Quinze ans de recherche appliquée en science des données

Sébastien Dejean

Université Toulouse III - Paul Sabatier

Mémoire

présenté pour l'obtention de

L'Habilitation à Diriger des Recherches

par

Sébastien Déjean

de l'Institut de Mathématiques de Toulouse

Quinze ans de recherche appliquée en science des données

Soutenance le 17 mai 2019 devant le jury composé de :

Rapporteurs :	Christophe Ambroise	Professeur, Université d'Evry
	Marie-Laure Martin-Magniette	Directrice de recherche, INRA AgroParisTech
	Fionn Murtagh	Professeur, University of Huddersfield
Examinatrices :	Laure Coutin	Professeure, Université Toulouse III
	Nathalie Vialaneix	Directrice de recherche, INRA Toulouse
Marraine :	Josiane Mothe	Professeure, Université Toulouse II
Invitée :	Nathalie Viguerie	Chargée de recherche, INSERM Toulouse

Résumé

Ce mémoire synthétise quinze ans d'activités scientifiques à l'Institut de Mathématiques de Toulouse. Il fait état de mon rôle dans des travaux de recherche interdisciplinaires autour de l'analyse de données. Dans ce cadre, au-delà de la mise en œuvre de méthodes statistiques, c'est toute une méthodologie que j'ai développée pour exploiter au mieux et valoriser des données. Ainsi, après avoir livré quelques réflexions sur la notion de donnée dans un premier chapitre, je consacre le deuxième chapitre à l'élaboration d'une méthodologie de travail dans le cadre de collaborations interdisciplinaires. J'illustre sa construction et sa mise en œuvre à travers plusieurs cas d'étude liés notamment à l'analyse de données issues de bio-technologies à haut-débit. Cette méthodologie s'étend de la formulation d'une question précise à l'interprétation des résultats d'une méthode statistique permettant potentiellement d'y répondre. Elle s'intègre naturellement dans ce qu'il est devenu courant d'appeler la science des données. Le troisième chapitre se focalise sur ma thématique privilégiée : l'intégration de données. Ce thème de recherche vise à développer des démarches ou des méthodes visant à extraire une information plus pertinente en analysant globalement plusieurs jeux de données plutôt qu'en les analysant séparément. Cette thématique est illustrée d'abord dans le cadre de la recherche d'information puis dans celui de l'analyse de données biologiques. Dans ce dernier cas, j'ai contribué au développement de nouvelles méthodes statistiques ainsi qu'à leur dissémination auprès de la communauté des biologistes. Pour cela, j'ai régulièrement supervisé la mise en œuvre de ces nouvelles méthodes dans des projets de recherche, j'ai encadré des étudiants en thèse et master et j'ai également contribué à la mise à disposition d'outils logiciels pour lesquels j'ai aussi assuré des actions de formation. Enfin, le quatrième chapitre est consacré à mes activités de soutien à la recherche.

Remerciements

Je souhaite d'abord remercier les deux premières personnes à qui j'ai parlé de mon projet de soutenir une habilitation à diriger des recherches : Josiane Mothe et Nathalie Vialaneix. Si je leur ai demandé de m'accompagner dans ce projet, c'est parce que je les connais depuis longtemps et que j'ai une très grande confiance en elle. Et pour tout dire, ce sont deux personnes qui m'impressionnent réellement par leur capacité de travail. Quand elles ont répondu avec enthousiasme à ma demande d'accompagnement, je savais qu'elles me pousseraient systématiquement dans la bonne direction.

Ensuite, c'est à Laure Coutin que j'ai parlé de ce projet lors de mon entretien professionnel annuel. La discussion que nous avons eue à ce moment-là et les encouragements que Laure a formulés ont été pour moi une source de motivation très importante tout au long de la rédaction de ce mémoire. Je la remercie pour cela et pour avoir accepté de participer au jury.

Mes remerciements s'adressent aussi très sincèrement à Christophe Ambroise, Marie-Laure Martin-Magniette et Fionn Murtagh pour avoir accepté d'être rapporteurs de mon mémoire. Au-délà de la pertinence de leurs rapports qui m'ont emmenée vers d'autres réflexions par rapport à mon travail, je leur suis également très reconnaissant d'avoir respecté les délais pour la remise de leur rapport facilitant ainsi grandement les démarches administratives.

La liste des personnes qui ont compté dans mon environ professionnel depuis mon arrivée à l'université Paul Sabatier en 2003 est très longue. Parmi ces personnes, je souhaite remercier plus particulièrement Alain Baccini et Philippe Besse qui m'ont tout de suite fait confiance et m'ont associé dès le début à des projets de recherche. J'ai énormément appris lors de nos nombreuses discussions. C'est aussi à mes collègues de bureau actuels, Nicolas Savy et Philippe Saint-Pierre, que j'adresse mes remerciements : pour la bonne humeur qui règne dans notre bureau et pour la liberté qu'ils m'offrent d'accueillir mes collaborateurs pour des réunions de travail ainsi que mes invités de passage pour des durées plus ou moins longues.

Ce mémoire fait état de nombreuses collaborations dans le domaine de la biologie. Et dans ce domaine là, j'aurais de nombreuses personnes à remercier également. J'en retiendrais deux, disons pour des raisons historiques, car ce sont les deux premiers biologistes que j'ai rencontrés professionnellement (c'était le 27 mai 2003 dans un train de retour de La Londe Les Maures où venait d'avoir lieu un atelier de formation de l'Inserm sur l'analyse statistique des biopuces !) : Nathalie Viguerie et Pascal Martin. Si j'arrive aujourd'hui à collaborer dans le cadre de projets interdisciplinaires, c'est en grande partie à eux que je le dois.

Certains projets de recherche bénéficient de l'implication très significative de doctorants et dans ce cadre là, il y a en deux qui m'ont particulièrement marqué : Kim-Anh Lê Cao et Ignacio González. Ils sont tous les deux à l'origine de l'aventure `mixOmics` qui entre dans sa dixième année d'existence. Qui aurait cru que ce package nous réunirait encore 10 ans après sa création ? Si c'est le cas aujourd'hui, c'est grâce à leur dynamisme et à leur implication sans faille dans ce projet. Avec eux aussi, j'ai énormément appris et j'apprends encore ; je les en remercie très sincèrement.

Dans la liste des personnes auprès de qui j'apprends énormément de choses, j'apprécie tout particulièrement le groupe informel des "ingénieurs statisticiens toulousains" qui se rencontrent 3 à 4 fois par an, au gré d'une organisation tournante en différents endroits de la place toulousaine. Ce groupe ne vivrait pas aussi bien sans les implications de collègues comme Thibault Laurent, Christophe Bontemps et de notre référence statistico-historico-culturelle qui détient le record de présentations lors de nos rencontres : Joseph Saint-Pierre. Je les remercie aussi très chaleureusement et j'espère que nos rencontres se poursuivront longtemps dans le même état d'esprit.

Dans le contexte de la recherche, il y a les chercheur.se.s ou assimilé.es, c'est à dire celles et ceux qui co-signent les articles ou qui apparaissent dans la rubrique Remerciements en fin d'article, et il y a les autres... Qu'on les appelle BIATSS, personnel de soutien ou de support à la recherche, personnel administratif, secrétaire, gestionnaire... (j'en fais partie en tant qu'ingénieur !) ils sont indispensables au bon fonctionnement de la recherche. Je ne citerai pas l'ensemble du personnel administratif que j'ai côtoyé ces dernières années, mais je souhaite remercier très sincèrement mes collègues Marie-Laure Ausset, Delphine Dalla-Riva pour leur professionnalisme, leur sens du collectif et pour nos nombreuses discussions

toujours enrichissantes. J'associe également à ces remerciements Françoise Michel dont l'efficacité n'a d'égal que la discrétion.

À l'IMT, il y a d'autres personnes qui contribuent au fait que je me sens bien dans mon environnement professionnel. Là aussi, il est délicat de ne citer que peu de noms, mais je me risque quand même à évoquer Guillaume Chèze, mon partenaire de cirque de Noël, entre autres, dont l'humour et la finesse d'esprit n'ont d'égal que la discrétion de Françoise Michel (vous me suivez ?) et Marcello Bernardara qui a contribué à la naissance de l'IMT Football Club.

Puisque que j'aborde le volet sportif, j'en profite pour évoquer mes co-équipiers de l'Athletics Coaching de Ramonville, club d'athlétisme que je fréquente dans l'espoir de progresser en course à pied. Les séances d'entraînement du jeudi soir sont pour moi une vraie bouffée d'oxygène dans des semaines de travail parfois difficiles. Alors merci aux coachs Riaucem et Javier et à toutes celles et ceux qui donnent une aussi bonne ambiance à nos entraînements. Si cela vous intéresse, contactez-moi, le club accueille toujours avec grand plaisir les nouvelles têtes (et jambes) !

Enfin, tout cela ne serait pas grand-chose si je n'étais pas aussi comblé dans ma vie privée.

Table des matières

What of the future ? The future of data analysis can involve great process, the overcoming of real difficulties, and the provision of a great service to all fields of science and technology. Will it ? That remains to us, to our willingness to take up the rocky road of real problems in preference to smooth road of unreal assumptions, arbitrary criteria, and abstract results without real attachments. Who is for the challenge ?

John Tukey, 1962
The Future of Data Analysis

Introduction

Ce mémoire est pour moi l'occasion de faire un bilan sur une quinzaine d'années d'activités professionnelles dans le monde de la recherche. Je travaille en effet depuis 2003 à l'Université Toulouse III Paul Sabatier. J'ai d'abord été affecté au Laboratoire de Statistique et Probabilités puis à l'Institut de Mathématiques de Toulouse (IMT) qui a fédéré, à sa création en 2007, les trois laboratoires de mathématiques de l'Université Paul Sabatier dont le LSP. C'est dans cet environnement que j'exerce le métier d'ingénieur de recherche en calcul scientifique. Le RÉFérentiel des Emplois-types de la Recherche et de l'ENseignement Supérieur (REFERENS [1]), avant sa modification en 2016, décrivait brièvement mon métier en ces termes : *L'ingénieur de recherche en calcul scientifique analyse, dans le cadre de projets de recherche, un problème théorique ou une situation d'expérience et d'observation. Il recherche les méthodes d'analyse, conçoit et optimise les outils permettant le traitement du problème. Il s'assure de la pertinence des résultats obtenus.*

Dans ce cadre, j'ai décliné les différentes missions associées à ce statut dans le contexte de projets de recherche associant les mathématiques à d'autres disciplines. Parmi les domaines avec lesquels j'ai le plus d'interactions, deux se détachent nettement ; il s'agit de la biologie et de la recherche d'information en informatique.

Biologie ou la révolution des omiques Pour collaborer au mieux avec les biologistes, j'ai réalisé une remise à niveau en biologie pour me rappeler globalement que : les organismes vivants eukaryotes sont composés de cellules, ces cellules ont un noyau, dans ce noyau, de l'ADN structuré en double hélice, cet ADN étant trop gros et trop important pour sortir du noyau, il est recodé en ARN messager (ARNm), molécule simple brin conçue pour sortir du noyau, et faire la rencontre de ribosomes qui vont décoder cet ARNm pour fabriquer des protéines en assemblant des briques élémentaires appelées acides aminés. On peut considérer en première approche que les gènes sont des morceaux de l'ADN. Précisons également que la transcription est le passage de l'ADN à l'ARNm, et la traduction, le passage de l'ARNm à la protéine. Ces éléments forment ce qui est communément appelé le dogme de la biologie moléculaire qui, même s'il reste globalement vrai, a été considérablement re-visité grâce aux avancées de la recherche. J'y reviendrai dans le chapitre 3. Dans ce contexte, la technologie a évolué pour ouvrir l'ère des omiques ou de la biologie dite à haut-débit. En effet, les mécanismes moléculaires menant de l'information génétique aux caractères observables d'un organisme via les protéines, peuvent désormais être étudiés à grande échelle par des technologies comme le séquençage de nouvelle génération. Cette technologie permet de déterminer l'ordre des composants d'une macromolécule (par exemple les nucléotides pour l'ADN et l'ARN) à des vitesses de plus en plus élevées pour des coûts de plus en plus faibles. Et c'est justement cette grande échelle qui a profondément bouleversé les habitudes des biologistes en les confrontant à des données de plus en plus volumineuses, ouvrant ainsi un vaste espace de collaborations avec des statisticiens.

Recherche d'information La recherche d'information est le domaine de la recherche qui s'intéresse à la structure, à l'analyse, à l'organisation, au stockage, à la recherche et à la découverte de l'information d'après la définition qu'en a donné, Gerard Salton, un des pionniers de ce domaine [79].

Très schématiquement, la recherche d'information peut se voir comme une série de questions :

- **quoi ?** quelle est l'information qu'un utilisateur recherche, en général, ce `quoi` se traduit par une requête qui peut prendre des formes différentes : quelques mots, une phrase, un texte, une image...
- **comment ?** à partir de la requête d'un utilisateur (on sait maintenant ce que l'on cherche), le `comment` est le cœur d'un système de recherche d'information (SRI), c'est la méthode que l'on va utiliser pour répondre au besoin exprimé par l'utilisateur via sa requête.
- **où ?** on sait ce que l'on cherche (`quoi`), on connaît la méthode à utiliser (`comment`), encore faut-il savoir où chercher. C'est le `où` qui est généralement défini par le terme de collection. C'est l'ensemble des ressources dans lequel des informations pertinentes vis-à-vis de la requête de l'utilisateur sont sensées se trouver.

1. referens.enseignementsup-recherche.gouv.fr

— **pertinence ?** on sait quoi chercher, où et comment ! On est donc en mesure de fournir des résultats à l'utilisateur. C'est très bien, mais ces résultats sont-ils pertinents ? Cette question peut être traitée dans le cadre de campagnes d'évaluation durant lesquelles requêtes et collection sont fournies à des compétiteurs désireux de confronter leur SRI à d'autres sur des bases communes. Les organisateurs connaissant de leur côté, les documents à retourner, ils sont à même d'évaluer les résultats des systèmes en utilisant de multiples indicateurs de performance majoritairement basés sur le rappel (rapport du nombre de documents pertinents retournés sur le nombre total de documents pertinents) et la précision (rapport du nombre de documents pertinents retournés sur le nombre de documents retournés).

D'autres domaines... Les deux domaines précédemment évoqués rassemblent la grande majorité des collaborations auxquelles j'ai participé activement, mais ce ne sont pas les seuls. De façon plus ponctuelle, j'ai des interactions avec des collègues dans d'autres domaines et dans des contextes différents. Notamment, je suis régulièrement en contact avec des étudiants en orthodontie afin de superviser les analyses statistiques qu'ils mettent en œuvre dans leurs recherches. Par exemple, j'ai ainsi suggéré l'utilisation de modèles de durée de vie pour la modélisation de la durée de traitement de canines incluses. Dans un autre contexte, je collabore avec des collègues sociologues et linguistes autour de la notion du jeu utilisé pour l'apprentissage en vue de développer la plateforme *Check Your Smile*[2]. C'est ainsi que j'ai participé à la journée *Genre, générations, structure sociale et territoires à l'ère du numérique* pour y co-présenter un travail intitulé *Jeu, jeunes et numériques : des relations à déconstruire* [73]. Les travaux menés dans ce contexte se concrétisent autour d'enquêtes menées auprès d'étudiants toulousains concernant leur rapport au jeu. Je supervise l'analyse des réponses à ces enquêtes qui est généralement traitée par des étudiants en projets tutorés dont j'assure l'encadrement. Plus loin dans ce mémoire, partie 2.2.3, j'aborde des travaux menés en collaboration avec des orthophonistes. Ces travaux ont été concrétisés par une publication dans le *Journal of Voice* [35] et ils sont à l'origine de contacts qui ont abouti à un projet financé par l'Agence Nationale de la Recherche sur l'aide au diagnostic de la maladie de Parkinson par analyse de la voix, projet que j'aborde dans la section 2.2.3. Plus récemment, j'ai initié des contacts avec la Fédération Régionale de Recherche en Psychiatrie et Santé Mentale (Ferrepsy [3]) pour accompagner des chercheurs et chercheuses en psychiatrie et les aider à améliorer les analyses statistiques habituellement réalisées dans cette communauté.

Un point commun Le point commun des projets auxquels j'ai participé dans ces différents contextes réside dans l'analyse de données. Qu'elles soient qualitatives ou quantitatives, volumineuses ou pas, précises ou pas, obtenues par sondage ou par mesures physiques... tous les projets auxquels j'ai contribué ont généré des données. Mon implication a pour objectif systématique de faire parler ces données par des méthodes statistiques. Cependant, dans de tels contextes, le travail ne se résume pas à mettre en œuvre des méthodes statistiques ; cela ne représente qu'une partie du travail à accomplir et pas forcément la plus difficile. Cette idée est exprimée notamment dans un livre de Jordan Ellenberg, professeur de mathématiques à l'université du Wisconsin, intitulé *How not be Wrong* et sous-titré *The hidden maths of everyday life* ou *The power of mathematical thinking* selon les éditions [33]. Je cite ici un extrait de la traduction française parue dans le numéro de février 2018 du magazine La Recherche.

Je n'aime pas les problèmes "issus de la vie courante". Ils donnent une image fausse du rapport entre mathématiques et réalité. "Bobby a 300 billes ; il en donne 30% à Jenny" [...] Mais les questions du monde réel ne ressemblent pas à ce type d'énoncés. Un problème dans le monde réel, c'est quelque chose comme : "La récession et ses suites ont-elles été particulièrement dures pour l'emploi des femmes, et dans ce cas, dans quelle mesure est-ce le résultat de l'administration Obama ?" Votre calculette n'a pas de bouton pour ça. Car pour donner une réponse vous devez connaître autre chose que de simples chiffres. [...] Ce n'est qu'après avoir formulé ces questions que vous pouvez prendre votre calculette. Mais à ce point, le véritable travail intellectuel est déjà terminé. Diviser un nombre par un autre, ce n'est que du calcul ; savoir ce que vous devez diviser par quoi, ça, c'est des mathématiques.

L'idée exprimée dans ce passage est également présente dans les travaux de Stella Baruk, mathématicienne, chercheuse en pédagogie des mathématiques et auteur de nombreux ouvrages sur le sujet, depuis *Échec et maths* en 1973 jusqu'à *Les chiffres ? Même pas peur* en 2016. Un article paru dans Le Monde en 2008 et intitulé *Stella Baruk, le goût des maths, une affaire de langue*[4] mentionne notamment le fait qu'elle regrette que pour enseigner les mathématiques dès le plus jeune âge, l'accent soit mis davantage sur la technique que sur le sens. On revient ainsi à l'idée précédemment évoqué que dans les mathématiques, la

2. www.checkyoursmile.fr
3. www.ferrepsy.fr
4. www.lemonde.fr/le-monde-2/article/2008/09/12/stella-baruk-le-gout-des-maths-une-affaire-de-langue_1094437_1004868.html

méthode n'est pas le seul élément de considération à prendre en compte, la compréhension d'un problème et sa reformulation en termes mathématiques font partie intégrante d'une démarche mathématique.

Cette notion illustre parfaitement mon travail en interaction permanente avec des scientifiques de disciplines différentes. Les difficultés ne résident pas seulement dans la mise en œuvre de méthodes statistiques (la calculette dont il est question dans l'extrait prend ici la forme d'un logiciel de statistique) mais dans les discussions préalables et dans la préparation des données. Ainsi, le fait de réaliser un test statistique sur des données bien organisées ne représente généralement pas de difficulté, mais le chemin à parcourir avant de conclure au fait que la réalisation de ce test apportera des éléments de réponse à la question posée est généralement beaucoup plus long et ardu.

Plan du mémoire Ainsi, dans ce mémoire intitulé en référence à l'article de David Donoho *50 years of data science* [21], il est question de travaux autour de l'analyse de données. Ainsi, je consacre un premier chapitre à des réflexions sur la notion de données. Cela me permet de définir le vocabulaire élémentaire et de discuter du chemin parfois long à parcourir entre des données brutes et des données susceptibles d'être soumises à une analyse statistique afin de fournir des réponses à un problème concret.

Ensuite, j'aborde dans un deuxième chapitre la méthodologie de travail que j'ai mise en place au gré de mes collaborations. Elle se caractérise par une feuille de route qui me sert de repère lorsque je m'implique dans un projet. Loin de se focaliser sur la donnée, elle positionne l'analyse de données au centre d'une démarche allant de la formulation d'une question à la réponse qui lui est apportée via le recours à une méthode statistique. J'illustre cette démarche sur quatre cas d'étude qui, à des degrés divers, m'ont permis d'aboutir à cette méthodologie.

Dans un troisième chapitre, j'aborde des travaux en lien avec mon thème méthodologique privilégié en statistique : l'intégration de données. Il est d'abord question d'une démarche générale d'intégration de données dans le cadre d'une collaboration suivie dans le domaine de la recherche d'information. Dans un second temps, je présente mes contributions au développement et à la diffusion sous forme logicielle de nouvelles méthodes pour l'intégration de données biologiques acquises via les nouvelles bio-technologies.

Enfin, dans un quatrième chapitre, je présente quelques éléments liés à des activités de support à la recherche soutenant ma volonté de postuler à une habilitation à diriger des recherches.

Chapitre 1

Quelques réflexions autour des données

Ce mémoire évoque largement des questions autour de l'analyse de données. Avant de discuter de la mise en œuvre de méthodes sur des données réelles dans des cas concrets, je souhaite livrer quelques éléments de réflexion autour de la notion de données. Qu'est-ce qu'une donnée ? Que pouvons-nous en faire ? Comment la manipuler, la nettoyer ? Comment en extraire de l'information, de la connaissance ? Que faire sans données ? sont des questions qui sont abordées dans ce chapitre.

1.1 Qu'est-ce qu'une donnée ?

Le dictionnaire de l'Académie Française dans sa 9ème édition [1] définit une donnée comme suit :

> *II. DONNÉE n. f. XIIIe siècle, au sens de « distribution, aumône » ; XVIIIe siècle, comme terme de* **mathématiques**. *Participe passé féminin substantivé de donner au sens de « indiquer, dire ».1. Fait ou principe indiscuté, ou considéré comme tel, sur lequel se fonde un raisonnement ; constatation servant de base à un examen, une recherche, une découverte. Les données de la science. Les données de l'expérience.* **Données statistiques**. *Ma théorie s'appuie sur des données précises. C'est un raisonnement fondé sur des données incertaines. Dans cette affaire, il est essentiel de connaître la donnée de départ. Par ext. Idée principale, thème constitutif qui est à l'origine d'une œuvre littéraire. La donnée d'une tragédie. 2. PSYCHOL. Ce qui est connu immédiatement par le sujet, indépendamment de toute élaboration de l'esprit, par opposition à ce qui est connu par induction ou déduction, par raisonnement, par calcul. Titre célèbre : Essai sur les données immédiates de la conscience, d'Henri Bergson (1889). 3.* **MATH.** *Souvent au pluriel. Chacune des quantités ou propriétés mentionnées dans l'énoncé d'un problème et qui permettent de le résoudre. 4.* **INFORM**. *Représentation d'une information sous une forme conventionnelle adaptée à son exploitation. Le traitement automatique des données. Une banque, une base de données.*

La définition ci-dessous positionne la notion de `donnée` dans le domaine scientifique et plus précisément, entre autres, en mathématiques et en informatique et il est également fait mention de *données statistiques*. L'idée principale étant qu'une donnée est un élément sur lequel on va s'appuyer pour développer un raisonnement, une démarche d'analyse.

Si on regarde du côté anglophone avec le Cambridge Dictionnary [2], la définition de *data* est la suivante :

> *Information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer*

Et dans le Merriam-Webster [3] :

> *Definition of data 1 : factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation*
> *the data is plentiful and easily available —H. A. Gleason, Jr.*
> *comprehensive data on economic growth have been published —N. H. Jacoby*
> *2 : information in digital form that can be transmitted or processed*
> *3 : information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful*

1. www.academie-francaise.fr/le-dictionnaire/la-9e-edition
2. dictionary.cambridge.org
3. www.merriam-webster.com

Dans ces deux définitions du terme *data*, une notion intéressante apparaît dans le troisième item du Merriam-Webster avec un accent plus technologique : il est en effet question d'une information qui intègre également du bruit (*irrelevant information*) et qui doit donc être traité pour être porteuse de sens (*must be processed to be meaningful*). Alors que la définition du dictionnaire de l'Académie Française évoque un *un fait ou un principe indiscuté, ou considéré comme tel.*

Ces éléments ne sont pas nécessairement contradictoires mais notent essentiellement des modifications de sens selon le domaine d'utilisation du terme.

Donc, à ce niveau, on considère qu'une donnée est un point de départ de quelque chose. Pour aller un peu plus loin, je cite ici un extrait d'une leçon donnée par Serge Abitboul au Collège de France [2] ; il y est question de données, d'information et de connaissance.

> *Des mesures de température relevées chaque jour dans une station météo, ce sont des données. Une courbe donnant l'évolution dans le temps de la température moyenne dans un lieu, c'est une information. Le fait que la température sur Terre augmente en fonction de l'activité humaine, c'est une connaissance. Ces trois notions sont très proches les unes des autres. Grossièrement, voici le sens que nous leur donnerons :*
>
> — *Une donnée est une description élémentaire, typiquement numérique pour nous, d'une réalité. C'est par exemple une observation ou une mesure.*
>
> — *À partir de données collectées, de l'information est obtenue en organisant ces données, en les structurant pour en dégager du sens.*
>
> — *En comprenant le sens de l'information, nous aboutissons à des connaissances, c'est-à-dire à des « faits » considérés comme vrais dans l'univers d'un locuteur, et à des « lois » (des règles logiques) de cet univers.*

L'enchaînement décrit ici fait le lien entre donnée et connaissance qui est un thème assez largement répandu sous la terminologie *Data2Knowledge*. En effet, toute démarche d'analyse de données vise à mieux comprendre un phénomène, une situation, un système au travers d'informations plus ou moins précises, éventuellement partielles, voire partiales, que l'on a pu acquérir. Pour cela, on peut noter que l'information émerge de données *organisées* et *structurées* et que la connaissance se situe dans *l'univers d'un locuteur* (pas dans le domaine de la statistique). En résumé, il s'agit donc bien de transformer des données en connaissance. Cette démarche et cette terminologie se retrouvent aussi bien dans des sociétés [4] que dans des offres de formation [5] ou des groupes de recherche [6].

Une autre façon de positionner les données au cœur d'une démarche scientifique est exprimée dans cette phrase de Jean-Paul Benzécri citée dans [57] : *Le modèle doit suivre les données.* En d'autres termes, ce sont les données qui conduisent et pour reprendre une autre terminologie relativement à la mode, on parle donc de démarche *data-driven*. Ce terme est défini de la façon suivante sur le site wikipedia [7] consulté le 19 avril 2018 : *The adjective data-driven means that progress in an activity is compelled by data, rather than by intuition or by personal experience.* Dans un contexte commercial, l'approche *data-driven* est définie en ces termes dans un glossaire du domaine de l'économie numérique [8].

> *Le Data Driven, également appelée Data Driven Marketing, se base sur une approche qui consiste à prendre des décisions stratégiques sur la base d'une analyse et d'une interprétation des données.*
> *L'approche Data Driven permet d'examiner et d'organiser la data dans le but de mieux cerner ses consommateurs et ses clients. Le « pilotage par la donnée » va donc permettre à une organisation de contextualiser et/ou de personnaliser le message à ses prospects ainsi qu'à ses clients.*

Poursuivons maintenant ce chapitre en évoquant comment interpréter des données.

1.2 Interpréter des données

Comme on l'a vu précédemment, une démarche d'analyse de données vise globalement à acquérir des connaissances sur un phénomène. On cherche ainsi à savoir ce que les données ont à nous apprendre dans un cadre spécifique. Cela étant, chaque méthode statistique va révéler un aspect des données sans forcément fournir un point de vue global, voire une vérité sur le phénomène étudié. C'est finalement l'idée que l'on retrouve dans la parabole des aveugles et de l'éléphant. Cette histoire qui semble prendre sa source en Inde dans la tradition bouddhiste discute les limites de la perception et de l'importance d'un contexte. Même si ce n'est pas l'objet initial, cette histoire illustrée par la figure 1.1 peut être interprétée

4. www.d2k.com - Transforming data2knowledge
5. Master Data & Knowledge de l'Université Paris Saclay par exemple
6. Data to Knowledge (D2K) Research Group, University of Bristol, Medical School
7. https ://en.wikipedia.org/wiki/Data-driven
8. www.atinternet.com/glossaire/data-driven

d'un point de vue statistique : chaque méthode ou indicateur statistique va fournir un point de vue sur les données. Et ce n'est que l'assimilation de l'ensemble qui est susceptible de fournir une interprétation cohérente d'un phénomène.

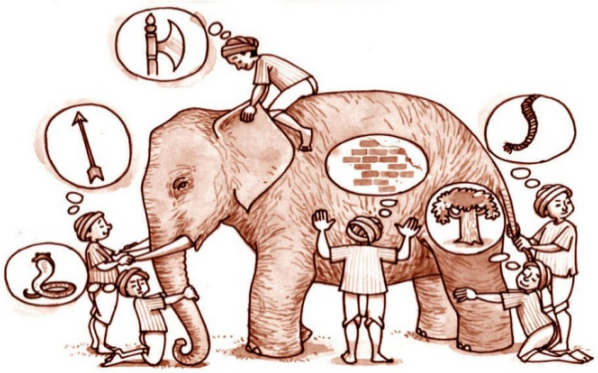

FIGURE 1.1 – Illustration de la parabole *Six aveugles et un éléphant*. Source inconnue.

C'est une idée assez similaire que l'on retrouve sous l'appelation d'effet Rashomon [9]. Cet effet est nommé ainsi en référence à un film japonais dans lequel quatre témoins racontent des versions différentes d'un crime en se basant sur les mêmes faits. Dans le contexte statistique, il est évoqué par L. Breiman [13] pour justifier le recours à des méthodes d'apprentissage par aggrégation de modèles. En combinant les différents points de vue compatibles avec les faits, on a plus de chances d'accéder à la vérité.

C'est finalement dans cette même veine que j'illustre parfois l'utilisation de certains indicateurs statistiques. Par exemple, dans cet article paru dans Le Monde du 7 juillet 2009 [10] intitulé *Les filles brillent en classe, les garçons aux concours*, la première phrase *Elles obtiennent de meilleurs résultats en cours de scolarité, mais réussissent moins bien les concours des meilleures grandes écoles que les hommes* présente une situation qui peut sembler contradictoire. Il suffit cependant de lire l'article jusqu'au bout pour avoir une explication de ce phénomène en termes statistiques : *Alors comment comprendre ce déséquilibre ? « D'un point de vue technique, il semble que la structure du concours HEC crée d'avantage d'hétérogénéité chez les hommes que chez les femmes », estime M. Peyrache. Si, « en moyenne », les performances des hommes et des femmes sont similaires, « les notes des femmes sont concentrées autour de la moyenne, tandis que celles des hommes sont très dispersées avec beaucoup de très bonnes notes et de très mauvaises. Mécaniquement, quand on sélectionne les 380 premiers résultats, on a un peu plus d'hommes ».* En l'interprétant dans la perspective de l'effet Rashomon, on peut considérer que les deux affirmations contenues dans le titre *Les filles brilles en classe, les garçons aux concours* apparaissent contradictoires alors qu'elles se basent sur les mêmes chiffres. Il convient ainsi de ne pas se contenter d'interpréter des indicateurs de position, mais de les compléter par des indicateurs de dispersion.

Pour finir sur le sujet, j'évoque ici le fameux paradoxe de Simpson par cet exemple tiré du numéro de Février-Mars 2018 de la revue *Pour la Science* [11] (tableau 1.1).

TABLE 1.1 – Données fictives illustrant le paradoxe de Simpson, extrait de Pour la Science, numéro Février-Mars 2018.

	Physique		Biologie		Cumul	
	G	F	G	F	G	F
Réussite	80	10	4	50	84	60
Echec	10	0	6	40	16	40
Total	90	10	10	90	100	100
% réussite	89%	100%	40%	55%	**84%**	**60 %**

Là aussi, deux affirmations apparemment contradictoires peuvent être tirées des chiffres du tableau 1.1. Les filles réussissent mieux que les garçons dans chacune des disciplines (taux de réussite de 100% et 55% contre 89% et 40%) ; les garçons réussisent globalement mieux que le filles (taux de réussite de 84% contre 60%).

9. en.wikipedia.org/wiki/Rashomon_effect
10. www.lemonde.fr/societe/article/2009/09/07/les-filles-brillent-en-classe-les-garcons-aux-concours_1236895_3224.html
11. Jean-Paul Delahaye citant *Statistiques Méfiez-vous !* de Nicolas Gauvrit

Ces quelques exemples illustrent les précautions à prendre dans une démarche d'analyse des données. Une interprétation de données et/ou de résultats de méthodes statistiques est à effectuer par un expert du domaine à l'origine des données accompagné dans cette démarche par un statisticien. Cet accompagnement doit intervenir le plus tôt possible dans une projet, avant même que toute donnée soit disponible. C'est le sujet de la planification expérimentale que j'aborde maintenant.

1.3 Des plans d'expérience

Pour planifier une future acquisition de données avec des collaborateurs, il y a deux phrases qui me servent généralement de préambule. La première qui est attribuée à R.A. Fisher stipule que le recours à un statisticien après l'acquisition de données équivaut à une autopsie par un médecin légiste : *To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination : he may be able to say what the experiment died of.* La seconde, plus récente est extraite de [50] : *While a good design does not guarantee a successful experiment, a suitably bad design guarantees a failed experiment—no results or incorrect results.* Elle établit aussi clairement le fait qu'une mauvaise planification de l'acquisition de données est une garantie d'échec de l'analyse de données.

Ainsi, lorsque j'ai la possibilité de discuter de la planification d'une expérience biologique, mon rôle revient assez fréquemment à contre-balancer l'attitude illustrée dans l'image 1.2 qui consiste à penser que *plus on a de données, plus on a d'information* ; phrase que je reformule parfois en *plus on a de données, plus on a de bruit.*

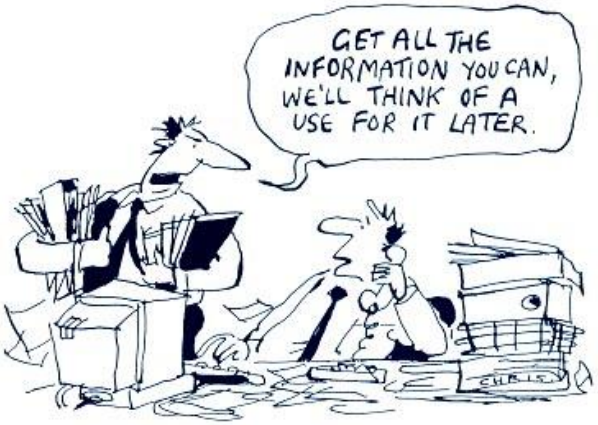

FIGURE 1.2 – Image illustrant une démarche opposée à la planification expérimentale : accumuler des données et aviser ensuite. Source initiale inconnue.

Par exemple, il m'est arrivé d'interroger des collègues biologistes sur la pertinence d'un plan d'expériences visant à croiser 6 lignées biologiques avec 6 traitements et 2 temps de mesure, soit 72 conditions différentes. Je ne prétends pas décider des expériences à mener, mais en ayant conscience des méthodes statistiques qui seront à utiliser pour analyser les données générées dans un tel plan (par exemple, une ANOVA à 3 facteurs dans le cas présent), je peux parfois susciter une réflexion qui visera, par exemple, à diminuer le nombre de modalités d'un facteur voire à supprimer un des facteurs pressentis afin de mieux cibler une question précise et d'acquérir les données les plus à même d'y répondre.

Souvent, tout l'enjeu de la planification expérimentale réside dans la faculté de poser ou de faire poser une question précise. Par exemple, *Je m'intéresse à l'effet du traitement sur les lignées* n'est pas une question. En insistant pour que ce soit réellement une question qui soit posée, on arrive petit à petit à voir les priorités de l'expérience envisagée et ainsi à mieux la planifier.

Ensuite, une fois les données générées, celles-ci ne sont pas forcément prêtes à être soumises à une analyse statistique : il convient de les organiser sous une forme propice.

1.4 Des données bien rangées

It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data. [93]. Cet extrait complète l'idée évoquée en introduction qui exprime qu'une fois que l'on sait quelle méthode appliquer, *le véritable travail intellectuel est terminé.* Il le complète sur un aspect plus pratique : l'organisation des données. En effet, on peut très bien se retrouver dans un cas où : une question précise a été formulée, des données ont été acquises, on sait quelle méthode utiliser, en utilisant quel logiciel,

mais ça ne marche pas immédiatement parce que la façon dont les données ont été stockées n'est pas celle attendue par le logiciel. Il convient donc de porter une attention particulière à l'organisation des données.

Ainsi, dans l'article [93], H. Wickham discute le principe de données bien rangées (*tidy data*) et le justifie au travers d'une véritable philosophie de la donnée ; l'état de données bien rangées se caractérisant par le fait que la structure physique du jeu de données (en ligne et colonne) est directement lié à la signification des données. Pour cela, trois règles sont établies pour caractériser des données bien rangées :

1. chaque variable forme une colonne ;
2. chaque observation forme une ligne ;
3. chaque type « d'unité observationnelle » forme une table.

Le fait d'avoir des données bien rangées implique une mise en œuvre plus aisée pour la manipulation, la visualisation et la modélisation. C'est l'ensemble de ces considérations qui a conduit à la réalisation du *tidyverse*[12] : un ensemble de packages pour le logiciel R dédiés à la science des données et partageant cette philosophie des données décrites dans [93] et reprise dans [94].

Il est maintenant établi que tous ces aspects liés à des démarches préalables à l'analyse de données propement dite font partie de la science des données. Donoho [21] les intègre dans les 6 points définissant une *Greater Data Science* :

1. Recueil, préparation et exploration de données
2. Représentation et transformation de données
3. Calculer avec des données
4. Modélisation de données
5. Visualisation et présentation de données
6. Science au sujet de la Data Science

En résumé : pas d'analyse de données, sans préparation des données. Ce point peut être d'autant plus complexe que les données sont volumineuses et qu'elles n'ont pas été recueillies selon un plan d'expérience mais plutôt selon la démarche illustrée par la figure 1.2 caractéristique de l'ère des *big data*.

1.5 Des données massives

Il est difficile de parler de données, sans parler de données massives et du fameux terme *big data*. Cela étant, entre le début et la fin de la rédaction de ce mémoire (grossièrement du début à la fin de l'année 2018), la popularité de ce terme semble avoir nettement diminué, au profit du terme d'intelligence artificielle, qui bien que n'étant pas du tout synonyme, l'a remplacé dans les termes en vogue dans le domaine de l'analyse de données.

Pour aborder le sujet des données massives, je vais illustrer mon propos par cette anecdote. Lors d'une conférence internationale en Statistique et Santé organisée à l'Institut de Mathématiques de Toulouse en janvier 2018[13], j'ai animé une table ronde sur le thème *Is Big Data in health a myth or a reality ?* Certains témoignages apportés durant cette table ronde étaient très évocateurs des multiples apparences du *big data*. Indépendamment des multiples V (Volume, Vélocité, Variété, Véracité...) fréquemment utilisés pour caractériser les *big data*, certains ont évoqué le fait que l'impossibilité d'importer un fichier de données dans son logiciel de statistique favori permettait d'obtenir le « label » *big Data* !

Les données massives sont une réalité et le fait de les exploiter au mieux représente généralement un défi important. Mais la problématique concrète ne doit pas être perdue de vue dans les développements suscités par les *big data*. Donoho consacre une partie de son article [21] (section 10 *The Next 50 Years of Data Science*, sous-section 10.3 *Scientific Data Analysis, Tested Empirically*) à évoquer des études montrant que des méthodes récentes et relativement sophistiquées n'améliorent pas, en pratique, les résultats de méthodes plus anciennes et plus rudimentaires. Il évoque notamment les travaux de [45], ses propres travaux avec Jin [22] ainsi que ceux de Zhao et al. [100]. Et les extraits de certaines de ces publications sont assez éloquents :

— Extrait de [45] : *The situation to date thus appears to be one of very substantial theoretical progress, leading to deep theoretical developments and to increased predictive power in practical applications. While all of these things are true, it is the contention of this paper that the practical impact of the developments has been inflated ; that although progress has been made, it may well not be as great as has been suggested.*

12. www.tidyverse.org
13. www.cimi.univ-toulouse.fr/mib/en/conference-statistics-and-health

— Extrait de from [22] repris dans [21] : *That is, every one of the more glamorous techniques suffers worse maximal regret. Boosting, random forests, and so on are dramatically more complex and have correspondingly higher charisma in the machine learning community. But against a series of preexisting benchmarks developed in the machine learning community, the charismatic methods do not outperform the homeliest of procedures...*
— Extrait de [100] : *We hope our work will help shift the emphasis of ongoing prediction modeling efforts in genomics from the development of complex models to the more important issues of study design, model interpretation, and independent validation.*

Ces études présentent à mon sens un léger biais car il s'agit de procéder à des méta-analyses de comparaison de méthodes sur plusieurs jeux de données. Cela peut donner un contexte plus favorable à une méthode élémentaire par rapport à des méthodes plus sophistiquées optimisées pour une problématique et un jeu de données spécifiques. Cela étant, ce que je souhaite mettre en avant ici, c'est le fait que le volume toujours croissant des données auxquelles on doit faire face ne doit pas faire oublier les méthodes élémentaires qui ont fait leur preuve depuis de nombreuses années. C'est un peu en ces mêmes termes que j'ai entendu lors d'une conférence consacrée au *big data*, un ingénieur de la société Criteo (*entreprise française de reciblage publicitaire personnalisé sur internet*[14]) mentionner le fait que les temps de calculs étant cruciaux dans le domaine de la publicité en ligne, les méthodes modernes de type *machine learning* (SVM, forêts aléatoires...), n'étaient pas utilisables en pratique et que seul un modèle de régression linéaire était susceptible de fournir des éléments de réponse dans les temps impartis qui se comptent en milli-secondes pour certains systèmes d'enchères pour l'allocation de bannière de publicité en ligne.

L'idée ici n'est pas de prôner un conservatisme systématique vis-à-vis des méthodes d'analyse de données mais d'insister à nouveau sur le fait que les méthodes d'analyse de données doivent avant tout servir à... analyser des données. Et donc, si une méthode, aussi rudimentaire soit-elle, permet d'exploiter pertinemment les données, il n'est peut-être pas utile de lui préférer une méthode plus récente, plus « glamour »(le terme est utilisé dans [21]) si celle-ci n'améliore pas l'interprétation concrète des données.

Un autre élément à prendre en compte dans cette discussion est le suivant : faut-il nécessairement traiter l'ensemble des données dont on dispose pour un problème donné ? Cette citation extraite de [57] : *Mais la statistique nous rappelle qu'il est parfois vain de vouloir traiter des millions d'observations lorsqu'il y a des possibilités d'échantillonnage.* C'est à partir de ce constat que je suggère parfois à des étudiants qui rencontrent des problèmes de place mémoire ou de temps de calcul de commencer leurs analyses sur une partie des données. En renouvelant l'expérience sur différents échantillons issus des données, on peut ainsi se rendre compte si les résultats obtenus sur plusieurs échantillons sont stables ou pas, et décider ensuite de la suite à donner à l'analyse.

Pour conclure à ce sujet et relier à nouveau la notion de donnée à celle d'information, je reprends ici l'extrait d'une entrevue donnée par C. Villani au journal du CNRS[15] : *[au sujet de l'imagerie médicale] Dans certains cas, avec seulement 2% des données, vous parvenez à reconstituer l'information utile. Avec des exemples comme celui-ci, on saisit l'un des drames du monde actuel : l'information est perdue dans les données, et le problème est de parvenir à trouver celle qui compte.*

La phrase *l'information est perdue dans les données* est pour moi très révélatrice du risque que l'on encourt à manipuler des *big data* ; on peut se retrouver submergé par des données sans pouvoir en extraire une information pertinente.

1.6 Des données manquantes

Qu'ils soit volumineux ou pas, un jeu de données peut être soumis au problème de données manquantes. Mais, qu'est-ce qu'une donnée manquante ? La réponse à cette question n'est pas forcément évidente. Se mettre d'accord sur ce point avec la personne à l'origine des données est déjà un grand pas vers l'élaboration d'une stratégie de gestion de ces données. S'agit-il d'une mesure dont on ignore la raison de l'absence ? S'agit-il d'une valeur à mesurer trop faible pour l'outil utilisé ? Le fait qu'une donnée soit manquante nous donne-t-il quand même des informations ? Autant de questions qui permettent de se situer dans le paysage des données manquantes en les qualifiant de *missing completely at random, MCAR*, *missing at random, MAR*, *missing not at random* et d'envisager les stratégies permettant de faire avec ou sans (imputation, interpolation, délétion partielle...). Ces quelques éléments caractérisent le fait que la gestion des données manquantes est un problème complexe qui est résumé dans cette citation attribuée à Gertrude Mary Cox, statisticienne américaine : *The best thing to do with missing values is not to have any.*

J'utilise parfois cet exemple pour inciter à résister à la tentation de remplacer les cases vides d'un tableau de données par zéro : considérons que l'on suit la masse d'un individu soumis à un régime

14. www.criteo.com/fr
15. lejournal.cnrs.fr/articles/avila-villani-deux-hommes-qui-comptent

alimentaire spécifique. Cet individu passe une visite médicale hebdomadaire durant laquelle il est pesé. Supposons maintenant pour une raison quelconque, cet individu ne se présente pas à la visite médicale. On ne peut donc pas peser cet individu et on a ainsi une donnée manquante dans la chronologie de ses mesures. On peut soit faire sans cette mesure (certaines méthodes statistiques ne s'offusquent pas de données manquantes), soit effectuer une imputation pour cette donnée manquante. Si on opte pour une stratégie d'imputation, le remplacement de la valeur inconnue par zéro n'est, bien évidemment dans ce contexte, absolument pas pertinente. Une méthode d'interpolation entre les mesures effectuées aux deux dates les plus proches de la visite manquée semblerait raisonnable.

Dans d'autre cas de figure, la situation n'est pas aussi limpide. Je m'en suis rendu compte en travaillant avec des chercheurs manipulant des données protéomiques acquises par spectrométrie de masse. Lorsqu'une donnée est manquante (en tout cas, caractérisée en tant que telle par le logiciel qui génère les données à la sortie du spectromètre), on n'en connaît pas toujours la raison. Le logiciel a-t-il considéré cette valeur comme peu fiable et donc a préféré la marquer comme manquante ? la protéine mesurée est-elle vraiment absente du fluide étudié ? ou sa concentration est-elle trop faible, inférieure au seuil de détection de la machine utilisée ? Selon la réponse à cette question, qu'il n'est pas toujours possible de donner, les stratégies d'imputation peuvent être radicalement différentes. Le sujet est toujours d'actualité et une revue récente peut-être trouvée dans [48].

Cette notion de données manquantes s'élargit à des cas de figures dans lequel on considère que des lignes d'un tableau de données sont manquantes. Dans un contexte d'intégration de données, cela peut correspondre à des individus pour lesquels on dispose d'un jeu de données (par exemple dosages sanguins) mais pas d'un second (par exemple données génomiques). Des méthodologies ont été récemment développées pour aborder ces questions [90, 47].

Pour extrapoler le sujet de cette section, je termine ce chapitre en évoquant l'absence totale de données.

1.7 Que faire sans données ?

À l'opposé de la section 1.5 qui aborde la gestion de données particulièrement volumineuses, et dans la lignée de la précédente, celle-ci s'intéresse à la question : que faire quand on n'a pas de données ? Ce qui représente en quelque sorte un cas extrême de données manquantes.

Une possibilité réside dans la modélisation. Par exemple, c'est sur ce sujet que G. Bobashev, chercheur au Center for Data Science de RTI International [16] est intervenu lors la conférence en statistique et santé déjà mentionnée [17] avec une présentation intitulée *Modeling of the data, with the data, and instead of the data* ; autrement dit, la modélisation comme une alternative à la donnée, dans un contexte bio-médical. Schématiquement, on peut en effet penser que si on dispose d'une connaissance suffisamment précise d'un phénomène pour le « mettre en équation », il n'est pas utile d'acquérir des données pour le caractériser. En revanche, on peut parfaitement envisager que la modélisation soit, *in fine*, un des objectifs de l'analyse de données ; c'est d'ailleurs parfois ce que nous écrivons dans des demandes de financements... Mais le chemin est extrêmement long avant d'y aboutir. Je n'en dis pas plus sur la modélisation comme alternative à la données car cela dépasse le cadre de ce mémoire.

Une autre solution lorsqu'on ne dispose pas de données, consiste à en fabriquer. Et je discute ici quelques éléments concernant la simulation de données. Le verbe *simuler* est défini ainsi dans le dictionnaire de l'Académie Française [18] :

> *SIMULER. v. tr. Feindre, imiter, faire paraître réelle une chose qui n'est pas. Simuler un combat. Simuler une attaque. Simuler une infirmité, la joie. Simuler une vente, une donation.*

Dans notre contexte, il est donc question de générer des données qui auront l'apparence de la réalité alors qu'elles ne sont pas réelles. Cela ouvre beaucoup de perspectives... Et finalement, un des moyens que l'on utilise pour donner l'apparence du réel à des données « artificielles » [19] consiste à utiliser des générateurs aléatoires. Même si le sujet est relativement bien balisé, le principe même des générateurs aléatoires peut être une préoccupation pour la reproductibilité des résultats de la recherche. Il peut être assez tentant de simuler des données tant que notre méthode n'obtient pas les meilleurs résultats. C'est pourquoi, j'ai tendance à rejeter toute démarche de simulation dès qu'il s'agit de montrer qu'une méthode est meilleure que les autres. En revanche, je soutiens sans problème une telle démarche lorsqu'il s'agit de montrer comment fonctionne une méthode. Et j'ai très largement recours à des données simulées lorsque j'interviens en formation.

16. www.rti.org/expert/georgiy-bobashev
17. www.cimi.univ-toulouse.fr/mib/en/conference-statistics-and-health
18. cette fois-ci dans sa 8ème version `academie.atilf.fr/8` vu que la 9 n'a pas encore atteint le mot simuler au moment où je la consulte en novembre 2018)
19. je me demande si la démarche de simulation aurait autant de succès si on parlait de mise en œuvre sur données artificielles plutôt que sur données simulées.

1.8 Conclusion

Dans ce chapitre, j'ai présenté quelques réflexions sur la notion de donnée. Il ne s'agit pas d'une monographie exhaustive sur le sujet mais d'éléments de discussion issus de mon expérience dans le contexte de la recherche en science des données. Le message essentiel réside dans la nécessaire prise en compte de multiples aspects liés à la donnée : de la raison de son recueil ou de son acquisition, aux (pré-)traitements auxquels elle doit être soumise pour devenir informative et source de connaissance. Le fait qu'elle constitue la matière première de la science des données ne signifie pas pour autant qu'elle est à l'origine de toute démarche de recherche. Le chapitre suivant aborde justement un cadre de travail qui place la donnée au cœur de tout un système et non pas au commencement d'une histoire.

Chapitre 2

Élaboration d'une méthodologie pour la recherche interdisciplinaire autour de données

La plupart de mes travaux de recherche s'effectuent en collaboration avec des chercheurs de disciplines différentes ; le point commun de ces travaux étant les besoins en analyse de données. Au fil de mes collaborations, j'ai développé une méthodologie que je déploie et adapte selon les contextes. C'est ce cadre que je détaille dans la première partie de ce chapitre. Ensuite, j'illustre à travers trois cas d'étude la mise en œuvre de ce cadre ainsi que son évolution au cours du temps. Dans la troisième partie, je donne quelques éléments d'appréciation de ce cadre de travail.

2.1 Cadre de travail

2.1.1 Pluridisciplinarité ou interdisciplinarité ?

Un premier point que je souhaite aborder dans ce chapitre concerne la dénomination des activités que je mène en collaboration avec des collègues de disciplines différentes. Dans [43] la pluridisciplinarité est vue comme une *juxtaposition de regards spécialisés* et cela est de même affirmé dans [24] qui parle de *superposition de points de vue éloignés les uns des autres, sans dégager de véritable unité, de lien, de liant entre les disciplines.* En revanche, d'après [43], l'interdisciplinarité *met en place un dialogue et des échanges entre les disciplines* et, d'après [24] *il résulte [de l'interdisciplinarité] un croisement fertile à la fois des démarches abordées et des résultats observés en vue de l'enrichissement des informations collectées, et par conséquent une compréhension plus complète, voire systémique, de l'objet étudié.*

Selon ces définitions, mes travaux se posent plus naturellement dans le champ de l'interdisciplinarité entre biologie, essentiellement, et statistique. Toujours selon [24] concernant l'interdisciplinarité, *Il s'agit, à partir d'une discipline considérée, de se demander et de voir ce que les disciplines connexes apportent de plus en termes de connaissance, de manière d'appréhender les choses.* C'est typiquement la façon dont naissent les collaborations dans lesquelles je m'implique. À partir d'une discipline (généralement la biologie), je contribue avec mes compétences en statistique à apporter de nouvelles connaissances au sujet étudié par le biais de l'analyse de données issues des bio-technologies. Une collaboration ne peut être efficace que si elle intègre des échanges entre les personnes impliquées et donc entre les disciplines concernées. La simple juxtaposition des compétences ne peut pas suffire.

2.1.2 Méthodologie d'analyse

Que ce soit dans un cadre interdisciplinaire ou pas, l'analyse statistique doit suivre une certaine démarche. Au gré de mes collaborations, j'ai défini une feuille de route que je présente régulièrement en séminaire ou dans des formations pour des étudiants et chercheurs en biologie. Cette feuille de route se compose de sept étapes :

1. énoncer clairement une question précise ;
2. prévoir les méthodes d'analyse des données ;
3. mettre en place un plan d'expérience ;
4. acquérir les données ;
5. analyser les données ;

6. interpréter les résultats ;
7. répondre à la question posée.

Dans cet enchaînement, j'attire l'attention sur le fait que les données sont au cœur de la démarche et non pas uniquement au début. En pratique, beaucoup de disciplines sollicitent des statisticiens au niveau de la cinquième étape, voire de la sixième. Dans ce dernier cas, il s'agit de collègues qui ont pu mener à bien l'analyse de données via un logiciel de statistique facile d'accès. C'est au moment d'interpréter les résultats que les questions commencent à se poser : *j'ai effectué une classification sur mes données, j'obtiens un arbre (dendrogramme) mais je ne sais pas quoi en faire et en plus quand je clique sur un autre bouton, j'obtiens un arbre différent.* Face à cela, je peux très bien jouer le jeu et accompagner mon collaborateur dans l'interprétation des résultats pour lui permettre de publier relativement rapidement. L'autre enjeu pour moi réside dans le fait de présenter cette feuille de route en sept points et d'inciter, autant que possible, les collègues à me contacter un peu plus tôt pour le prochain projet. Et je suis ravi de constater que cela fonctionne car je suis de plus en plus sollicité dès les prémices d'un projet, au moment où on s'interroge sur la question prioritaire à traiter dans le projet à venir.

Cette feuille de route que j'ai développée petit à petit est assez proche des modèles que l'on peut trouver dans la littérature. Le plus répandu semble être le modèle PPDAC [63] dont le sigle signifie : Problem - Plan - Data - Analysis - Conclusion ; sigle qui fonctionne aussi bien en français : Problème - Planification - Données - Analyse - Conclusion. Le détail de chacun de ces points est fourni dans [63]. J'en retiens une sythèse ici :

— Problème : comprendre ce qui doit être appris d'une expérience.
— Plan : l'objectif de cette étape est de concevoir un plan pour le recueil et l'analyse des données.
— Données : l'objectif de cette étape est d'exécuter le *Plan* et de s'assurer de la qualité des données en vue de l'analyse.
— Analyse : l'objectif de l'étape d'*Analyse* est d'utiliser les données collectées et les informations du *Plan* pour traiter les questions formulées dans l'étape *Problème*.
— Conclusion : l'objectif de l'étape de *Conclusion* est de reporter les résultats de l'analyse dans le langage utilisé lors de l'étape *Problème*.

Tout comme dans la feuille de route que j'utilise, les données ne sont pas au début de la démarche, mais au cœur de celle-ci. Et on retrouve finalement les mêmes éléments autour :

— *Problem* équivaut à l'énoncé d'une question précise ;
— *Plan* englobe les deux étapes relatives aux méthodes à utiliser et à la planification expérimentale ;
— *Analysis* et *Conclusion* recouvrent les trois étapes d'analyse, d'interprétation et de conclusion selon des recoupements sensiblement différents.

La définition du problème étant la base de la démarche, il convient d'y porter une grande attention. L'extrait de [63] concernant l'étape *Problem* se poursuit ainsi *Understanding what is to be learned from an investigation is so important that it is surprising that it is rarely, if ever, treated in any introduction to statistics.* C'est à ce point initial de la démarche que le document [4] est spécifiquement consacré. Ces éléments font écho à une phrase de J. Tukey que l'on retrouve dans [88] : *Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.*

J'insiste également avec mes interlocuteurs sur le fait que la démarche doit aboutir à une réponse au problème posé en des termes de la discipline d'application (biologie par exemple), ce qui est décrit comme objectif de l'étape de conclusion en tant que *report the results of the study in the language of the Problem.* En d'autres termes, la fin de l'histoire n'est pas une p-value inférieure à 0.05.

D'autres modèles ou processus analytiques existent. On trouve par exemple dans [84] des références à [68] et [23] proposant les processus d'analyse suivants :

— D'après [68] : *Frame the question, Understand your data, Choose a method, Calculate the statistic, Interpret the statistic, Test the significance of the statistic, Question the results.*
— D'après [23] : *Objective (Problem specification), Setting and subjects (Planning and data), Design and measures (Analysis), Results and conclusions (Outcome).*

Le second [23] est assez proche de ma feuille de route et du modèle PPDAC ; en revanche le premier [68] me semble trop mettre en avant la notion de significativité or c'est une tendance contre laquelle j'essaie de m'opposer pour ne pas inciter les collègues à se focaliser sur un test statistique et une p-value.

Pour clore cette partie, je reviens sur le modèle PPDAC et sur le fait qu'il est généralement présenté sur un schéma circulaire pour mettre en évidence le fait qu'au bout de la démarche, une fois que l'on a répondu au problème posé, il est fort probable qu'un nouveau problème se pose... Un autre élément réside dans le fait que, quelle que soit la façon dont on la formalise, la démarche d'analyse n'est pas linéaire, car on est généralement amené à remettre en question une étape précédente en fonction de l'étape en cours, ce que [63] décrit en ces termes : *Bouncing back and forth between stages is common in the development of the complete PPDAC structure.* C'est aussi pour cette raison que je n'adhère pas

du tout au terme de *pipeline* que l'on rencontre parfois et qui incite à penser que l'analyse se déroule linéairement sans possibilité d'exploration de voies parallèles ni de retour en arrière. J'y reviendrai dans le chapitre 3 consacré à l'intégration de données.

2.2 Cas d'étude

Il s'agit ici de présenter trois études révélatrices de ces interactions nécessaires pour la mise en œuvre de la méthodologie statistique décrite précédemment. La première concerne l'analyse statistique de données transcriptomiques acquises pour contribuer à la recherche sur le cancer pancréatique ; c'est la première étude dans le domaine de la biologie dans laquelle j'ai été impliqué. La seconde concerne également la recherche en biologie et elle a la particularité de s'appuyer sur des données cinétiques qui requièrent des traitements adaptés et permettent également de tester la méthode statistique évoquée précédemment dans un cadre moins conventionnel. La troisième étude est en fait double : elle concerne deux projets différents mais dans le même domaine (le traitement de la voix) et me permet d'illustrer le rôle que je peux assurer pour donner de l'ampleur à une collaboration naissante.

2.2.1 Exploration de données d'expression pour l'étude du cancer pancréatique

Les données étudiées ici ont été acquises afin de mieux comprendre les mécanismes biologiques impliqués dans le cancer du pancréas. J'ai pris part à cette étude en collaborant avec Henrik Laurell, chercheur en biologie à l'Institut des Maladies Métaboliques et Cardiovasculaires (I2MC [1]). Les résultats ont été publiés dans la revue *World Journal of Gastroenterology* [55]. Le résumé de cet article est reproduit ci-dessous.

AIM : To compare gene expression profiles of pancreatic adenocarcinoma tissue specimens, human pancreatic and colon adenocarcinoma and leukemia cell lines and normal pancreas samples in order to distinguish differentially expressed genes and to validate the differential expression of a subset of genes by quantitative real-time RT-PCR (RT-QPCR) in endoscopic ultrasound-guided fine needle aspiration (EUS-guided FNA) specimens.
METHODS : Commercially dedicated cancer cDNA macroarrays (Atlas Human Cancer 1.2) containing 1176 genes were used. Different statistical approaches (hierarchical clustering, principal component analysis (PCA) and SAM) were used to analyze the expression data. RT-QPCR and immunohistochemical studies were used for validation of results.
RESULTS : RT-QPCR validated the increased expression of LCN2 (lipocalin 2) and for the first time PLAT (tissue-type plasminogen activator or tPA) in malignant pancreas as compared with normal pancreas. Immunohistochemical analysis confirmed the increased expression of LCN2 protein localized in epithelial cells of ducts invaded by carcinoma. The analysis of PLAT and LCN2 transcripts in 12 samples obtained through EUS-guided FNA from patients with pancreatic adenocarcinoma showed significantly increased expression levels in comparison with those found in normal tissues, indicating that a sufficient amount of high quality RNA can be obtained with this technique.
CONCLUSION : Expression profiling is a useful method to identify biomarkers and potential target genes. Molecular analysis of EUS-guided FNA samples in pancreatic cancer appears as a valuable strategy for the diagnosis of pancreatic adenocarcinomas.
Keywords : Pancreas, Colon, Adenocarcinoma, Gene expression profiling, Endoscopic ultrasonography, Ultrasound, Fine needle aspiration

Résumé de l'article [55]

Ma contribution à ces travaux et à cet article se situent naturellement dans l'analyse statistique : de la justification des méthodes employées à l'interprétation des résultats obtenus en passant par la mise en œuvre.

Les données

Les données analysées dans cette étude correspondent à l'expression de 871 gènes mesurée sur 22 échantillons biologiques répartis ainsi :

- Lignées pancréatiques (7 échantillons) codées : ASPC1, Bx-PC3, Capan 1, Capan 2, Mia-PaCa2, NP 29, Panc1
- Lignées coliques (5 échantillons) codées : CaCo2, HCT116, HT29, SW480, SW620
- Lignée leucémique (1 échantillon) codéee : K562

1. www.i2mc.inserm.fr

— Pièces tumorales (6 échantillons) codées : PT1, PT2, PT3, PT4, PT5, PT6
— Pancréas normal (3 échantillons) codées : PancNorm1, PancNorm2, PancNorm3

Par rapport aux 1176 gènes annoncés dans le résumé, 305 ont été supprimés de la liste finale des gènes analysés pour des raisons diverses (qualité insuffisante de la mesure, valeur trop faible, données manquantes...).

Sans rentrer dans les détails des spécificités des différentes catégories d'échantillons biologiques, je signale qu'elles sont de deux types nettement différents : les lignées cellulaires issues de culture in vitro et les tissus prélevés chez des individus.

Même si je n'ai pris part à cette étude qu'après la génération de ces données, je n'ai pas eu de remarque particulière à formuler sur d'éventuels biais qui auraient pu être corrigés par une meilleure planification expérimentale.

Un aperçu global des données

L'acquisition de données d'expression étant relativement novatrice en 2004-2005 dans ce contexte, il était nécessaire de fournir un aperçu global des données. En effet, il n'était pas forcément courant pour un biologiste de se retrouver confronté à un tableau de plusieurs milliers de chiffres ; de l'ordre de 20000 dans ce cas précis. Ce n'est pas encore du *big data*, mais suffisamment important pour déstabiliser quelqu'un qui n'est pas familier avec l'analyse statistique et qui a l'habitude de gérer ces données avec un tableur. Des indicateurs statistiques usuels sont bien sûr toujours pertinents, mais calculer l'expression moyenne de chaque gène, par exemple, ne rend pas forcément mieux compte de l'état des données disponibles. Le recours à une méthode exploratoire multivariée s'est donc imposé et dans ce cadre, c'est une Analyse en Composantes Principales (ACP) que j'ai réalisée.

L'ACP mise en œuvre permet de donner un aperçu des relations entre gènes et échantillons. Parmi les différentes options possibles pour la représentation des résultats d'une ACP (représentation des individus et des variables), c'est vers le biplot que nous nous sommes orientés pour son côté synthétique. Une analyse élémentaire de ce biplot (Figure 2.1), révèle que le premier axe, qui représente 39% de la variance globale des données, oppose les lignées cellulaires (à gauche) et les échantillons de tissu (à droite). Le deuxième axe, quant à lui, met en évidence une opposition entre les échantillons de pancréas malins (tumeurs pancréatiques et lignées cellulaires pancréatiques) en bas et les autres échantillons en haut.

Afin d'améliorer l'interprétation des résultats, seuls les 28 individus-gènes contribuant le plus aux deux premières composantes principales sont représentés. Les interprétations biologiques approfondies de ces résultats sont dans l'article [55]. J'en extrais un exemple qui illustre le côté rassurant qu'a eu cette analyse sur la qualité des données et sur les conclusions pouvant être tirées de cette étude.

Interestingly, CTRL (chymotrypsin-like), was found to be specifically expressed in normal pancreas and absent in PT or PCL. CTRL is a poorly characterized serine protease with chymotrypsin- and elastase-2-like activities, which is expressed as a pro-protease in normal pancreatic tissue. The absence of CTRL expression in adenocarcinoma samples is in agreement with the absence of other serine protease digestive enzyme, such as chymotrypsin and trypsin, present in normal pancreatic tissue and known to be down-regulated in pancreatic cancer.

En lisant cela, on peut se dire que l'ACP a illustré des évidences ; les grandes lignes de l'interprétation révèle en effet des phénomènes déjà connus. Je suis d'accord avec cette affirmation et j'ai tendance à m'en féliciter. En effet, comment envisager de poursuivre une collaboration si au premier contact avec les données (ce que nous donne une ACP sans a priori), les phénomènes mis en évidence ne sont pas des éléments déjà connus. On peut raisonnablement se dire que le biologiste pourrait douter, d'une part de la méthode statistique qu'il ne connaît pas forcément, d'autre part de ses données si le statisticien le persuade que, si la méthode montre certains éléments et pas d'autres, c'est parce qu'ils sont les plus importants dans les données. Et donc, enfoncer des portes ouvertes en débutant par une ACP est une bonne base sur laquelle on peut construire une collaboration.

La suite de l'analyse de ces données met en évidence le rôle particulier des gènes KRT7, LCN2 et PLAT qui ont été validés ensuite par d'autres expériences biologiques. Notons que ces gènes font partie des 28 gènes représentés sur le biplot et qu'ils se situent dans le cadran inférieur-droit dans la direction pointée par les échantillons de tumeur pancréatique.

J'insiste sur le fait que ce qui est raconté ici est le résultat d'interactions régulières avec Henrik Laurell. Afin de le convaincre de l'intérêt de la méthode, je l'ai accompagné dans l'interprétation des résultats de l'ACP (figure 2.1 et d'autres non présentées ici) : comment interpréter les pourcentages de variance expliquée, pourquoi représenter des flèches, comment interpréter les proximités et les éloignements... C'est au cours de ces discussions que l'histoire biologique que les données avaient à raconter a vu le jour petit à petit.

Quelques années après la publication de cet article, je suis toujours en contact avec H. Laurell. Il a grandement contribué à ma montée en compétences en biologie me permettant ainsi de mieux appréhender

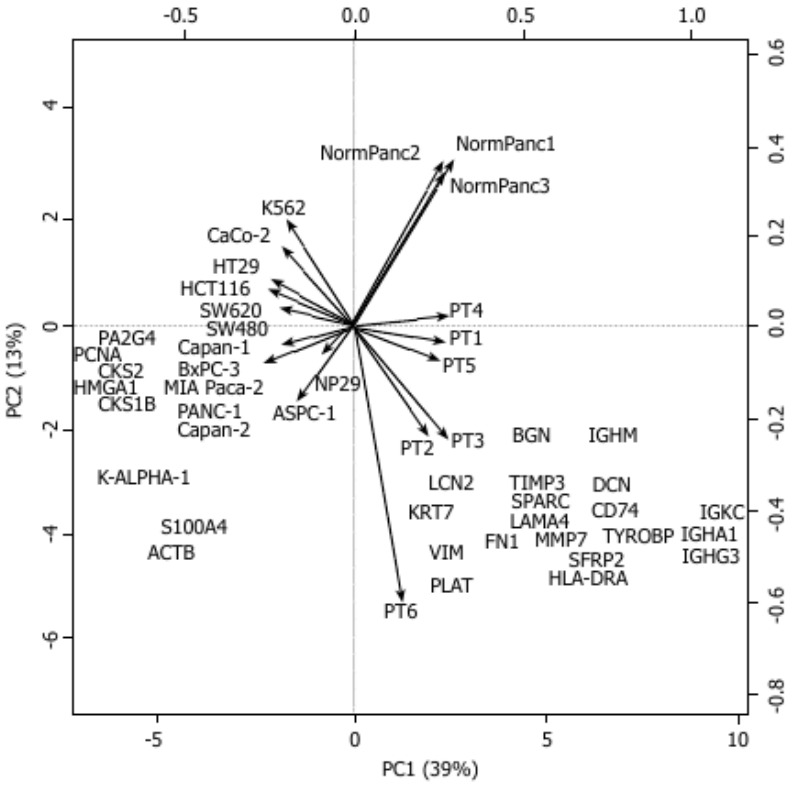

Figure 3 Principal component analysis (PCA) of gene expression data. Biplot resulting from a PCA of the line and column centered data containing 871 genes (individuals) in lines and 22 samples in columns (variables). The 28 genes contributing the most to the total variability are shown. The two principal components (PC1 and PC2) contribute to more than 50% (39% and 13%) of the total variability and resolve four biological sample categories: PC1 on the horizontal x-axis distinguish between cell lines (left) and tissue samples (right) whereas PC2 on the vertical y-axis distinguish between malignant pancreas samples (bottom) and other sample categories (top).

FIGURE 2.1 – Représentation biplot des résultats d'une ACP accompagné de la légende issue de l'article [55].

de nouvelles collaborations. Il a également accepté de faire partie du comité de pilotage de la plateforme de biostatistique GenoToul que j'ai mis en place en 2011.

2.2.2 Classification de cinétiques d'expression de gènes

Le projet évoqué ici s'intègre dans une collaboration régulière avec Pascal Martin de l'unité de Toxicologie Alimentaire (ToxAlim [2]) de l'Inra de Toulouse. Dans le chapitre 3, il sera question d'une autre étude qui a orienté de nombreux développements méthodologiques autour de l'intégration de données (voir 3.3.4). Ces liens ont également été consolidés dans le cadre du projet Plast-Impact (Impacts métabolique et endocrinien de deux contaminants de la chaîne alimentaire issus de la plasturgie). Les travaux de ce consortium, financé par l'ANR entre 2006 et 2009 et impliquant ToxAlim et l'IMT ainsi que d'autres partenaires de l'Inserm, font partie des études qui ont contribué à la mise en évidence de la dangerosité des phtalates. J'ai contribué à la rédaction d'un article [42] dans lequel nous associons une analyse multi-résolution par ondelettes à une méthode discriminante pour exploiter au mieux des données métabolomiques obtenues par spectrométrie de masse.

P. Martin fait également partie, comme H. Laurell, des biologistes avec qui j'entretiens des relations privilégiées. Les problèmes apportés par P. Martin ont toujours un intérêt statistique indéniable et elles ont donné lieu à plusieurs collaborations concrétisées par des publications [28, 41, 42, 40, 66]. P. Martin et moi-même avons également, à plusieurs reprises, fonctionné en binôme pour des actions de formation durant lesquelles notre complémentarité méthodologique a toujours été appréciée.

Les données

Les travaux présentés dans cette partie illustrent un cas d'analyse de données cinétiques. Ce type de données nécessite un traitement adapté que je souhaite décrire ici. Ma motivation première à le détailler vient du fait qu'à mon sens il s'agit de la combinaison astucieuse de méthodes existantes qui reflète bien mes compétences.

Les données dont il est question ici ont été acquises durant une expérience de jeûne chez la souris. Il s'agit des mesures de l'expression de 120 gènes acquises à 11 temps différents allant de 0 à 72 heures. Il ne s'agit pas à proprement parler de données longitudinales car ce ne sont pas les mêmes individus qui sont suivis au cours du temps ; chaque point de mesure nécessitant le sacrifice de la souris. Le terme de *time-course study* est parfois utilisé pour évoquer ces études pour lesquelles des individus différents sont observés à chaque temps.

La question biologique posée revient à proposer une méthode permettant de regrouper les gènes selon leur profil d'expression au cours du temps. Il s'agit donc de mettre en œuvre une méthode de classification de courbes.

Microarray data acquired during time-course experiments allow the temporal variations in gene expression to be monitored. An original postprandial fasting experiment was conducted in the mouse and the expression of 200 genes was monitored with a dedicated macroarray at 11 time points between 0 and 72 hours of fasting. The aim of this study was to provide a relevant clustering of gene expression temporal profiles. This was achieved by focusing on the shapes of the curves rather than on the absolute level of expression. Actually, we combined spline smoothing and first derivative computation with hierarchical and partitioning clustering. A heuristic approach was proposed to tune the spline smoothing parameter using both statistical and biological considerations. Clusters are illustrated a posteriori through principal component analysis and heatmap visualization. Most results were found to be in agreement with the literature on the effects of fasting on the mouse liver and provide promising directions for future biological investigations.

Résumé de l'article [28]

Les données sont présentées sur la figure 2.2 issue de l'article [28]. Chaque point correspond à l'expression d'un gène pour une souris. Une couleur correspond à un gène ; une ligne joint les valeurs moyennes à un temps donné pour un gène donné. Selon les temps, on dispose de 2 à 4 points de mesures pour un même gène.

Stratégie de classification

La démarche qui a été mise en œuvre afin de répondre à la question posée est la suivante :

1. modélisation de chaque profil d'expression par des splines de lissage ;

2. www6.toulouse.inra.fr/toxalim

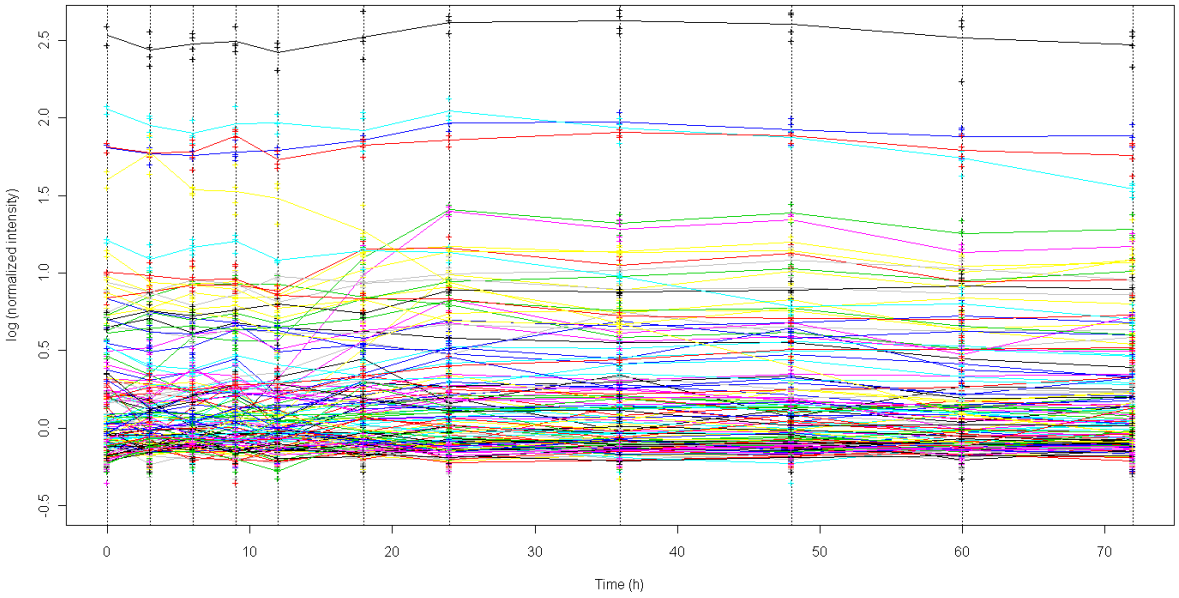

FIGURE 2.2 – Représentation des données relatives à l'expression de 120 gènes (convertie en log) durant une période de jeûne chez la souris [28]. Les lignes verticales indiquent les temps de mesure. Chaque point est la mesure d'un point pour une souris. Les lignes colorées joignent la valeur moyenne d'un gène à un temps donné.

2. calcul des dérivées premières de chaque profil ;
3. combinaison de classification ascendante hiérarchique et d'algorithme k-means pour regrouper les gènes sur la base des dérivés de leur profil ;
4. représentation graphique des résultats pour faciliter l'interprétation des groupes.

Comme je l'ai annoncé précédemment, il s'agit de l'enchaînement de méthodes relativement élémentaires, mais, encore une fois, la difficulté et l'intérêt de la démarche ne réside pas simplement dans la mise en œuvre des méthodes mais dans toute la démarche qui l'a précédée.

Aparté sur le lissage

Modéliser un profil par un lissage spline revient à trouver la fonction f qui minimise le critère suivant :

$$\sum [y_i^j - f(t_j)]^2 + \lambda \int [f''(u)]^2 du \tag{2.1}$$

Cette formule synthétise remarquablement ce que Henri Poincaré écrivait dans *La science et l'hypothèse* (1902).

> *Je veux déterminer une loi expérimentale ; cette loi, quand je la connaîtrai, pourra être représentée par une courbe ; je fais un certain nombre d'observations isolées ; chacune d'elles sera représentée par un point. Quand j'ai obtenu ces différents points, je fais passer une courbe entre ces points en m'efforçant de m'en écarter le moins possible et, cependant, de conserver à ma courbe une forme régulière, sans points anguleux, sans inflexions trop accentuées, sans variation brusque du rayon de courbure. Cette courbe me représentera la loi probable, et j'admets, non seulement qu'elle me fait connaître les valeurs de la fonction intermédiaires entre celles qui ont été observés, mais encore qu'elle me fait connaître les valeurs observées elles-mêmes plus exactement que l'observation directe (c'est pour cela que je la fais passer près de mes points et non pas par ces points eux-mêmes). [...] Les effets ce sont les mesures que j'ai enregistrées ; ils dépendent de la combinaison de deux causes : la loi véritable du phénomène et les erreurs d'observation.*

Extrait de *La science et l'hypothèse*, Henri Poincaré, 1902.

J'apprécie beaucoup cet extrait car il permet aussi de discuter de l'utilisation des formules mathématiques avec des personnes qui peuvent ressentir un blocage à la moindre apparition d'un début de formule. Ainsi, grâce à cette citation, on peut clairement faire le parallèle entre :

— $\sum[y_i^j - f(t_j)]^2$ et *... je fais passer une courbe entre ces points en m'efforçant de m'en écarter le moins possible...*, d'une part et

— $\lambda \int [f''(u)]^2 du$ et *de conserver à ma courbe une forme régulière, sans points anguleux, sans inflexions trop accentuées, sans variation brusque du rayon de courbure* d'autre part.

et accorder à la formule le mérite de la concision.

Un des points essentiels de cette approche consiste à définir au mieux un paramètre de lissage. Après quelques premières tentatives, il a été convenu avec P. Martin d'utiliser un paramètre de lissage commun à tous les gènes. Ensuite, afin de guider au mieux ce choix du paramètre de lissage λ dans la formule 2.1, nous nous sommes non seulement appuyés sur des indices statistiques (l'ACP sur les données lissées en est un) mais aussi sur l'interprétabilité biologique des résultats. En effet, il est indispensable que les résultats parlent au biologiste. Pour cela, il n'était pas question d'utiliser un paramètre de lissage trop faible ne pénalisant pas assez le manque de régularité de la fonction car cela aurait eu pour effet de proposer des profils trop irréguliers, ce qui n'est pas biologiquement pertinent. L'étude porte sur l'expression de gènes et il n'est pas cohérent d'observer des variations trop brusques des profils temporels. Si variations il y a, elles sont dues aux imprécisions de mesure. Quant à des variations de plus grande amplitude, elles pourraient être dues au rythme circadien des souris étudiées ; ce point est un fait connu et il n'était pas intéressant de le caractériser dans le cadre de cette étude.

Un autre extrait de la *La science et l'hypothèse*, appuie cette démarche.

> *[...] le choix ne peut être guidé que par des considérations de simplicité. Prenons le cas le plus banal, celui de l'interpolation. Nous faisons passer un trait continu, aussi régulier que possible, entre les points donnés par l'observation. Pourquoi évitons-nous les points anguleux, les inflexions trop brusques ? Pourquoi ne faisons-nous pas décrire à notre courbe les zigzags les plus capricieux ? C'est parce que nous savons d'avance, ou que nous croyons savoir que la loi à exprimer ne peut pas être si compliquée que cela.*

Extrait de *La science et l'hypothèse*, Henri Poincaré, 1902.

Cette dernière phrase illustre de façon très explicite le fait que les méthodes statistiques que nous mettons en œuvre doivent s'appuyer sur les connaissances a priori des phénomènes qu'elles sont destinées à analyser. Dans notre cas, le choix du paramètre de lissage et des courbes qui en découlent doit correspondre au phénomène biologique étudié.

Pour accompagner ce choix, voici un exemple de planche (Figure 2.3) que j'ai produite pour que nous puissions discuter du choix du paramètre de lissage.

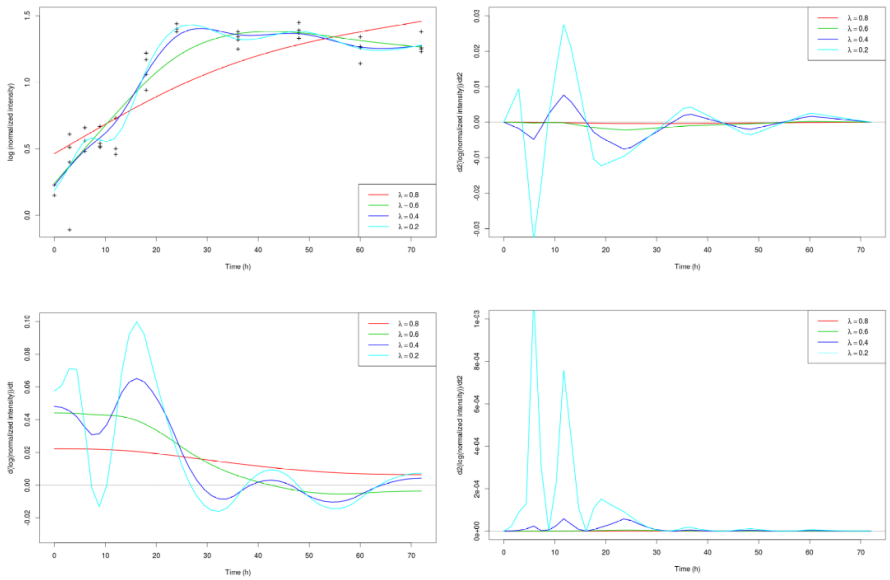

FIGURE 2.3 – Illustration de l'influence du paramètre de lissage pour un gène donné sur : les courbes lissées (en haut à gauche), les dérivées premières des courbes lissées (en bas à gauche), sur les dérivées secondes des courbes lissées (en haut à droite), sur le carré des dérivées secondes des courbes lissées (en bas à droite).

Ces graphiques permettent également de mieux faire comprendre le rôle de curseur que joue le paramètre de lissage dans la recherche d'un compromis entre la proximité avec les données observées et la régularité de la courbe lissée.

Représentation des résultats

Comme je l'ai déjà évoqué, la plus grande partie de mon travail évoqué dans ce chapitre concerne ce qui se passe autour de la mise en œuvre de méthodes statistiques. Dans cette partie, j'évoque ce qui se passe une fois les résultats de classification obtenus. L'objectif de la représentation des résultats consiste à les rendre facilement interprétables. Si cet objectif relève d'une évidence, la démarche pour l'atteindre est loin de l'être et je présente par la suite trois représentations graphiques différentes.

Les résultats obtenus pour la classification des profils d'expression sont présentés dans les figures 2.4 et 2.5. Dans la première, les profils sont représentés par les courbes lissées séparément pour chaque groupe. On y distingue assez nettement les quatre profils :

- **km1** : gènes présentant une croissance d'expression jusqu'à environ 30 heures puis une stabilisation voire une décroissance ;
- **km2** : gènes présentant un profil stationnaire ;
- **km3** : gènes dont l'expression diminue au cours du temps ; à noter un gène très spécifique qui connaît une forte décroissance ;
- **km4** : trois gènes dont l'expression connaît une forte croissance durant la première moitié de l'étude.

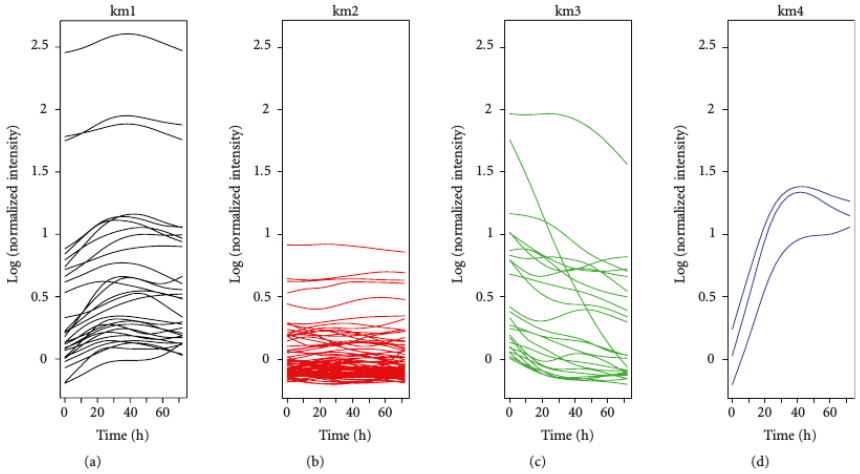

FIGURE 7: Representation of the smooth curves distributed in 4 clusters determined after hierarchical and *k*-means classification.

FIGURE 2.4 – Représentation des 4 groupes issus de la classification des profils temporels d'expression. Extrait de l'article [28].

La représentation graphique proposée sur la figure 2.5 a été ajoutée à l'article afin de s'adresser au mieux à la communauté biologiste. Ce type de représentation [32] est devenu très populaire au début des années 2000 dans le cadre de l'analyse de données de microarrays utilisant une technologie en double-couleur (fluorophores rouge et vert) et il est apparu opportun d'utiliser cet outil dans nos travaux.

Nous en avons profité pour aborder un souci concernant la représentation d'images colorées en présentant deux images. La différence entre les deux réside dans la non prise en compte de 4 gènes spécifiques dans l'image de droite. Il s'agit des 3 gènes à forte croissance du groupe km4 (3 lignes horizontales rouges en haut de l'image) et du gène à forte décroissance du groupe km3 (1 ligne horizontale verte vers le bas de l'image). Dans l'image de gauche, ces gènes spécifiques ont tendance à saturer l'espace des couleurs et à ne laisser que des teintes sombres à l'ensemble des autres gènes. En les mettant à l'écart, la représentation de droite redonne quelques couleurs notamment aux profils à faible croissance (zone rouge en haut à gauche) qui apparaissent très sombres sur la carte de gauche.

Ce point me permet d'aborder ici une question qui m'est fréquemment posée : que dois-je faire des individus atypiques ? Dans ce cas, et après avoir écarté le fait qu'une donnée atypique n'est pas due à une erreur de mesure, je suggère de réaliser la même analyse ou représentation graphique avec et sans ces individus, comme cela vient d'être illustré sur la figure 2.5. Les premiers résultats (avec tous les individus) mettront certainement en évidence des résultats évidents et essentiellement liés à un ou des individus atypiques ; les seconds (sans les individus atypiques) permettront de se rendre compte si une forêt ne se cache pas derrière l'arbre.

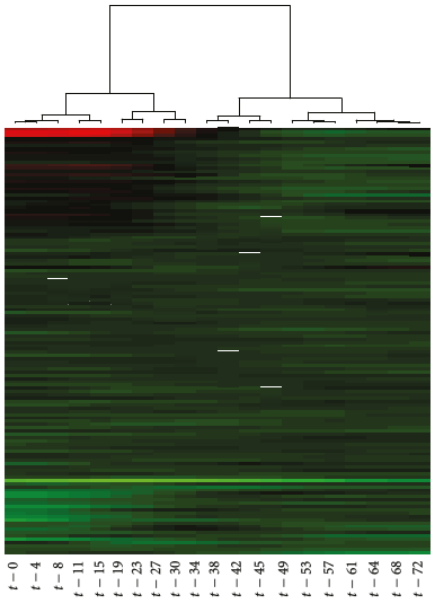

FIGURE 9: Heatmap of smoothed gene expression profiles for the whole dataset. Genes are ordered according to their cluster determined by the *k*-means algorithm. Horizontal blue lines separate the 4 clusters. Values increase from green to red *via* black.

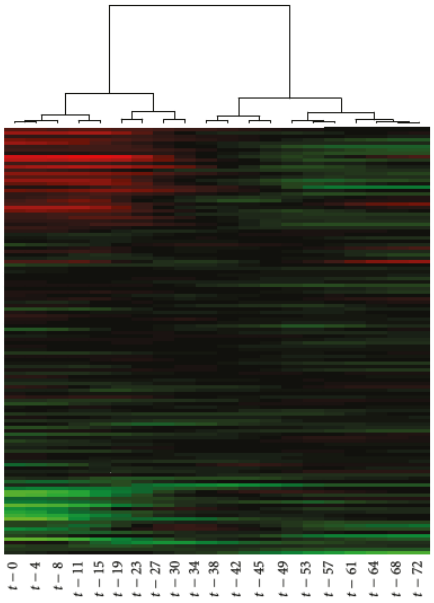

FIGURE 10: Heatmap of smoothed gene expression profiles without SCD1 and km4-genes. Graphical features are the same as Figure 9.

FIGURE 2.5 – Représentation des profils d'expression temporels sous forme de heatmap. Extrait de l'article [28] repris dans [29].

Aparté sur la visualisation

La visualisation des résultats d'une méthode statistique est toujours un élément clé. Le point relevé ci-dessus n'est qu'un infime point de vue sur ce sujet auquel je me suis davantage intéressé par la suite. Je me suis ainsi régulièrement servi de l'article [99] qui propose de nous « échapper » [3] du *RGBland* (RGB pour Red-Green-Blue) pour permettre notamment une meilleure lecture aux personnes atteintes de daltonisme. Je suis également devenu un utilisateur régulier du package `RColorBrewer` [69] de R qui propose des palettes spécifiques selon les besoins :

- *séquentielles* pour la représentation de données quantitatives ordonnées avec un minimum et un maximum.
- *divergentes* pour la représentation de données quantitatives présentant un intérêt spécifique pour des valeurs autour d'une référence. La corrélation est un indicateur qui se prête bien à des représentations par palette divergente : une couleur « neutre » au milieu du nuancier pour les valeurs proches de zéro et deux couleurs différentes selon que l'on s'écarte de zéro par valeurs positives ou négatives.
- *qualitatives* pour la représentation de variables catégorielles sur des diagrammes en bâtons par exemple.

De plus, le package `RColorBrewer` propose une fonction d'affichage des palettes disponibles assortie d'une option `colorblindFriendly` qui permet de n'afficher que des palettes dont les différentes teintes sont distinguables par les personnes souffrant de daltonisme. Certaines revues scientifiques exigent parfois de tenir compte de cet aspect dans les figures des articles.

Plus généralement, la visualisation de données est une démarche à laquelle j'attache de plus en plus d'importance que ce soit dans mes travaux de recherche ou dans les formations que j'assure. Dans le cadre des formations, j'attire l'attention sur le fait qu'une représentation n'est pas objective et qu'elle supporte un point de vue subjectif pris par la personne qui a produit le graphique. En tant que lecteur, il convient d'être particulièrement vigilant afin de ne pas se laisser emporter par un effet visuel trop marqué qui masquerait une information plus importante. C'est un commentaire dans le même esprit que j'ai retrouvé dans une chronique de Nicolas Bourriaud dans Beaux-Arts Magazine, numéro de juin 2018, *Dans ce monde dominé par le « visuel », on ne regarde pas on visionne* et plus loin citant Serge Daney, critique de cinéma français (1944-1992), *le visuel pourrait aini se définir comme « ce qui nous dispense d'aller voir »*. Et donc, c'est avec cette idée en tête qu'il convient de considérer une représentation graphique ; sans nous dispenser de « regarder » et « d'aller voir ».

3. d'après le titre de l'article [99] *Escaping RGBland : Selecting colors for statistical graphics.*

Un projet interdisciplinaire menant à une méthodologie d'analyse

Cette étude illustre à mon sens parfaitement les différents points mentionnés dans la première partie de ce chapitre. D'abord, il s'agit clairement d'un travail interdisciplinaire plutôt que pluridisciplinaire de par les échanges indispensables entre les deux disciplines qui ont conduit au choix d'une technique de lissage, puis ce choix étant fait, à la manière de déterminer un paramètre de lissage pertinent d'un point de vue biologique. Ensuite, elle illustre également les aller-retours indispensables dans la méthodologie de travail. La stratégie de classification qui a été mise en œuvre n'a pas été déterminée spontanément à l'issue de la première réunion de travail. Elle est le fruit d'un parcours durant lequel nous avons fait plusieurs tentatives d'analyse des données, évaluer les résultats et remis en cause notre démarche. De même, la question initialement posée ne l'était pas forcément en termes de classification de courbes, mais c'est finalement en ces termes qu'elle s'est imposée et que la stratégie d'analyse a été déployée. Je n'avais pas encore clairement en tête les éléments d'une méthodologie d'analyse mais cette étude est certainement l'une de celles qui m'ont conduit à formaliser les différentes étapes de la démarche présentée en début de chapitre.

2.2.3 Traitement de la parole

Les projets présentés dans cette partie illustre parfaitement mon rôle dans la recherche au sein de l'IMT. En effet, les premiers contacts dans le domaine du traitement de la parole ont été pris dans le cadre de la cellule CampuStat (voir partie 4.3.4) pour une demande de conseil de collègues de l'Irit concernant une analyse statistique. Ils se sont ensuite concrétisés par l'article [35], consacré aux troubles vocaux chez les entraîneurs sportifs, dont je discute ici en premier lieu. Par le biais du réseau ainsi formé, ils ont permis par la suite de consolider une collaboration dans le cadre d'un projet financé par l'Agence Nationale de la Recherche, sur l'amélioration du diagnostic différentiel entre la maladie de Parkinson et l'atrophie multisystématisée par analyse numérique de la parole ; j'en discute dans un second temps.

Mieux comprendre l'apparition de troubles vocaux chez les entraîneurs sportifs

Ce projet a été mené en collaboration avec Lionel Fontan, alors chercheur à l'Irit spécialisé en traitement de la parole. En 2015, il encadrait les travaux de deux étudiantes en orthophonie sur l'apparition de troubles vocaux chez les entraîneurs sportifs. La question centrale de l'étude vise à mettre en évidence les facteurs de risque auxquels sont soumis les entraîneurs sportifs concernant l'utilisation de leur voix. On peut en effet imaginer qu'une personne animant des séances sportives à longueur de journée (en musique, face à un grand groupe, dans une environnement bruyant comme une piscine...) puisse rencontrer des soucis avec sa voix. Ce genre de souci peut être apprécié par un Indice d'Handicap Vocal (*Voice Handicap Index*, VHI).

L'intérêt des *Sport and Fitness Instructors* (SFI) pour ce travail se voit au travers des commentaires libres qu'ils ont pu formuler en remplissant le questionnaire qui leur a été soumis. j'ai par exemple relevé celui-ci qui est particulièrement marquant : *Je suis heureuse que ce test existe, comme quoi des gens se préoccupent de la voix des coachs. Si j'avais fait les tests il y a un an, mes réponses auraient été bien différentes (difficultés à parler du lundi soir au samedi, puis au dimanche). Une collègue à même dû se faire opérer des cordes vocales à cause de ce métier. Pourtant personne ne parle de ce problème ! Merci.*

Le résumé, repris ci-dessous, illustre l'intérêt de l'étude qui se conclut par des propositions de mesures de prévention à destination des entraîneurs sportifs. L'analyse statistique que j'ai supervisée a clairement contribué à ces propositions. À noter que si les SFI sont une population à risque pour les troubles vocaux, les enseignants en sont une autre...

Objectives. Sports and fitness instructors (SFIs) are known for being a high-risk population for voice difficulties (VD). However, past studies have encountered various methodological difficulties in determining prevalence and risk factors for VD in SFIs, such as limited population, gender and selection biases, or poor statistical power, because VD were studied as a binary variable. The present research work addresses these issues and aims at studying the prevalence of vocal problems and risk factors in French SFIs, a population in which no such study was conducted yet. Another objective is to survey the French SFIs' habits and expectations regarding vocal prevention and care.
Study design. This is a cross-sectional study.
Methods. Three hundred and twenty SFIs answered a questionnaire, whether in an online (n = 267) or a paper (n = 53) version. The questionnaire consisted of 31 items addressing self-reported vocal difficulties, supposed risk factors, and personal health-care history, followed by the Voice Handicap Index assessment.
Results. Prevalence of self-reported vocal difficulties is 55%. The Voice Handicap Index is significantly associated with gender, age, and variables related to work environment (noise and music) and habits (shouting, frequency of classes), as well as with daily sleeping time. Results also indicate that a minority of the SFIs (37%) received information on vocal difficulties, whereas a majority (80%) declares being interested in participating in prevention programs.
Conclusions. This work confirms that SFIs are a high-risk population for VD, underlines the need for specific information programs in France, and provides relevant data for driving such preventive actions.
Key Words : sports and fitness instructors–prevalence of vocal problems–risk factors–Voice Handicap Index–professional voice use.

Résumé de l'article [35]

Les données Afin de contribuer à une meilleure connaissance des risques liés à la voix, les étudiantes orthophonistes ont soumis 320 entraîneurs sportifs à un questionnaire composé de 61 questions. Les 31 premières questions de cette étude abordent 6 grands thèmes :

- démographie : âge et sexe ;
- expérience et habitudes d'enseignement : années d'expérience, nombre d'heures de pratique par semaine, durée moyenne d'une séance... ;
- environnement d'enseignement : taille des salles, présence de la climatisation, type de bruit ambiant... ;
- habitudes liées à la voix : crier, parler et montrer les exercices en même temps, intensité de la musique, usage intensif de la voix pour une autre activité... ;
- style de vie et hygiène vocale : fumeur, sujet à reflux gastrique, consommation d'alcool, de jus de fruit, d'eau, durée moyenne du sommeil, asthme...
- historique des difficultés vocales et autre soin de santé : maux de gorge ou perte de voix sans maladie ORL, consultation d'un orthophoniste, traitement médical...

Les 30 questions suivantes sont des questions « certifiées » destinées au calcul du VHI de chaque répondant [49]. Cet article mentionne explicitement une considération statistique comme objectif de l'étude : *The aim of the present investigation was the development of a statistically robust Voice Handicap Index (VHI).* Cet objectif a été atteint par les auteurs en proposant d'abord un questionnaire de 85 questions dont la robustesse a été évaluée pour finalement aboutir à une version à 30 questions qui est reconnue aujourd'hui dans le domaine des troubles vocaux.

Ce sont donc les réponses à ces 320 questionnaires composés de 61 questions qu'il s'agit d'analyser ici.

Traitement statistique Une des particularités de ce jeu de données réside dans la nature des variables qui le composent. On y trouve des variables qualitatives (Sexe), des variables qualitatives ordinales (par exemple taille des salles codée en très petite, petite, moyenne, grande) et des variables quantitatives (Age, expérience, durée d'un cours...).

Ainsi, dans l'article [35], nous avons d'abord effectué des analyses statistiques ad hoc en fonction de la nature des variables considérées avec un rôle particulier joué par la variable VHI ; variable quantitative calculée à partir des 30 questions spécifiques. Il s'agit par exemple de calculs de corrélation entre variables quantitatives (VHI et âge par exemple), de tests statistiques pour évaluer la différence entre les VHI de différents sous-groupes d'individus (Homme/Femme, Crie/Ne crie pas...).

Au-delà de ces analyses élémentaires visant à illustrer certaines caractéristiques spécifiques du jeu de données, je suis satisfait d'avoir pu convaincre mes co-auteurs (et les relecteurs de l'article) d'intégrer une analyse en composantes principales afin de fournir un point de vue global sur les relations entre les différents thèmes abordés dans le questionnaire. Cette analyse n'a pu se faire qu'en considérant les variables comme quantitatives en les convertissant, quand cela était possible, en score.

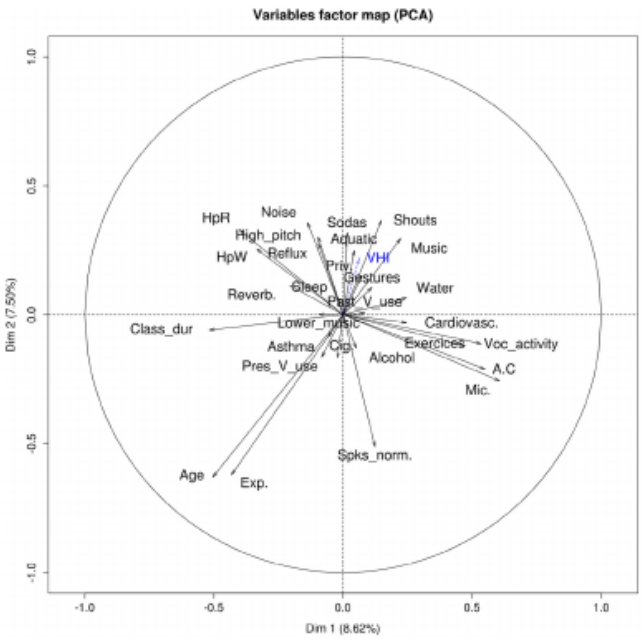

FIGURE 2. Variable representation from PCA performed on the whole data set. Active (resp. supplementary) variables are represented in black (resp. blue) arrows. The first two principal components (PCs) represent 16% of the variability. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Abbreviation	Extended description of variable	Abbreviation	Extended description of variable
A.C	Air-conditioning in teaching rooms	Mic.	Use of a microphone
Age	Age	Music	Music loudness
Alcohol	Alcohol consumption	Noise	Number of competing noise sources
Aquatic	Teaching in aquatic environments	Past_V_use	Past other vocal activities
Asthma	Asthma or allergies	Pres_V_use	Current other vocal activities
Cardiovasc.	Cardiovascular training classes	Priv.	Heavy vocal use in private situations
Cig.	Cigarette packs-years	Reflux	Gastric reflux
Class_dur	Classes duration	Reverb.	Teaching in reverberant rooms
Exercises	Performing exercises while teaching	Shouts	Shouting
Exp.	Work experience	Sleep	Sleep duration
Gestures	Performing gestures when teaching	Sodas	Sodas/juices consumption
High_pitch	Speaking with a higher pitch	Spks_norm.	Speaking normally
HpR	Max. teaching hours in a row	Voc_activity	Frequency of vocal activities
HpW	Hours per week	Water	Water consumption
Lower_music	Lowering the music when speaking		

FIGURE 2.6 – Représentation des variables issue de l'ACP des scores fournis par les SFI soumis à l'enquête [35].

Représentation des résultats La figure 2.6 illustre ainsi les relations entre l'ensemble des scores fournis par les SFI. Une version interactive 3D a également été produite et rendue disponible en ligne [4]. La variable VHI a été traitée ici comme variable supplémentaire. Cette figure met logiquement en évidence certaines banalités (et c'est tant mieux !) comme une association assez marquée entre un VHI élevé et le fait de crier (*Shouts*) et d'autres qui le sont peut-être moins comme le fait que les SFI les plus âgés ont des VHI plus faibles ; certainement un effet de l'expérience.

Si les tests statistiques sont fréquemment utilisés dans ce genre de contexte, ce n'est en revanche pas le cas, des analyses multivariées et je suis assez fier d'avoir su intégrer une telle analyse dans ce travail.

J'ai également été mis à contribution pour répondre aux relecteurs de notre article. Lors de la première soumission, nous avons en effet reçu le commentaire suivant de la part d'un relecteur :

Second, the statistical methods performed are confusing and interpretation of significance is improper.[...] Commenting on the level of statistical significance based on the p-value is incorrect. A p-value is either less than alpha value (rejecting null hypothesis) or it is not (retaining null hypothesis) ; a smaller p-value does not indicate that something has greater or stronger significance. Please delete adjectives (i.e. slightly, strongly, etc.) accordingly.

Ce sont des propos avec lesquels je suis en total désaccord car une p-value porte une information qui va au-delà de sa simple comparaison à un seuil. Nous avons exprimé ce désaccord très clairement à l'éditeur de la revue en question en nous appuyant sur une note [92] de l'American Statistical Association et cela l'a semble-t-il convaincu vu que l'article a été accepté assez rapidement après notre réponse.

4. www.irit.fr/recherches/SAMOVA/FONTAN/pca3D.html

Diagnostic différentiel entre la maladie de Parkinson et l'atrophie multisystématisée par analyse numérique de la parole

Contexte du projet Dans la lignée du projet précédent, j'évoque ici un projet en cours visant à améliorer le diagnostic de la maladie de Parkinson. Ce projet, financé par l'ANR pour la période 2016-2019, est intitulé *Diagnostic différentiel entre la maladie de Parkinson et l'atrophie multisystématisée par analyse numérique de la parole* et porte l'acronyme Voice4PD-MSA. J'assure la responsabilité scientifique de ce projet pour l'IMT porté par Khalid Daoudi, chercheur à l'Inria de Bordeaux, et impliquant des chercheurs de l'Irit ainsi que des centres hospitalo-universitaires de Toulouse et Bordeaux.

La démarche envisagée dans ce projet consiste d'abord à enregistrer la voix de patients atteints ou pas des maladies citées dans le titre du projet. À partir de ces enregistrements, les collègues de l'Irit et de l'Inria, spécialisés dans le traitement de la parole, extraient des caractéristiques susceptibles de mener à une partition des individus en fonction de leur pathologie. C'est sur ces caractéristiques que l'équipe de l'IMT travaille afin de déterminer celles qui sont susceptibles de poser le plus précocement possible le diagnostic des maladies en question.

Les détails sur la procédure et les méthodes envisagées sont donnés dans le descriptif de la sous-tâche 3.4 *Feature selection and classification* de la tâche 3 *Speech processing methodology* que j'ai contribué à rédiger et que je reprends ici.

> Exploring a novel approach to diagnose PD and MSA-P is the central motivation of our project. It is therefore preliminary from addressing the diagnosis problem in a clinical framework, where a deep understanding of the clinical context would be necessary to determine an experimental design in which the number of patients in each group would be one of the core parameters. For this reason, and consistently with the positioning of our project as a pilot study, we will favor exploratory approaches. We will follow two directions, each of them quantifying a specific aspect of the acquired data :
> As the differential diagnosis between PD and MSA-P is subtle, we will first use unsupervised clustering methods such as hierarchical clustering or k-means algorithm in order to identify potential clusters of patients, based on acoustic speech parameters. Such analyses do not take into account the status of the patients during the analysis. They instead consist in an exploratory analysis of the parameters that would highlight, for instance, the relative behavior of the parameters from linear and nonlinear speech analysis.
> Then, the most important part of the statistical analysis will focus supervised techniques to address the key questions of our project : differentiating HC from PD and PD from MSA-P. Unlike what will have been done using unsupervised techniques, the methods here will rely on the status of the patients. In this scope, three directions are usually followed : Univariate analyses based on statistical testing will be performed to detect acoustic speech parameters which highlight different features depending on the patients status. Multivariate analyses will also be used to complete univariate ones as speech parameters can be irrelevant individually but can be of major importance when used in combination. Linear Discriminant Analysis (LDA) and Partial Least Squares regression in its Discriminant variant (PLS-DA) are two methods highly used to achieve this purpose. The selection of the most important speech parameters will be automatically processed using sparse versions these methods [56]. The sparsity will be obtained in this context through a LASSO penalty term that enforces a trade-off between the data closeness and the model robustness following the pseudo-rule : the lower the parameters number, the more robust the model. Machine learning methods will then be used to complete the supervised framework. Unlike the linear methods of the previous step, machine learning methods such as Support Vector Machines (SVM) and Random Forest (RF) can handle nonlinear discrimination problems. In this context, the selection of relevant parameters will be made using importance scores (Gini or permutation score for RF) or ad-hoc elimination algorithms [30]. For multivariate and machine learning methods, the results will be strengthened by integrating a cross-validation procedure to calibrate optimally our various models.
> Once applied, these methods will make clear the potentiality of improving the discrimination between patients having two diagnoses problems (HC/PD and PD/MSA-P). Note finally, that the validation of the potential models will be performed on different data as those used in the previous steps.

Extrait de la présentation scientifique du projet Voice4PD.

La difficulté et l'intérêt pour moi de ce projet résident dans son interdisciplinarité et dans la nécessité de se confronter à un nouveau domaine avec un nouveau vocabulaire et de nouvelles problématiques.

La démarche entreprise dans le cadre du projet est assez similaire à celle décrite dans [3] : elle vise à *objectiver les troubles de la parole mis en évidence lors des autres examens [cliniques]*. Concernant la maladie de Parkinson, les troubles de la parole qu'elle peut causer ont été discutés par exemple dans [71].

Élaboration d'une stratégie pour l'analyse des données Dans ce projet, la difficulté ne réside pas dans la mise en œuvre des méthodes de discrimination mais dans la préparation des données. En effet, la matière première disponible dans le cadre de ce projet sont des enregistrements de la voix d'individus. Concrètement, les données sont disponibles au travers de plusieurs dizaines de fichiers au format WAV [5]. Lorsque ces fichiers de données auront été traités, analysés, pour en extraire des caractéristiques, le problème de discrimination sera alors traitable de façon quasi-routinière en mettant en œuvre et en comparant les méthodes usuelles (Analyse Factorielle Discriminante, méthode PLS-DA, réseaux de neurones, SVM, CART, forêts aléatoires) associées à des techniques de validation croisée pour s'assurer de la robustesse des résultats obtenus.

Des études préliminaires ont été menées dans le cadre du stage de Robin Vaysse étudiant en M1 de la formation Statistique et Informatique Décisionnelle d'avril à juillet 2018. Les fichiers mis à sa disposition étaient des enregistrements de la voix d'individus soit sains, soit atteints de la maladie de Parkinson (PD), soit atteints d'une atrophie multisystématisée (MSA-P). Les enregistrements en question concernent deux exercices spécifiques : le /a/ tenu et le /pa/-/ta/-/ka/. Dans le premier cas, il s'agit pour l'individu de maintenir le plus longtemps possible le son /a/. Dans le second, il s'agit de répéter les syllabes /pa/-/ta/-/ka/ le plus de fois possible en un temps donné. Cet enchaînement de syllabes est un exercice qui sollicite différentes zones de la bouche lorsqu'on les prononce : les lèvres pour /pa/, la langue et les dents pour /ta/, la gorge pour /ka/. Cela rend cet exercice révélateur de troubles potentiels chez des individus.

Ces premiers travaux ont permis d'esquisser la démarche d'analyse qui sera susceptible d'être mise en œuvre sur les enregistrements qui seront acquis dans le cadre du projet. Cette démarche est constituée des étapes suivantes :

1. pouvoir écouter les enregistrements et visualiser les signaux correspondants à l'aide d'un logiciel comme Wavesurfer [6] ;
2. extraire des paramètres des signaux sonores (énergie des signaux, tempogrammes, impulsions locales...) à partir des enregistrements avec des logiciels comme Essentia (*Open-source library and tools for audio and music analysis, description and synthesis* [7]) ou Opensmile (*The Munich Versatile and Fast Open-Source Audio Feature Extractor* [8]) ;
3. mener une analyse supervisée afin de discriminer au mieux les individus sains des individus malades ;
4. identifier les paramètres permettant de caractériser au mieux le statut des individus.

Ainsi, la démarche ne fait pas apparaître de verrous particuliers concernant la mise en œuvre de méthodes statistiques. La difficulté de ce projet et son intérêt résident dans le travail préparatoire à l'analyse statistique pour passer d'enregistrements de la voix d'individus à un tableau de données.

D'un projet pluridisciplinaire à un projet interdisciplinaire

Les deux projets que je viens d'évoquer sont clairement à des niveaux différents. Le premier, pour lequel j'ai contribué à l'analyse des réponses d'un questionnaire, peut être qualifié de pluridisciplinaire car il n'y a pas eu d'échanges sur le fond entre les disciplines. Ma contribution est intervenue après celle des orthophonistes sans que je n'ai eu d'apports, par exemple, dans la réalisation du questionnaire. En revanche, le projet Voice4PD se positionne clairement dans l'interdisciplinarité car plusieurs échanges ont déjà eu lieu notamment entre informaticiens spécialistes du traitement de la parole et statisticiens afin que les données extraites des enregistrements vocaux soit analysables. Selon moi, il est clair que l'IMT n'aurait pas été impliqué dans le projet Voice4PD si je n'avais pas accompagné précédemment le projet concernant la voix des entraîneurs sportifs.

2.3 Conclusion

Dans ce chapitre, j'ai mis en évidence une méthodologie que j'ai pu élaborer grâce à différentes collaborations scientifiques dans lesquelles je me suis impliqué. En pratique, cette méthode n'est pas à considérer de manière rigide. Chaque projet ayant son mode de fonctionnement propre en fonction des partenaires et des domaines scientifiques abordés, je peux choisir selon les cas de m'en écarter. Cette feuille de route sert de repère et permet de savoir de quoi on s'écarte. Elle permet également de s'impliquer dans

5. Waveform Audio File Format : Le Waveform Audio File Format (WAVE, ou WAV en rapport avec son extension de fichier), est un format conteneur destiné au stockage de l'audio numérique mis au point par Microsoft et IBM. Wikipedia, fr.wikipedia.org/wiki/Waveform_Audio_File_Format consulté le 20 avril 2018
6. sourceforge.net/projects/wavesurfer
7. essentia.upf.edu/documentation
8. www.audeering.com/technology/opensmile

un projet sans forcément passer par les premières étapes, mais dans ce cas, cela se fait en connaissance de cause, et en informant les partenaires.

Les études de cas que j'ai choisies de présenter ici, sont pour les deux premières à l'origine de mes réflexions sur le sujet. Dans les deux cas, les données étaient déjà générées lorsque les premiers contacts ont eu lieu. Dans le premier cas, l'étude sur le cancer pancréatique, il s'agissait surtout d'assurer une mise en valeur de données qui n'étaient pas encore courantes à l'époque. J'ai commencé dans ce cadre à apprendre à dialoguer avec des biologistes. Dans le second cas, la classification de cinétique d'expression de gènes, la stratégie finalement mise en œuvre n'a émergé qu'après des premières tentatives infructueuses (sans lissage, avec lissage sans dérivation). Ce projet a profondément contribué à attirer mon attention sur l'importance de la question posée et sur la nécessité de traduire une question biologique en question statistique.

Le double projet concernant le traitement de la parole me permet d'illustrer comment, entre deux projets avec des partenaires en partie identiques, je contribue à positionner le partenaire statisticien, plus en amont. Dans le premier cas, concernant les entraîneurs sportifs, les données étaient présentes quand j'ai été sollicité. Dans le second cas, j'ai été impliqué dès le début des discussions et j'ai ainsi contribué à la rédaction et au montage financier du projet Voice4PD déposé à l'ANR.

Pour compléter, j'ajoute quelques éléments sur la chronologie d'une collaboration. Entre le premier contact et, pour être concret, une publication en commun, il peut s'écouler plusieurs années ! C'est un élément dont il faut être conscient lorsque l'on envisage une vraie collaboration interdisciplinaire. Par exemple, les articles [55] et [28] ont été publiés respectivement 2 et 3 ans environ après la première discussion en commun. Avant cela, le fait de pouvoir travailler ensemble avec un objectif commun peut s'installer au bout de quelques mois, une fois que le dialogue est rendu possible par l'utilisation d'un vocabulaire commun, ou, a minima, compréhensible par l'autre.

C'est finalement dans ce type de collaborations que j'ai acquis ma spécificité scientifique : celle d'un expert assurant l'interface entre le domaine de la statistique et un autre domaine scientifique avec la volonté de s'attaquer à des questions concrètes.

Chapitre 3

Contribution à des développements pour l'intégration de données

Si je devais définir mon centre d'intérêt principal en recherche, je mentionnerais l'intégration de données. Ce terme étant relativement à la mode et utilisé dans des sens parfois sensiblement différents, je discute cette notion dans la première partie de ce chapitre afin de mieux positionner les deux problématiques que j'aborde ensuite. Il s'agira d'abord de travaux menées en Recherche d'Information (RI) pour lesquels la notion d'intégration est davantage liée à une démarche combinant différentes analyses statistiques. Puis j'évoquerai des travaux en biologie pour lesquels, j'ai contribué au développement et au déploiement de nouvelles méthodes d'analyse.

3.1 Qu'est-ce que l'intégration de données ?

Le terme d'intégration de données comme beaucoup d'autres en ce moment dans le domaine des données, révèle avant tout un caractère commercial. En effet, beaucoup de sociétés se positionnent sur ce créneau et proposent des produits et services de *data integration* ; ces produits et services pouvant aller de solutions techniques (fabrication de capteurs...) à des solutions logicielles souvent orientées bases de données. En règle générale, et quel que soit le domaine, l'intégration de données est reconnue comme un ensemble de moyens mis en œuvre pour rassembler des données de sources diverses et variées en vue d'en extraire une information pertinente.

De ce point de vue, l'illustration de la parabole des six aveugles et de l'éléphant que j'ai déjà utilisée en introduction et que je représente ici pour eviter aux lecteurs de revenir en arrière, s'accorde parfaitement avec cette notion d'intégration de données.

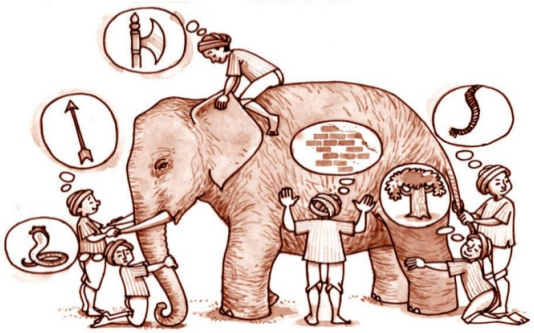

FIGURE 3.1 – Illustration de la parabole *Six aveugles et un éléphant*. Source inconnue.

En effet, chaque individu n'a à sa disposition qu'un point de vue restreint sur l'animal et serait dans l'incapacité de conclure à la présence d'un éléphant. Ce n'est qu'en se concertant, et en combinant leur différent niveau d'observation que ces six personnes peuvent, peut-être, en déduire qu'elles font face à un éléphant. Autrement dit :

$$\text{serpent} + \text{fleche} + \text{hache} + \text{mur} + \text{corde} + \text{tronc} = \text{éléphant}$$

Pour justifier le recours à une stratégie d'intégration, en plus de cette image de l'éléphant, j'utilise également le petit exemple ci-dessous (tableau 3.1 et figure 3.2). Supposons que l'on étudie l'effet d'une

variable qualitative à 2 modalités sur 2 variables quantitatives Vx et Vy. Les données fictives sont présentées dans le tableau 3.1.

TABLE 3.1 – Données fictives illustrant l'intérêt d'intégrer des données afin de répondre au mieux à un problème de discrimination des catégories de la variable `fact`.

	Vx	Vy	fact
1	2.00	2.30	A
2	3.00	2.10	B
3	4.50	3.50	A
4	5.00	3.10	B
5	5.50	3.30	B
6	6.00	4.30	A
7	7.00	4.00	B
8	8.00	5.10	A
9	8.50	4.80	B
10	9.00	5.00	B
11	10.00	6.00	A
12	11.00	6.50	A

En considérant chaque variable séparément, rien ne laisse supposer que la variable qualitative a une quelconque influence sur les variables quantitatives. Quelle que soit la méthode que l'on mettrait en œuvre, rien ne tendrait à conclure à un effet du facteur `fact` sur les variables `Vx` et `Vy` considerées individuellement (figure 3.2, gauche).

En revanche, le nuage de points représentant les données de `Vy` en fonction de `Vx` avec des symboles différents selon la modalité de la variable qualitative (figure 3.2, droite) révèle une distinction nette des deux groupes obervés. Ainsi, tirer parti des deux variables simultanément permet de fournir un point de vue différent sur les données et de considérer le problème de discrimination des deux groupes de façon plus optimiste.

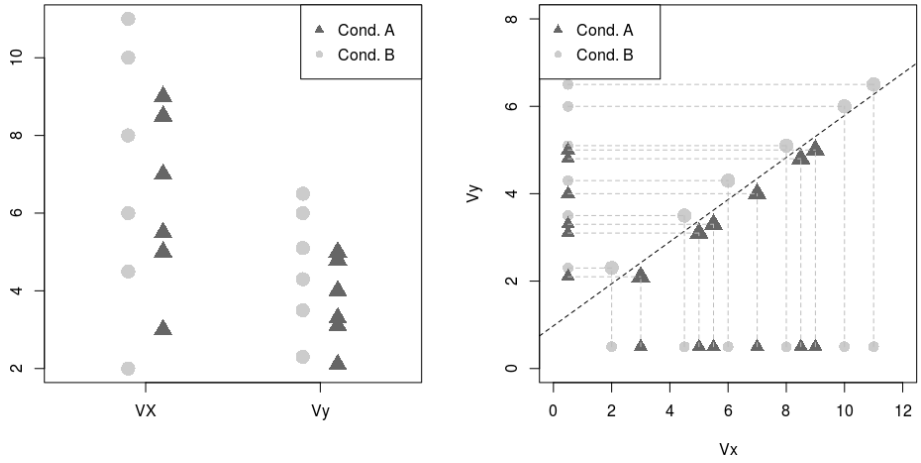

FIGURE 3.2 – Illustration de l'intérêt du couplage de méthodes statistiques et de la représentation graphique.

En élargissant la réflexion autour de ce petit exemple, on peut être convaincu du fait que l'analyse simultanée de plusieurs jeux de données est susceptible d'apporter des informations que l'analyse séparée de chaque jeu de données ne permet pas d'avoir.

Comme il est mentionné dans cet extrait tiré de [81], qui date de 2012 mais ce constat reste toujours d'actualité – *There is no universal approach to data integration, and many techniques are still evolving* – il n'y a pas d'approche universelle d'intégration de données. Pour illustrer ce propos, j'ai choisi de le décliner ici dans deux contextes différents. Le premier se positionne en recherche d'information ; dans ce cadre je décris comment un problème relativement simple à exprimer *Quel système utiliser pour répondre au mieux à une requête donnée ?* conduit à une démarche d'intégration de données. Dans la seconde partie, je présente mes contributions au développement de méthodes pour l'intégration de données biologiques acquises pour les mêmes échantillons biologiques. J'aborde en premier lieu les aspects méthodologiques de méthodes d'intégration développées pour ces données biologiques. Ensuite, je consacre une partie au package R `mixOmics` développé pour mettre ces méthodes à la disposition de la communauté scientifique.

Enfin, je présente plusieurs exemples d'intégration de données biologiques sur lesquels j'ai travaillé.

3.2 Une démarche intégrative en Recherche d'Information

Dans cette partie, après avoir présenté la problématique générale de l'intégration dans le contexte de la RI, je propose une vue schématique des données susceptibles d'être recueillies pour traiter cette problématique. Ensuite, je présente les travaux auxquels j'ai contribué dans ce domaine.

3.2.1 Problématique

De façon schématique, le problème central de la RI peut se résumer en une question relativement simple : *Quel système utiliser pour répondre au mieux à une requête donnée ?* Et cette question peut assez naturellement mener à une démarche d'intégration de données. Décryptons cette question :

— *Quel système utiliser...* : un système de RI (SRI) est, très grossièrement, un assemblage de briques élémentaires relatives à, entres autres, un modèle de recherche, une méthode de reformulation de requête, un nombre de termes à considérer... Déterminer quel système revient à définir quelle configuration de ces paramètres est la plus adaptée ;

— *... pour répondre au mieux...* : signifie que l'on souhaite utiliser un SRI qui obtient les meilleures performances, et il existe de multiples indicateurs susceptibles de caractériser ces performances ;

— *...à une requête donnée ?* : l'intérêt ne réside pas seulement dans le traitement d'une requête en particulier mais dans le traitement de requêtes futures qui lui ressembleront, qui partageront des caractéristiques similaires. D'où la nécessité de pouvoir caractériser une requête.

Ainsi, la démarche d'intégration qui en découle revient à extraire de l'information d'un ensemble de données dont une représentation schématisée est donnée sur la figure 3.3. Il convient de combiner l'ensemble des données afin de répondre au mieux au problème posé. Dans ce contexte, il n'est, a priori, pas raisonnable d'envisager une méthodologie susceptible de prendre toutes les données en entrée et de fournir en sortie des règles de conduites du type : *Pour une requête de tel type, utiliser un SRI ayant telle configuration pour maximiser tel indicateur de performance.* La stratégie d'intégration de données revient ainsi à développer une démarche plus qu'une méthode. Les différents travaux menés en collaboration avec l'équipe Systèmes d'Informations Généralisés de l'Irit, visent à aborder différents aspects de cette démarche intégrative.

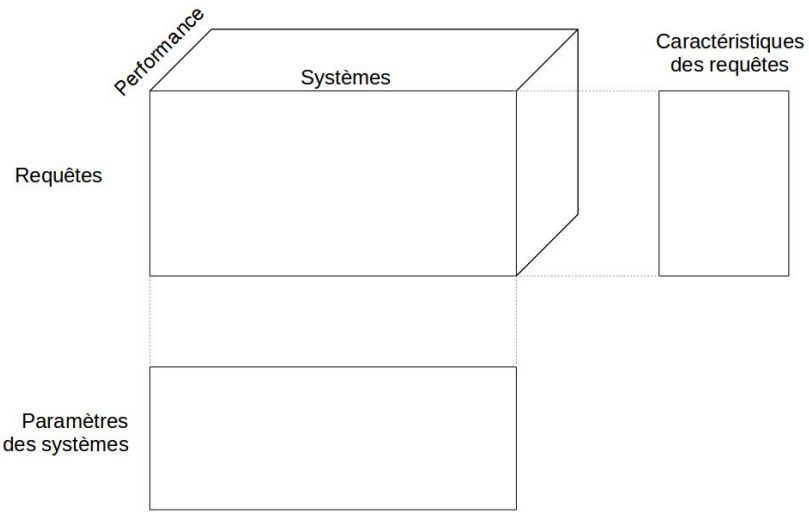

FIGURE 3.3 – Représentation schématique de l'ensemble des données disponibles dans le cadre d'une étude en recherche d'information.

3.2.2 Données

Afin de préciser les contours de la démarche intégrative, j'illustre ici les données susceptibles d'être recueillies dans chacun des blocs du schéma de la figure 3.3 afin de traiter la problématique décrite précédemment.

Ainsi :

— le tableau à 3 dimensions croisant Requêtes (en ligne), Systèmes (en colonne) et Performances (en strate) contient dans chaque cellule la valeur d'un indicateur de performance obtenu par un système pour une requête donnée.
— Le tableau *Caractéristiques des requêtes* peut contenir par exemple pour chaque requête (en ligne) des paramètres linguistiques (en colonne) la caractérisant.
— Le tableau *Paramètres des systèmes* pour chaque système (en colonne), un certain nombre de paramètres (en ligne) le caractérisant.

Chacun de ces éléments est davantage détaillé ci-dessous :

— Performance : il existe de nombreux indicateurs de performance. Le programme `trec_eval` [1] largement répandu dans la communauté RI permet d'en calculer environ 130. Ils sont généralement des déclinaisons et combinaisons du rappel et de la précision. Le rappel est le rapport du nombre de documents pertinents retournés sur le nombre total de documents pertinents. La précision est le rapport du nombre de documents pertinents sur le nombre de documents retournés. Ces indicateurs peuvent être calculés pour un nombre variable de documents retournés (précision et rappel à 5, précision et rappel à 10...). La section 3.2.3 aborde la question de la redondance de ces indicateurs de performance.
— Requêtes : une requête est une demande formulée à un SRI. Dans le cadre d'une campagne d'évaluation en RI comme la *Text REtrieval Conference* (TREC), une requête n'est pas constituée seulement de quelques mots-clés mais présente une structure avec un titre, une description et une partie narrative. En voici un exemple :

```
<title> Topic: Corporate Pension Plans/Funds
<desc> Description: Document will report on problems associated with pension
plans/funds such as fraud, skimming, tapping or raiding.
<narr> Narrative: A relevant document will report on problems associated with
pension plans/funds and also U.S. Government regulatory controls on pension
plans. Examples of problems that are considered relevant are fraud, skimming,
tapping or raiding.
```

  Ainsi, étant donné son contenu relativement complexe et riche, une requête peut être soumise à une analyse linguistique afin d'en extraire des caractéristiques potentiellement pertinentes en vue d'optimiser un SRI. Il s'agit par exemple de caractéristiques :
  — morphologiques : nombre d'acronymes, de termes avec un suffixe, de noms propres...
  — syntaxiques : nombre de conjonctions, de prépositions, de pronoms...
  — liées à l'ambiguïté des termes de la requête : nombre moyen des sens des mots déterminés à partir de WordNet [2]...
  — liées au potentiel discriminant des termes : l'*Inverse Document Frequency* (IDF), mesure par exemple la popularité ou la rareté d'un terme.
  — ...
— Systèmes : Les paramètres d'un SRI sont par exemple :
  — méthodes d'indexation : suppression de mots-outils, radicalisation, lemmatisation, représentation vectorielle...
  — modèles de recherche d'information : ensembliste, algébrique, probabiliste...
  — méthodes d'extension de requêtes : retour de (pseudo-)pertinence (*(pseudo-)relevance feedback*), pertinence positionnelle, pertinence de proximité...
  — ...

  Les travaux présentés dans la section 3.2.4 reprennent l'article [19] qui aborde la question de l'influence de ces paramètres sur les performances d'un SRI.

3.2.3 Étude de la redondance des mesures de performance

Afin d'évaluer la qualité de systèmes de recherche d'information, des campagnes de test sont régulièrement organisées. Dans ce contexte, les participants à la campagne ont accès à une collection de documents

1. trec.nist.gov/trec_eval
2. wordnet.princeton.edu

et à un ensemble de requêtes. Les organisateurs du concours ont, eux, de leur côté, la liste des documents pertinents pour chaque requête. En fonction des documents retournés par chaque système de recherche d'information concourant, les organisateurs ont la possibilité d'évaluer les systèmes et de les classer sur la base de mesures de performances. Le programme utilisé dans le cadre de la campagne TREC, `trec_eval` calcule un ensemble de 130 mesures de performances. Dans le travail dont il est question ici, nous nous sommes interrogés sur la pertinence d'avoir recours à autant d'indicateurs de performance et nous avons mené des analyses statistiques afin de permettre une évaluation globale des systèmes en ayant recours à un nombre limité de mesures de performance.

Ce travail publié dans la revue *Knowledge and Information System* [6] s'appuie sur l'analyse d'un jeu de données composé de 23518 lignes et 130 colonnes. Les lignes correspondent au nombre de couples système × requête qui ont été évalués via les 130 indicateurs de performance.

Evaluating effectiveness of information retrieval systems is achieved by performing on a collection of documents, a search, in which a set of test queries are performed and, for each query, the list of the relevant documents. This evaluation framework also includes performance measures making it possible to control the impact of a modification of search parameters. The program trec_eval calculates a large number of measures, some being more used like the mean average precision or recall-precision curves. The motivation of our work is to compare all measures and to help the user to choose a small number of them when evaluating different information retrieval systems. In this paper, we present the study we carried out from a massive data analysis of TREC results. Relationships between the 130 measures calculated by trec_eval for individual queries are investigated, and we show that they can be clustered into homogeneous clusters.

Résumé de l'article [6]

La stratégie mise en œuvre afin de réduire le nombre de mesures de performance à utiliser pour évaluer des systèmes de recherche d'information est la suivante :

1. analyser la matrice de corrélation des 130 mesures afin d'avoir une vue globale des similarités entre mesures ;
2. proposer une classification non supervisée des mesures de performance afin de déterminer un nombre raisonnable de groupes dans lesquels rassembler les mesures ;
3. interpréter les différents groupes retenus ;
4. identifier un représentant pour chaque groupe.

La figure 3.4 est directement issue de l'article [6]. La représentation d'une matrice de corrélation en niveaux de gris n'est pas du tout optimale. J'en propose ici (Figure 3.5) une nouvelle version utilisant une palette divergente du package `RColorBrewer` plus adaptée à la visualisation de valeurs de corrélation.

L'information essentielle qu'elle apporte est la forte corrélation positive exprimée par les teintes globalement très claires de l'image de la figure 3.4 et vert foncé dans la figure 3.5. La structure de l'image a été obtenue en procédant à une réorganisation optimale des données afin d'en mettre en évidence plus distinctement les grandes structures. J'apprécie beaucoup cette approche de réorganisation optimale décrite dans *La graphique et le traitement graphique de l'information* de Jacques Bertin [8]. J'ai été marqué par l'apparente simplicité de cette approche permettant de mettre visuellement en évidence « à la main » des structures dans des jeux de données. J'ai retrouvé cette approche dans le logiciel PermutMatrix [3] développé par Gilles Caraux et Sylvie Pinloche [15] et plus récemment dans le package `seriation` [44] pour le logiciel R. C'est le logiciel PermutMatrix que j'ai utilisé dans cette étude.

La mise en œuvre d'une classification ascendante hiérarchique ne pose aucun souci particulier dans ce contexte. Nous avons utilisé la distance euclidienne comme distance inter-individus et le critère de Ward comme critère d'agglomération pour obtenir le dendrogramme de la figure 3.6. La détermination d'un nombre de groupes raisonnable a ensuite été menée selon une règle pragmatique : retenir le plus grand nombre de groupes ayant une interprétation pertinente en termes de recherche d'information. Parmi les choix possibles, c'est vers 7 groupes que nous nous sommes orientés. C'est en mettant en œuvre un algorithme de k-means avec 7 groupes (algorithme initialisé aux centres des classes issues de la classification hiérarchique) que nous avons obtenus notre classification finale.

Je n'ai découvert qu'après l'article [17] qui s'intéresse spécifiquement à la classification de variables. Il aurait été intéressant de confronter notre approche à celle-ci et de voir notamment l'influence du choix de la distance euclidienne sur les résultats finaux.

Afin d'interpréter au mieux les 7 groupes de mesures ainsi obtenus et détaillés dans le tableau 3.2, nous avons représentés les mesures de performance sur le premier plan d'une ACP en les identifiant selon

3. www.atgc-montpellier.fr/permutmatrix/

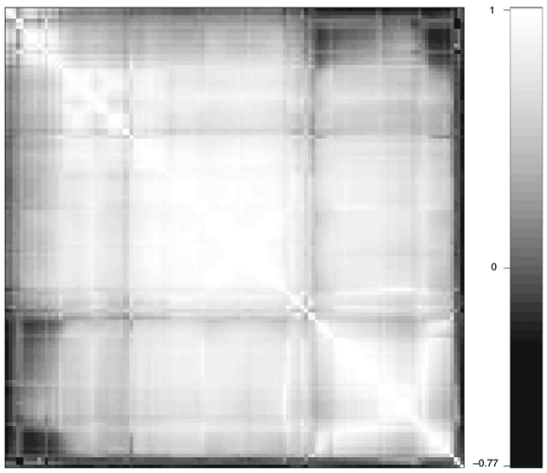

Fig. 1 Correlation matrix of the performance measures displayed using different *gray level shades* (from *black* −0.77 to *white* 1)

FIGURE 3.4 – Représentation de la matrice de corrélation entre les 130 mesures issue de l'article [6].

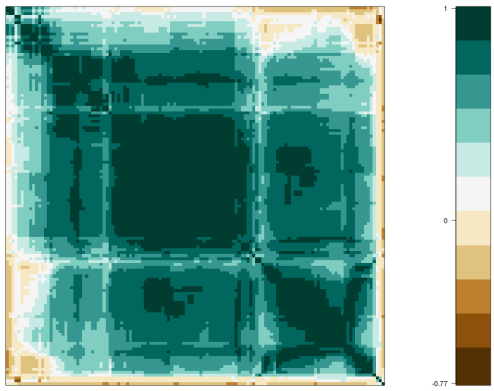

FIGURE 3.5 – Nouvelle représentation de la figure 3.4 en utilisant une palette divergente du package `RColorBrewer`.

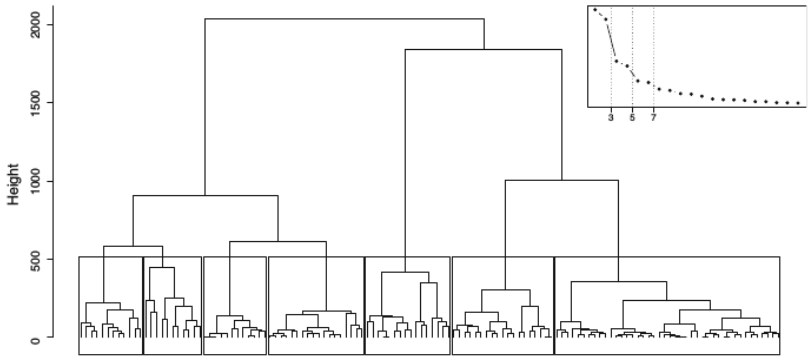

Fig. 3 Dendrogram representing the hierarchical clustering of the performance measures with a relevant pruning at 7 clusters. The sub-plot in the upper-right corner represents the height of the nodes of the dendrogram; it suggests relevant pruning at 3, 5, or 7 clusters

FIGURE 3.6 – Représentation de la classification hiérarchique des 130 mesures issue de l'article [6].

TABLE 3.2 – Groupes d'indicateurs résultant de la mise en œuvre successive de la classification ascendante hiérarchique et d'un algorithme k-means.

Cluster 1 (23 measures)

relative_unranked_avg_prec30 - relative_unranked_avg_prec20 - relative_prec30 - map_at_R - relative_unranked_avg_prec15 - relative_prec20 - P30 - relative_prec15 - int_0.20R.prec - relative_unranked_avg_prec10 - X0.20R.prec - ircl_prn.0.10 - P20 - P15 - relative_prec10 - bpref_10 - P10 - relative_unranked_avg_prec5 - relative_prec5 - P5 - bpref_5 - recip_rank - ircl_prn.0.00

Cluster 2 (16 measures)

P100 - P200 - unranked_avg_prec500 - unranked_avg_prec1000 - bpref_num_ret - P500 - bpref_num_all - P1000 - num_rel_ret - exact_unranked_avg_prec - num_rel - exact_prec - bpref_num_correct - utility_1.0_.1.0_0.0_0.0 - exact_relative_unranked_avg_prec - bpref_num_possible

Cluster 3 (12 measures)

bpref_top10Rnonrel - bpref_retnonrel - relative_unranked_avg_prec500 - avg_relative_prec - recall500 - relative_prec500 - bpref_allnonrel - relative_unranked_avg_prec1000 - exact_recall - recall1000 - relative_prec1000 - exact_relative_prec

Cluster 4 (45 measures)

X1.20R.prec - ircl_prn.0.30 - X1.40R.prec - int_map - X1.00R.prec - R.prec - int_1.20R.prec - exact_int_R_rcl_prec - int_1.00R.prec_infAP - avg_doc_prec - map - X11.pt_avg - X1.60R.prec - int_0.80R.prec - int_1.40R.prec - X0.80R.prec - old_bpref_top10pRnonrel - ircl_prn.0.40 - X1.80R.prec - int_1.60R.prec - X3.pt_avg_bpref - X2.00R.prec - bpref_top25p2Rnonrel - old_bpref - bpref_top10pRnonrel - int_1.80R.prec - int_0.60R.prec - int_2.00R.prec - bpref_top25pRnonrel - X0.60R.prec - bpref_top50pRnonrel - bpref_top5Rnonrel - ircl_prn.0.20 - ircl_prn.0.50 - int_0.40R.prec - X0.40R.prec - int_map_at_R - ircl_prn.0.60 - unranked_avg_prec30 - ircl_prn.0.70 - ircl_prn.0.80 - unranked_avg_prec200 - unranked_avg_prec100

Cluster 5 (18 measures)

bpref_topnonrel - fallout_recall_42 - fallout_recall_28 - fallout_recall_56 - rcl_at_142_nonrel - fallout_recall_71 - fallout_recall_85 - relative_unranked_avg_prec100 - fallout_recall_99 - fallout_recall_113 - relative_prec100 - fallout_recall_127 - relative_unranked_avg_prec200 - fallout_recall_142 - recall100 - relative_prec200 - recall200 - bpref_retall

Cluster 6 (13 measures)

fallout_recall_14 - unranked_avg_prec20 - unranked_avg_prec15 - ircl_prn.0.90 - fallout_recall_0 - unranked_avg_prec10 - recall30 - ircl_prn.1.00 - recall20 - recall15 - unranked_avg_prec5 - recall10 - recall5

Cluster 7 (3 measures)

rank_first_rel - num_nonrel_judged_ret - num_ret

leur groupe précédemment défini. Nous avons exclu de cette analyse le groupe de 3 mesures (Cluster 7) qui contient des indicateurs atypiques et peu pertinents comme le nombre de documents retournés (*num_ret*) et le rang du premier document pertinent parmi les documents retournés (*rank_first_rel*).

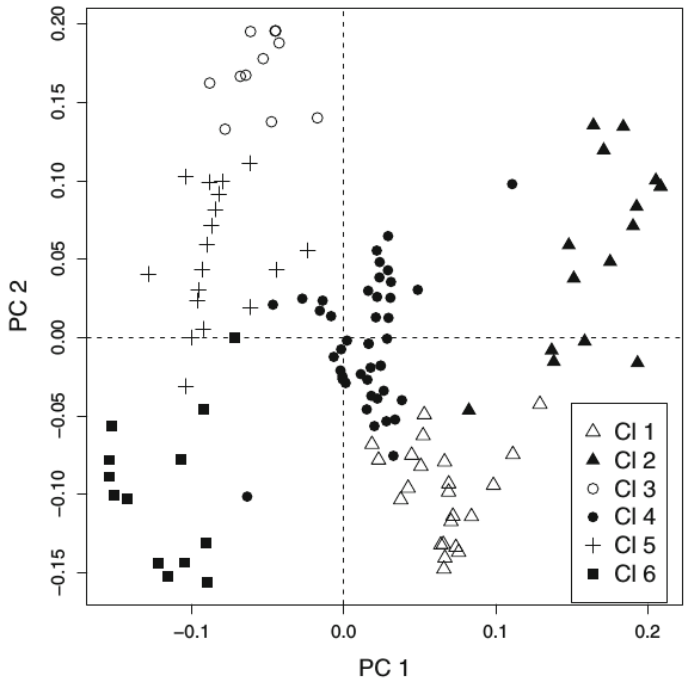

Fig. 4 Representation of variables on the first two principal components PC1 and PC2, respectively, explaining 67% and 13% of the total variance. *Symbols* reveal the cluster the variables belong to

FIGURE 3.7 – Représentation des résultats de l'ACP issue de l'article [6].

Cette représentation met ainsi en évidence une opposition sur l'axe horizontal entre, à gauche, les mesures basées sur le rappel (groupes 3 (○), 5 (+) et 6 (■)) et à droite, celles basées sur la précision (1 (△) et 2 (▲)). Selon l'axe vertical, l'opposition est essentiellement dirigée par le nombre de documents utilisés dans le calcul de l'indicateur : peu de documents (moins de 30) pour les groupes 1 (△) et 6 (■) en bas, davantage de documents pour les groupes 2 (▲) et 3 (○) en haut. De façon assez naturelle, le groupe 4 (●) majoritairement composé de mesures « moyennes »comme la *Mean Average Precision* (MAP) se positionne de façon centrale en intermédiaire entre les autres groupes.

Ainsi, en réponse au titre de l'article [6], *How many performance measures to evaluate Information Retrieval Systems ?*, nous affirmons que 6 paramètres suffisent à caractériser la performance globale d'un SRI. Concrètement, en considérant le tableau à 3 dimensions de la figure 3.3, nous venons de ramener son nombre de strates de 130 à 6. Nous suggérons également une démarche afin de déterminer un représentant pertinent pour chacune des 6 classes identifiées et montrons que le classement de systèmes sur la base de 6 indicateurs est très similaire à celui basé sur les 130 indicateurs fournis par `trec_eval`.

Cette étude a permis d'établir les résultats suivants :

— Les mesures de performance évaluées pour la tâche ad hoc de la campagne TREC sont quasiment toutes fortement corrélées positivement. Autrement dit, quand un système est performant, toutes les mesures de performance le montrent.

— Pour affiner ce premier résultat, nous avons proposé une classification des mesures de performances en groupes cohérents. Sans pour autant affirmer que les mesures d'un groupe sont les mêmes, nous avons montré qu'elles caractérisent des comportements similaires.

— Dans chacun des 7 groupes, nous avons proposé une mesure d'homogénéité du groupe ainsi qu'une stratégie visant à identifier un représentant. Pour des groupes très homogènes, n'importe quel indicateur peut être considéré comme un représentant ; dans le cas des groupes plus hétérogènes, un représentant est défini en tenant compte d'éléments statistiques (sa position à l'intérieur du groupe projeté par ACP) et d'aspects informatiques (popularité de la mesure dans la communauté RI).

Enfin, un dernier point de conclusion rappelle et insiste sur le fait que les travaux ont été mené dans un cadre interdisciplinaire statistique / recherche d'information. Voici un extrait de la conclusion qui met ce point en exergue : *We characterize some of the clusters not only on a mathematical point of view but on an information retrieval point of view as well.*

J'apprécie particulièrement ce point qui met encore une fois en évidence l'attention que je porte à ne pas perdre le lien avec le contexte des analyses statistiques mises en œuvre. Ainsi, le point de vue de la recherche d'information n'a jamais été perdu de vue dans cette analyse et ceci afin de fournir des résultats pertinents à la communauté cible de cette étude.

3.2.4 Identification des paramètres d'un SRI les plus influents sur les performances

Par rapport à notre schéma de référence de la figure 3.3, nous nous intéressons ici à l'exploitation du tableau des paramètres des systèmes que nous associons à des indicateurs de performances afin de déterminer les paramètres les plus influents sur la performance.

Le travail présenté ici a fait l'objet d'une communication que j'ai assurée lors de la conférence internationale *The First International Conference on Advances in Information Mining and Management* (IMMM 2011) [19]. Des étudiants du département Génie Mathématique et Modélisation de l'Institut National des Sciences Appliquées ont contribué à ce travail dans le cadre d'un projet tutoré dont j'ai assuré le co-encadrement avec J. Mothe.

This paper presents the results mining a large set of information retrieval results. The objective of this study is to determine which parameters significantly affect search engine performances. We focus on the main features of information retrieval : indexing parameters and search models. Statistical analysis identify the retrieval model as the most important parameter to be tuned to improve the performance of an information retrieval system. We also show that the significant parameters depend on the topic difficulty.
Keywords-component ; information retrieval ; data mining ;parameter analysis ; performance prediction

Résumé de l'article [19]

Le problème abordé consiste à identifier les caractéristiques d'un système de recherche d'information qui influe le plus sur ses performances. Pour cela, la plateforme Terrier [4] a été utilisée. Elle permet d'implémenter de nombreuses fonctionnalités existantes pour l'indexation et la recherche d'information. Nous avons pu ainsi générer un jeu de données à partir de 100 requêtes pour plusieurs méthodes d'indexation, plusieurs modèles de recherche d'information et plusieurs méthodes d'extension de requêtes. Les différents paramètres sont présentés dans le tableau 3.3 issu de l'article publié dans les actes de la conférence IMMM 2011.

TABLE 1. PARAMETERS AND VALUES USED FOR A SEARCH.

Parameters	Meaning	Values
Top	Topic number	351, ..., 450
Field	Topic field	T; T+D; T+D+N
Bloc	Size of the indexing bloc	1, 5, 10
Idf	Inverse document frequency	FALSE, TRUE
Ref	Query reformulation	None, Bo1bfree, Bo2bfree, KLbfree
Model	Retrieval model	BB2c1, BM25b0.5, DFRBM25c1.0, IFB2c1.0, InexpB2c1.0, InexpC2c1.0, InL2c1.0, PL2c1.0, TFIDF
DocNb	Number of documents (reformulation)	0, 3, 10, 50, 100, 200
qe_md	Minimum number of documents in which the term should appear to be used in the query epension	0, 2
qe_t	Number of terms used in the query epension	0, 1

TABLE 3.3 – Paramètres utilisés pour l'exécution d'un *run*.

Chaque combinaison des paramètres présentés dans le tableau 3.3 conduit à un SRI dont les résultats obtenus sur les 100 requêtes peuvent être évalués par le calcul de mesures de performances.

4. www.terrier.org

Les données ainsi générées sont composées de 98650 observations en ligne (chaque ligne représentant un *run* c'est à dire une requête suivi de la chaîne des modules utilisés) et de 8 variables en colonnes : 7 variables pour les paramètres utilisés et 1 variable pour la mesure de performance (*Mean Average Precision*, MAP) sur laquelle nous nous sommes focalisés dans ce travail.

Des analyses bivariées ont d'abord été menées pour évaluer l'influence des différents paramètres sur la MAP. Elles sont illustrées sur la figure 3.8. Elles montrent par exemple que le recours à une technique de reformulation de requête (Ref) a un effet significatif sur la MAP obtenue par un système (graphique au milieu à droite) ; en revanche, l'utilisation de l'IDF (graphique au milieu à gauche) n'en a pas.

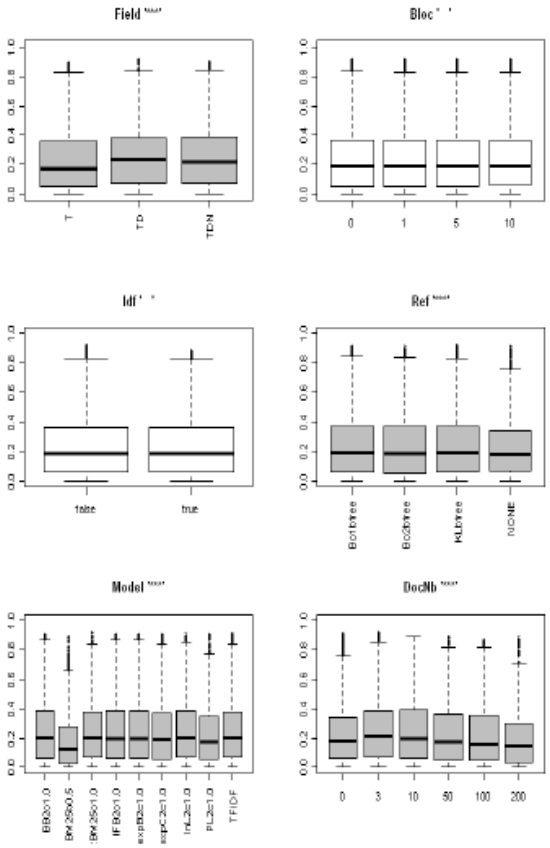

Figure 2. Boxplots representing MAP according to the different levels of each parameter (Field – 3 levels, Bloc – 4 levels, Idf – 2 levels, Ref – 4 levels, Model – 9 levels, DocNb – 6 levels). The symbol near the title indicates the p-value of the test according to the code: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. Grey boxplots highlights cases where the parameter is highly significant in the ANOVA model (p-value < 0.001).

FIGURE 3.8 – Représentation des boxplots des valeurs de MAP en fonction des modalités de chaque paramètre du système de recherche d'information utilisé. Figure extraite de l'article [19].

Afin de compléter cette première impression, nous avons souhaité avoir une approche mutivariée afin de mieux identifier les paramètres susceptibles d'influencer la performance du système. Pour cela, nous avons choisi d'utiliser les *Classification And Regression Trees* (CART) [14, 59]. Dans la première approche présentée ici, nous avons opté pour la variante *Classification* de cette méthode et avons pour cela recodé la variable MAP en 3 classes : *Bad* (MAP inférieure au premier quartile), *Average* (MAP comprise entre le premier quartile et le troisième quartile) et *Good* (MAP supérieure au troisième quartile).

Ces travaux sont à l'origine d'autres développements. Je mentionne par exemple, les travaux de thèse d'Anthony Bigot [9] qui a approfondi les résultats précédents en proposant une stratégie visant à sélectionner le SRI le plus apte à traiter une requête selon sa difficulté et, en complément, à définir la meilleure configuration d'un système en s'appuyant sur un sous-ensemble de documents dont la pertinence est connue [10].

Dans l'article [5], nous abordons en complément la définition de classes de requêtes en fonction de leur difficulté. Le résumé de cet article est repris ici :

Search engines are based on models to index documents, match queries and documents and rank documents. Research in Information Retrieval (IR) aims at defining these models and their parameters in order to optimize the results. Using benchmark collections, it has been shown that there is not a best system configuration that works for any query, but rather that performance varies from one query to another. It would be interesting if a meta-system could decide which system configuration should process a new query by learning from the context of previous queries. This paper reports a deep analysis considering more than 80,000 search engine configurations applied to 100 queries and the corresponding performance. The goal of the analysis is to identify which configuration responds best to a certain type of query. We considered two approaches to define query types : one is post-evaluation, based on query clustering according to the performance measured with Average Precision, while the second approach is pre-evaluation, using query features (including query difficulty predictors) to cluster queries. Globally, we identified two parameters that should be optimized : retrieving model and TrecQueryTags process. One could expect such results as these two parameters are major components of IR process. However our work results in two main conclusions : 1) based on post-evaluation approach, we found that retrieving model is the most influential parameter for easy queries while TrecQueryTags process is for hard queries ; 2) for pre-evaluation, current query features do not allow to cluster queries to identify differences in the influential parameters.

Résumé de l'article [5]

Cet article met en évidence la difficulté de la détermination de classes de requêtes. Ce problème peut être abordé de deux principales façons différentes selon que l'on considère ou pas les résultats obtenus par des SRI. On parle ainsi de critères :

— pré-évaluation, si on essaie de caractériser une requête avant de la soumettre à un SRI. Il s'agit ici par exemple de considérer des paramètres linguistiques des requêtes ; on intègre ainsi le tableau des caractéristiques des requêtes de notre schéma de référence 3.3. Par exemple, on peut penser qu'une requête qui contient des mots populaires (avec un IDF faible) et avec des sens différents (ambiguïté des termes) sera plus difficile à traiter par un SRI qu'une requête avec des mots rares ayant un unique sens.

— post-evaluation, si on détermine la difficulté d'une requête en fonction des résultats obtenus par des SRI. Une requête pour laquelle la plupart des SRI obtiennent de mauvaises performances est à considérer comme difficile.

Une des conclusions de ce travail publié en 2015 montre la difficulté de caractériser une requête en pré-évaluation de façon suffisamment pertinente pour que cela soit utile dans la détermination des paramètres d'un SRI. Et c'est précisément sur ce sujet que des travaux sont en cours. Nous nous intéressons notamment à l'étude de combinaisons de paramètres pré- et post- en vue de quantifier les apports de chacune des deux sources. Ces travaux nous orientent vers la problématique du choix de modèles sur laquelle travaille actuellement un collègue en contrat post-doctoral.

3.2.5 Conclusion

Ce qui positionne ces travaux dans ce chapitre 3 dédié à l'intégration de données c'est la démarche globale qui est par nature intégrative. La démarche intégrative est liée au problème central qui nous intéresse ici et qui est formulé par la question écrite au début de la section 3.2 : *Quel système utiliser pour répondre au mieux à une requête donnée ?*. Chaque étude présentée ici s'insère dans une problématique globale, on est réellement face à un puzzle. On essaie de compléter un coin de l'image en rassemblant quelques pièces mais on ne sait pas encore le raccrocher à l'ensemble ; on essaie un autre coin, on regroupe quelques pièces et on essaie de regrouper les deux ensembles...

Par ailleurs, ce que le schéma de la figure 3.3 page 33 ne met pas en évidence, c'est la quantité de données et les ressources de calcul nécessaires pour arriver à des données analysables. L'utilisation de la plateforme Terrier sur différents jeux de données a parfois nécessité le recours au méso-centre de calcul Calmip [5] pour traiter les giga-octets de données brutes nécessaires à l'élaboration des tableaux analysables statistiquement.

Enfin, je souhaite insister sur l'interdisciplinarité de ces travaux. Ce n'est que grâce à une recherche à l'interface entre statistique et informatique que ces travaux ont pu se concrétiser et que d'autres sont encore en cours. J'évoque simplement un aspect de ces travaux qui consiste à appréhender le problème de la difficulté d'une requête sur la base de critères automatiques, en comparaison avec une évaluation humaine de la difficulté. Pour aborder cette problématique, le champ de l'interdisciplinarité devra nécessairement

5. www.calmip.univ-toulouse.fr

s'étendre au-delà du couple statistique-informatique pour aller également vers les sciences humaines. Et cela sera certainement une nouvelle source de complexité et de stimulation.

3.3 Analyse intégrative de données biologiques

Les éléments présentés dans cette partie s'insèrent dans le domaine de la biologie intégrative (*integrative biology*) ou biologie systémique (*systems biology*). Ces deux termes sont très à la mode en ce moment et, même s'ils revêtent parfois des sens sensiblement différents, l'idée principale reste la même. On en trouve des définitions dans [81] – *Generally, data integration can be defined as the process of combining data residing in diverse sources to provide users with a comprehensive view of such data. There is no universal approach to data integration, and many techniques are still evolving* – ou dans [74] – *Statistical methods in integrative genomics aim to answer important biology questions by jointly analyzing multiple types of genomic data (vertical integration) or aggregating the same type of data across multiple studies (horizontal integration)*– ou encore dans [75] –*Data integration : The incorporation of multi-omic information in a meaningful way to provide a more comprehensive analysis of a biological point of interest.*

Ce sont des termes que l'on retrouve également de plus en plus dans des organismes ou départements de recherche. Notons par exemple, *Institute for Systems Biology*[6] organisme de recherche dans le domaine biomédical basé à Seattle qui pose comme base de la biologie systémique la phrase suivante *Systems biology is based on the understanding that the whole is greater than the sum of the parts,* qui résonne comme un slogan ; le Centre de Biologie Intégrative, unité de recherche qui vient d'être créée à Toulouse et réunissant 5 anciennes unités de recherche en biologie ; et le *Melbourne Integrative Genomics*[7] structure de l'Université de Melbourne dans laquelle j'ai eu le plaisir de passer quelques semaines durant l'été 2018 et où j'ai en partie rédigé ce mémoire.

Dans ce contexte, je présente dans un premier temps, le type de problématiques qui se posent en biologie intégrative ainsi que les données recueillies pour y répondre. J'aborde ensuite les méthodes statistiques qui ont été développées et auxquelles j'ai parfois contribué pour exploiter au mieux ces données et fournir des éléments de réponse aux problèmes biologiques posés. Je consacre ensuite une partie au logiciel `mixOmics` que j'ai contribué à développer pour fournir un outil de mise en œuvre de ces méthodes et répondre aux besoins de la communauté. Je conclue cette partie en présentant deux projets de recherche dans lesquels je me suis impliqué.

3.3.1 Données

Dans le domaine de la biologie, il est de plus en plus fréquemment admis que le recueil et l'analyse d'un seul jeu de données ne fournit pas suffisamment d'informations pour offrir une compréhension approfondie d'un système ou d'un phénomène biologique. Il convient donc d'acquérir plusieurs jeux de données si possible sur les mêmes unités expérimentales et d'en assurer une analyse la plus intégrée possible. Pour cela, ces dernières années ont vu un essor considérable des technologies susceptibles de traiter un échantillon biologique pour en extraire des informations quantitatives variées et volumineuses. Ces technologies dites à haut-débit et réunies parfois sous la terminologie 'omique en raison de leur terminaison commune (génomique, transcriptomique, protéomique, métabolomique...) peuvent ainsi générer un nombre important de données pour un seul échantillon biologique faisant ainsi entrer la biologie dans l'ère des *big data*.

La figure 3.9 est une vue synthétique des différents niveaux d'entrée dans une approche de biologie intégrative (génome, épigénome, transcriptome, protéome et métabolome) accompagnés dans le bandeau supérieur des types d'information susceptibles d'être acquis par les nouvelles bio-technologies.

Ce sont, entre autres, les avancées de ces technologies qui ont profondément remis en perspective le dogme de la biologie moléculaire que j'ai évoqué en introduction de ce mémoire. Le passage de l'information contenue dans les cellules d'un organisme, son ADN (premier bloc vertical de la figure 3.9), à son protéome (quatrième bloc) qui confère à chaque cellule sa structure et sa fonction, est soumis à bien plus de facteurs que ne le laissait supposer les connaissances scientifiques il y a une trentaine d'années environ. L'épigénétique, par exemple (deuxième bloc), étudie l'influence de facteurs environnementaux sur le fonctionnement du génome. Des modifications chimiques, comme la méthylation de l'ADN, peuvent en effet modifier la transcription sans modifier l'information génétique et conduire ainsi à des phénotypes différents malgré des génomes identiques. Au niveau du transcriptome (troisième bloc), les technologies de séquençage permettent également d'identifier des phénomènes d'épissage alternatif permettant à une même séquence d'ARN de produire plusieurs protéines.

Au-delà des aspects biologiques, j'attire l'attention sur le fait que les données qui sont générées par ces technologies sont de natures très variées. Les SNP (pour *single-nucleotide polymorphism*) du bloc génome

6. www.systemsbiology.org
7. https ://research.unimelb.edu.au/integrative-genomics

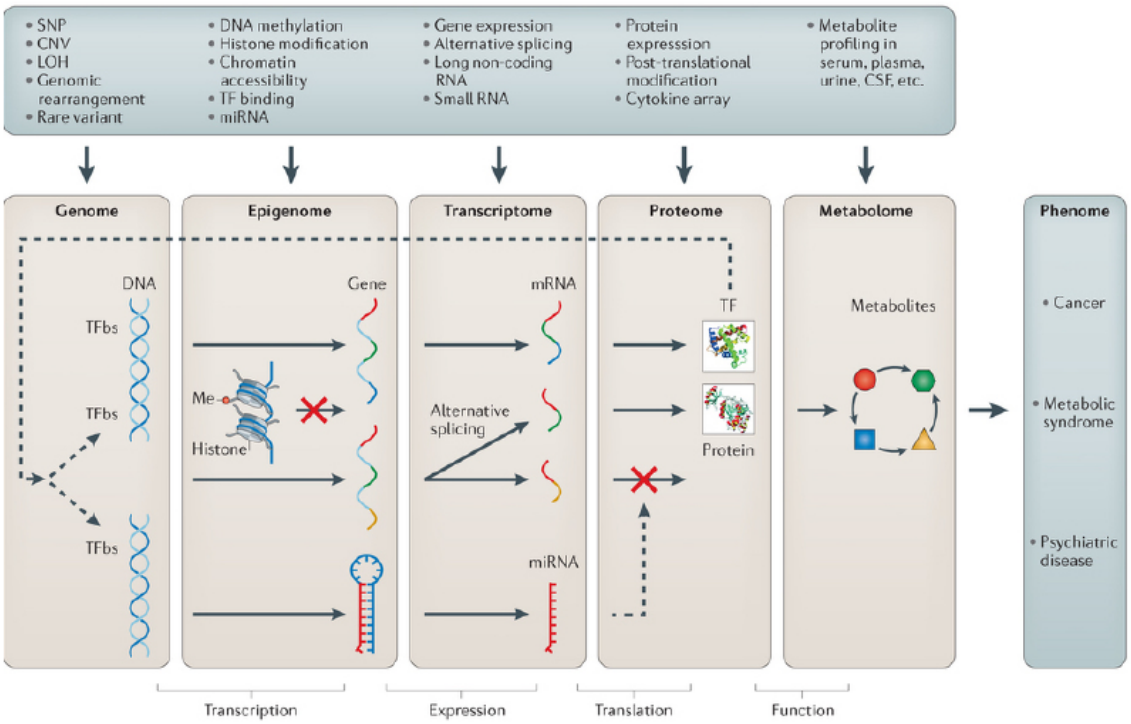

FIGURE 3.9 – Représentation des différents niveaux d'étude dans une approche de biologie intégrative : de l'information génétique à l'ensemble des caractères observables (phénome) de l'organisme étudié. Figure extraite de [75].

de la figure 3.9 traduisant la variation d'une seule paire de bases du génome, sont généralement recueillies selon une variable binaire. Les données relatives au transcriptome d'abord acquises par la technologie des puces à ADN ont été des mesures de fluorescence et donc assimilées à des valeurs continues réelles. Elles sont maintenant plus largement recueillies sous forme de comptage dans le cadre de l'utilisation de techniques de séquençage. Les données protéomiques et métabolomiques (quatrième et cinquième bloc) peuvent quant à elles être recueillies par spectrométrie de masse donnant lieu à des données sous forme de spectres. On voit ainsi que pour exploiter au mieux ces technologies, il faut développer des méthodes adaptés à la spécificité des données générées.

Par rapport à la partie précédente concernant une démarche intégrative en recherche d'information, les données dont il est question dans ce cadre ont la particularité d'être acquises sur les mêmes individus ; c'est en tout cas le cadre dans lequel je me place dans cette partie. On dispose ainsi de jeux de données que l'on peut schématiquement positionner côte-à-côte et c'est pour cela que l'on parle parfois d'intégration horizontale ou de N-intégration pour caractériser le fait que l'intégration se fait sur les mêmes observations d'un jeu de données rangé dans un tableau $n \times p$. Il peut s'agir de données quantitatives correspondant par exemple à des expressions de gènes (transcriptomiques), des abondances de protéines (protéomique), des dosages de certaines molécules (métabolomiques) ou encore de paramètres phénotypiques (longueur des feuilles d'une plante) ou paramètres cliniques (dosages sanguins chez des individus) à moins grande échelle et plus couramment mesurées. Il peut également s'agir de données qualitatives pouvant correspondre à des traitements différents ou à des génotypes différents (sauvage *Wild-Type* ou mutant).

L'intégration verticale est également un sujet d'intérêt biologique pour combiner les mesures des mêmes variables (par exemple des mêmes gènes) mais lors d'expérimentations différentes. Ce point a été abordé par exemple dans Rohart et al. [76] et l'implémentation de la méthodologie développée dans cet article fait partie du package `mixOmics` sous la terminologie Multivariate INTegrative method (MINT). Il s'agit ici d'identifier des variables (gènes par exemple) potentiellement intéressantes sans que cette conclusion ne repose que sur les données relatives à une seule expérience. Le problème à prendre en compte avec ce type de données réside dans un effet potentiellement important des différentes sources de données acquises dans des endroits différents et à des instants différents également. Ces sources de variation peuvent en effet masquer les phénomènes d'intérêt biologique. Le développement de méthodes d'intégration dites verticales permet également d'envisager une ré-utilisation de données existantes et parfois disponibles dans des bases de données publiques afin d'extraire des bio-marqueurs plus robustes dans la mesure où il ne dépendent pas d'une expérience en particulier. C'est avec cette ligne directrice

que des développements sont en cours autour du module GEM2net [98] de la base de données *a Complete Arabidopsis Transcriptome database* (CATdb [8]) [36]. Ce module fournit l'accès à des données transcriptomiques relatives à des études de multiples stress chez la plante modèle *Arabidopsis thaliana*. La mise en œuvre de modèles de mélange de graphes tend à mettre en évidence une réponse globale au stress des plantes quelle que soit l'origine de ce stress [67].

Cela étant, les notions de verticalité et d'horizontalité concernant l'intégration de données dépendent de la façon dont les données sont rangés dans les tableaux ; ainsi, ces notions sont évoquées de façon transposée dans [74] par rapport au référentiel que j'utilise dans ce mémoire.

Même si le cadre idéal est donné par l'acquisition de multiples variables sur un même ensemble d'observations, en pratique, cela n'est pas toujours le cas, et certains jeux de données peuvent être acquis sur un nombre restreint d'échantillons biologiques. Ce cas de figure a donné lieu à des développements méthodologiques. Les travaux de Voillet et al. [90] et de Imbert et al. [47] par exemple abordent ce problème en considérant les échantillons non soumis à l'ensemble des mesures comme des lignes manquantes d'un tableau. Et ils développement une stratégie d'imputation en créant un ensemble de donneurs potentiels susceptibles de remplacer les lignes manquantes d'un tableau de données.

Dans la suite, je m'intéresse exclusivement au problème de l'intégration horizontale avec plusieurs jeux de données acquis sur les mêmes échantillons biologiques, le nombre de variables dépassant largement le nombre d'observations.

3.3.2 Méthodes

Analyse des Corrélations Canoniques

Les méthodes dont il est question dans cette partie vise à démêler les relations pouvant exister entre plusieurs jeux de données acquis sur les mêmes individus. Historiquement, une des premières méthodes à aborder ce problème pour deux jeux de données est l'analyse des corrélations canoniques (ACC) avec une première référence sur le sujet pouvant être [46].

L'ACC de deux matrices X $(n \times p)$ et Y $(n \times q)$ suppose d'abord que $p \leq n$ et que $q \leq n$, ensuite que les matrices X et Y sont de plein rang p et q respectivement. Je présente ici le principe de l'ACC afin de m'en servir plus loin pour aborder les extensions qui ont été proposées. Une façon de voir l'ACC consiste à considérer un problème résolu par un algorithme itératif.

La première étape consiste à trouver les deux vecteurs $a^1 = (a_1^1, \ldots, a_p^1)'$ et $b^1 = (b_1^1, \ldots, b_q^1)'$ qui maximisent la corrélation entre les combinaisons linéaires U_1 et V_1 définies par :

$$U_1 = Xa^1 = a_1^1 X^1 + a_2^1 X^2 + \ldots + a_p^1 X^p$$

et

$$V_1 = Yb^1 = b_1^1 Y^1 + b_2^1 Y^2 + \ldots + b_q^1 Y^q,$$

en supposant les vecteurs a_1 et b_1 normalisés $\text{var}(U_1) = \text{var}(V_1) = 1$.

En d'autres termes, cela revient à résoudre

$$\rho_1 = \text{cor}(U_1, V_1) = \max_{a,b} \text{cor}(Xa, Yb),$$

sous la contrainte $\text{var}(Xa) = \text{var}(Yb) = 1$.

Les composantes U_1 et V_1 ainsi créées sont appelées les premières composantes canoniques et ρ_1 la première corrélation canonique.

Les composantes et corrélations canoniques d'ordres supérieurs peuvent ensuite être trouvées selon une démarche pas à pas. Ainsi, pour $s = 1, \ldots, p$, on peut successivement trouver les corrélations $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_p$ et les vecteurs correspondants $(a_1, b_1), \ldots, (a_p, b_p)$, en maximisant

$$\rho_s = cor(U_s, V_s) = \max_{a^s, b^s} \text{cor}(Xa^s, Yb^s) \text{ sous la contrainte } \text{var}(Xa^s) = var(Yb^s) = 1,$$

et sous la contrainte additionnelle

$$\text{cor}(U^s, U^t) = cor(V^s, V^t) = 0 \text{ pour } 1 \leq t < s \leq p.$$

D'un point de vue géométrique, notons

$$P_X = X(X'X)^{-1}X' = \frac{1}{n} X S_{XX}^{-1} X' \text{ et } P_Y = Y(Y'Y)^{-1}Y' = \frac{1}{n} Y S_{YY}^{-1} Y'$$

les projecteurs orthogonaux dans les espaces engendrés par les colonnes de X et Y respectivement. Alors il est établi (voir par exemple [64]) que

8. urgv.evry.inra.fr/CATdb

— les corrélations canoniques ρ_s sont les racines carrées positives des valeurs propres λ_s de $P_X P_Y$ (qui sont les mêmes que celles de $P_Y P_X$) : $\rho_s = \sqrt{\lambda_s}$;
— les vecteurs $U_1, \ldots, U_p$ sont les vecteurs propres standardisés correspondant aux valeurs propres décroissantes $\lambda_1 \geq \ldots \lambda_p$ de $P_X P_Y$;
— les vecteurs $V_1, \ldots, V_p$ sont les vecteurs propres standardisés correspondant aux mêmes valeurs propres de $P_Y P_X$.

Cette méthode nécessite un nombre d'observations supérieur au nombre de variables du jeu de données qui en compte le plus ($n \geq \max(p, q)$). En effet, S_{XX} et S_{YY} sont singulières quand, respectivement, $n < p$ et $n < q$. De plus, même si $n \geq p$ (resp. $n \geq q$), la matrice S_{XX} (resp. S_{YY}) peut être mal conditionnée si les variables de la matrice X (resp. Y) présentent de fortes colinéarités. En pratique, quand le nombre de variables augmente tout en restant inférieur au nombre d'observations, les premières corrélations canoniques sont très proches de 1 à cause d'un recouvrement des sous-espaces canoniques ; ce qui n'est pas du tout informatif vis à vis de l'interprétation des résultats. Ainsi, il a été recommandé [31] de respecter la condition $n \geq p + q + 1$ pour mettre en œuvre une analyse des corrélations canoniques. Qu'elle soit théorique ou pratique, une telle contrainte concernant un nombre minimal d'observations est impossible à respecter dans le cas des données biologiques issues des technologies à haut-débit. Cela reviendrait à disposer de plus d'échantillons biologiques (plantes, souris...) que, par exemple, de gènes dont l'expression est mesurée par puces à ADN ou par séquençage. Or, les gènes ou transcrits se comptent par milliers voire dizaines de milliers et le nombre d'échantillons biologiques ne se compte que de façon exceptionnelle par centaines. Dans ces conditions, de nombreux développements méthodologiques ont visé à aborder ce problème qui se résume en *Que faire quand $n \ll p$?*

Extensions pour le cas $n \ll p$

Analyse des corrélations canoniques régularisée Une façon de gérer le problème $n \ll p$ consiste à inclure une étape de régularisation dans les calculs. Une telle approche dans le contexte de l'ACC a été proposée d'abord dans [89], puis développée dans [58]. Ignacio González a poursuivi ces travaux dans le cadre de sa thèse [38].

Le principe de la régularisation consiste à considérer une pénalisation l_2 (*ridge*) dans le problème d'optimisation. Cela aboutit au remplacement de S_{XX} et S_{YY} par respectivement $\Sigma_{XX}(\lambda_1)$ et $\Sigma_{XX}(\lambda_1)$ définis par :

$$\Sigma_{XX}(\lambda_1) = S_{XX} + \lambda_1 I_p \text{ et } \Sigma_{YY}(\lambda_2) = S_{YY} + \lambda_2 I_q.$$

L'ajout des valeurs λ_1 et λ_2 sur la diagonale des matrices S_{XX} et S_{YY} les rend les matrices $\Sigma_{XX}(\lambda_1)$ et $\Sigma_{YY}(\lambda_2)$ inversibles même si S_{XX} et S_{YY} ne le sont pas. Les alternatives à cette approche résident dans des variantes de la pénalisation utilisée. Par exemple, [96, 70, 95] utilise une contrainte l_1 (*lasso*) pour réaliser une sélection de variables. L'article [91] utilise la pénalité *elastic net* combinant les pénalités *ridge* et *lasso*. Il est à noter que ces travaux ont tous été motivés par une problématique biologique liée à l'analyse de données 'omiques et publié dans des revues à forte connotation biologique : *Statistical Applications in Genetics and Molecular Biology*, *Biostatistics* et *Biometrical Journal*. Cela illustre le fait que des questions biologiques peuvent être à l'origine de développements méthodologiques en statistique.

Le travail d'Ignacio a notamment consisté à proposer une démarche de validation croisée pour régler au mieux les paramètres de régularisation λ_1 et λ_2 [40]. J'ai contribué au développement et à la mise en œuvre de ces travaux dans le package `CCA` pour le logiciel R [40]. Ce package qui a servi de prototype pour le package `mixOmics` dont je parlerai plus loin (section 3.3.3) n'est plus maintenu même s'il continue à être utilisé car nous recevons encore des messages d'utilisateurs que nous ré-orientons vers le, toujours actif, package `mixOmics`. J'ai également contribué au déploiement de ces méthodes dans le cadre de plusieurs projets de recherche [41, 18].

Méthodes PLS Un premier élément au sujet des méthodes PLS consiste à discuter de la signification du sigle PLS. Deux définitions co-habitent : *Partial Least Squares* et *Projection to Latent Structures* et si la première a, semble-t-il, était utilisée à l'origine, la seconde semble s'imposer de plus en plus [1]. Et personnellement, c'est aussi vers *Projection to Latent Structure* que je penche car cela me semble plus explicite de considérer la méthode comme une méthode de projection... sur des structures latentes.

Je ne souhaite pas présenter de façon détaillée la régression PLS ; la quasi-totalité des éléments que l'on peut souhaiter connaître sur le sujet peut être trouvée dans [86] ou dans cet ouvrage collaboratif plus récent [34]. En revanche, je reprends ici quelques idées sur le principe dont je me sers parfois lorsque j'interviens en formation à ce sujet.

Un des intérêts de la régression PLS réside dans le fait qu'elle peut s'appliquer même quand le nombre de variables explicatives est plus grand que le nombre d'observations ce qui n'est pas le cas de la régression linéaire multiple. Afin de comprendre comment cela peut fonctionner, j'utilise le principe de la régression

sur composantes principales. Notons, X la matrice $n \times p$ des variables explicatives et y la variable à expliquer et considérons $p > n$. Le recours à l'analyse en composantes principales permet de transformer la matrice X en une matrice $T = XW$ de dimension $n \times k$ avec $k < n$ en ne conservant que les k premières composantes principales. Ensuite, la régression expliquant la variable y peut être menée en considérant ces k premières composantes principales comme variables explicatives.

Voici que cela donne vu sous un angle itératif :

— Contruction d'une première composante t_1

$$t_1 = w_{11}x_1 + ... + w_{1p}x_p$$

— Régression simple de y sur t_1

$$y = c_1 t_1 + y_1$$

— D'où l'on peut extraire l'expression de y en fonction de la totalité des variables x_i

$$y = c_1 w_{11} x_1 + ... + c_1 w_{1p} x_p + y_1$$

— Pour ajouter une deuxième composante t_2 non corrélée à t_1

$$t_2 = w_{21}x_{11} + ... + w_{2p}x_{1p}$$

où les x_{1j} sont les résidus des régressions des variables X_j sur t_1.

— Nouvelle régression : $y = c_1 t_1 + c_2 t_2 + y_2$

— ...

Cette façon de voir les choses présente également l'avantage de faire apparaître la double combinaison linéaire via la notation *c.w* que l'on retrouve sur les graphiques proposés par le logiciel SIMCA-P de la société Umetrics. Ce logiciel est assez répandu dans le domaine de la chimiométrie où la régression PLS s'est largement développée pour pallier les soucis liés à des variables en quantité importante et potentiellement colinéaires. L'article [97] intitulé *PLS-regression : a basic tool of chemometrics* a été cité plus de 5700 fois d'après Google Scholar [9] consulté le 26 octobre 2018.

À la différence de la régression sur composantes principales, la régression PLS tient compte de la variable y dans le calcul de la matrice T. Le cas évoqué ici est parfois appelé PLS1 ; il correspond à l'explication d'une seule variable quantitative par d'autres variables quantitatives. La régression PLS s'applique aussi au cas de l'explication de plusieurs variables quantitatives (cas parfois nommé PLS2). Dans ce cas, elle présente une analogie certaine avec l'analyse des corrélations canoniques dans la mesure où elle revient à calculer des combinaisons linéaires des variables de deux paquets X et Y ayant une covariance maximale (c'est la corrélation qui est maximisée par l'ACC). Parmi les nombreuses méthodes développées autour ce principe (voir le nombre de sigles déclinés autour de PLS : PLS-DA, PLS-SVD, O-PLS, PLS-PM, L-PLS, PLS-SEM...), signalons les variantes en modes *canonique* et *régression* de la méthode. Elles diffèrent dans le déroulement de l'algorithme par un calcul différent des résidus utilisés à l'étape suivante. Dans le mode régression, à l'étape i, les résidus de la matrice Y sont calculés en fonction de l'information extraite de la matrice X à l'étape $i-1$. Dans le mode canonique, la matrice Y est déflatée compte tenu de l'information extraite de la matrice Y elle-même. Ces deux variantes permettent d'aborder deux questions biologiques différentes. En effet, dans le cas du mode canonique, chaque jeu de données joue un rôle symétrique, sans un aucun lien directionnel supposé entre les deux. Cela peut se prêter à l'analyse de jeux de données issus de technologies différentes mais visant à mesurer la même information biologique (par exemple, des données transcriptomiques mesurées par puces à ADN et séquençage). Dans le cas du mode régression, les matrices X et Y jouent des rôles asymétriques et cela peut correspondre à un lien de causalité supposé entre par exemple l'expression de gènes et la quantification de protéines dans un tissu.

Sélection de variables Un des objectifs de l'analyse intégrative de données biologiques consiste à sélectionner parmi les nombreuses variables mesurées (potentiellement plusieurs milliers voire dizaines de milliers dans le cas des données 'omiques) les variables les plus « pertinentes ». Selon la méthode et notamment son caractère supervisé ou non, la pertinence d'une variable revêt une définition différente. Il peut s'agir d'une variable particulièrement impliquée dans la relation avec d'autres variables d'un autre jeu de données pour une analyse non supervisée, ou d'un biomarqueur potentiel dans le cadre d'une analyse supervisée.

Parmi les travaux qui ont abordé la sélection de variables dans le cadre de l'analyse intégrative de données biologiques figurent la thèse de Kim-Anh Lê Cao [60] et l'article [62] spécifiquement consacré à l'élaboration d'un algorithme définissant une variante *sparse* de la méthode PLS. Elle s'appuie pour

9. scholar.google.fr

cela sur la décomposition en valeurs singulières (SVD) utilisable en PLS et sur les travaux de Shen et Huang [82] proposant une version *sparse* de l'ACP par une pénalisation *lasso*. Cette pénalisation appliquée sur chaque paire de vecteurs de coefficients des combinaisons linéaires (*loadings*) permet de sélectionner des variables d'intérêt en annulant les coefficients des variables les moins intéressantes.

Ces travaux ont été ensuite généralisés à plusieurs cas de figures de l'analyse intégrative qu'elle soit supervisée, Sparse PLS-DA [56] et/ou multiblocs [83].

Extensions pour le cas multi-blocs Le cas multi-blocs correspond à l'analyse simultanée de plus de deux jeux de données. Ce sujet dispose d'une littérature abondante ; voir par exemple cet ouvrage collectif [34]. Le chapitre dû à Tenenhaus et Hanafi [87] recense 16 méthodes publiées entre les années 1961 et 2006. Un extrait du tableau de synthèse présenté dans cet article est repris ici (tableau 3.10).

Method	Criterion
(1) SUMCOR (Horst 1961)	$Max \sum_{j,k} Cor(\boldsymbol{F}_j, \boldsymbol{F}_k)$ or $Max \sum_j Cor\left(\boldsymbol{F}_j, \sum_k \boldsymbol{F}_k\right)$
(2) MAXVAR (Horst 1961) or GCCA (Carroll 1968)	$Max \{\lambda_{first}[Cor(\boldsymbol{F}_j, \boldsymbol{F}_k)]\}$ **(a)** or $Max \sum_j Cor^2(\boldsymbol{F}_j, \boldsymbol{F}_{j+1})$
(3) SsqCor (Kettenring 1971)	$Max \sum_{j,k} Cor^2(\boldsymbol{F}_j, \boldsymbol{F}_k)$
(4) GenVar (Kettenring 1971)	$Min \{\det[Cor(\boldsymbol{F}_j, \boldsymbol{F}_k)]\}$
(5) MINVAR (Kettenring 1971)	$Min \{\lambda_{last}[Cor(\boldsymbol{F}_j, \boldsymbol{F}_k)]\}$ **(b)**
(6) Lafosse (1989)	$Max \sum_j Cor^2\left(\boldsymbol{F}_j, \sum_k \boldsymbol{F}_k\right)$
(7) Mathes (1993) or Hanafi (2005)	$Max \sum_{j,k} \lvert Cor(\boldsymbol{F}_j, \boldsymbol{F}_k)\rvert$
(8) MAXDIFF (Van de Geer, 1984 & Ten Berge, 1988)	$Max_{all\ \lVert w_j\rVert=1} \sum_{j\neq k} Cov(\boldsymbol{X}_j w_j, \boldsymbol{X}_k w_k)$
(9) MAXBET (Van de Geer, 1984 & Ten Berge, 1988)	$Max_{all\ \lVert w_j\rVert=1} \sum_{j,k} Cov(\boldsymbol{X}_j w_j, \boldsymbol{X}_k w_k)$
(10) MAXDIFF B (Hanafi and Kiers 2006)	$Max_{all\ \lVert w_j\rVert=1} \sum_{j\neq k} Cov^2(\boldsymbol{X}_j w_j, \boldsymbol{X}_k w_k)$
(11) (Hanafi and Kiers 2006)	$Max_{all\ \lVert w_j\rVert=1} \sum_{j\neq k} \lvert Cov(\boldsymbol{X}_j w_j, \boldsymbol{X}_k w_k)\rvert$
(12) ACOM (Chessel and Hanafi 1996) or Split PCA (Lohmöller 1989)	$Max_{all\ \lVert w_j\rVert=1} \sum_j Cov^2(\boldsymbol{X}_j w_j, \boldsymbol{X}_{j+1} w_{j+1})$ or $Min_{F,p_j} \sum_j \left\lVert \boldsymbol{X}_j - F p_j^T \right\rVert^2$
(13) CCSWA (Hanafi et al., 2006) or HPCA (Wold et al., 1996)	$Max_{all\ \lVert w_j\rVert=1, Var(F)=1} \sum_j Cov^4(\boldsymbol{X}_j w_j, \boldsymbol{F})$ or $Min_{\lVert F\rVert=1} \sum_j \left\lVert \boldsymbol{X}_j \boldsymbol{X}_j^T - \lambda_j F F^T \right\rVert^2$
(14) Generalized PCA (Casin 2001)	$Max \sum_j R^2(\boldsymbol{F}, \boldsymbol{X}_j) \sum_h Cor^2\left(x_{jh}, \hat{F}_j\right)$ **(c)**
(15) MFA (Escofier and Pagès 1994)	$Min_{F,p_j} \sum_j \left\lVert \frac{1}{\sqrt{\lambda_{first}\left[Cor(x_{jh}, x_{jl})\right]}} \boldsymbol{X}_j - F p_j^T \right\rVert^2$
(16) Oblique maximum variance method (Horst 1965)	$Min_{F,p_j} \sum_j \left\lVert \boldsymbol{X}_j \left(\frac{1}{n} \boldsymbol{X}_j^T \boldsymbol{X}_j\right)^{-1/2} - F p_j^T \right\rVert^2$

FIGURE 3.10 – Tableau de synthèse des méthodes multi-blocs extrait de [87].

Même si des méthodes existent et certaines depuis longtemps pour l'analyse simultanée de plus de deux groupes de variables acquises sur les mêmes individus que ce soit dans le cas non supervisé [85] ou le cas supervisé [83] ; et même si l'implémentation de ces méthodes est déjà effective dans les versions récentes du package `mixOmics` [77], leur mise en œuvre sur des données réelles d'origines variées est encore à accentuer pour améliorer notre savoir-faire. Et je prends une part active à cet aspect de la recherche pour populariser ces méthodes multi-blocs (voir 3.3.4) et les rendre opérationnelles : c'est à dire utilisables et interprétables. J'en reviens ainsi au cœur de mes activités de recherche : gérer tout ce qui se passe autour de la mise de en œuvre des méthodes.

Synthèse En résumé, les méthodes qui ont été évoquées précédemment :
- sont des méthodes linéaires ;
- permettent d'analyser simultanément plusieurs tableaux de données quantitatives ;
- peuvent être mise en œuvre dans le cadre d'une analyse supervisée ou non ;
- peuvent intégrer une stratégie de sélection de variables.

Autrement dit, chacune des méthodes fournit les coefficients de combinaisons linéaires des variables initiales ; ces coefficients étant optimisés au regard d'un critère d'optimalité spécifique à chaque démarche. La stratégie de sélection a pour conséquence d'annuler certains des coefficients de la combinaison linéaire.

3.3.3 Développement du package R `mixOmics`

Comme tout produit fini, le logiciel a l'avantage de diffuser et l'inconvénient de figer [57].

Cette citation, en tout cas sa seconde partie concernant le fait de figer, est à remettre en question avec le développement des logiciels libres. Le logiciel R connaît en effet des mises à jour régulières (environ tous les 6 mois) et chaque package peut évoluer de façon indépendante et des mises à jour plus fréquentes sont

tout à fait envisageables et relativement courantes selon l'activité des développeurs de packages. Donc, dans la citation introductive de cette partie, je retiens ici surtout le fait que le logiciel *a l'avantage de diffuser*. Il est certain en effet, que les travaux méthodologiques menées dans le domaine de l'intégration de données biologiques par mes collègues et moi n'auraient pas eu le même impact sans la réalisation du package `mixOmics`. Le tableau 3.4 présente quelques articles en lien avec le développement du package, ainsi que le thème principal de l'article (logiciel ou méthodologique) et le nombre de citations de cet article d'après Google Scholar consulté le 27 novembre 2018.

TABLE 3.4 – Référence, type et nombre de citations (d'après Google Scholar consulté le 27 novembre 2018) pour quelques articles liés au développement du package `mixOmics`. (*) *highly accessed* quelques temps après sa parution.

Article	Année de publication	Type	Nb citations
[61]	2009	logiciel `mixOmics`	263
[40]	2008	logiciel `CCA`	176
[41]	2009	méthode (R)CCA	48
[62]	2008	méthode SPLS	278
[77]	2017	logiciel `mixOmics`	70
[39]	2012	logiciel/méthode	94 (*)

Même si la comparaison brute de ces chiffres n'a pas vraiment de sens car le nombre de citations d'un article dépend bien entendu de nombreux facteurs (la revue dans laquelle l'article il est publié notamment ou encore la renommée des auteurs), on peut par exemple constater que l'article consacré au package `CCA` [40] est beaucoup plus cité que celui consacré à la mise en œuvre des méthodes dans un contexte biologique [41]. Et dans cet esprit que j'ai contribué à la diffusion de méthodes statistiques par la mise à disposition d'outils logiciels pour la communauté scientifique.

Ma contribution la plus significative concerne le package `mixOmics` pour le logiciel R. L'histoire de ce package commence en 2004 avec le début de la thèse d'Ignacio González [38] et se poursuit en 2005 avec le début de la thèse de Kim-Anh Lê Cao [60]. Dans les deux cas, il y est question de l'intégration de données d'origine biologique acquises via les technologies de la biologie dite à haut-débit. Ignacio s'est concentré sur le développement de l'analyse des corrélations canoniques pour des données comportant plus de variables que d'observations ; Kim-Anh a, de son côté, étudié des extensions parcimonieuses (*sparse*) de la méthode PLS dans le même contexte. Convaincus du fait que la meilleure des méthodes n'a d'intérêt pour la communauté qu'à condition de fournir un outil permettant de la mettre en œuvre, nous avons créé un package pour le logiciel R. Une première version du package `mixOmics` a donc été diffusée en 2009 sur le dépôt international *Comprehensive R Archive Network* [10].

Depuis, le package a bien grandi. Il contient aujourd'hui environ 150 fonctions pour 19 méthodologies implémentées et 10 jeux de données (figure 3.11). Il est accompagné par un site web `www.mixomics.org` contenant un tutoriel et des études de cas, une adresse de contact et une *newsletter*. À partir de 2014, nous avons entrepris des actions de formation et fin 2018, ce sont plus de 500 participants qui ont assisté à l'un des 18 ateliers que nous avons proposés dans différents contextes : formation permanente, école chercheur, tutoriel sélectionné lors de *The 13th European Conference on Computational Biology*, Strasbourg 2014, *Short course* sélectionné lors de *XXIXth International Biometric Conference*, Barcelone 2018...

Voici par exemple le retour d'un participant à la formation organisée en juillet 2018 à Melbourne. Il donne la teneur des nombreux retours positifs que nous avons eus suite à nos interventions.

Thanks a lot for the very interesting (and extremely useful) workshop. [...] We are back in [our lab] and already thinking of all the data we could analyze using the skills we learned at the workshop.

L'aspect utile des méthodes est très souvent mis en avant et le fait que nous associons systématiquement des séances pratiques (parfois selon le principe *Bring your own data*) rend les participants confiants quant à leur possibilité d'utiliser le logiciel de façon autonome, sinon, ils savent également que le « service après-vente »est assuré grâce à l'adresse de contact `mixomics@math.univ-toulouse.fr`.

Une de mes satisfactions liées aux actions de formation que nous avons menées réside dans le fait que certains participants sont, à leur tour, devenus intervenants. L'exemple d'Olivier Chapleur, chercheur à l'Institut national de Recherche en Sciences et Technologies pour l'Environnement et l'Agricuture (Irstea) est certainement le plus révélateur. Olivier a participé à une formation en 2016. Nous avons gardé le contact et il est venu à Toulouse pour que l'on travaille ensemble de février à mai 2017. Il a ensuite participé au *Mixomics Advanced Workshop* que nous avons organisé à Toulouse en octobre 2017. Nous le

10. cran.r-project.org

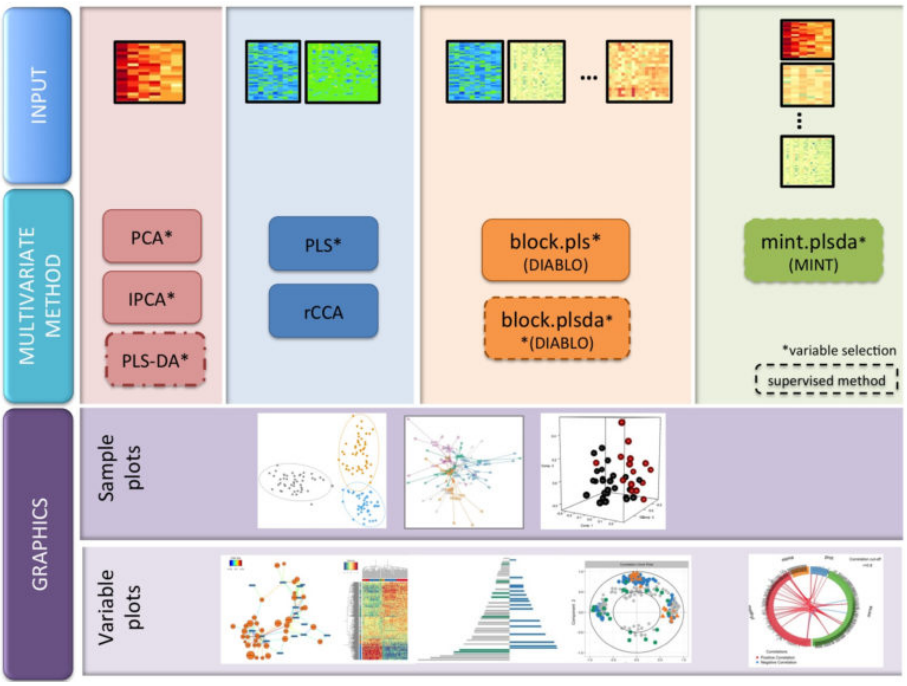

FIGURE 3.11 – Représentation schématique du contenu du package `mixOmics`.

sollicitons maintenant pour intervenir en formation. Et les recherches menées en commun ont abouti à une publication courant 2018 [72].

Pour conclure au sujet de `mixOmics`, j'évoque également le volet « Communication »de ce projet. Nous avons en effet quelques outils promotionnels à notre disposition : un logo dont la forme rappelle un réseau pour caractériser les liens que l'outil cherche à nouer entre différents jeux de données, des *flyers*, une petite cuillère que nous distribuons aux participants aux formations, formations durant lesquelles les intervants revêtent la tenue estampillée `mixOmics`. Plus scientifiquement, ces éléments sont complétés par une palette de couleur *color blind friendly* propre au package. Cette palette est utilisée par défaut dans les représentations graphiques et elle fournit en quelque sorte une signature du package.

FIGURE 3.12 – Les outils promotionnels du package `mixOmics` : un logo, une plaquette de présentation, une cuillère pour les participants aux formations, la tenue des intervenants et une palette de couleur dédiée.

3.3.4 Deux cas d'études

J'ai contribué à la mise en œuvre des méthodes évoquées précédemment dans le cadre de nombreux projets de recherche. Certains ont abouti à des publications dans des domaines variés. Par exemple, une étude visant à mettre en relation des notes données par un panel de goûteurs et des mesures physico-chimiques a été publiée dans la revue *Meat Science* [18]. Plus récemment, c'est dans le domaine de la micro-biologie que nous nous sommes intéressés à la mise en relation de données d'expression de gènes avec des mesures bio-physiques et bio-chimiques acquises par microscopie à force atomique [80]. Enfin, l'article [72] paru dans la revue *Water research* présente un travail intégrant une analyse de données

visant à mieux comprendre les phénomènes en jeu dans les processus microbiens intervenant dans le décomposition de déchets organiques.

J'ai choisi de présenter plus en détail deux études qui couvrent chronologiquement mes activités dans ce domaine. En effet, l'étude `nutrimouse` est la première qui a suscité l'intérêt de quelques membres de l'équipe de statistique et probabilités de l'Institut de Mathématiques de Toulouse (autour notamment d'Alain Baccini et de Philippe Besse) pour la problématique de l'intégration de données. Elle était menée par Pascal Martin et Thierry Pineau, de l'unité Inra devenue ToxAlim [11] et s'est concrétisée entre autres par l'article [41]. La seconde concerne des travaux toujours en cours en biologie végétale ; elle m'a permis de me confronter pour la première fois à des données multi-blocs et s'est concrétisée par un article [27] disponible pour le moment sur le serveur bioRxiv [12].

Relations entre expression de gènes et acides gras hépatiques dans une étude de nutrition chez la souris

Dans le cadre d'une étude de nutrition chez la souris, 40 animaux sont répartis selon un plan complet équilibré à 2 facteurs : génotype à 2 modalités (`WT` et `PPAR`α) et régime à 5 modalités (`ref, coc, sun, lin, fish`) à 4 répétitions. Les 5 régimes diffèrent par leur teneur en acides gras. Ainsi, le régime `ref` est considéré comme un régime de référence, il est composé d'un mélange 50/50 d'huile de maïs et d'huile de colza. Le régime `coc` est à base d'huile de coco hydrogénée et représente une régime à base d'acides gras saturés. Le régime `sun`, à base d'huile de tournesol, est une régime enrichi en acides gras ω_6. Les régimes `lin` et `fish`, respectivement composés d'huile de lin pour l'un et huiles de maïs, de colza et à base de poisson pour l'autre, sont des régimes enrichis en acides gras ω_3. Le projet visait à mieux comprendre les réactions de l'organisme de la souris (en tant qu'animal modèle) et du foie en particulier face à des régimes alimentaires aux compositions en acides gras variées. C'est un sujet d'intérêt majeur en termes de santé publique compte tenu de l'évolution des habitudes alimentaires susceptibles d'être à l'origine de maladies chroniques (obésité, diabète, maladies cardio-vasculaires).

Pour ces 40 souris, nous disposons de deux ensembles de données, auxquelles nous faisons maintenant référence sous le terme `nutrimouse`. Ils sont constitués de l'expression de 120 gènes d'une part et de la concentration de 21 acides gras d'autre part, le tout au niveau du foie. La question posée est la suivante : *existe-t-il des liens entre les variations de certains acides gras et les variations de l'expression de certains gènes ?*

Face à cette question, l'analyse des corrélations canoniques ne peut pas être appliquée car le nombre d'observations (40) est inférieur au nombre de variables (120+21). Face à ce constat, des développements méthodologiques ont dû être entrepris. Comme je l'ai évoqué précédemment, c'est le travail de thèse d'Ignacio González [38] qui a permis de fournir des éléments de réponse à cette question en développant l'analyse des corrélations canoniques régularisée (ACCR). C'est la mise en œuvre de l'ACCR sur les données `nutrimouse` (et sur d'autres données publiques que je ne présente pas ici) qui est décrite dans [41].

Biological data produced by high throughput technologies are becoming more and more abundant and are arousing many statistical questions. This paper addresses one of them ; when gene expression data are jointly observed with other variables with the purpose of highlighting significant relationships between gene expression and these other variables. One relevant statistical method to explore these relationships is Canonical Correlation Analysis (CCA). Unfortunately, in the context of postgenomic data, the number of variables (gene expressions) is usually greater than the number of units (samples) and CCA cannot be directly performed : a regularized version is required. We applied regularized CCA on data sets from two different studies and show that its interpretation evidences both previously validated relationships and new hypothesis. From the first data sets (nutrigenomic study), we generated interesting hypothesis on the transcription factor pathways potentially linking hepatic fatty acidsand gene expression. From the second data sets (pharmacogenomic study on the NCI-60 cancer cell line panel), we identified new ABC transporter candidate substrates which relevancy is illustrated by the concomitant identification of several known substrates. In conclusion, the use of regularized CCA is likely to be relevant to a number and a variety of biological experiments involving the generation of high throughput data. We demonstrated here its ability to enhance the range of relevant conclusions that can be drawn from these relatively expensive experiments.

Résumé de l'article [41]

11. www6.toulouse.inra.fr/toxalim
12. www.biorxiv.org

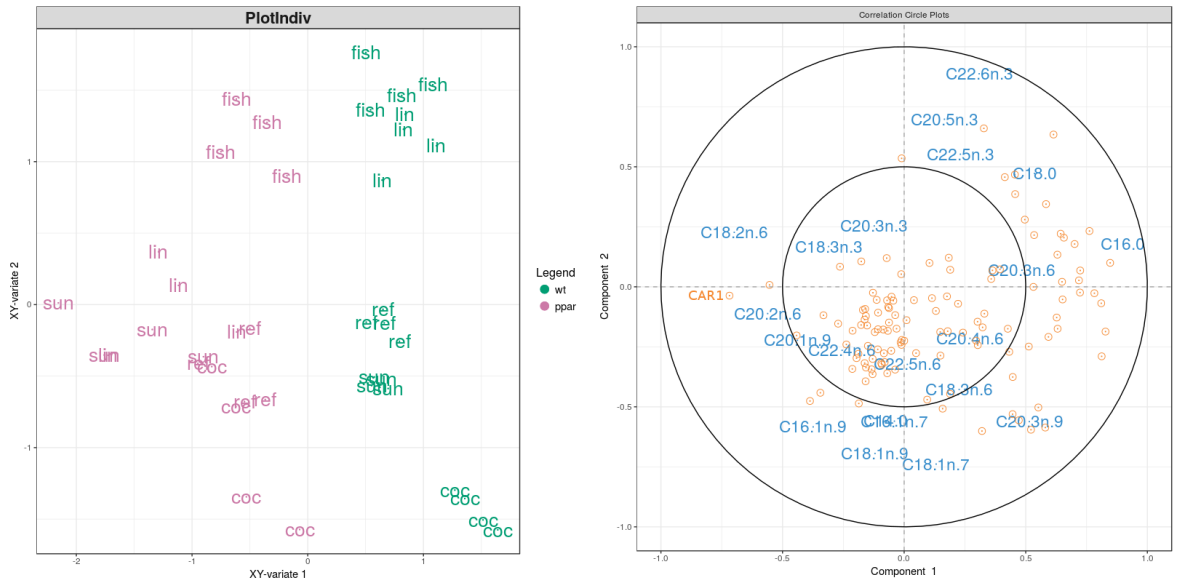

FIGURE 3.13 – Représentations sur les deux premières composantes obtenues par analyse des corrélations canoniques régularisée : à gauche, représentation des individus identifiés par leur génotype (`WT` en vert et `PPAR`α en rose) et leur régime (`ref, coc, sun, lin, fish`) ; à droite, représentation des variables (lipides identifiés par leur nom, gènes représentés par des cercles oranges, sauf le gène CAR1 identifié par son nom). Ces figures apparaissent avec un visuel différent dans [41].

Sans rentrer dans les détails de l'interprétation biologique que l'on peut retrouver dans [66], signalons par exemple le fait que les individus WT se séparent mieux que les individus PPARα selon le régime alimentaire qu'ils ont suivi (Figure 3.13 gauche). Un autre élément marquant de cette analyse réside dans la position particulière du gène CAR1 (figure 3.13 droite), comme gène ayant la coordonnée la plus négative sur la première composante canonique. Il est associé à une accumulation de l'acide linoléïque (C18 :2ω6) chez les souris `PPAR`α ; ce qui est un autre résultat biologique important discuté dans [66].

Le jeu de données `nutrimouse` fait désormais partie intégrante du package `mixOmics` ; il nous sert régulièrement de jeu test pour évaluer de nouvelles méthodes et c'est le jeu de données que nous utilisons de façon privilégiée en formation.

Ce travail illustre le fait que des questions biologiques peuvent être à l'origine de développements méthodologiques en statistique. Cela est possible à condition de savoir se parler entre biologistes et statisticiens et c'est exactement la mission que j'assure dans des projets pluri- ou interdisciplinaires : permettre à des scientifiques de domaines différents de se parler et de se comprendre.

Analyse multi-omiques en biologie végétale

Le travail présenté ici a été entrepris dans le cadre du projet WallOmics. Ce projet co-financé par la Région Midi-Pyrénées (devenue Occitanie) et l'Université Fédérale Toulouse Midi-Pyrénées vise à étudier la réponse des plantes au réchauffement climatique. Le modèle de la plante utilisé pour cela est *Arabidopsis thaliana*. Cette plante est largement utilisée comme modèle pour les études de biologie moléculaire et de grandes quantités de ressources technologiques et de données expérimentales sont disponibles (voir la base de données CATdb dévà évoquée). Elle est de plus distribuée dans le monde entier, dans des conditions de croissance différentes en température, altitude et humidité. L'objectif du projet réside dans l'intégration de données hétérogènes issues de l'écologie, de la génétique, des techniques 'omiques ainsi que des données phénotypiques afin d'évaluer les réponses des plantes face au réchauffement climatique en termes de composition de la paroi cellulaire, de l'expression des gènes et de la fonction des protéines de la paroi cellulaire.

La thèse d'Harold Duruflé [25] [13], que j'ai contribué à encadrer, se positionne dans ce projet. Après une étude pilote concernant deux écotypes `Col` et `Sha` [26], nous nous sommes intéressés à des données plus volumineuses issues de 5 écotypes : 4 provenant des Pyrénées `Roch, Grip, Hern, Hosp`, vivant à des altitudes différentes et un écotype de référence, `Col`, vivant en Pologne à une altitude relativement basse. Des graines issus de ces 5 écotypes ont poussé dans des chambres de culture à deux températures différentes. La disposition des différents pots dans lesquels ont poussé les plantes a été le sujet de discussion en amont du projet. Il convenait d'éviter tout biais potentiels liés à d'éventuels gradients d'éclairage, de

13. Récompensé du Prix Henri Gaussen 2018 de l'Académie des Sciences, Inscriptions et Belles-Lettres de Toulouse.

température, de courant d'air... Le fait de changer régulièrement les pots de place dans les barquettes étant trop complexe à gérer en pratique, c'est un changement d'orientation des barquettes contenant plusieurs pots qui a été retenue. Le déroulement de l'expérience est ainsi le résultat d'un compromis ménageant les contraintes techniques et les biais statistiques éventuels. Il était primordial de pouvoir discuter de ce sujet en amont, vu le volume et le coût représentés par les données qui allaient être générées sur différentes plateformes technologiques. En effet, pour l'ensemble de ces plantes, quatre ensembles de données omiques ont été recueillis :

- **Phénome** : 9 variables phénotypiques ont été mesurées pour la tige ou pour la rosette des plantes (masse, diamètre, nombre de feuilles...).
- **Metabolome** : l'identification et la quantification de 7 sucres ont été menées dans la paroi des cellules.
- **Proteome** : l'identification et la quantification de protéines pariétales ont été réalisées par spectrométrie de masse pour aboutir aux concentrations de 364 protéines pour la rosette et 414 pour la tige.
- **Transcriptome** : l'expression de 19763 transcrits pour la rosette et 22570 pour la tige a été évaluée selon la technologie de séquençage de l'ARN.

Se confronter à des données aussi volumineuses et variées étant quelque chose de relativement nouveau dans la communauté végétaliste, nous avons décidé de décrire dans une publication le protocole d'étude statistique que nous avons élaboré [27] ; ma position de dernier auteur dans cet article atteste de mon rôle de superviseur dans ce travail. Cet article est pour le moment disponible sur le serveur de pré-publications bioRxiv en l'attente de la validation de certains résultats biologiques qui feront l'objet d'une autre publication plus orientée biologie. Nous envisageons, en effet, de soumettre ces deux articles simultanément dans la même revue selon le principe des articles *batck-to-back* ; chacun abordant deux aspects différents d'un même projet.

The high-throughput data generated by new biotechnologies used in biological studies require specific and adapted statistical treatments. In this work, we propose a novel and powerful framework to manage and analyse multi-omics heterogeneous data to carry out an integrative analysis. We illustrate it using the package `mixOmics` for the R software as it specifically addresses data integration issues. Our work also aims at confronting the most recent functionalities of `mixOmics` to real data sets because, even if multi-block integrative methodologies exist, they still have to be used to enlarge our know how and to provide to biologists an operational framework. Natural populations of the model plant Arabidopsis thaliana are employed in this work but the framework proposed is not limited to this plant and can be deployed whatever the organisms of interest and whatever the biological question. Four omics data sets (phenomics, metabolomics, cell wall proteomics and transcriptomics) have been collected, analysed and integrated in order to study the cell wall plasticity of plants exposed to a sub-optimal temperature growth condition. The methodologies presented starts from basic univariate statistics and leads to multi-block integration analysis, and we highlight the fact that each method is associated to one biological purpose. To make use of this powerful framework led us to novel biological conclusions that could not have been shown using standard statistical approaches.

Résumé de l'article [27]

La richesse de ces données fait que nous pouvons mettre en œuvre et illustrer la plupart des méthodes statistiques exploratoires. C'est l'idée que nous souhaitons faire passer avec la figure 3.14 présentée dans l'article [27].

Les cercles et ellipses fléchés sur la droite de la figure 3.14 illustrent la nécessité des va-et-vient entre les différentes méthodes. Chaque méthode est susceptible d'apporter un éclairage nouveau sur les résultats d'une autre méthode mise en œuvre précédemment et peut ainsi nécessiter de renouveler certaines analyses en mettant de côté quelques individus ou variables susceptibles de masquer des phénomènes d'intérêt. C'est également en ce sens que je me méfie des formules du type : mettre en place un *pipeline* d'analyse, comme je l'ai aussi évoqué dans le chapitre 2. Cette terminologie peut laisser penser que l'on peut démarrer une analyse et en sortir après un déplacement rectiligne, sans retour en arrière. Je ne crois pas à cela dans le cadre de l'analyse de données surtout dans le domaine de la recherche.

Sans revenir sur l'ensemble des analyses statistiques qui ont été menées pour cette étude ni détailler les résultats biologiques qui sont encore en cours d'exploitation, je présente ici quelques résultats d'analyses multi-blocs en partie menées par Merwan Selmani, étudiant dans le département Génie Mathématique et Modélisation de l'Insa de Toulouse, pendant son stage de 4ème année.

La représentation des individus sur les deux premières composantes multi-blocs (non présentée ici) illustre la nette distinction des deux groupes de plantes ayant poussé à des températures différentes et

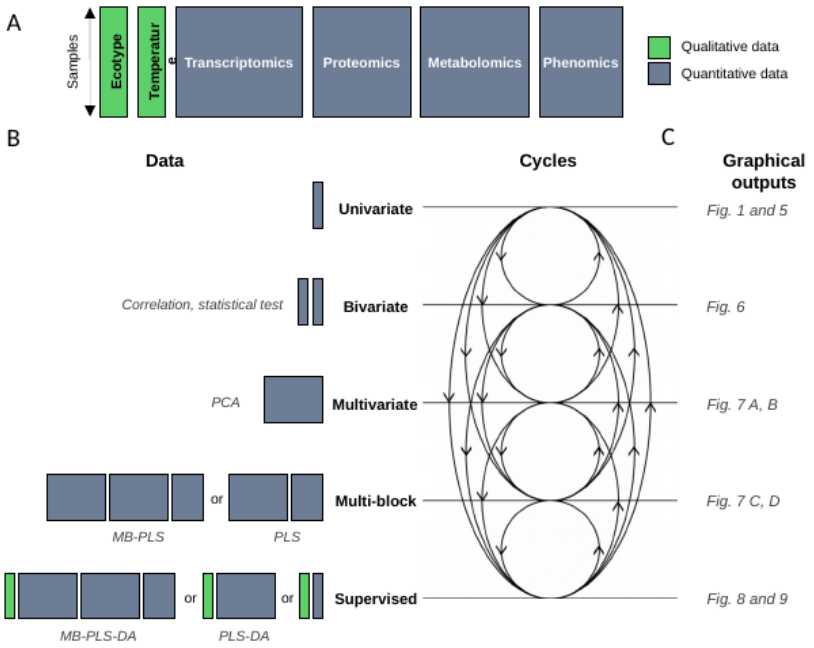

Figure 4. One purpose, one method to analyse qualitative and quantitative blocks. A) Schematic representation of the different blocks (or data sets) co-analysed in this study. The samples are represented in rows and the variables in columns. B) Schematic overview of the methods implemented represented by cycles within an integrative study. C) Examples of graphical outputs detailed in the results section. PCA: Principal Component Analysis; MB: Multi-Blocs; PLS: Partial Least Squares regression; DA: Discriminant Analysis. Qualitative and quantitative blocks are represented in green and grey respectively.

FIGURE 3.14 – Représentation schématique des données et des analyses statistiques à mettre en œuvre sur les données du projet WallOmics. Figure extraite de [27].

ceci dans les 4 sous-espaces engendrés par chaque ensemble de variables. Du côté des variables, on ne tire rien de plus que le fait qu'il est nécessaire d'associer une stratégie de sélection de variables à l'analyse multi-blocs pour permettre une interprétation des résultats. Ce sont les résultats d'une telle analyse qui sont présentées dans la figure 3.15.

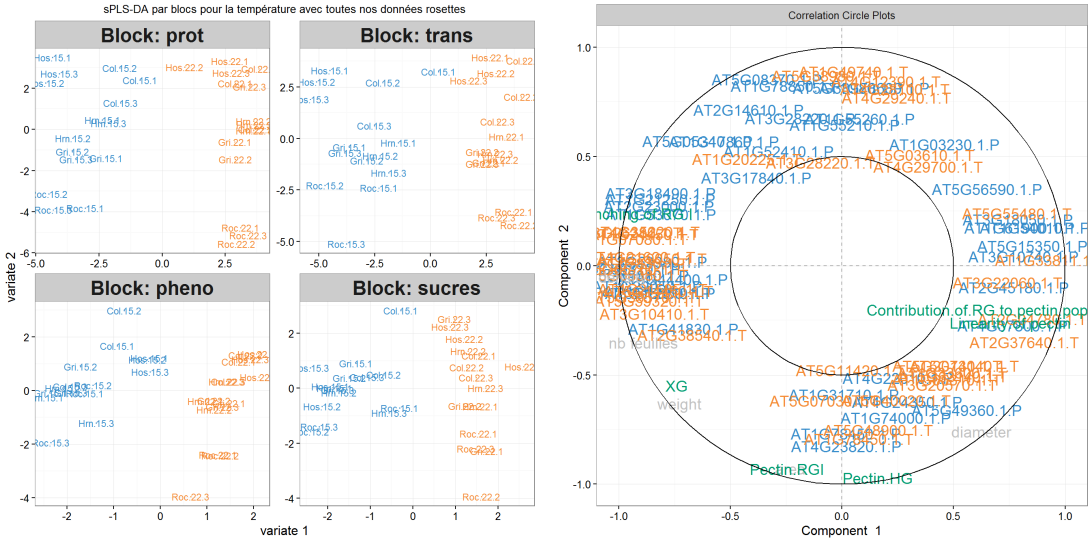

FIGURE 3.15 – Représentation des individus (à gauche) et des variables (à droite) à l'issue d'une analyse multi-blocs parcimonieuse supervisée par le facteur température.

Concernant les représentations des individus, la discrimination des deux groupes de plantes selon la température est toujours nettement visible malgré le recours à un nombre nettement inférieur de variables. Du côté des variables, donc, certaines associations entre des variables issues de chacun des blocs commencent à se dessiner.

Afin de mieux les identifier, on peut avoir recours à d'autres représentations graphiques. Par exemple, la figure 3.16 propose une représentation de type *heatmap* pour visualiser des groupes de variables issues des 4 blocs associées à des groupes d'individus.

Le travail visant à interpréter biologiquement certains des éléments mis en évidence par ces analyses intégratives est très exigeant et il est encore en cours au moment où je rédige ce mémoire.

Même si les méthodes existent, même si des logiciels permettent de les mettre en œuvre sur des données, le chemin peut être encore long avant d'en extraire des connaissances biologiques nouvelles. C'est également sur la base de ces interprétations que les méthodes existantes pourront être remises en cause et que d'autres pourront être développées. Cela étant, les travaux que nous avons menés dans ce cadre et que nous présentons dans [27] établissent une preuve de concept illustrant que l'on dispose des outils statistiques permettant de mener à bien un projet de biologie intégrative.

3.3.5 Conclusion

Les travaux de cette section, tout comme ceux évoqués dans le domaine de la recherche d'information (section 3.2), sont par nature interdisciplinaires. Ils se structurent autour du triplé biologie-bioinformatique-biostatistique afin de garder le rythme des évolutions technologiques et des développements méthodologiques sans perdre de vue l'interprétation biologique des résultats. Ils s'insèrent dans un domaine en plein essor et les sollicitations sur le sujet ne manquent pas. J'ai par exemple été invité récemment pour participer au Nordic Precision Medicine Forum [14] qui aura lieu en mars 2019. Contribuer à étendre le domaine d'application des méthodes d'intégration de données au domaine bio-médical et plus spécifiquement à la médecine personnalisée est un vrai défi que j'envisage avec beaucoup de motivation.

14. precisionmedicineforum.com/nordic-2019

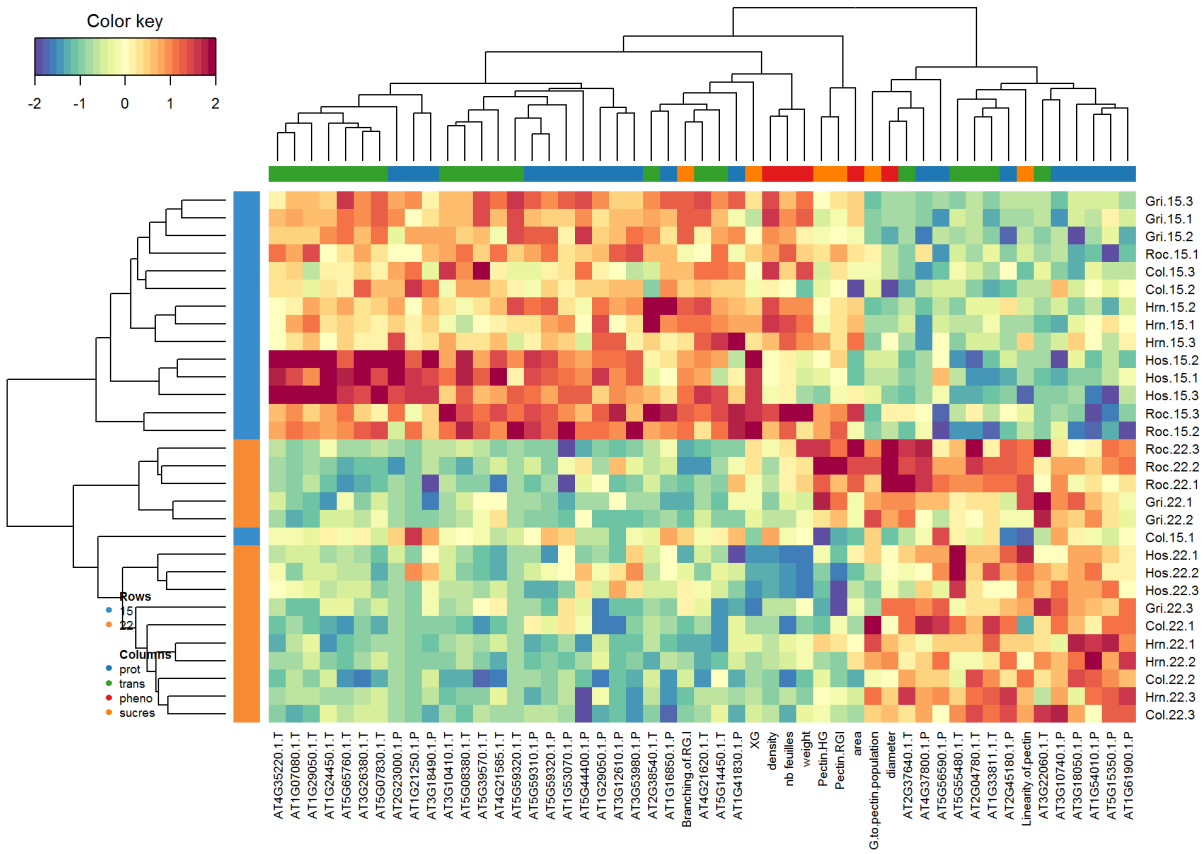

FIGURE 3.16 – Représentation de type *heatmap* des résultats d'une analyse multi-blocs parcimonieuse supervisée par le facteur température.

Chapitre 4

Activités de support à la recherche

Après les premiers chapitres plus techniques, je souhaite donner ici quelques éléments complémentaires appuyant ma volonté de prétendre à une habilitation à diriger des recherches. J'y évoque d'abord mes activités d'encadrement et mes responsabilités dans le cadre de projets de recherche. Je présente ensuite mes activités en enseignement et formation ainsi que mes contributions à l'animation scientifique. Après cela, je consacre une partie au rôle d'expert que j'assure dans différents contextes. Et je termine par un sujet qui me tient à cœur, la vulgarisation scientifique, en présentant quelques actions que je mène dans ce cadre.

4.1 Encadrement de la recherche

Mon expérience dans l'encadrement de la recherche se caractérise dans les conditions habituelles de cet exercice, à savoir l'encadrement de personnes et la responsabilité scientifique de projets de recherche. J'illustre ces deux aspects respectivement dans les paragraphes suivants.

4.1.1 Encadrement, accompagnement et suivi de personnes

J'ai été officiellement co-directeur de deux thèses en informatique avec Josiane Mothe, professeure à l'Irit. Il s'agit des thèses de Léa Laporte [51] et Anthony Bigot [9]. Les deux thèses avaient en commun une interaction nécessaire entre le domaine de la recherche d'information, domaine d'origine des projets, et celui des mathématiques.

Thèse de Léa Laporte

La thèse de Léa Laporte, soutenue le 18 novembre 2013 a été co-financée par la Région Occitanie et la société Nomao [1]. Elle visait à développer des méthodes de sélection de variables dans le cadre de l'apprentissage d'ordonnancement (*Learning-to-rank*). Les travaux de Léa ont abouti à plusieurs publications [52, 54, 53]. Léa est actuellement maître de conférence à l'Insa de Lyon (Laboratoire d'InfoRmatique en Image et Systèmes d'information, LIRIS). Si je remonte un peu dans le temps, ce projet de thèse a germé lorsque j'ai encadré le stage de Léa lorsqu'elle était en 4ème année à l'Insa de Toulouse dans le département Génie Mathématique et Modélisation. J'avais remarqué ses grandes qualités, notamment rédactionnelles, et lorsqu'elle m'a dit qu'elle souhaitait poursuivre ses études en thèse, j'ai commencé à en parler à Josiane avec qui nous avions des projets en cours, et le projet de thèse s'est finalement concrétisé en collaboration avec la société Nomao. Dans ce contexe, assez similaire à celui d'une thèse sous Convention Industrielles de Formation par la Recherche (CIFRE), nous, encadrants, avons dû rester vigilants à l'indispensable équilibre à maintenir entre, d'une part les besoins et les attentes de l'entreprise, et d'autre part, les exigences académiques liées à l'obtention d'un diplôme de doctorat. La suite de la carrière de Léa en tant qu'enseignant-chercheur dans une école d'ingénieur semble montrer que nous n'avons pas échoué dans cet exercice.

Raconter cette histoire dans le cadre de ce mémoire illustre le fait que je suis conscient de la nécessité d'intéresser relativement tôt des étudiants et des étudiantes à une éventuelle poursuite de leurs études jusqu'au doctorat. Car cela permet d'anticiper et d'assurer plus facilement la conjonction entre un sujet de recherche, un-e étudiant-e et un financement.

1. fr.nomao.com

Thèse d'Anthony Bigot

La thèse d'Anthony, soutenue le 7 octobre 2014, s'est déroulée dans un contexte différent, avec notamment un financement académique. Ses travaux visaient à étudier les méthodes de fusion de résultats de systèmes de recherches d'information. Ils ont nécessité d'importantes ressources de calcul du méso-centre de calcul Calmip et ont abouti à plusieurs publications [10, 11, 12] ainsi qu'à la constitution d'un jeu de données test que nous avons continué à exploiter après la thèse.

À l'issue de sa thèse, Anthony n'a pas souhaité poursuivre dans le monde de la recherche académique ; il est aujourd'hui consultant *Business Intelligence* pour la société Halys Conseil [2].

Contributions significatives à l'encadrement de thèse

Harold Duruflé : la thèse d'Harold que j'ai déjà évoquée précédemment (partie 3.3.4) a été dirigée par Christophe Dunand du Laboratoire de Recherche en Sciences Végétales et Philippe Besse de l'IMT. J'ai pris une part active dans l'encadrement en supervisant les analyses statistiques effectuées sur les données du projet et en contribuant à la formation d'Harold en statistique et en R. Vers la fin de sa thèse, Harold a même participé, en tant qu'intervenant, à une formation `mixOmics`.

Céline Bougel : la thèse de Céline Bougel a débuté en octobre 2017 ; elle est dirigée par Sandrine Andrieu (UMR Inserm 1027, Épidémiologie et analyses en santé publique : risques, maladies chroniques et handicaps), Nicolas Savy et Philippe Saint-Pierre de l'IMT. Elle porte sur l'évolution des fonctions cognitives de patients dans le cadre d'essais de prévention de la maladie d'Alzheimer. La première année a été consacrée à l'analyse de données issues d'un essai de prévention de la maladie d'Alzheimer afin de dégager des profils d'évolution des fonctions cognitives. Mon rôle consiste à conseiller Céline sur la mise en œuvre de méthodes de classification de courbes et à valider avec elle et une psychologue clinicienne les résultats obtenus.

Participation à des comités de thèse

Dans le registre « accompagnement, suivi de personnes », je souhaite également mentionner mon implication dans huit comités de thèse (six listés ci-dessous en plus des deux évoqués précédemment). Ces comités dont le rôle et le fonctionnement peut varier selon les écoles doctorales, a globalement pour but de s'assurer que la thèse se déroule bien et, dans le cas contraire, de discuter d'alternatives pour la poursuite des travaux de recherche.

J'ai ainsi participé au comité de thèse des doctorantes et doctorants listés ci-dessous. J'y mentionne le titre, parfois provisoire, de leur thèse afin d'illustrer à la fois les différences (domaines applicatifs) et les points communs (génération de données) de ces différents projets.

- Marine Deshors (période 2016-2018) : *Design of optimal enzymatic cocktails for deconstruction and degradation of raw materials from plant origin by the development and the use of in and ex-situ physical measurements and ex-situ biochemical measurements*
- Lucas Marmiesse (2014-2016) : Modélisation mathématique du réseau transcriptionnel contrôlé par MYB30 et MYB96, deux facteurs de transcription impliqués dans la réponse de défense de la plante modèle *Arabidopsis thaliana*
- Mathilde le Sciellour (en cours) : Mise en évidence de relations entre la composition du microbiote intestinal de l'hôte et l'efficacité alimentaire du porc
- Lucie Khamvongsa-Charbonnier (en cours) : Integrative analysis of genomic regulation combining cistrome, epigenome and transcriptome
- Mathieu Arnal (en cours) : Développement d'une évaluation génomique pour l'analyse de données longitudinales : Application aux contrôles élémentaires chez les caprins laitiers
- Franck Boizard (en cours) : Du protéome urinaire vers le protéome tissulaire rénal : analyse des réseaux d'interaction protéines-protéines

Dans la plupart de ces cas, j'ai été sollicité pour apporter mon expertise dans les méthodes statistiques pour l'analyse de données biologiques, avec une attention particulière portée sur les problématiques d'intégration de données.

Je ne positionne pas, bien entendu, une participation à un comité de thèse au même niveau qu'une direction de thèse, mais je sais que ces expériences me permettent d'avoir un regard critique sur le déroulement d'un thèse et contribuent à ma volonté de prétendre à une habilitation à diriger des recherches.

2. www.halysconseil.com

Ingénieurs et contrats post-doctoraux

En plus de l'accompagnement de travaux de thèse, j'ai contribué à l'encadrement de personnes recrutées pour des durées déterminées soit dans le cadre de contrats post-doctoraux, soit sur des profils d'ingénieur :

— Alexandre Eveillard recruté sur un contrat post-doctoral dans le cadre du projet Plast-Impact. Ce projet qui s'est déroulé de 2006 à 2009 visait à étudier les impacts métabolique et endocrinien de deux contaminants de la chaîne alimentaire issus de la plasturgie : le Bisphénol A et le DEHP (diéthylhexyl phtalate). Il a été financé par l'Agence Nationale de la Recherche (ANR) dans le cadre du Programme National de Recherches en Alimentation et Nutrition Humaine.

— François Bartolo recruté en tant qu'ingénieur d'études dans le cadre du projet SynTHACS : Biologie synthétique pour la synthèse de molécules chimiques à haute valeur ajoutée à partir de ressources carbonées renouvelables. Ce projet qui s'est déroulé de 2012 à 2016 était financé par l'ANR dans le cadre du Programme Biotechnologies et Bioressources des Investissement d'Avenir.

— Myriam Badawi recruté en tant qu'ingénieur de recherche dans le cadre du projet Systomics *Systems approaches to study microbial consortia : integrating meta-omics data to elucidate the functioning of lignocellulolytic microbial consortia* (2016-2018) mené en collaboration avec le Laboratoire d'Ingénierie des Systèmes et des Bio-Procédés et financé par l'Inra via les meta-programmes *Meta-Omics and Microbial Ecosystems*.

Dans les trois cas, les personnes recrutées avaient des profils de biologistes avec des affinités plus ou moins marquées pour la biostatistique et/ou la bioinformatique. En travaillant dans le cadre de projets impliquant l'IMT et à mon contact, ils ont acquis de nouvelles compétences en bio-statistique ce qui a contribué à faciliter leur insertion professionnelle.

Les travaux d'Alexandre Eveillard ont porté sur l'analyse de données métabolomiques disponibles sous forme de spectres. Il a pu ainsi se former à de nouvelles techniques mathématiques notamment les transformées en ondelettes et les analyses statistiques supervisées. Ses travaux, consolidés ensuite par Ignacio Gonález, ont donné lieu à une publication [42]. Alexandre a mis fin à son contrat avant son terme afin d'accepter une offre d'emploi dans un société attirée par son profil de biologiste avec des compétences avérées en statistique.

François Bartolo a souhaité rejoindre l'IMT afin de compléter son profil de biologiste-bioinformaticien par une montée en compétences en statistique. Il a assuré l'analyse des données biologiques générées dans le cadre du projet SynTHACS et a également contribué au développement du package `mixOmics`. À l'issue de son contrat, il n'a au aucun mal à trouver un emploi dans une société de conseil et service en biostatistique vivement intéressée par sa polyvalence en bioinformatique et biostatistique.

Myriam Badawi a travaillé dans le cadre du projet Systomics sur un profil hybride « post-doc / ingénieur de recherche ». Elle a notamment assuré l'analyse statistique des données biologiques et s'est ainsi formée à des méthodes statistiques comme les méthodes PLS ou les modèles mixtes. Elle a été recrutée en 2017 en tant que maître de conférence à l'Université du Maine (Le Mans) dans le département de Biologie et Géosciences.

À travers ces trois exemples, je souhaite mettre en évidence le fait que j'ai su accompagner des collègues dans des situations différentes, avec des objectifs différents, afin de compléter leur formation et de faciliter la poursuite de leur carrière professionnelle.

4.1.2 Responsabilité scientifique dans des projets de recherche

L'autre aspect de l'encadrement de la recherche que je souhaite aborder ici est la responsabilité scientifique que j'assume pour mon laboratoire dans plusieurs projets scientifiques collaboratifs. Dans ce cadre, je prends en charge les actions habituelles liées à ce rôle : contribution à la rédaction du projet, définition d'un budget pour mon équipe, recrutement et encadrement de personnes sur des contrats à durée déterminée.

Voici 3 projets pour lesquels j'ai assuré la responsabilité scientifique du partenaire Institut de Mathématiques de Toulouse.

— Voice4PD-MSA – Differential diagnosis between Parkinson's disease and Multiple System Atrophy using digital speech analysis, 2016-2020. Financement : Agence Nationale de la Recherche (ANR), AAP Technologies pour la santé.

— WallOmics – Production et traitement de données « omics » hétérogènes en vue de l'étude de la plasticité de la paroi chez des écotypes pyrénéens de la plante modèle A. thaliana, 2014-2017. Financement : Région Midi-Pyrénées et Université Fédérale Toulouse Midi Pyrénées. Co-responsabilité avec Philippe Besse.

— « SYNTHACs » – Biologie synthétique pour la synthèse de molécules chimiques à haute valeur ajoutée à partir de ressources carbonées renouvelables, 2012-2016. Financement : Agence Nationale

de la Recherche (ANR), Investissement d'avenir, Programme Biotechnologies et Bioressources. Co-responsabilité avec Philippe Besse.

4.2 Formation

4.2.1 Pour des publics variés

Une des clés de la réussite d'un projet interdisciplinaire réside dans le fait de pouvoir communiquer. Et donc pour acquérir un minimum de vocabulaire commun, les actions de formation sont indispensables. Cela fait maintenant plus d'une dizaine d'années que je contribue à des actions de formation à destination de publics variés. Ma plus grande contribution se fait dans le cadre de la formation professionnelle. Je suis ainsi intervenu à de nombreuses reprises pour former, essentiellement mais pas seulement, des biologistes à l'analyse statistique et au logiciel R. Pour une majorité de ces formations, j'ai assuré non seulement l'intervention auprès des participants mais aussi la définition du contenu de la formation et la rédaction de supports de formation.

Récemment, j'ai commencé à impliquer de nouveaux collègues afin de les former eux-mêmes en tant que formateur et de pouvoir leur passer la main sur des contenus qui me stimulent un peu moins. Ainsi, lorsque j'entreprends une action de formation, je le fais toujours avec un très grand plaisir et une grande motivation et les retours des participants sont en général très bons.

Plus ponctuellement, j'assure quelques interventions auprès d'étudiants en formation initiale, souvent en biologie. Afin de ne pas m'installer dans une certaine routine, j'ai accepté en 2015 d'assurer un enseignement à des élèves ingénieurs en 3ème année de l'Institut Supérieur de l'Aéronautique et de l'Espace (Isae-Supaero [3]) ; c'est un enseignement que j'assure toujours. Il s'agissait pour moi d'un défi car c'est un type de public que je connaissais pas. Ce sont globalement des étudiants qui comprennent très bien les mathématiques et avec des inclinaisons vers la physique. Il a fallu que j'adapte ma pédagogie pour ce public assez radicalement différent de mes publics habituels plus frileux envers les mathématiques. Par exemple, plutôt que de leur faire analyser des données transcriptomiques issues d'une étude de nutrition chez la souris, je leur ai proposé un sujet de bureau d'études autour de la mise en œuvre de méthodes statistiques appliquées à l'imagerie : segmenter une image par des méthodes de classification, compresser une image par une analyse en composantes principales. À la lecture des rapports, j'ai été ravi de constater plusieurs fois qu'ils avaient apprécié l'exercice et la conclusion de l'un d'entre eux se terminaient par ces mots : *Ce projet nous a permis de mettre en œuvre les méthodes de statistique exploratoire étudiées en cours, dans un cadre différent nous permettant d'avoir un bref aperçu du domaine de l'image. Nous avons pu notamment appréhender le lien entre les analyses et les phénomènes qu'elles mettent en évidence sur les images, et donc l'apport de l'analyse statistique sur de tels objets. Et en prime, c'était fun.* Je reconnais que le *en prime, c'était fun*, a été pour moi la cerise sur le gâteau !

Pour illustrer différemment le fait que je m'adapte au public, je présente ici différentes façons que j'utilise pour parler de la variance lors d'une introduction à la statistique. Ayant souvent constaté un blocage voire un rejet de la formule mathématique, je ne peux pas me contenter d'afficher la formule de la variance, de dire « c'est bon pour tout le monde » et de passer à la suite. Un des aspects de la formation consiste alors à lever ce blocage. Par exemple, une approche consiste à affirmer d'abord, de façon un peu provocatrice, qu'une formule mathématique est tellement plus simple à retenir que les mêmes choses écrites avec des mots. N'est-il pas plus simple de retenir que l'écart-type s'écrit $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ plutôt que de retenir la phrase *L'écart-type est la racine carrée de la moyenne des carrés des écarts à la moyenne* ? À cet instant-là, tout le monde n'est peut-être pas encore convaincu, alors on enchaîne en passant à la phase *Dessine moi une variance (et un écart-type).*

Pour cela, je me sers de la figure 4.1 que je présente petit à petit pour afficher d'abord, la moyenne, puis les écarts à la moyenne (des segments), puis les carrés des écarts à la moyenne (que dessine-t-on ? des carrés bien sûr), puis la moyenne des carrés des écarts à la moyenne (on constate au passage que la variance est un carré), et on termine avec un écart-type qui a le bon goût d'être un segment et donc qui s'exprime dans la même unité que les données (et c'est pour cela qu'on l'aime bien et qu'on le préfère souvent à la variance).

Ainsi, en décortiquant quelques formules comme celle de la variance ou celle associée au lissage spline avec la citation de H. Poincaré 2.2.2, j'arrive généralement à démystifier l'usage des formules mathématiques et à aborder la suite dans un contexte moins crispé.

C'est la même ligne directrice que je suis dans mes activités de recherche afin de pouvoir discuter et interagir avec des collègues experts dans leur domaine mais pas toujours à l'aise face aux mathématiques et à la statistique.

3. www.isae-supaero.fr

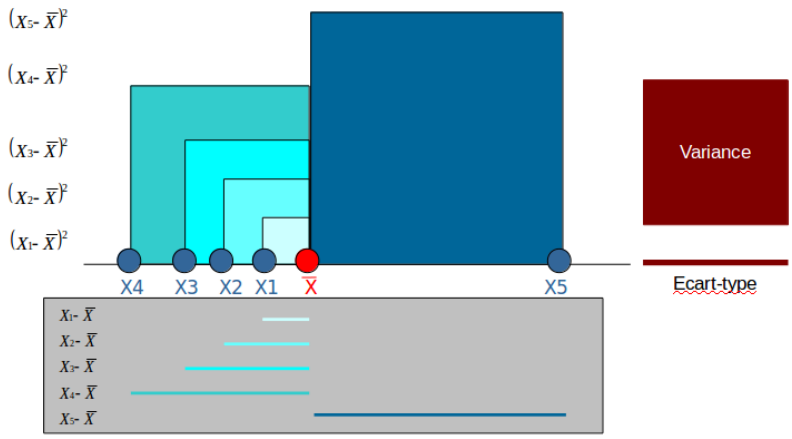

FIGURE 4.1 – Illustration du calcul de la variance et de l'écart-type.

4.2.2 Avec des objectifs différents

Selon le public et la durée, les objectifs d'une action de formation sont nécessairement à adapter. J'en évoque quelques-uns ici.

- **Comprendre le principe des méthodes statistiques**, et pas forcément la base mathématique sous-jacente. Il n'est pas nécessaire de mentionner à un public de biologistes par exemple, qu'une ACP peut se réaliser de façon équivalente soit par une décomposition en éléments propres de la matrice de covariance des données, soit par une décomposition en valeurs singulières de la matrice de données. En revanche, comprendre, par exemple, en quoi l'ACP est une méthode de réduction de dimension semble plus important. Je pense également indispensable de faire comprendre le principe d'un test statistique sans pour autant parler du lemme de Neyman-Pearson.
- **Connaître les conditions d'utilisation des méthodes**, afin de les utiliser de façon légitime et pas parce que d'autres articles les mentionnent.
- **Savoir se passer de méthodes statistiques**. Un autre aspect de la formation consiste aussi à faire passer le message, plagiant le slogan d'une campagne de l'Assurance Maladie (Les antibiotiques c'est pas automatique. Parlez-en avec votre médecin.) : *La statistique, c'est pas automatique. Parlez-en à un-e statisticien-ne.* Cela fait toujours réagir notamment lorsque je suis face à des médecins. C'est une idée assez similaire qui est exprimée dans l'article [20] qui se focalise sur les barres d'erreurs qui sont fréquemment ajoutées aux diagrammes en bâtons : *However, if n is very small (for example n = 3), rather than showing error bars and statistics, it is better to simply plot the individual data points.* J'ai ainsi fréquemment essayé de dissuader des collègues d'utiliser un test statistique car ils n'étaient pas dans les conditions requises pour l'effectuer avec certains des arguments développés dans cet article *An unhealthy obsession with p-values is ruining science*[4] ou encore dans le livre de Sylviane Gasquet-More [37] d'où est extrait le passage suivant : *Après nous avoir convaincu de leur objectivité fondamentale, il ne reste plus aux chiffres qu'à nous amener doucement à penser qu'ils en déterminent le monopole. Dès lors, une forme de hiérarchie gagne l'argumentation et le raisonnement : contenir quelques chiffres qualifie automatiquement votre discours, même si personne ne prend la peine de comprendre vraiment ce qu'ils signifient, voire même s'ils sont sans rapport avec le sujet traité ! A contrario, de ce fait, toute argumentation purement textuelle semble dépréciée [...] comme si le raisonnement et la rigueur ne pouvaient exister hors des chiffres.*
- **Mettre en œuvre avec un logiciel de statistique**. J'ai pris part à de nombreuses formations au logiciel R et au package `mixOmics` et une des premières difficultés consiste à convaincre des non spécialistes que les logiciels en ligne de commande (comme R) c'est très bien. Cela limite considérablement le risque d'erreurs, surtout au début, car si on ne connaît rien au langage, on ne peut pas faire d'erreur car on ne peut tout simplement rien faire. Cette affirmation est à modérer compte tenu du fait que l'on peut très bien copier-coller des commandes sans les comprendre, mais cela reste à mon avis limité par rapport à l'usage d'un logiciel avec une interface graphique plus conviviale. C'est une idée assez similaire qui est exprimée dans [57] : *Les logiciels accessibles et faciles à utiliser permettront une large diffusion des méthodes, mais donneront parfois lieu à des utilisations inconsidérées dans des domaines où une réflexion minutieuse et une grande prudence seraient de mise. La médiation des logiciels est un nouveau paramètre dont il faut tenir compte.*

4. www.vox.com/2016/3/15/11225162/p-value-simple-definition-hacking

— **Savoir interpréter les résultats**. Comme je l'ai déjà mentionné à plusieurs reprises, l'histoire ne commence pas et ne se termine pas avec l'analyse des données. Il convient donc d'interpréter les résultats des méthodes statistiques afin de fournir des éléments de réponse dans le domaine d'origine des données. L'article [7] que j'ai déjà cité précédemment cite en préambule une phrase attibuée à Claude Bernard *Je ne rejette pas l'usage de la statistique mais je condamne le fait de ne pas essayer d'aller au délà.* Je m'approprie cette idée en formation en insistant systématiquement sur le fait que, dans le domaine de la biologie, la statistique n'est qu'un outil sur lequel on ne doit pas s'arrêter. À une question biologique, on doit proposer une réponse biologique.

4.2.3 Par des moyens inhabituels

Partant du principe que les émotions renforcent la mémoire, je me risque généralement à user de moyens inhabituels pour générer des émotions lorsque j'interviens en enseignement et formation. Dans le registre des émotions, celles qui me semblent le plus accessibles en formation sont la surprise et la gaieté plutôt que la peur, la tristesse et la colère... Et donc, j'essaie de générer des émotions par l'usage d'un humour que l'on peut parfois considérer comme douteux, mais que j'assume complètement. Un exemple d'humour douteux ? Quand on parle de moyenne et médiane, je cite souvent la phrase de l'humoriste Gustave Parking, *La majorité des français sont plus cons que la moyenne.* Sans discuter le fond de cette affirmation, s'interroger sur sa véracité permet d'éclairer la différence entre moyenne et médiane. Un autre exemple ? Pour tout savoir sur la variance, il ne faut pas se rendre sur le site `www.variance.fr` car Variance est aussi... une marque de lingerie. Une fois cette digression passée, on peut parler de la variance dans une ambiance plus détendue.

J'ai également recours à des illustrations dont le site `xkcd.com`, *A webcomic of romance, sarcasm, math, and language*, est une source remarquable. Les images de la figure 4.2 en sont issues. Elles peuvent permettre d'illustrer les dangers de l'extrapolation en régression linéaire, le principe des barres d'erreur et la différence entre corrélation et causalité (la planche en 3 vignettes est certainement ma préférée !).

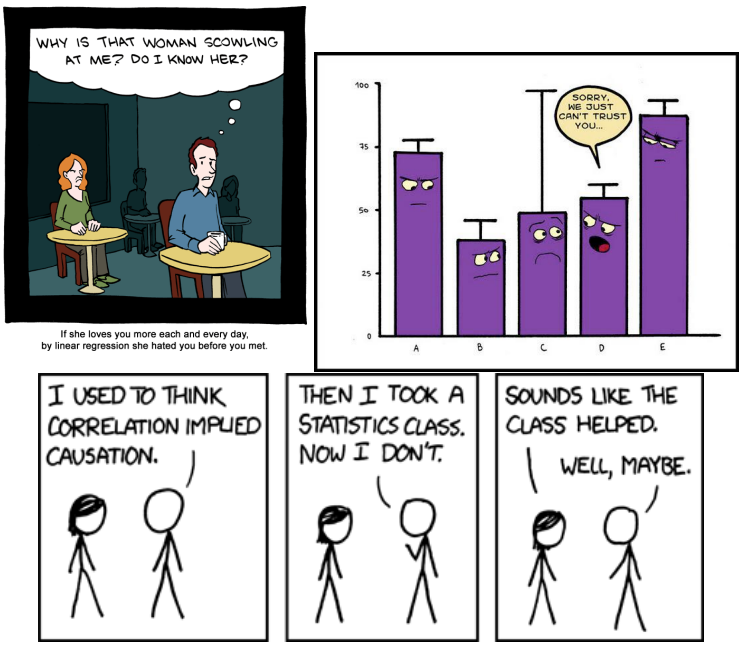

FIGURE 4.2 – Quelques illustrations de notions statistiques issues du site `xkcd.com`.

La presse propose parfois ce type d'illustrations humoristiques. Celle de la figure 4.3 accompagnait un article d'Anne Eveno paru dans le journal Le Monde et intitulé *Salaires : l'exception française*. Elle illustre une interprétation possible d'une augmentation du salaire moyen : quand les personnes touchant un salaire bas perdent leur emploi...

Par ailleurs, au gré de mes lectures, j'ai recueilli quelques citations que je mentionne sur mes supports de formation pour animer les séances. En voici quelques-unes :

— Pour me présenter : *Selon le profil que j'ai établi en me basant sur les deux fax, le plus vieux des ravisseurs est très certainement universitaire. Il doit être ingénieur ou mathématicien. Je ne serais pas étonné si on me disait que c'est un statisticien ou un spécialiste du calcul des probabilités.* Extrait de *Le Carré de la vengeance*, Pieter Aspe, (traduction Emmanuèle Sandron).

FIGURE 4.3 – Illustration d'une interprétation possible d'une augmentation du salaire moyen. Source Colcanopa `colcanopa.com` *Dessins d'actualité et illustrations* publié dans un article d'Anne Eveno *Salaires : l'exception française* paru dans Le Monde, décembre 2013.

— Pour illustrer le principe de la statistique inférentielle : *J'ai 26 ans, je travaille dans le département du contrôle des marchandises [...]. Il serait impossible de les contrôler soigneusement une à une [...]. Par conséquent, on se borne à tirer sur quelques boucles de chaussures, à grignoter quelques gâteaux à titre d'échantillon.* Extrait de *Le communiqué du kangourou*, nouvelle tirée du recueil *L'éléphant s'évapore*, Haruki Murakami (traduction Corinne Atlan).
— Pour parler de corrélation : *Ma vie et celle du commissaire Flores avaient suivi des lignes à la fois divergentes et concomitantes : il montait et je descendais dans une corrélation non fortuite, attendu que ses mérites se fondaient généralement sur mes échecs.* Extrait de *Les égarements de mademoiselle Baxter*, Eduardo Mendoza (traduction de Delphine Valentin).
— Pour parler de corrélation partielle : *Votre hypothèse à vous, dis-je, serait qu'entre ces deux phénomènes il n'y aurait pas de relation de cause à effet mais une simple situation de parallélisme derrière laquelle il y aurait un autre et mystérieux facteur.* Extrait de *La course au mouton sauvage*, Haruki Murakami (traduction Patrick De Vos).

La bande dessinée offre aussi quelques pistes intéressantes. Je me sers notamment d'un extrait de la bande dessinée *Imbattable, le seul véritable super-héros de bande dessinée* de Pascal Jousselin [5]. Les vignettes présentées sur la figure 4.4 montrent le super-héros aux prises avec un individu malfaisant *Two-D boy, le boy X-Y qui se rit du Z* dont le super-pouvoir consiste à s'affranchir de la troisième dimension et à interagir avec des objets qui se situent loin dans la perspective. C'est parfois une tentation que l'on peut avoir au moment d'interpréter certaines proximités entre individus projetés sur un plan formé par des composantes principales sans tenir compte des pourcentages de variance expliquée. L'analyse en composantes principales me donne aussi l'occasion de passer la musique angoissante de la série *The twilight zone*, série de science-fiction des années 60-70, dont le titre français est *La quatrième dimension.* Et puisque la quatrième dimension est angoissante (sans parler des suivantes), le recours à des méthodes de réduction de dimension comme l'ACP est un moyen de calmer notre angoisse.

Pour finir ce tout d'horizon de mes ressources inhabituelles, je présente dans la figure 4.5 une photo de l'ancien président des USA, J.F. Kennedy, qui me sert à attirer l'attention des participants à une formation sur les échelles des axes des représentations graphiques.

Cette photo montre en effet des graphiques avec des échelles dont aucune ne commence à zéro. Il est ainsi relativement facile d'imaginer le discours qui accompagnait cette présentation : tous les voyants socio-économiques sont au vert !

5. pjousselin.free.fr

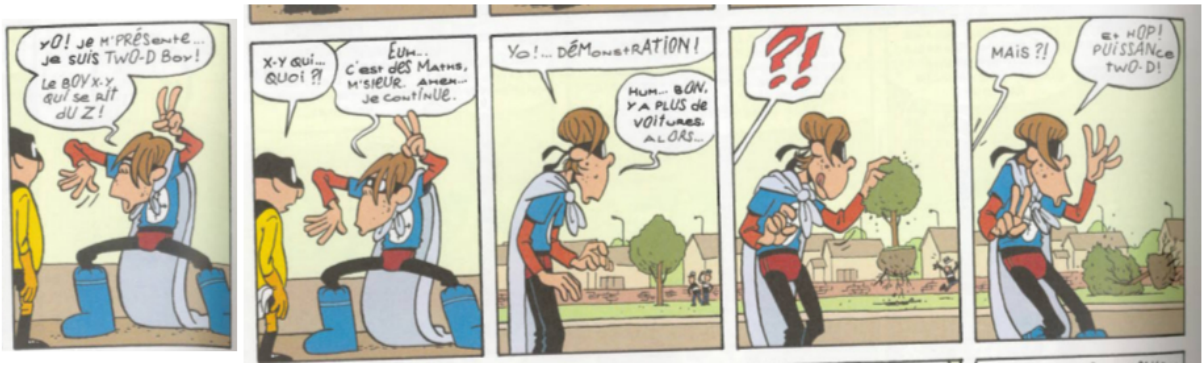

FIGURE 4.4 – Extrait de la bande dessinée *Imbattable* que j'utilise pour attirer l'attention sur les risques d'interprétation trop hâtive des résultats des méthodes de projection.

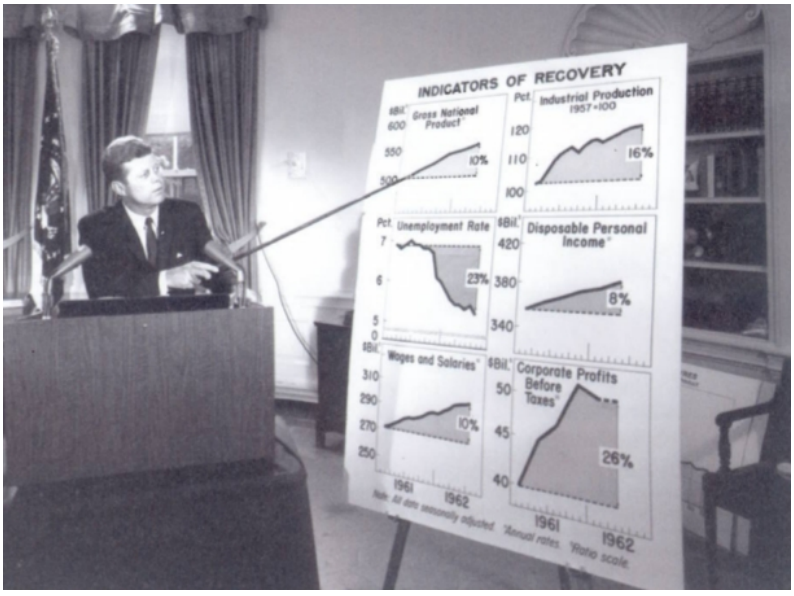

FIGURE 4.5 – Photo de l'ancien président des USA, John F. Kennedy, extraite de l'article *C'était...l'Amérique de Kennedy*, un hors-série des magazines L'Histoire et Paris-Match que j'utilise pour attirer l'attention sur les échelles des axes de certaines représentations graphiques.

4.3 Contribution à l'animation scientifique

4.3.1 Organisation de congrès

Participer à la recherche se caractérise aussi par des contributions à l'animation scientifique. Parmi les manifestations scientifiques auxquelles j'ai participé, je mentionne ici celles qui me semblent les plus significatives.

- implication régulière dans l'organisation des Journées annuelles en bioinformatique et biostatistique organisées par GenoToul ;
- membre du comité d'organisation, trésorier des 45èmes Journées de Statistique, Toulouse, mai 2013.
- membre du comité d'organisation des 5èmes Rencontres R, Toulouse, juin 2016.
- membre du comité de pilotage du semestre thématique *Mathématiques et Informatique pour les sciences du vivant*[6] organisé par le Centre International de Mathématiques et d'Informatique de Toulouse (Labex Cimi), et membre du comité d'organisation de trois événements organisés dans ce cadre.
- actuellement membre du comité d'organisation de la conférence internationale UseR ![7] qui aura lieu à Toulouse du 8 au 12 juillet 2019.

Au delà des aspects organisationnels, la contribution à l'organisation de ces manifestations, d'envergure internationale pour certaines, me permet de consolider mon réseau professionnel dans le domaine académique et ailleurs.

4.3.2 La plateforme GenoToul Biostatistique

GenoToul est un réseau de plateformes en sciences du vivant (Fig. 4.6). La plupart des plateformes ont un volet technologique très développé permettant à la communauté scientifique régionale, et au-delà, d'avoir accès à des technologies modernes inaccessibles à l'échelle d'un laboratoire.

FIGURE 4.6 – Genotoul : un réseau de plateformes en sciences du vivant

Dans ce contexte, la plateforme de biostatistique a été créée en 2008 et j'en assure la co-responsabilité depuis 2011. Le terme qui semble décrire au mieux ses activités est celui d'un carrefour de compétences en statistique pour la biologie. Pour illustrer ce fonctionnement, faisons un peu d'étymologie avec le dictionnaire de l'Académie française[8].

CARREFOUR n. m. XIIe siècle, quarefoz. Du bas latin quadrifurcus, « qui a quatre fourches ». 1. Endroit où se croisent deux ou plusieurs rues, routes ou chemins. Par ext. Voie publique. Expr. fig. Se trouver, être à un carrefour, au moment de prendre une décision engageant l'avenir. Vous voici à un carrefour. Par analogie. Lieu où se rencontrent des civilisations, des cultures différentes. 2. Réunion où l'on échange, où l'on confronte librement des idées sur un thème donné.

6. www.cimi.univ-toulouse.fr/mib
7. www.user2019.fr
8. 9ème édition, version informatisée `atilf.atilf.fr/academie9.htm`

C'est donc effectivement dans ce rôle de lieu d'échanges que la plateforme s'est installée dans le paysage scientifique local. Pour poursuivre la métaphore, on peut considérer que ce carrefour est à l'embranchement des quatre voies : formation, co-encadrement de stages et projets tuteurés, recherche (partenariat dans le cadre de projets) et animation scientifique.

4.3.3 Groupe d'ingénieurs statisticiens

De façon plus informelle, avec Thibault Laurent, ingénieur à l'unité TSE-R (Toulouse School of Economics - Recherche), je suis à l'origine d'un groupe qui n'a aucune existence réglementaire. Ce groupe, qui n'a pas de nom plus abouti que *ingé-stat* , rassemble des collègues d'origines diverses tant sur le plan statuaire (fonction publique, société privée) que thématique (biologie, économétrie, industrie...) qui se reconnaissent plus ou moins sous la désignation d'ingénieur statisticien. L'idée de ce groupe est partie de dicussions avec Thibault sur les points communs des activités que nous menons dans nos affectations respectives. Nous avons pensé, chacun de notre côté, à des personnes susceptibles d'être intéressées par des échanges autour de nos activités. Les premiers échanges ont eu lieu fin 2011 pour une première réunion le 8 mars 2012. Aujourd'hui, ce sont 44 personnes qui sont inscrites sur la liste de diffusion ; toutes ne participent pas systématiquement, mais l'affluence à chaque réunion oscille entre 12 et 20 personnes. Les présentations faites lors de ces rencontres (environ 1 par saison, 4 par an) sont disponibles ici : `www.thibault.laurent.free.fr/ingestat.html`. Le fonctionnement de ce groupe est assez proche de celui des groupes Partage d'Expérience et de Pratiques en Informatiques instaurés par l'Inra [9].

4.3.4 CampuStat

Derrière l'acronyme CampuStat se cache la signification suivante : Cellule d'AccoMPagnement pour l'Usage de la STATistique. Cette structure informelle met une vitrine sur une activité que nous menons ponctuellement avec Laurent Risser, ingénieur de recherche dans l'équipe de statistique et probabilités de l'IMT. Dans ce cadre, notre démarche consiste à conseiller des personnes scientifiques ou pas provenant du monde académique ou d'ailleurs, confrontées à un problème de statistique. Cette démarche est motivée par la création de nouvelles collaborations et la valorisation de l'activité de recherche à l'IMT, ainsi que la diffusion de méthodes mathématiques dans la société et l'industrie. Notre contribution peut tout à fait aboutir à un simple conseil ou à une recommandation de formation. Les collaborations initiées dans le cadre de CampuStat peuvent aussi conduire à la réponse à des appels à projets interdisciplinaires, la rédaction de projets ANR, le (co-)financement de thèses... Ce type de démarches s'insère depuis peu dans les activités de la cellule de valorisation de l'IMT, dont je fais partie depuis sa création en juin 2017.

C'est dans ce cadre, que nous avons eu quelques échanges fascinants avec une historienne travaillant sur les différentes traductions du livre *Le devisement du monde* retraçant les aventures de Marco Polo lors de son voyage vers la Chine. Il s'agissait de mener une analyse textuelle de différents documents et nous avons accompagné cette collègue pour valider la démarche statistique. D'autres contacts avec des sociologues du sport ont donné lieu à des projets tutorés visant à analyser les résultats d'une enquête sur la place des femmes dans les unes du journal *L'Equipe* pendant les Jeux Olympiques de Londres en 2012.

C'est également suite à un contact pris dans ce cadre qu'une collaboration s'est développée autour du traitement de la voix. Ce mémoire en fait état dans les parties 2.2.3 et 2.2.3.

4.4 Expertise

4.4.1 Pour la relecture d'article

Je suis régulièrement sollicité en tant que *referee* pour des articles soumis à des revues internationales. Dans ce cadre, mon rôle consiste généralement à évaluer l'emploi de méthodes statistiques pour répondre à une question précise dans un contexte appliqué (en biologie : *BMC Medical Research, Human Heredity, Journal of Biological System, PlosOne, Analytica Chimica Acta, Analytical Chemistry, Chemometrics and Intelligent Laboratory Systems*, en recherche d'information : *Information Retrieval Journal*). J'ai aussi contribué à des évaluations sur des aspects plus méthodologiques essentiellement en lien avec les méthodes d'intégration de données (*BMC Bioinformatics, Computational Statistics and Data Analysis, International Journal of Tomography and Statistics*) ainsi qu'à des publications concernant des développemetnts de logiciel (*Computers and Geosciences, R journal*).

9. www6.inra.fr/pepi

4.4.2 Du métier d'ingénieur

Ma volonté de préparer ce mémoire en vue d'obtenir une habilitation à diriger des recherches ne s'oppose en rien à mon métier d'ingénieur. La direction de travaux de recherche est une des multiples facettes du métier d'ingénieur. J'ai pu me rendre compte de cela en participant pendant 5 ans, entre 2013 et 2017, en tant qu'expert à la Commission d'Évaluation des Ingénieurs (CEI) de l'Inra, Méthodes pour la recherche (MPR), domaine Informatique, bio-informatique, statistiques et calcul scientifique. Une participation à cette commission implique une analyse de dossiers (fiche, rapport d'activités et synthèse des réalisations), la rédaction de messages de type « evaluation conseil » à destination de l'ingénieur évalué, et la participation aux réunions plénières annuelles de la commission. La lecture et l'analyse des dossiers que j'ai eu à évaluer, m'ont clairement montré que le métier d'ingenieur était totalement compatible avec la direction de recherches. J'ai ainsi eu connaissance de dossiers d'ingénieurs ayant obtenu une habilitation à diriger des recherches selon un parcours et une évolution dans le métier d'ingénieur tout à fait cohérents et dans lesquels je me retrouve. Et c'est certainement une des raisons qui m'a conduit à envisager la rédaction de ce mémoire.

Par ailleurs, depuis 2016, je fais partie de la liste officielle des experts de la Branche d'Activités Professionnelles (BAP) E Informatique, Statistiques et Calcul scientifique pour les jurys de concours Ingénieurs et personnels Techniques de Recherche et de Formation (ITRF). En ayant participé à plusieurs jurys de recrutement (phases d'admissibilité ou d'admission), je me suis là aussi rendu compte de la diversité des cadres d'exercice du métier d'ingénieur et de sa totale adéquation avec la direction de recherches.

Pour conclure cette partie sur le métier d'ingénieur dans le cadre d'un mémoire d'habilitation à diriger des recherches, je souhaite insister sur le fait que les projets de thèse que je serais susceptible de diriger auront comme fil rouge sous-jacent ce métier d'ingénieur vers lequel je souhaite engager des jeunes docteurs à (envisager de) se tourner.

4.5 Vulgarisation scientifique

... face au déploiement de la société des calculs, il est nécessaire d'encourager la diffusion d'une culture statistique vers un public plus large que celui des seuls spécialistes.

Cet extrait du livre de Dominique Cardon, *À quoi rêvent les algorithmes. Nos vies à l'heure des big data* [16] illustre pour moi l'intérêt de s'adresser à des personnes au-delà de notre sphère professionnelle habituelle. C'est ce que je réalise dans divers cadres que je présente ci-dessous.

4.5.1 Les Cafés de l'IMT

FIGURE 4.7 – Une des bannières des Cafés de l'IMT réalisée par Marie-Laure Ausset.

Le travail d'interface que j'assure dans le cadre de projets de recherche m'a conduit progressivement à mener ponctuellement des actions de vulgarisation. Le lien entre les deux activités vient de la nécessité de faire comprendre à quelqu'un qui n'est pas spécialiste de notre domaine, l'intérêt de nos méthodes et de nos outils.

Dans ce cadre, je suis particulièrement fier d'avoir contribué, avec mes collègues Marie-Laure Ausset et Guillaume Cheze, à installer dans le paysage de l'IMT, le séminaire *Les Cafés de l'IMT*[10]. Ce séminaire accueille des présentations de mathématiques à destination d'un public de non spécialistes. Il permet à l'ensemble des membres de l'IMT y compris les personnels non scientifiques d'entendre parler de mathématiques dans un cadre convivial et décontracté. Les thèmes mis en avant ont souvent des co-notations pratiques voire ludiques, mais ce n'est pas systématique. Pour en illustrer la diversité, voici quelques titres de présentations ayant eu lieu dans le cadre des Cafés de l'IMT : *Quand les girafes, tortues, libellules, carottes, feuilles d'érable... fabriquent des maillages de Voronoi* par Raphaël Loubère, *Des impressions 3D pour comprendre la 4D* par Arnaud Chéritat, *Loyauté des décisions algorithmiques*

10. perso.math.univ-toulouse.fr/les-cafes-de-l-imt

par Philippe Besse, *Tsunamis, Vagues scélérates, Mascarets : une déferlante de maths* par Pascal Noble, *Infini, art et mathématiques ou qu'est-ce que ce truc dans le métro ?* par Marcello Bernardara, *Votes, paradoxes et mathématiques* par Guillaume Cheze, *Magie et mathématiques* par Laurent Miclo, *Jonglage et permutations* par Stéphane Lamy...

Ce séminaire qui a commencé par des discussions informelles entre doctorants et personnels administratifs et techniques de l'IMT fait maintenant partie intégrante des séminaires de l'IMT et ses annonces sont relayées par les services de communication de l'Université Paul Sabatier et du CNRS. Marie-Laure Ausset et moi avons même été invités à Radio Campus en novembre 2018 pour en parler.

4.5.2 Liens avec le secondaire

Forum des métiers

Il est un exercice que j'apprécie particulièrement : la rencontre avec des collégiens ou des lycéens. Je participe régulièrement à des forums des métiers. J'adore notamment répondre à leurs questions (y compris celle concernant le salaire qui arrive généralement en premier !) du style :

Q. À quoi ça sert les maths ?
R. j'aurais plus de mal à dire à quoi les maths ne servent pas
Q. Mais monsieur, vous aimez les maths ? sérieux ?
R. Et oui, depuis tout petit, mais j'aime aussi le sport...

En général, pour les mettre tout de suite en contact avec des mathématiques « utiles », je prends le pari, que j'ai toujours gagné jusqu'à présent, qu'au moins une personne dans la salle a un appareil dentaire dans la bouche. En m'adressant à des personnes entre 12 et 16 ans, il est vrai que je prends assez peu de risque.

Et donc, je peux leur montrer une radiographie dentaire (Fig 4.8 à gauche), en général ils reconnaissent et ensuite montrer que cela ressemble beaucoup à une parabole (Fig 4.8 à droite). Les élèves de lycée font en général le lien avec des polynômes de degré 2, pour les élèves de collège, il suffit de leur dire qu'après $y = ax + b$ qui est représenté par une droite, ils verront $y = ax^2 + bx + c$ qui sera représenté par une parabole (*oui, comme l'appareil que l'on met sur le toit pour avoir plein de chaînes de télévision, notez bien que c'est la même forme, mais en 3D...*).

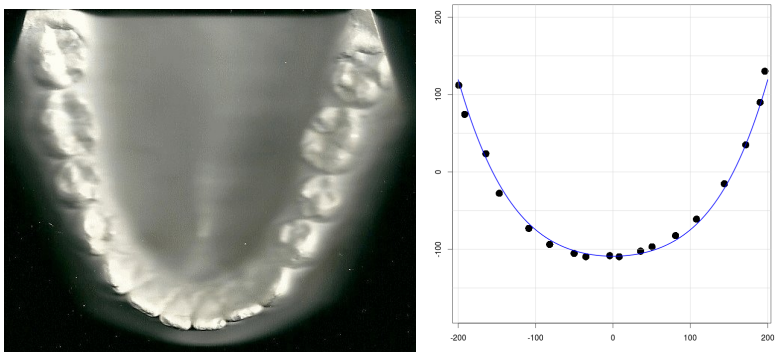

FIGURE 4.8 – Radiographie dentaire fournie par Dr M. Rotenberg et modélisation par une polynôme de degré 4 de la position des dents représentées par les pointes cuspidiennes.

Ainsi, je peux leur montrer qu'un outil mathématique qu'ils connaissent déjà peut être directement utile dans le cadre d'un sujet qui les préoccupe directement : l'éclat de leur sourire !

Ces ajustements de courbe sur des formes d'arcade dentaire ont été étudiées par Dr M. Rotenberg (orthodontiste à Ramonville Saint-Agne, 31 et maître de conférence à la faculté dentaire de l'Université Paul Sabatier) dans sa thèse [78].

La glace étant brisée par cet exemple qui leur parle, on peut ensuite élargir la discussion à d'autres applications plus ou moins éloignées de leurs préoccupations.

Le Cercle Sofia Kovalevskaïa

Le Cercle Sofia Kovalevskaïa (CSK [11]) est un club de mathématiques et d'informatique destiné à tous les élèves de la 3ème à Bac+2 de la région Occitanie. Il a été créé par Christophe Barré, enseignant de mathématiques au Lycée Pierre de Fermat de Toulouse. J'ai tout de suite apprécié l'idée de ce club et j'y

11. www.animath.fr/spip.php ?article2706

suis intervenu chaque année depuis sa création en 2014-2015 sur des sujets comme l'ingénierie mathématique, les probabilités et la statistique. Ma dernière intervention dans ce cadre date de décembre 2018. J'ai donné une conférence intitulée *La statistique dans tous ses états* lors d'une séance spéciale organisée en soirée dans une salle du centre-cille de Toulouse en partenariat avec l'association AssoSciences [12]. Depuis début 2018, je fais partie du bureau de l'association CSK.

Accueil d'élèves de collège

Dans le cadre des liens avec le secondaire, j'accueille régulièrement des élèves de collège en stage d'observation à l'Institut de Mathématiques de Toulouse. Afin de les sensibiliser à la statistique et à l'analyse de données, je les fais réfléchir par exemple à l'utilisation des indicateurs statistiques sur la base d'articles parus dans la presse. Quand un article mentionne, *le salaire moyen des salariés a progressé permettant ainsi un gain de pouvoir d'achat*, je leur demande d'imaginer des situations dans lesquelles on peut effectivement augmenter le salaire moyen et améliorer le pouvoir d'achat et d'autres dans lesquelles le salaire moyen peut augmenter dans une entreprise sans que le pouvoir d'achat de quiconque n'augmente. On peut en venir à s'amuser avec des données dans un tableur simulant une entreprise de quelques salariés avec pour différents objectifs : augmenter le salaire moyen mais pas le salaire médian, augmenter le salaire médian mais pas le salaire moyen, augmenter les deux en n'augmentant personne... J'ai souvent eu de bons retours au sujet de ces petits jeux mathématico-statistique et certains enseignants du secondaire m'ont parfois demandé de reproduire ces petits exercices. Selon l'intérêt des élèves, je peux également les initier à la programmation avec R voire à LaTeXpour celles et ceux qui apprécient les belles formules.

12. www.assosciences.com

Conclusion

Lors de la rédaction de ce mémoire, j'ai trouvé de nombreux éléments de réflexion dans deux articles qui jalonnent l'histoire de l'analyse de données. Le premier, *The future of Data Analysis* est dû à John Tukey et date de 1962 [88] ; le second, *50 year of data science* a été écrit plus de cinquante ans plus tard par David Donoho comme une réponse aux prévisions du premier [21]. C'est en référence à ces articles qui m'ont largement inspiré que j'ai intitulé ce mémoire. J'ai eu l'occasion de mentionner et discuter plusieurs éléments tirés de ces articles et j'en extrait ici un dernier de [88] pour initier ma conclusion : *We must face up to the fact that, in any experimental science, our certainty about what will happen in a particular situation does not come from directly applicable experiments or theory, but rather comes mainly through analogy between situations which are not know to behave similarly. Data analysis has, of necessity, to be an experimental science, and needs therefore to adopt the attitude of experimental science [...]. Finally, we need to give up the vain hope that data analysis can be founded upon a logico-deductive system like Euclidean plane geometry [...] and to face up the fact that data analysis is intrinsically an empirical science.* Ainsi, J. Tukey présente l'analyse de données comme une science expérimentale qui ne peut pas raisonnablement se positionner selon un système logico-déductif. C'est une formulation à laquelle j'adhère entièrement car cela correspond au savoir-faire que j'ai acquis au cours de mes activités de recherche. Face à des données, je réagis souvent selon l'attitude : *Essayer, analyser et aviser*. L'histoire de l'analyse d'un jeu de données n'est pas écrite à l'avance en fonction de théorèmes ou propriétés théoriques des méthodes mises en œuvre et c'est pour moi le grand intérêt de mes activités et notamment dans le cadre de collaborations pluri- ou interdisciplinaires.

Dans ce contexte, je considère que ma thématique privilégiée est l'intégration de données. Derrière ce terme un peu fourre-tout, comme j'ai pu le discuter au début du chapitre 3, se cache un ensemble de démarches et méthodes visant à obtenir des informations à partir de plusieurs sources de données. Au delà des exemples que j'ai présentés dans ce mémoire, l'intégration de données peut revêtir des aspects beaucoup plus complexes et c'est précisément vers un cadre plus complexe que je m'oriente en ce moment dans le cadre du projet OptimiPS-TC (Optimiser le Parcours de Soins des patients Traumatisés Crâniens). Ce projet, encore en phase très préliminaire, a pour objectif de sauver des vies. J'ai d'abord été surpris par la formulation de cet objectif par les cliniciens à l'origine du projet, mais leur raisonnement est très clair : les données contenues dans les différents systèmes d'information d'un centre hospitalo-universitaire contiennent des informations susceptibles de mieux connaître les patients et leur parcours de soins. Mieux connaître les patients et leur parcours de soins, c'est mieux les soigner eux-mêmes ainsi que les futurs patients, et mieux les soigner, c'est potentiellement augmenter leur durée de vie. Donc effectivement, contribuer à exploiter au mieux des données médicales peut contribuer à sauver des vies. La spécificité, hormis les difficultés réglementaires et législatives concernant l'accès et la manipulation des données personnelles et médicales ainsi que des considérations éthiques, sont liées à l'hétérogénéité des données. En effet, rassembler les informations relatives à un patient pour le soigner au mieux risque fort de ressembler *in fine* à un inventaire à la Prévert : données numériques issues d'analyses biologiques, données textuelles issues de compte-rendu de visites ou d'examens, images (2D, 3D) issues d'examens radiologiques, scanner, IRM... données spectrales provenant d'enregistrement type électro-cardiogrammes, électro-encéphalogrammes... Dans ce contexte, caractériser des profils-types de patients afin de définir, par exemple, des parcours de soins optimaux, représente un vrai défi. A priori, j'ai tendance à positionner ce défi au niveau de l'utilisation astucieuse et combinée de méthodes existantes. Il existe, en effet, des méthodes efficaces pour traiter :

- des données numériques et en extraire des informations pertientes ; ce mémoire en fait état ;
- des images et en extraire des caractéristiques ;
- des données textuelles ;
- des données spectrales pour détecter des pics, identifier des motifs, repérer des tendances...

Chacun de ces champs apportera, bien sûr, son lot de questions intéressantes et de développements méthodologiques visant à améliorer l'existant. Mais, combiner ces données, combiner les résultats des analyses statistiques de chacun de ces types de données est un domaine dans lequel je souhaite poursuivre des activités. Parmi les pistes existantes, les méthodes à noyaux étudiées par J. Mariette dans sa thèse

[65] semblent prometteuses pour l'intégration de données hétérogènes.

Pour finir sur note plus personnelle, je souhaite évoquer le fait que mon moteur dans les multiples activités que je mène dans mon travail est le plaisir. Plaisir de contribuer à répondre à des questions concrètes, plaisir de contribuer à la montée en compétences de collaborateurs, plaisir de voir la petite étincelle dans les yeux d'un collaborateur, stagiaire, étudiant qui vient de comprendre quelque chose, plaisir encore de faire collaborer des collègues issus de disciplines différentes et d'apprendre de nouvelles choses à leur contact. Travailler avec des biologistes, des médecins, des psychiatres, des informaticiens, des agronomes, des vétérinaires, des orthodontistes, des sociologues, des orthophonistes... est d'une richesse exceptionnelle. Alors, même s'il est de bon ton dans une carrière d'afficher une mobilité géographique, j'affirme que je voyage bien plus en collaborant avec des personnes qui viennent d'horizons aussi variés même sans quitter la région toulousaine. Cela ne signifie pas pour autant que je ne bougerai jamais de ma situation actuelle. Sincèrement, je n'en sais rien, mais si je dois bouger ce sera parce que je serai persuadé que ma nouvelle situation me maintiendra dans cette dynamique.

Bibliographie

[1] H. Abdi : Partial least squares regression and projection on latent structure regression (PLS regression). *Wiley Interdisciplinary Reviews : Computational Statistics*, 2(1):97–106, 2010.

[2] S. Abitboul : *Sciences des données : de la logique du premier ordre de la toile.* Collège de France, 2012. Leçon inaugurale prononcée le jeudi 8 mars 2012.

[3] C. Amosse, F. Vannier, L. Cabrejo, P. Auzou et D. Hannequin : Les troubles de la parole. *NPG Neurologie - Psychiatrie - Gériatrie*, 4(19):11 – 14, 2004.

[4] P. Arnold : What about the p in the ppdac cycle ? an initial look at posing questions for statistical investigation. 03 2012.

[5] J. Ayter, A. Chifu, S. Déjean, C. Desclaux et J. Mothe : Statistical Analysis to Establish the Importance of Information Retrieval Parameters. *Journal of Universal Computer Science, Information Retrieval and Recommendation*, 21(13 (2015)):1767–1789, décembre 2015.

[6] A. Baccini, S. Déjean, L. Lafage et J. Mothe : How many performance measures to evaluate information retrieval systems ? *Knowledge and Information System*, 30:693–713, 2012.

[7] J. Berkson : Smoking and lung cancer : Some observations on two recent reports. *Journal of the American Statistical Association*, 53(281):28–38, 1958.

[8] J. Bertin : *La graphique et le traitement graphique de l'information.* Flammarion, 1977.

[9] A. Bigot : *Analyse et catégorisation des systèmes de recherche d'information pour la sélection et l'adaptation aux besoins en information.* Thèse de doctorat, Université de Toulouse, Toulouse, France, octobre 2014. (Soutenance le 07/10/2014).

[10] A. Bigot, S. Déjean et J. Mothe : Choisir la meilleure configuration d'un système de recherche d'information. *Document numérique, Document Numérique 1/2014*, Hors-série(2):125–147, 2014.

[11] A. Bigot, S. Déjean et J. Mothe : Learning to Choose - Automatic Selection of the Information Retrieval Parameters (student paper). *In Spanish Conference on Information Retrieval, Coruña, 18/06/2014-20/06/2014*, page (on line), http ://www.springerlink.com, avril 2014. Springer.

[12] A. Bigot, S. Déjean et J. Mothe : Learning to Choose the Best System Configuration in Information Retrieval : the case of repeated queries. *Journal of Universal Computer Science, Information Retrieval and Recommendation*, 21(13):1726–1745, décembre 2015.

[13] L. Breiman : Statistical modeling : The two cultures. *Statistical Science*, 16(3):199–231, 2001.

[14] L. Breiman, J. Friedman, C.J. Stone et R.A. Olshen : *Classification and Regression Trees.* The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984.

[15] G. Caraux et S. Pinloche : Permutmatrix : a graphical environment to arrange gene expression profiles in optimal linear order. *Bioinformatics*, 21(7):1280–1281, 2005.

[16] D. Cardon : *A quoi rêvent les algorithmes - Nos vies à l'heure des big data.* La République des idées. Seuil, 2015.

[17] M. Chavent, V. Kuentz-Simonet, B. Liquet et J. Saracco : Clustofvar : An r package for the clustering of variables. *Journal of Statistical Software, Articles*, 50(13):1–16, 2012.

[18] S. Combes, I. González, S. Déjean, A. Baccini, N. Jehl, H. Juin, L. Cauquil, B. Gabinaud, F. Lebas et C. Larzul : Relationships between sensory and physicochemical measurements in meat of rabbit from three different breeding systems using canonical correlation analysis. *Meat Science*, 80(3):835 – 841, 2008.

[19] J. Compaoré, A.M. Gueye, S. Déjean, J. Mothe et J. Randriamparany : Mining information retrieval results ; significant ir parameters. *In Proceedings of the The First International Conference on Advances in Information Mining and Management, IARIA*, 2011.

[20] G. Cumming, F. Fidler et D.L. Vaux : Error bars in experimental biology. *The Journal of Cell Biology*, 177(1):7–11, 2007.

[21] D. Donoho : 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4): 745–766, 2017.

[22] D. Donoho et J. Jin : Higher criticism thresholding : Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, 105(39):14790–14795, 2008.

[23] G. Draper, T. Vincent, M. E. Kroll et Swanson J. : Childhood cancer in relation to distance from high voltage power lines in england and wales : a case-control study. *British Medical Journal*, 330(4):1–5, June 2005.

[24] L. Dupuy : Co, multi, inter, ou trans-disciplinarité ? la confusion des genres... Document de travail à destination des étudiants du CIEH (Certificat International d'Ecologie Humaine, web.univ-pau.fr/RECHERCHE/CIEH/documents/La confusion des genres.pdf, 2018.

[25] H. Duruflé : *Production et traitement de données omiques hétérogènes en vue de l'étude de la platicité de la paroi chez des écotypes de la plate modèle Arabidopis thaliana provenant d'altitudes contrastées*. Thèse de doctorat, Université de Toulouse, 2017.

[26] H. Duruflé, V. Hervé, P. Ranocha, T. Balliau, M. Zivy, J. Chourré, H. San Clemente, V. Burlat, C. Albenne, S. Déjean, E. Jamet et C. Dunand : Cell wall modifications of two arabidopsis thaliana ecotypes, col and sha, in response to sub-optimal growth conditions : An integrative study. *Plant Science*, 263:183 – 193, 2017.

[27] H. Duruflé, M. Selmani, P. Ranocha, E. Jamet, C. Dunand et S. Déjean : A powerful framework for an integrative study with heterogeneous omics data : from univariate statistics to multi-block analysis. Soon submitted to e-life journal, 2018.

[28] S. Déjean, P. Martin, A. Baccini et P. Besse : Clustering time-series gene expression data using smoothing spline derivatives. *EURASIP Journal on Bioinformatics and Systems Biology*, 2007(1):70561, Jun 2007.

[29] S. Déjean et J. Mothe : Visual clustering for data analysis and graphical user interfaces. *In* C. Hennig, M. Meila, F. Murtagh et R. Rocci, éditeurs : *Handbook of Cluster Analysis*, chapitre 30, pages 679 – 702. Chapman & Hall, CRC Press, 2015.

[30] R. Díaz-Uriarte et S. Alvarez de Andrés : Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(3), 2006.

[31] M.L. Eaton et M.D. Perlman : The non-singularity of generalized sample covariance matrices. *The Annals of Statistics*, 1(4):710–717, 1973.

[32] M.B. Eisen, P.T. Spellman, P.O. Brown et D. Botstein : Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25):14863–14868, 1998.

[33] J. Ellenberg : *How not to be wrong. The power of mathematical thinking*. Penguin Press, 2014.

[34] V. Esposito Vinzi, W.W. Chin, J. Henseler et H. Wang, éditeurs. *Handbook of Partial Least Squares*. Handbooks of Computational Statistics. Springer, 2010.

[35] L. Fontan, M. Fraval, A. Michon, S. Déjean et M. Welby-Gieusse : Vocal problems in sports and fitness instructors : A study of prevalence, risk factors and need for prevention in france. *Journal of Voice*, 31(2):261.e33–261.e38, 2016.

[36] S. Gagnot, J.-P. Tamby, M.-L. Martin-Magniette, F. Bitton, L. Taconnat, S. Balzergue, S. Aubourg, J.-P. Renou, A. Lecharny et V. Brunaud : CATdb : a public access to arabidopsis transcriptome data from the URGV-CATMA platform. *Nucleic acids research*, 36(Database issue): D986 – D990, 2008.

[37] S. Gasquet-More : *Plus vite que son nombre*. La République des idées. Seuil, 1999.

[38] I. González : *Analyse Canonique Régularisée pour des données fortement multidimensionnelles*. Thèse de doctorat, Université de Toulouse, 2007.

[39] I. González, K.-A. Lê Cao, M.J. Davis et S. Déjean : Visualising associations between paired 'omics' data sets. *BioData Mining*, 5(1):19, Nov 2012.

[40] I. González, S. Déjean, P. Martin et A. Baccini : CCA : An R package to extend canonical correlation analysis. *Journal of Statistical Software*, 23(12), 2008.

[41] I. González, S. Déjean, P. Martin, O. Gonçalves, P. Besse et A. Baccini : Highlighting relationships between heteregeneous biological data through graphical displays based on regularized canonical correlation analysis. *Journal of Biological Systems*, 17(02):173–199, 2009.

[42] I. González, A. Eveillard, C. Canlet, A. Paris, T. Pineau, P. Besse, P. Martin et S. Déjean : Selecting the good level of details in undecimated wavelet transform improves the classification of samples from metabolomic data. *JP Journal of Biostatistics*, 10(2):61–79, 2009.

[43] E. Goupy : Croiser les disciplines, croiser les arts. Les dossiers pédagogiques. Site du musée des Abattoirs, 2018.

[44] M. Hahsler, K. Hornik et C. Buchta : Getting things in order : An introduction to the R package seriation. *Journal of Statistical Software*, 25(3):1–34, March 2008.

[45] D.J. Hand : Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–14, 2006.

[46] H. Hotelling : Relations between two sets of variates. *Biometrika*, 28(3-4):321–377, 1936.

[47] A. Imbert, A. Valsesia, C. Le Gall, C. Armenise, G. Lefebvre, P. Gourraud, N. Viguerie et N. Villa-Vialaneix : Multiple hot-deck imputation for network inference from RNA sequencing data. *Bioinformatics*, 34(10):1726–1732, 2018.

[48] A. Imbert et N. Vialaneix : Décrire, prendre en compte, imputer et évaluer les valeurs manquantes dans les études statistiques : une revue des approches existantes. *Journal de la Société Française de Statistique*, 159(2):1–55, 2018.

[49] B.H. Jacobson, A. Johnson, C. Grywalski, A. Silbergleit, G. Jacobson, M.S. Benninger et C.W. Newman : The voice handicap index (vhi)development and validation. *American Journal of Speech-Language Pathology*, 6(3):66–70, 1997.

[50] K.M. Kerr : Experimental design to make the most of microarray studies. *In* Arkady B. Brownstein, Michael J.and Khodursky, éditeur : *Functional Genomics : Methods and Protocols*, pages 137–147. Humana Press, Totowa, NJ, 2003.

[51] L. Laporte : *La sélection de variables en apprentissage d'ordonnancement pour la recherche d'information : vers une approche contextuelle.* Thèse de doctorat, Université Paul Sabatier, Toulouse, France, novembre 2013. (Soutenance le 18/11/2013).

[52] L. Laporte, S. Déjean et J. Mothe : Multiple Clicks Model for Web Search of Multi-clickable Documents (short paper). *In International Conference on Enterprise Information Systems (ICEIS), Angers, 04/07/2013-07/07/2013*, page (electronic medium), http ://www.scitepress.org/, juin 2013. SciTePress.

[53] L. Laporte, S. Déjean et J. Mothe : Sélection de variables en apprentissage d'ordonnancement : évaluation des SVM pondérés. *Document numérique, Evaluation en Recherche d'Information*, 18/1 - 2015:99–121, 2015.

[54] L. Laporte, R. Flamary, S. Canu, S. Déjean et J. Mothe : Non-convex Regularizations for Feature Selection in Ranking with Sparse SVM. *IEEE Transactions on Neural Networks and Learning Systems*, 25(6):1118–1130, juin 2014.

[55] H. Laurell, M. Bouisson, P. Berthélemy, P. Rochaix, S. Déjean, P. Besse, C. Susini, L. Pradayrol, N. Vaysse et L. Buscail : Identification of biomarkers of human pancreatic adenocarcinomas by expression profiling and validation by gene expression analysis in endoscopic ultrasound-guided fine needle aspiration samples. *World Journal of Gastroenterology*, 12(21):3244–3351, 2006.

[56] K.-A. Lê Cao, S. Boitard et P. Besse : Sparse PLS discriminant analysis : biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, 12(1):253, Jun 2011.

[57] L. Lebart, M. Piron et A. Morineau : *Statistique exploratoire multidimensionnelle : visualisation et inférences en fouilles de données.* Dunod, 2006.

[58] S.E. Leurgans, R.A. Moyeed et B.W. Silverman : Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society B*, 55(3):725–740, 1993.

[59] W.-Y. Loh : Classification and regression tree methods. *In* F. Ruggeri, R. Kenett et F. W. Faltin, éditeurs : *Encyclopedia of Statistics in Quality and Reliability*, pages 315 – 323. John Wiley & Sons, Ltd, 2008.

[60] K.-A. Lê Cao : *Outils statistiques pour la sélection de variables et l'intégration de données « omiques »*. Thèse de doctorat, Université de Toulouse, 2008.

[61] K.-A. Lê Cao, I. González et S. Déjean : integromics : an R package to unravel relationships between two omics datasets. *Bioinformatics*, 25(21):2855–2856, 2009. the package first called 'integrOmics' has been renamed 'mixOmics'.

[62] K.-A. Lê Cao, D. Rossouw, C. Robert-Granié et P. Besse : A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 2008. Article 35.

[63] R. J. MacKay et R. W. Oldford : Scientific method, statistical method and the speed of light. *Statist. Sci.*, 15(3):254–278, 08 2000.

[64] K.V Mardia, J.T Kent et J.M Bibby : *Multivariate Analysis*. Academic Press, 1979.

[65] J. Mariette : *Apprentissage statistique pour l'intégration de données de sources et de natures multiples*. Thèse de doctorat, Université de Toulouse, 2017.

[66] P.G. Martin, H. Guillou, F. Lasserre, S. Déjean, A. Lan, Pascussi J.-M., M. San Cristobal, P. Legrand, P. Besse et T. Pineau : Novel aspects of PPARalpha-mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study. *Hepatology*, 45(3):767 – 777, 2007.

[67] M.-L. Martin-Magniette : Une méta-analyse transcriptomique identifie une réponse globale aux stress chez la plante modèle Arabidopsis. Séminaire IMABS, Toulouse, Janvier 2019.

[68] A. Mitchell : *The ESRI Guide to GIS Analysis, Volume 2 : Spatial Measurements and Statistics*. ESRI Press, Redlands, CA, USA, 2005. ISBN 9781589482951.

[69] E. Neuwirth : *RColorBrewer : ColorBrewer Palettes*, 2014. R package version 1.1-2.

[70] E. Parkhomenko, D. Tritchler et J. Beyene : Sparse canonical correlation analysis with application to genomic data integration. *Statistical applications in genetics and molecular biology*, 8:Article 1, 02 2009.

[71] S. Pinto, A. Ghio, B. Teston et F. Viallet : La dysarthrie au cours de la maladie de parkinson. histoire naturelle de ses composantes : dysphonie, dysprosodie et dysarthrie. *Revue Neurologique*, 166(10):800 – 810, 2010. Journées Internationales de la SFN - 7 et 8 octobre 2010.

[72] S. Poirier, S. Déjean et O. Chapleur : Support media can steer methanogenesis in the presence of phenol through biotic and abiotic effects. *Water Research*, 140:24 – 33, septembre 2018.

[73] V. Potier, M. Lalanne, N. Yassine-Diab et S. Déjean : Jeu, jeunes et numériques : des relations à déconstruire. Genre, générations, structure sociale et territoires à l'ère du numérique - Méthodes et analyses croiées. Journée d'étude pluridisciplinaire.

[74] S. Richardson, G.C. Tseng et W. Sun : Statistical methods in integrative genomics. *Annual Review of Statistics and Its Application*, 3(1):181–209, 2016.

[75] M. Ritchie, E. Holzinger, R. Li, S. Pendergrass et D. Kim : Methods of integrating data to uncover genotype-phenotype interactions. *Nature reviews. Genetics*, 16:85–97, 2015.

[76] F. Rohart, A. Eslami, N. Matigian, S. Bougeard et K.-A. Lê Cao : Mint : a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC Bioinformatics*, 18(1):128, Feb 2017.

[77] F. Rohart, B. Gautier, A. Singh et K.-A. Lê Cao : mixomics : An R package for 'omics feature selection and multiple data integration. *PLoS Computational Biology*, 13(11), 2017.

[78] M. Rotenberg : *Modélisation de la forme d'arcade dentaire de jeunes adultes*. Thèse de doctorat, Université de Toulouse, 1996.

[79] G. Salton : *Automatic Information Organization and Retrieval*. McGraw Hill Text, 1968.

[80] M. Schiavone, S. Déjean, N. Sieczkowski, M. Castex, E. Dague et J.-M. François : Integration of biochemical, biophysical and transcriptomics data for investigating the structural and nanomechanical properties of the yeast cell wall. *Frontiers in Microbiology*, 8:1806, 2017.

[81] M.V. Schneider et R.C Jimenez : Teaching the fundamentals of biological data integration using classroom games. *PLoS Computational Biology*, 8(12), 2012.

[82] H. Shen et J.Z. Huang : Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6):1015 – 1034, 2008.

[83] A. Singh, B. Gautier, C.P. Shannon, M. Vacher, F. Rohart, S.J. Tebutt et K.-A. Le Cao : Diablo - an integrative, multi-omics, multivariate method for multi-group classification. *bioRxiv*, 2016.

[84] M. de Smith, M. Goodchild et P. Longley : *Geospatial Analysis : A Comprehensive Guide to Principles, techniques and software tools*. Winchelsea Press, 6th édition, 2018. www.spatialanalysisonline.com.

[85] A. Tenenhaus, C. Philippe, V. Guillemot, K.-A. Lê Cao, J. Grill et V. Frouin : Variable selection for generalized canonical correlation analysis. *Biostatistics*, 15(3):569–583, 2014.

[86] M. Tenenhaus : *La régression PLS : théorie et pratique*. TEHCNIP, 1998.

[87] M. Tenenhaus et M. Hanafi : *A Bridge Between PLS Path Modeling and Multi-Block Data Analysis*, pages 99–123. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[88] J.W. Tukey : The future of data analysis. *Ann. Math. Statist.*, 33(1):1–67, 03 1962.

[89] H.D. Vinod : Canonical ridge and econometrics of joint production. *Journal of Econometrics*, 4(2):147–166, 1976.

[90] V. Voillet, P. Besse, L. Liaubet, M. San Cristobal et I. Gonzalez : Handling missing rows in multi-omics data integration : Multiple imputation in multiple factor analysis framework. *BMC Bioinformatics*, 17, 10 2016.

[91] S. Waaijenborg, P. De Witt Hamer et A. Zwinderman : Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis. *Statistical applications in genetics and molecular biology*, 7:Article3, 02 2008.

[92] Ronald L. Wasserstein et Nicole A. Lazar : The ASA's statement on p-values : Context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.

[93] H. Wickham : Tidy data. *Journal of Statistical Software, Articles*, 59(10):1–23, 2014.

[94] H. Wickham et G. Grolemund : *R for Data Science*. O'Reilley, 2017.

[95] I. Wilms et C. Croux : Sparse canonical correlation analysis from a predictive point of view. *Biometrical Journal*, 57(5):834–851, 2015.

[96] D. Witten, R. Tibshirani et T. Hastie : A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics (Oxford, England)*, 10:515–34, 05 2009.

[97] S. Wold, M. Sjöström et L. Eriksson : PLS-regression : a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109 – 130, 2001.

[98] R. Zaag, J. P. Tamby, C. Guichard, Z. Tariq, G. Rigaill, E. Delannoy, J. P. Renou, S. Balzergue, T. Mary-Huard, S. Aubourg, M. L. Martin-Magniette et V. Brunaud : GEM2Net : from gene expression modeling to -omics networks, a new CATdb module to investigate arabidopsis thaliana genes involved in stress response. *Nucleic acids research*, 43(Database issue): D1010 – D1017, 2014.

[99] A. Zeileis, K. Hornik et P. Murrel : Escaping RGBland : Selecting colors for statistical graphics. *Computational Statistics & Data Analysis*, 53(9):3259–3270, 2009.

[100] S. D. Zhao, G. Parmigiani, C. Huttenhower et L. Waldron : Más-o-menos : a simple sign averaging method for discrimination in genomic data analysis. *Bioinformatics*, 30(21):3062 – 3069, 2014.