter à notre jeu des questions-réponses ! Nous avons l'honneur de dresser le portrait de la récipiendaire d'une belle distinction du CNRS, en l'occurrence, le Cristal 2016, qu'il décerne à ses ingénieurs et techniciens. Nous y reviendrons.*

[1]Laboratoire d'Informatique de Grenoble : CNRS / Grenoble INP / Inria / Université Grenoble Alpes, Saint-Martin-d'Hères et Grenoble, France, nadine.mandran@imag.fr

[2]GREGHEC : CNRS / HEC Paris, Jouy-en-Josas, France, stoltz@hec.fr

Pour l'heure, je voudrais commencer par le commencement : où en étais-tu vers 20 ans, comment se déroulaient tes premières années d'études supérieures, quels étaient tes projets d'avenir à l'époque, quelle profession et dans quel univers te voyais-tu exercer ?

NM : A 20 ans, ce sont pour moi la fin du DUT[3] « statistiques et techniques quantitatives de gestion[4] » et une seule envie : travailler pour être indépendante. Mais en 1983, la statistique, qui sait ce que c'est ? Quels emplois existe-t-il dans ce domaine ? Mes camarades de promotion ont décidé de partir du côté de l'informatique. A l'époque, c'était plus facile : la connaissance de l'algorithmique et d'un langage de programmation suffisait à se faire embaucher comme informaticien. L'informatique ne me tentait toutefois pas, c'est ainsi que j'ai postulé en sciences de gestion et en statistique...

J'ai rédigé et envoyé des dizaines de lettres de candidatures. Un jour, une annonce correspondant au profil que j'attendais a été publiée : c'était au bout de six mois de candidatures. Même en 1983 il était déjà difficile de décrocher un emploi ! Il s'agissait d'un poste de technicien à l'AgroParisTech[5], pour faire des analyses statistiques et de la gestion de bases de données dans un laboratoire de zootechnie.

C'est ainsi que j'ai commencé à mettre en application la statistique. Cela n'a pas toujours été facile !

GS : *Quelles étaient plus précisément tes fonctions, que devais-tu réaliser dans ton premier poste ? Ta formation en DUT t'y avait-elle préparée ? Je crains que non, vu la manière dont tu t'es exclamée...*

NM : L'intitulé exact de mes fonctions était « technicienne en analyses statistiques et gestion de données ». Et non, on ne peut pas dire que j'étais vraiment préparée à ces fonctions. En particulier, à l'IUT, nous faisions les analyses à la machine à calculer et avec des abaques. A l'AgroParisTech, il y avait des outils de traitements statistiques que je ne connaissais pas du tout ; ma formation en informatique à l'IUT m'a heureusement aidée à comprendre rapidement comment ces outils fonctionnaient.

Ma chance a également été d'avoir d'une part, un chef de service qui avait, lui, une très bonne connaissance de ce domaine, et d'autre part un collègue informaticien.

GS : *J'ai tant de questions à te poser sur cet environnement statistique « vintage », et également sur le détail des analyses statistiques à mener. Je les garde pour une partie à part de cet entretien.*

J'ai pour l'heure une rapide question subsidiaire, toujours dans l'objectif de mieux cerner ton parcours. Dois-je donc comprendre que tu as essentiellement appris la statistique sur le tas ?

NM : La statistique appliquée, oui ; quant à la théorie, cela allait mieux, les cours que j'avais suivis étaient suffisants.

Je crois d'ailleurs qu'aujourd'hui, le problème est toujours le même, hélas ! J'ai pu m'en rendre compte il y a deux ans, avec des stagiaires de niveau master ayant effectué auparavant un parcours en IUT. Les étudiants connaissaient la statistique d'un point de vue théorique, mais la

[3]Diplôme universitaire de technologie, formation post-baccalauréat de deux ans, que l'on prépare dans un IUT (institut universitaire de technologie).

[4]Qui correspond à l'actuel DUT STID dont il est question à la note en bas de page 14.

[5]Sauf mention contraire, nous parlerons des établissements d'enseignement et de recherche en utilisant leur dénomination actuelle.

pratique métier leur manquait. Peut-être que voir plus d'études de cas réels durant leur formation leur permettrait de mieux aborder le métier de statisticien.

GS : *Là encore, j'aimerais bien t'en faire dire davantage sur les cours, formations et angles d'attaque que tu juges nécessaires dans un cursus de statistique. Mais bien sûr, et surtout au vu du titre de notre revue, nous en parlerons longuement dans une partie à part !*

Avançons pour l'heure dans ta carrière : combien de temps es-tu restée dans ce laboratoire de zootechnie, quel est le chemin qui t'a menée ensuite jusqu'au laboratoire d'informatique de Grenoble (le LIG) ?

NM : Je suis restée à l'AgroParisTech de 1983 à 1991. Ensuite, mon mari a trouvé du travail en région grenobloise, j'ai moi-même obtenu ma mutation à l'INRA (Institut national de la recherche agronomique) de Grenoble, dans un laboratoire d'économie et sociologie rurales, et nous avons quitté la région parisienne. Ma mutation s'est effectuée au prix d'un changement de fonctions : sur un poste d'administratrice systèmes et réseaux. Au bout de trois ans sur ce poste, heureusement, le directeur du laboratoire m'a proposé de travailler sur un projet européen, et là, j'ai retrouvé la gestion des données. Je me suis formée aux bases de données relationnelles. Ensuite, je suis revenue vers les statistiques en travaillant sur un autre projet, qui concernait le suivi des doctorants (en biologie) : il s'agissait de comprendre leurs trajectoires professionnelles, les déterminants professionnels à l'œuvre, et à travers cela, les transferts de la recherche vers l'industrie par le biais de ces doctorants.

Le vrai tournant a eu lieu un beau matin, de manière tout à fait aléatoire : une de mes collègues me demande pourquoi je ne suivrais pas à nouveau une formation. C'était visiblement la question à me poser ce jour-là, car six mois plus tard, je commençais un DESS[6] en production et analyse de données en sciences politiques. Ce fut une superbe année, partagée avec de jeunes étudiants en formation initiale (ils avaient 22 ans... et moi 37).

GS : *Pourquoi avoir décidé, après tout ce temps à l'AgroParisTech et à l'INRA, de suivre un DESS orienté vers les sciences politiques ? Il devait bien y avoir des DESS de statistique orientés vers les sciences du vivant...*

NM : A Paris je travaillais certes en zootechnie, donc en sciences du vivant. Mais ensuite, à l'INRA de Grenoble, j'étais, comme je te le disais, dans un laboratoire de sciences économiques et rurales. J'avais travaillé avec des données d'Eurostat (l'organisme de statistique publique de l'Union Européenne) sur les transferts de la PAC (politique agricole commune) aux agriculteurs, sur le RGA (recensement général de l'agriculture), et j'avais également réalisé une étude sur l'insertion professionnelle des doctorants de l'INRA dans les entreprises du domaine des biotechnologies. Ces derniers sujets étaient orientés vers les sciences humaines et sociales (SHS), c'est pour cela que j'avais besoin de me former en production et analyse de données en SHS.

GS : *Revenons à ton parcours professionnel. Comment ce DESS a-t-il changé (ou pas) ta carrière ?*

NM : Dans un premier temps, le changement était inexistant : malgré ce nouveau diplôme, je n'avais pas de progression de carrière à l'INRA. J'ai donc passé un concours d'ingénieur au CNRS et y ai été recrutée. En 2001, j'ai ainsi été affectée au laboratoire de recherche de Sciences Po Grenoble. J'ai principalement travaillé sur la conception d'enquêtes d'opinion, et en particu-

[6]DESS : diplôme d'études supérieures spécialisées, correspondant dans nos terminologies actuelles à un master à visée professionnelle.

lier, l'*European social survey*[7]. C'était une étude européenne, comme son nom l'indique, donc de très grande envergure, c'était un projet intéressant... mais je ne me sentais pas très à l'aise dans cet environnement.

C'est ainsi, qu'après trois ans à Sciences Po Grenoble, j'ai intégré le LIG, pour mener des expérimentations avec des utilisateurs de dispositifs informatiques – c'est-à-dire de la production et de l'analyse de données ! Mon travail dans ce poste prend aujourd'hui d'autres dimensions : d'une part, je suis en train de rédiger une thèse de doctorat, et d'autre part, je gère un projet pour onze équipes du laboratoire sur cette thématique de la production et de l'analyse des données.

GS : *Je ne résiste pas : quel est le titre de ta thèse ? Et comment trouves-tu le temps de mener des travaux de recherche au sein de toutes tes fonctions opérationnelles ?*

NM : Je n'en ai pas encore choisi le titre... Mais je peux indiquer en revanche les questions auxquelles j'apporte une contribution. Quels processus de conduite de la recherche faut-il appliquer dans la recherche en informatique centrée sur l'humain ? Quel processus expérimental[8] suivre ? Quels indicateurs de traçabilité retenir pour cette recherche ? Je fournis un modèle que j'ai appelé THEDRE, pour *Traceable Human Experimental Design Research*. Cette thèse de doctorat est une synthèse des travaux que j'ai menés en accompagnant les doctorants pour faire leurs propres expériences en cours de thèse. Elle n'est pas déconnectée de mon travail quotidien, au contraire ! En revanche, pour l'important travail de rédaction, j'ai pris cinq semaines sur mes congés annuels et j'ai également droit à 120h de congés de formation, soit presque quatre semaines supplémentaires.

GS : *Ah, justement, ces jours-ci, quand je t'écris, je reçois une notification automatique par retour de courriel me disant que tu es en congés de formation. Ne serait-ce pas la dernière ligne droite de rédaction ?*

NM : Oui, c'est tout à fait ça !

GS : *Ta modestie t'en empêche, mais pour conclure la description de ton parcours, je me dois de révéler à nos lecteurs que tu es lauréate en 2016 d'une médaille de Cristal du CNRS :*

> La médaille de cristal du CNRS distingue des ingénieurs, des techniciens et des administratifs. Elle récompense celles et ceux qui, par leur créativité, leur maîtrise technique et leur sens de l'innovation, contribuent aux côtés des chercheurs à l'avancée des savoirs et à l'excellence de la recherche française.

Ta notice de lauréate te présente ainsi (de manière moins vivante et plus technique que ce que tu nous as écrit plus haut), je cite :

> Ingénieure d'études en traitement et analyse de bases de données
>
> Grâce à une double compétence en sciences humaines et sociales et en statistiques, Nadine Mandran joue un rôle pivot au sein du Laboratoire d'informatique de Gre-

[7] « Enquête européenne sur les attitudes, les croyances et les comportements des Européens menée tous les deux ans depuis 2002 à l'initiative de la Fondation européenne de la science » selon Wikipédia.

[8] Question technique posée après l'entretien. GS : *Ces processus expérimentaux, dont il est question ici et plus tard dans l'entretien, sont-ils en lien avec la théorie statistique des plans d'expérience ?* NM : Quand je parle de processus expérimental, c'est au sens large. C'est à dire comment intégrer l'humain pour construire et évaluer de la connaissance scientifique. Pour cela, il faut avoir les moyens de réaliser des protocoles adaptés et qui intègrent des méthodes issues des sciences humaines et sociales.

noble (LIG). Experte en évaluation expérimentale des systèmes qui impliquent une interaction homme-machine, elle étudie les processus de démarche expérimentale favorisant l'intégration de l'humain dans la construction et l'évaluation des systèmes. Ce travail l'a conduit naturellement à gérer une plateforme pour la capitalisation et l'analyse des données expérimentales. En tant qu'évaluatrice des contributions de la recherche, elle est très régulièrement associée aux publications comme co-auteure. Attentive à la transmission des savoirs, Nadine Mandran accompagne, par ailleurs, les doctorants du laboratoire dans la mise en œuvre de ces expérimentations. Enfin, elle pilote la cellule « support et qualité » de son laboratoire, avec pour objectif l'amélioration continue des services. Prenant appui sur les compétences des personnels administratifs et informatiques, elle veille ainsi à formaliser les procédures, capitaliser et archiver les documents.

Qu'as-tu ressenti en recevant cette distinction ? Qui aimerais-tu remercier ?

NM : Pas mal d'émotion ! J'ai été heureusement surprise, car mon parcours pluridisciplinaire est difficilement valorisable avec les critères standards des évaluations dans les concours, c'est donc une très belle reconnaissance. J'en suis très fière.

Remercier quelqu'un ou plusieurs personnes en particulier est très difficile, car dans mon parcours, j'ai eu la chance de rencontrer de très nombreux collègues qui m'ont tendu des perches pour progresser... De peur d'oublier d'en citer quelques-uns si je commençais à en dresser la liste, je préfère les remercier en bloc !

Questionnaire à la Proust

GS : *Pour continuer de faire connaissance, Nadine, je te propose un moment de détente : un questionnaire de Proust (ou en tout cas, sa version statistique). La règle du jeu est qu'il faut donner une réponse courte, à tout le moins pour la plupart des réponses...*

Quel est ton résultat de statistique préféré (théorème ou application) ?

NM : Sans aucun doute les travaux de visualisation de données de Charles Minard puis de Jacques Bertin ! Et aussi ces petits tests non-paramétriques qu'on applique quotidiennement. Quoique les outils de modélisation soient très utiles également. Et puis il y a les analyses factorielles et les classifications. Tous ces outils me plaisent bien quand utilisés à bon escient !

GS : *Qu'entends-tu par « ces petits tests non-paramétriques » ?*

NM : Je pense surtout aux tests de comparaison de deux échantillons, test de Mann-Whitney quand ils sont indépendants et test des rangs de Wilcoxon quand ils sont appariés.

GS : *Quel est ton manuel de statistique préféré ?*

NM : Les *Méthodes statistiques* de Snedecor et Cochran [4].

GS : *Fréquentiste, bayésienne, ou autre ?*

NM : Il faut utiliser le cadre qui convient au moment venu !

GS : *Statistique ou statistiques ?*

NM : La statistique pour les statistiques appliquées, donc les deux.

GS : *Comment situes-tu la statistique par rapport à la* data science *?*

NM : Je dirais : la théorie versus un métier.

GS : *Quelle est la faiblesse principale de la statistique ?*

NM : De ne pas être reconnue comme un domaine en soi. Si tu développes des méthodes de statistique théorique, les mathématiciens te disent que ce ne sont pas des mathématiques ; et si tu fais des statistiques appliquées, on te dit que tu ne fais pas de statistique, et encore moins des mathématiques. Vive la statistique libre !

GS : *Quelle est ta vertu préférée en statistique ?*

NM : D'être à la recherche du sens procuré par les données, d'être des outils d'aides à la décision, et non pas de manipulation... mais la frontière est mince.

GS : *Quelle est notre principal défaut en tant que statisticiens ?*

NM : La comparaison : nous sommes toujours en train de comparer pour comprendre.

GS : *Quel est le rêve de tout statisticien ?*

NM : Le mien : ne plus traiter de données !

GS : *Et son cauchemar ?*

NM : Le mien : ne plus traiter de données ! Et d'ailleurs je me demande si cela ne va pas me manquer à la retraite...

GS : *Pourquoi la recherche en mathématique et en informatique est-elle masculine et le monde de la statistique est-il plus mixte ?*

NM : C'est bien ce que je disais, la statistique est un domaine à part...

GS : *Connais-tu la SFdS, que t'apporte-t-elle ?*

NM : Je la connais, bien sûr : c'est une communauté qui comprend ce que je fais et c'est rare !

L'environnement statistique et informatique dans les laboratoires, du début des années 1980 à nos jours

GS : *Dans cette partie, je voudrais que nous revenions sur l'équipement et les outils utilisés pour mettre en œuvre la statistique. Tu as mentionné dans ton portrait qu'à l'IUT, vous utilisiez des machines à calculer et des abaques et qu'à l'AgroParisTech, l'environnement était plus riche.*

Commençons par l'IUT. Vous y aviez appris la mise en œuvre statistique d'avant l'ère informatique, c'est ça ? Peux-tu m'en dire davantage ? J'imagine par exemple que les jeux de données devaient être tout petits. Ne faisiez-vous pas sans arrêt des erreurs de calcul ? Comment expliquerais-tu ces techniques à la génération qui est née après la généralisation des micro-ordinateurs et ne sait même pas ce qu'est une abaque ?

NM : De 1981 à 1983, pendant mon cursus à l'IUT, les analyses de données que j'ai pu mener étaient « manuelles », avec pour seul outil une calculatrice de marque Texas Instruments. Les méthodes abordées couvraient les tests d'hypothèses et la régression linéaire simple. Dans le cours de recherche opérationnelle, nous avions également abordé l'analyse factorielle avec trois variables. je n'ai pas souvenir d'avoir mené de traitements statistiques avec un quelconque

ordinateur. Mais comme tu le dis, les jeux de données étaient de petites tailles. Je ne suis pas certaine que l'on faisait plus d'erreurs de calculs qu'aujourd'hui. Le travail quasi-manuel demande de préparer les plans de traitement pour ne pas travailler en mode essai–erreur comme aujourd'hui...

Pour expliquer ce temps à la génération actuelle des jeunes statisticiens : je dirais que c'est compliqué de décrire l'utilisation des abaques ! De même, comment revenir à la lecture de ces « vieillles tables » de quantiles et de fonctions de répartition alors que tous les logiciels de statistique nous donnent désormais des P-valeurs avec 6 chiffres après la virgule ? Voire des sorties du type $p < 0.0000$ (ah, ah !)...

Dans certaines écoles d'ingénieur en mécanique, les étudiants réalisent encore en première année des dessins avec planche, feuille, papier et crayon, à l'« ancienne », pour apprendre la rigueur et les notions de vue dans l'espace. Bien évidemment, ils passent ensuite rapidement aux logiciels de dessin. De même, peut-être qu'enseigner deux ou trois études de cas à l'ancienne, et notamment avec les « vieilles tables » pourrait apporter un réel sens aux données et à ce fameux α d'erreur de première espèce ou à cette fameuse P-valeur.

En tout cas, pour ma part, j'enseigne la pratique métier de la statistique en licence 3 et en master 2 professionnels, et ce que j'essaie surtout de transmettre, c'est que pratiquer la statistique, ce n'est pas simplement utiliser un logiciel, mais étudier des données, appliquer des méthodes en étant conscient de leurs limites, et en réalité, être en recherche de sens (sans tordre les données).

GS : *Pour des raisons pragmatiques (pas d'ordinateur à l'examen !), j'apprends à mes étudiants à lire les tables de fonctions de répartition, afin de calculer des P-valeurs... Mais ne parlons pas pour l'instant des enseignements, je compte t'interroger en détails à ce sujet plus tard !*

Passons plutôt à l'AgroParisTech. Quel était l'environnement statistique en place là-bas, fin 1983 ? Comment a-t-il évolué durant le temps que tu y as passé ? Et comment viviez-vous, toi mais aussi tes collègues, ces évolutions ?

NM : L'environnement informatique, c'était, côté machines, des Mini 6 ; il fallait réserver sa place pour travailler sur un des six terminaux[9] à disposition pour l'école. Côté systèmes d'exploitation, Multics[10] était installé à Jouy-en Josas ; la connection s'y effectuait *via* modem. La messagerie existait déjà. Bien entendu, tout cela fonctionnait en ligne de commande.

Côté statistique, pour faire de l'analyse des données, nous utilisions les bibliothèques de procédures FORTRAN nommées amance, acomp et acobi. Ces algorithmes avaient été développés à partir de l'ouvrage de référence de Lebart, Morineau et Tabard [3], publié en 1977.

Ensuite, vers 1985, un PC a fait son apparition au laboratoire : c'était un Victor avec des disquettes $5''\ 1/4$... une très grande avancée ! A l'exception de mon chef de service de l'époque et des jeunes ingénieurs, personne ne comprenait vraiment l'intérêt de ces outils pour la recherche en zootechnie.

Le début des années 1990 a vu l'arrivée des environnements graphiques. J'étais tellement à l'aise avec les lignes de commandes que je ne voyais pas l'intérêt de ce mode d'interaction.

[9]Selon Wikipédia, consulté le 4 octobre 2016 : « Le Mini 6 est l'un des mini-ordinateurs commercialisés par la Compagnie internationale pour l'informatique (CII), qui permettait de fonctionner en relation avec un grand système [...]. »

[10]Selon Wikipédia, consulté le 4 octobre 2016 : « Multics (acronyme de MULTiplexed Information and Computing Service) est le nom d'un système d'exploitation en temps partagé. »

Mais j'ai vite changé d'avis quand j'ai compris à quel point c'était utile aux personnes non familiarisées avec l'informatique. Coté statistique, c'est la période où SAS (*Statistical Analysis System*) est arrivé dans les laboratoires. Là encore, c'était une réelle avancée !

GS : *Mais SAS n'a-t-il pas été conçu et développé dès la seconde partie des années 1970 ? C'est également un environnement assez peu graphique. Dans un autre genre, R également a été développé dans la première partie des années 1990, quand son usage s'est-il généralisé ? Et qu'en est-il des autres logiciels, comme SPSS, Stata, etc. ?*

Je rappelle simplement à nos lecteurs qu'en 1991, au début de la période que nous évoquons ici, tu as quitté l'AgroParisTech pour l'INRA de Grenoble, où tu es restée jusqu'en 2001. Ma question ici vise à te faire parler des mutations informatiques et logicielles vécues dans cette tranche de vie...

NM : Pour tout avouer, je ne connais pas bien l'histoire des logiciels de statistique. SAS, au départ, n'était pas graphique du tout, mais assez rapidement, des compléments comme SAS/Insight sont apparus pour pallier cela. Ils ont aidé à populariser SAS dans les entreprises et à en faire un logiciel de référence pour la fouille de données. Entre-temps, l'INRA a acquis des licences pour les logiciels S et S-Plus : je ne sais toutefois pas s'ils ont été beaucoup utilisés. La politique d'achat des logiciels de l'INRA étant malgré tout focalisée sur SAS, c'est avec lui que j'ai essentiellement travaillé jusqu'en 2001, c'est-à-dire jusqu'au jour où j'ai quitté l'INRA. Je recourais également à l'occasion à Statgraphics et SPAD.

GS : *Tu devines mes questions suivantes...*

NM : En effet ! Quand j'ai intégré Sciences Po Grenoble en 2001, j'ai utilisé Converso pour gérer les enquêtes de terrains, SPSS pour en traiter les données, ainsi que SPAD. J'ai également eu la chance de me former aux logiciels d'analyse lexicale ; mon préféré, c'était Alceste.

En 2005, quand je suis arrivée dans mon poste actuel, au laboratoire d'informatique, j'ai été affectée dans une équipe qui n'était pas familiarisée avec le traitement de données. Le logiciel de référence était donc... Excel ! J'ai proposé que des licences SAS et SPAD soient acquises, pour que je puisse travailler. La réponse a été : « Nous (sous-entendu, nous, les informaticiens) allons te développer un logiciel ». J'ai eu de la peine à les convaincre du contraire mais ai finalement décroché ma première licence SAS dès l'année de mon arrivée. (Pour la petite histoire, la licence SAS ne coûtait que 200 euros par an à l'époque !) En revanche, je n'ai jamais retrouvé SPAD...

Je me suis ensuite mise à R : c'était gratuit (je n'avais plus personne à convaincre), puis, utiliser R fait sérieux : on s'embête à ré-écrire du code... Aujourd'hui, je n'effectue plus de traitements statistiques moi-même, sauf exception, mais je supervise ceux des étudiants et des doctorants. Et je leur conseille de recourir à R, car c'est un logiciel puissant, et la communauté qui est derrière est solide. Pour des informaticiens, la prise en main de R n'est pas très compliquée ; ce n'est pas le cas en revanche pour une bonne partie des chercheurs et étudiants des sciences humaines et sociales.

GS : *Un dernier défi, Nadine, pour clore cette partie : pourrais-tu nous fournir une petite mise en perspective de toutes ces évolutions ?*

NM : L'évolution de l'informatique a été colossale, depuis que j'ai commencé à travailler en 1983. Mes tâches consistaient alors, tout à la fois, en de la gestion de système, de la programmation, de la gestion de bases de données et en des traitements de données. Aujourd'hui, je pense

que personne ne peut plus maîtriser simultanément tous ces aspects-là de l'informatique. Les métiers se sont vraiment spécialisés : administrateur système et réseau, développeur, concepteur, gestionnaire de bases de données, etc.

Au niveau des logiciels, j'ai l'impression que l'offre est de plus en plus riche, mais de plus en plus floue également. Régulièrement, un collègue prétend avoir trouvé le nouveau logiciel miracle qui fait tout... Le risque aussi de nombres de logiciels, c'est que leur pérennité n'est pas garantie. C'est sans doute ce qui fait la force de R : la communauté qui le soutient, le développe, et le met en œuvre.

Recommandations sur les formations en statistique

GS : *Je te propose de parler maintenant des enseignements et formations, ceux que tu dispenses toi-même, mais aussi de l'enseignement de la statistique en général.*

Pour situer ton propos, peux-tu commencer par nous décrire les enseignements que tu donnes en ce moment et des enseignements passés qui t'ont particulièrement marquée ?

NM : Pour ne parler que des enseignements les plus récents, donc grenoblois, et en commençant par ceux dans les cursus de sciences humaines : j'ai assuré des travaux pratiques sur la programmation en SAS dans la seconde année du master Progis[11] de Sciences Po Grenoble, de 2002 à 2005. Ont suivi des cours de fouille des données (*data mining*) en master 2 d'analyse économique à la faculté d'économie de Grenoble, de 2009 à 2014.

Côté formations en informatique : d'une part, j'assure un module de formation à l'école doctorale MSTII[12] de Grenoble, focalisé sur la démarche expérimentale pour la recherche en informatique centrée humain. D'autre part, le petit dernier est, depuis cette année, un cours de remise à niveau en statistique en licence professionnelle[13] « Big Data ».

GS : *Pour mieux situer ton propos : pourrais-tu nous raconter une expérience marquante en enseignement ?*

NM : Sans hésiter : la première année où j'ai assuré des cours de fouille des données en faculté d'économie. Je devais enseigner l'utilisation de l'analyse en composantes principales (ACP), de l'analyse des correspondances multiples (ACM), et la classification ascendante hiérarchique (CAH). J'ai eu recours à des fichiers de données très « propres », sur lesquels les étudiants ne devaient finalement que répliquer les méthodes que je leur présentais. Le cours s'est évidemment déroulé sans encombre. Mais quelques mois plus tard, lors de la période de stages en entreprise, une des tutrices me téléphone pour me signaler que sa stagiaire ne savait pas traiter les données. Quelle surprise pour moi ! Elle parlait d'une étudiante parmi celles et ceux qui s'étaient révélés les plus brillants dans mon cours.

En échangeant avec la tutrice, j'ai compris la chose suivante, tout à fait fondamentale : au bout de cinq ans d'études, les étudiant-e-s ont la tête pleine de méthodes mais ont des difficultés pour la mise en œuvre de ces méthodes sur des données réelles dans un contexte professionnel.

[11] Qui « forme des spécialistes des études dans le domaine de l'opinion, du marketing et des médias » ; voir http://www.masterprogis.fr/

[12] Ecole doctorale mathématiques, sciences et technologies de l'information, informatique ; voir http://edmstii.ujf-grenoble.fr/

[13] Décrite dans un numéro précédent de notre revue, voir [2].

J'ai donc changé radicalement ma méthode d'enseigner ! Mon module durait 32 heures, ce qui est un temps suffisamment long pour mettre en place un enseignement par l'erreur, renouvelé pendant cinq ans. Cela a été une belle réussite, je crois. J'ai ré-utilisé cette approche cette année, pour mon cours de remise à niveau en licence professionnelle... mais je n'ai eu que 12 heures, ce qui est trop court.

GS : *Qu'appelles-tu un « enseignement par l'erreur » ? Et peux-tu nous expliquer au passage les points sur lesquels les étudiant-e-s n'étaient pas formés précédemment, et qui se révèlent cruciaux pour la mise en œuvre de méthodes statistiques en entreprise ?*

NM : Je dis « apprentissage par l'erreur » mais je pourrais également parler de « résolution de problèmes ». Le cadre est le suivant : je propose aux étudiants un objectif, qui est d'étudier une problématique sociale, leur indique un résultat, qui doit être une présentation projetée sur un écran, et leur donne un jeu de données support, *via* l'adresse du dépôt de données de l'*European social survey* dont il a été question plus haut.

Leur seule contrainte est d'utiliser une méthode d'ACP, d'ACM ou de CAH sur ces données (afin d'illustrer le contenu du cours) ; pour le reste, libre à eux de procéder comme ils le veulent, de rajouter tous les traitements statistiques qu'ils désirent. De manière générale, en deux heures, ils ont fini l'exercice. Et c'est là que je leur pose une première question : « Avez-vous vérifié la validité de vos données ? » Et une seconde question, encore plus déstabilisante : « Quelle est la problématique ? » C'est ici qu'il est intéressant d'observer comment les étudiants réagissent... car souvent, ils ne se sont jamais posé cette question de la validité des données et ils ne savent pas vraiment formuler une problématique à laquelle une étude statistique doit répondre.

Tu l'auras compris, ils repartent à la case départ, mais de manière plus prudente et en se posant dès lors un peu plus de questions. Ensuite toutes les heures, je leur demande de s'arrêter pour faire avec eux le bilan de leur progression, de préférence sous forme de questions. « Vous avez des données manquantes sur cette question, quelle a été votre approche pour pallier ce fait ? » S'ils m'expliquent alors qu'ils ont supprimé cette variable, je leur en demande la raison. Et je procède ainsi de suite, jusqu'à la question : « Pourquoi avez-vous mis en œuvre une ACP, pourquoi pas une ACM ? » Et finalement, au bout de 32 heures, ils sont capables d'élaborer une présentation qui est correcte sur le fond, établie à partir de données elles-mêmes correctement traitées, et avec un niveau de recul suffisant pour que cette présentation soit abordable par des non-spécialistes de la science des données.

Mon objectif fondamental est de les aider à comprendre en quoi consiste le métier de *data scientist* : avant tout, des prises de décisions à chaque instant du traitement de données, décisions qui sont garantes d'un bon traitement de données et qui évitent des sur-interprétations.

Chaque année de cette nouvelle formule, les étudiants de master ont vraiment apprécié ce cours... et moi-même, de mon coté, j'étais satisfaite de ne pas (seulement) leur expliquer la statistique théoriquement, mais de faire œuvre de statistique appliquée, et *in fine* de les former à la déontologie de notre métier. A ce propos, j'ai dernièrement trouvé un article [1] écrit par Deming en 1965 sur les pratiques métiers et la déontologie des statisticiens : un petit bijou ! Que nous devrions lire et relire, faire lire à nos étudiants, et aussi aux femmes et hommes politiques, qui manipulent si bien les chiffres.

GS : *L'approche que tu utilises ici fait écho à un vœu que tu formulais en page 51 : enseigner aux étudiants que la pratique métier de la statistique, c'est « en réalité, être en recherche de*

sens (*sans tordre les données*) *». Tu suggérais également, en page 46, que les étudiants devraient « voir plus d'études de cas réels durant leur formation ».*

Ton approche d'enseignement est fantastique, mais n'est-elle pas très personnelle et difficilement reproductible ? Pour dire les choses autrement (et avec un manque d'humilité qui va te faire bondir sans doute) : comment les formations qui n'ont pas une Nadine Mandran sous la main peuvent-elles s'y prendre ? Je voudrais que tu formules maintenant, si tu le veux bien, quelques conseils et recommandations pour l'évolution de nos formations en statistique, à l'ère de la science des données...

NM : Cette méthode est utilisable et utile dans deux cas de figure : premièrement, lorsque les étudiants disposent de connaissances théoriques, parfois vastes, mais qu'ils ne savent pas mettre en pratique (c'est le cas du cours de master que j'ai décrit en détails ci-dessus) ; deuxièmement, quand on souhaite enseigner des pratiques métiers ne sollicitant pas trop de connaissances théoriques.

Je crois que je peux enseigner ainsi, par erreur et questions, grâce à ma propre expérience et à mon propre questionnement : j'analyse des données depuis près de trente ans, j'ai par conséquent intégré des automatismes, et je me suis déjà suffisamment posé de questions sur le fonctionnement de ces méthodes et leurs limites.

Il me semble que tous les praticiens de l'analyse des données doivent pouvoir suivre une telle approche pour leurs cours. Le seul pré-requis est une bonne expérience « métier » du sujet. Il faut en effet savoir répondre à toutes les difficultés rencontrées par les étudiants, notamment au niveau des étapes de préparation des données, et pouvoir leur expliquer ce qui sera attendu d'eux en entreprise (en termes d'autonomie de décisions et de compte-rendus).

En réalité, cette méthode correspond également à une approche suivie dans le cas de projets en IUT STID[14] par exemple... mais peut-être pas dans des formations plus théoriques « à la fac ».

Cela étant, s'il ne faut pas perdre de vue que la pratique, c'est bien et incontournable, il reste indispensable que les étudiants disposent de fondements théoriques solides. Cette année, j'ai été stupéfaite de découvrir des étudiants de L3 ne connaissant pas les notions de « loi de distribution » ni de loi normale. Comment enseigner la statistique dans ces conditions ?

Enfin, avec la montée en puissance de l'analyse de données massives (*big data*), les personnes non averties pensent à tort que l'analyse de données, ce sont des données qui sont traitées par des algorithmes. De nombreux étudiants en informatique sont convaincus d'être des analystes de données, car ils savent créer des algorithmes pour les appliquer sur des données. Une manière de corriger ces tendances serait de travailler sur les profils métiers des statisticiens (ou des *data scientists*, pour utiliser une terminologie à la mode) : identifier les valeurs ajoutées de ces métiers par rapport aux informaticiens, et communiquer de manière large sur ces métiers d'expertise. Les métiers de la statistique attirent peu d'étudiants car la statistique est associée aux mathématiques ; certes, c'est vrai, mais cela ne recouvre pas toute la dimension d'analyse et d'interprétation, qui devrait être enseignée et valorisée.

[14] IUT : Institut universitaire de technologie ; STID : Statistique et informatique décisionnelle, le parcours où la statistique et l'analyse de données sont les plus présentes parmi l'ensemble des parcours des IUT.

Le mot de la fin

GS : *Y a-t-il un sujet sur lequel tu aurais voulu t'exprimer et que j'ai oublié ? Ou peut-être as-tu un message à faire passer à nos lecteurs ? Bref, que veux-tu écrire comme fin à cet entretien ?*

NM : Je voudrais souligner que l'analyse des données n'est pas nouvelle, même les pharaons avaient leurs *data scientists* en un sens. Certes, c'était du simple dénombrement de populations, mais le recueil de données était déjà là pour donner du sens et aider à la prise de décision.

Aujourd'hui, parce que nous sommes dans une ère de données massives, le métier de statisticien apparaît comme tout nouveau, il devient central. Or, vouloir l'automatiser en le réduisant à un simple processus automatique, qu'une machine pourrait effectuer, est à mon avis une erreur. Dans ce contexte, il me semble important de former et d'armer les étudiants qui se destinent à être statisticiens (pour ne pas écrire *data scientists*) : il faut insister auprès d'eux sur le fait que les résultats de l'analyse des données doivent être reçus par un être humain, doué d'intelligence, qui seul peut y mettre du sens.

Je voudrais également souligner que certes, nous sommes deux communautés, les statisticiens et les informaticiens, à nous préoccuper de l'exploitation des données massives. Mais je pense que notre travail de statisticiens doit se faire main dans la main avec les informaticiens, car d'une part ils se préoccupent de l'optimisation des algorithmes, ce qui rend notre métier plus souple et flexible. D'autre part, les travaux de formalisation des pratiques métiers des statisticiens qu'ils conduisent vont permettre d'augmenter l'utilisabilité des plateformes d'analyse. C'est un travail pluridisciplinaire coûteux mais riche. Ces travaux[15], auxquels je participe, sur cette formalisation des pratiques métiers sont aussi une réelle opportunité d'identifier la manière dont nous travaillons et d'offrir ainsi aux étudiants de meilleures pratiques. Je plaide avec force pour la pluridisciplinarité : pour un enrichissement mutuel !

GS : *Merci, Nadine, d'avoir accepté de répondre à mes questions et de partager ton expérience entre tous ces mondes, statistique, informatique et sciences politiques. J'espère ne pas t'avoir trop retardée dans la dernière ligne droite de rédaction de ton manuscrit de thèse...*

NM : Non, ne t'inquiète pas : répondre à tes questions a constitué un exercice très agréable !

GS : *Tant mieux ! La quatrième édition de cette chronique, dans le numéro du premier semestre de 2017, recueillera à nouveau le témoignage d'un praticien en entreprise, pour continuer notre alternance praticien-universitaire.*

Références

[1] Deming, W. (1965), Principles of professional statistical practice, *Annals of Mathematical Statistics*, **36**(6), 1883–1900.

[2] Dupuy-Chessa, S., S. Lambert-Lacroix, et G. Blanco-Lainé (2016), Un parcours Big Data en alternance dans une licence professionnelle, *Statistique et Enseignement*, **7**(1), 121–126.

[3] Lebart, L., A. Morineau, et N. Tabard (1977), *Techniques de la description statistique : méthodes et logiciels pour l'analyse de grands tableaux*, Dunod-Bordas.

[15] NM : Je peux citer la plateforme Sakura du LIG, le projet ANR (agence nationale de la recherche) Hubble en *e-learning*, le PIA (projet d'investissements d'avenir) Ikats, etc.

[4] Snedecor, G. et W. Cochran (1971), *Méthodes statistiques*, Association de coordination technique agricole (ACTA), traduit par l'ACTA à partir de la sixième édition originelle en anglais, qui date de 1957.