Comment la modélisation statistique contribue-t-elle à la recherche biomédicale et à la R&D de l'industrie ?

Myriam Maumy-Bertrand

Habilitation à diriger des recherches

INSTITUT DE RECHERCHE MATHÉMATIQUE AVANCÉE

UMR 7501

Strasbourg

Université de Strasbourg
Spécialité MATHÉMATIQUES APPLIQUÉES

Myriam Maumy-Bertrand

Comment la modélisation statistique contribue-t-elle à la recherche biomédicale et à la R&D de l'industrie ?

Soutenue le 18 septembre 2020
devant la commission d'examen

Hervé Abdi, rapporteur
Stéphanie Allassonnière, rapporteuse
Jean Bérard, garant d'habilitation
Marianne Clausel, rapporteuse
Véronique Maume-Deschamps, examinatrice
Gilbert Saporta, examinateur

https://irma.math.unistra.fr

cnrs dépasser les frontières

IRMA Institut de Recherche Mathématique Avancée

Université de Strasbourg

À Anaëlle, David, Frédéric et Jacqueline.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Bilan

1.1 Expériences en matière de recherche

Au cours des quinze dernières années passées au sein de l'Institut de Recherche Mathématique Avancée de l'Université de Strasbourg, plus connu sous le nom de l'IRMA, (qui est également une unité mixte de recherche, UMR 7501) mes activités de recherche se sont concentrées autour de plusieurs thématiques fédératrices fortes émanant de la statistique, évidemment.
Pendant ces quinze années, mes thèmes de recherche ont au fur et à mesure évolué et d'autres sujets éloignés de mon sujet de thèse (le processus empirique et ses dérivés) m'ont intéressé. Je dois reconnaître que tous ces travaux de recherche que j'ai entrepris ont été formateurs pour moi et enrichissants pour mon expérience professionnelle.
Actuellement le sujet de recherche qui me préoccupe la plupart du temps est l'inférence statistique au sens large du terme à savoir petite ou grande dimension du jeu de données d'étude, avec ou sans données manquantes, et parfois même avec ou sans censure.
Grâce à mes différentes collaborations au sein de l'Université Louis Pasteur (devenue le 1er janvier 2009 l'Université de Strasbourg) ou avec les entreprises qu'elles soient de petite, moyenne ou de grande taille, je me suis intéressée aussi bien au cas de la régression linéaire, de la régression linéaire généralisée qu'à d'autres contextes de régression, comme la régression bêta ou le modèle de Cox, pour lesquels avec mes co-auteurs, nous avons non seulement produit de nouveaux critères de choix de modèles mais aussi de sélection de variables. Les contextes d'applications nous ont incité à nous appuyer sur des approches de régression pénalisée, comme la régression *ridge*, le *lasso* ou la méthode *elastic net*, ou encore des approches de régression par moindres carrés partiels parcimonieuse ou non.

Depuis presque quatre ans maintenant, il est à noter, qu'à travers la thèse d'Emmanuelle Claeys (Claeys (2019a)), intitulée « Clusterisation incrémentale, multicritères de données hétérogènes pour la personnalisation d'expérience utilisateur » et soutenue le 12 novembre 2019 à l'IRMA, je m'intéresse à une nouvelle thématique liée aux bandits manchots et au test A/B, une technique de marketing particulièrement employée dans la communication en ligne. Je développerai cette thématique dans le chapitre 4 et je présenterai le sujet de thèse dans le paragraphe1.2.2 de ce chapitre.

Ma première thématique de recherche est le processus empirique et les processus qui en dérivent. Le sujet que m'a proposé le Professeur Paul Deheuvels en septembre 1999, pour ma thèse de doctorat, s'intitule « Étude du processus empirique composé », (Maumy (2002a)). Ce dernier m'a permis de comprendre et de maîtriser les outils probabilistes incontournables pour pouvoir établir des résultats de convergence de processus et d'étudier des propriétés fondamentales de ces derniers qui permettent de résoudre des problèmes en statistique mathématique. Ma thèse m'a aussi permis d'écrire trois articles scientifiques seule (Maumy (2001), Maumy (2002b), Maumy (2004)). Ces trois années de thèse m'ont appris l'indépendance et l'autonomie, qualités qui m'ont beaucoup aidé dans ma carrière au sein de l'IRMA. Je dois noter que lorsque je suis arrivée en 2004 au sein de l'équipe de statistique, aucun membre ne pouvait collaborer avec moi. En effet, leurs sujets de recherche étaient très éloignés des miens.

Au début de ma carrière de chercheur en statistique mathématique, carrière que j'ai commencée à l'Université Louis Pasteur, il m'est apparu important d'avoir une vision précise de la manière dont la statistique mathématique dont je suis issue était appliquée par des utilisateurs ou par des praticiens. C'est une des façons de parvenir à proposer des résultats issus des mathématiques appliquées qui serviront à résoudre des problèmes concrets. J'ai donc continué, dès mon poste de permanent à me confronter à des problèmes que les expérimentateurs considéraient comme ouverts, c'est-à-dire des problèmes qu'aucune des techniques statistiques existantes ne leur avait permis de résoudre.

J'emploie le mot « continuer » car effectivement pendant mon premier poste d'ATER (Assistant Temporaire d'Enseignement et de Recherche) à l'Université de Rennes 2, j'ai eu l'opportunité de me confronter à un problème issu du monde économique : dénombrer les touristes en Bretagne afin de savoir si le parc hôtelier était adapté à la demande. En effet, avec l'ouverture des frontières européennes, les enquêtes qui étaient faites aux frontières auparavant ne pouvaient plus avoir lieu. La méthodologie qui a été développée pour répondre à cette problématique a été présentée dans plusieurs colloques francophones et internationaux, Deville et Maumy (2004a,b); Deville *et al.* (2005); Deville et Maumy (2005)

et appliquée à l'étude MORGOAT (Morgoat (2005); Deville et Maumy (2006)). La technique qui émane de nos réflexions avec Jean-Claude Deville (à l'époque Professeur à l'ENSAI qui se situe à Bruz,) et les acteurs de l'observatoire du tourisme a été présentée dans l'article intitulé « Extensions de la méthode d'échantillonnage indirect et son application aux enquêtes dans le tourisme. »(Deville et Maumy-Bertrand (2006)). Je ne me suis pas arrêtée de m'intéresser à cette thématique depuis lors. Je continue régulièrement à participer à la mise en place et à réfléchir à de nouvelles méthodologies. J'ai d'ailleurs fait partie pendant huit ans (de mars 2009 à 2017) du bureau du Groupe Enquêtes, Modèles et Applications de la Société Française de Statistique en tant que secrétaire. J'interviens toujours dans ce groupe en tant que membre associée. Je citerai également à titre d'exemple les travaux de Jean-Marc Philip (Philip et Maumy (2008)) qui concerne les enquêtes dans les pays en développement et ceux de l'enquête Corse réalisés par Fabio Rendina et Jimmy Armoogum de l'institut français des sciences et technologies des transports, de l'aménagement et des réseaux, plus connu sous l'abréviation IFSTTAR, (Rendina *et al.* (2016, 2017, 2018)).

J'ai également eu l'opportunité de participer ou de réaliser l'analyse de jeux de données complexes ayant souvent une composante temporelle voire spatio-temporelle comme ceux récoltés par l'observatoire national des maladies du bois de la vigne (Kobes *et al.* (2007), Bertrand *et al.* (2007), Bertrand *et al.* (2008), Kuntzmann *et al.* (2013)), ceux consacrés à l'étude des assemblages d'espèces de phytoplanctons (Rolland *et al.* (2009), Bertrand et Maumy (2010)), ou ceux qui étudient des problématiques d'éthologie (Jacobs *et al.* (2008), Bourjade *et al.* (2014)) ou encore ceux de médecine (Grenèche *et al.* (2011a), Grenèche *et al.* (2013), Vallat *et al.* (2013), Carapito *et al.* (2016), Nengsih *et al.* (2019), Tetsi *et al.* (2019), Carapito *et al.* (2020)).

Ainsi, depuis le début de mes travaux de recherche, j'ai abordé plusieurs problématiques aussi bien d'un point de vue de la statistique théorique ou de la statistique appliquée que d'un point de vue de la statistique computationnelle. Je vais maintenant ci-dessous préciser ce terme. La **statistique computationnelle** peut constituer, pour certains chercheurs en statistique, l'interface entre la statistique et l'informatique. C'est le domaine de la science informatique spécifique à la science mathématique de la statistique. Cette zone se développe aussi rapidement, ce qui conduit à des appels qu'un concept plus large de l'informatique devrait être enseignée dans le cadre général de l'éducation statistique. Le terme de **statistique computationnelle** est souvent utilisé pour référer à des calculs intensifs des méthodes statistiques, y compris les méthodes de ré-échantillonnage, les méthodes de chaînes de Markov et de Monte-Carlo, la régression locale, l'estimation de la densité du noyau, les modèles additifs généralisés et depuis quelques temps les réseaux de neurones artificiels.

Comme je l'exprimais auparavant, je présenterai ci-dessous quelques sujets qui relèvent des quatre problématiques citées et définies ci-dessus : la régression pénalisée (Vallat *et al.* (2013), Jung *et al.* (2014), la pré-publication Aouadi *et al.* (2018)), la régression des moindres carrés partiels (PLS) (Magnanensi *et al.* (2016a), Magnanensi *et al.* (2017) et Nengsih *et al.* (2019)) et certaines de leurs extensions (Meyer *et al.* (2010), Bertrand *et al.* (2013b), Bastien *et al.* (2015) et la pré-publication Magnanensi *et al.* (2016b)) ainsi qu'une extension aléatoire de l'ajustement multidimensionnel, la pré-publication Alawieh *et al.* (2018). Ces quatre thématiques restent actives à ce jour et, depuis 2016, j'y ai ajouté deux composantes : la première s'appuie sur la théorie des bandits-manchots que nous utilisons dans les tests A/B présentée dans l'article de Claeys *et al.* (2017) et plus généralement de l'apprentissage statistique (Claeys *et al.* (2020b)). La deuxième se base sur les réseaux bayésiens (Kenett *et al.* (2020)). Mes principaux domaines d'application de ces deux nouvelles thématiques sont le monde industriel et le domaine médical. Un stage de fin d'étude sur l'application des réseaux bayésiens en radiothérapie, réalisé par Mélanie Piot (élève à l'ESIEA Paris, école d'ingénieurs du monde du numérique), a été financé par un PEPS1 (Projets Exploratoires Premier Soutien) de l'AMIES. Nous reviendrons sur ce stage par la suite dans ce chapitre. Il se prolongera à partir de septembre 2020 par une thèse de doctorat sur le même sujet, intitulée « Exploitation des données cliniques en radiothérapie. Approches multicentriques et causales ». Cette thèse sera co-financée par la région Grand-Est et l'Université de Technologie de Troyes (UTT) et encadrée par Frédéric Bertrand et moi-même à partir d'octobre 2020.

Comme je l'ai mentionné ci-dessus, les impulsions à l'origine de mes recherches peuvent se classer en deux catégories :

1. soit dans un but d'améliorer une méthodologie existante mais qui a montré ses limites. Je donne quelques exemples ci-dessous :

   (a) traitement des valeurs manquantes et des problèmes de colinéarité en régression logistique (Meyer *et al.* (2010)) ou pour les modèles de Cox (Bastien *et al.* (2015)),

   (b) influence des valeurs manquantes en régression des moindres carrés partiels (Nengsih *et al.* (2019)),

   (c) critère de choix de variables ou du nombre de composantes en régression des moindres carrés partiels et ses extensions aux modèles linéaires généralisés (Magnanensi *et al.* (2016a, 2017)),

   (d) cas des modèles parcimonieux (Magnanensi *et al.* (2016b); Bastien *et al.* (2015)),

2. soit même dues à la nécessité de concevoir des outils spécifiques pour des expériences innovantes pour lesquelles il faut inventer une solution faite pour ce nouveau jeu de données issu d'une nouvelle expérience. Je cite comme exemple :

   (a) l'inférence temporelle de réseaux de gènes (Vallat *et al.* (2013), Jung *et al.* (2014)),

   (b) l'inférence conjointe de réseaux (Bertrand et Maumy-Bertrand (2020e)),

   (c) la détermination de cibles optimales pour prédire une intervention dirigée fiable dans un réseau biologique (Aouadi *et al.* (2018)),

   tous ces outils tournent autour de la collaboration avec le PU-PH Seiamak Bahram (membre senior de l'institut universitaire de France) et le MCU-PH Laurent Vallat du nouvel hôpital civil de Strasbourg.

L'objectif que je me suis fixé, en tant qu'enseignant-chercheur est de maintenir un échange bi-directionnel entre les développements théoriques et les applications de ces derniers dans les projets nterdisciplinaires auxquels je participe. D'ailleurs, je me suis rendue compte rapidement que les hypothèses communément faites pour permettre une exploitation scientifiquement rigoureuse des modèles statistiques ne sont qu'exceptionnellement compatibles avec des jeux de données réels, même s'ils sont collectés avec les meilleurs protocoles expérimentaux puisque les problèmes rencontrés tiennent généralement à la nature même des observations et non à la méthodologie de mesure. Or ce sont ces hypothèses qui permettent une gestion rigoureuse des risques d'erreur qui apparaissent dans la théorie des tests de significativité ou des niveaux de confiance présents lors de la construction de régions de confiance. En outre, les séries statistiques ou les échantillons réels présentent souvent des problématiques additionnelles comme l'absence de certaines valeurs, présence de valeurs manquantes avec des mécanismes d'apparition plus ou moins complexes en fonction de l'appareil ou de la méthodologie de collecte des mesures, ou la présence de valeurs atypiques qui peuvent grandement influencer les outils statistiques utilisés.

Mon projet de recherche, voir le chapitre 7, propose des axes qui me permettront de continuer de s'intéresser aux thématiques citées ci-dessus (régression pénalisée, régression des moindres carrés partiels, inférence de réseaux biologiques, apprentissage statistique). Plusieurs demandes de financement sur ces axes ont été reçues positivement et ont remporté plusieurs appels à projets (PEPS1 et une future allocation doctorale comme je l'ai déjà mentionné, voir le chapitre **??**).

1.2 Expériences en matière d'encadrement

1.2.1 Encadrement de stages

Depuis mon arrivée à l'unité de formation et de recherche (UFR) de mathématique et d'informatique en septembre 2004 de l'Université de Strasbourg, j'ai pris l'habitude d'accompagner principalement des étudiants dans leur stage de fin de deuxième année de master. J'ai également, mais moins fréquemment, eu l'opportunité de suivre des étudiants de première année de master. De 2005 à 2007, j'encadrai principalement un ou deux étudiants du DESS de mathématiques discrètes - mathématiques appliquées à l'informatique dont le responsable était à l'époque le Professeur Dominique Collombier. Puis le DESS de Mathématiques discrètes - Mathématiques appliquées à l'informatique s'est éteint et a donné lieu à la naissance de deux masters : celui de biostatistique et de statistique industrielle et celui de mathématiques discrètes. J'ai continué d'encadrer des étudiants de master et plus particulièrement issus de celui de biostatistique. Ensuite, en 2010, Nicolas Poulin est arrivé au sein de l'équipe de statistique de l'IRMA en tant qu'ingénieur de recherche pour le CeStatS (centre de statistique de Strasbourg) et a pris en charge l'encadrement de tous les mémoires des étudiants sur les deux années du master de biostatistique.
Puis, de février 2013 jusqu'à juin 2019, j'ai accompagné quelques étudiants dans leur mémoire de deuxième année de master des métiers de l'éducation et de la formation, diplôme indispensable pour l'obtention du CAPES. Je citerai pour exemple le mémoire de Vincent Einsetler en juin 2013 sur l'utilisation de l'intervalle de confiance dans la vie quotidienne et celui d'Aline Girardet en juin 2019 sur la visualisation de données statistiques au collège.
J'encadre depuis plus de dix ans des étudiants dans le magistère, devenu diplôme d'université d'actuariat de Strasbourg (DUAS). Ce magistère comporte la dernière année d'une licence et du master d'actuariat. Ce DU, à vocation professionnelle, a la particularité de conférer, à l'issue des trois années de formation, le titre d'actuaire, reconnu par l'institut des actuaires, à chaque étudiant qui se présente devant un jury formé d'universitaires et de responsables de l'institut des actuaires. Depuis 2016, la dernière année du DUAS peut être faite sous forme de formation en alternance. J'ai aussi été la tuteure académique de 14 étudiants alternants. En tant que tuteure académique, je devais aller visiter au moins deux fois les étudiants sur leur lieu de travail en alternance, souvent situé à Paris.
Enfin, j'encadre depuis plus d'une dizaine d'années, un ou deux élèves ingénieurs issus de l'ESIEA (école d'ingénieurs) dans leur stage de fin d'étude.
Par exemple, pour cette année universitaire, j'encadre :

- Ambre-Tiffany Etame (01 Octobre 2019 - 01 Octobre 2020). Mémoire d'alternance de fin d'étude pour obtenir le DUAS et le master 2 d'Actuariat à

l'AG2R LA MONDIALE. Le titre du mémoire est la construction de lois statistiques de mortalité des autonomes et d'entrée en dépendance (du risque dépendance).

- Vadym Hadetskyi (16 Mars 2020 – 16 Septembre 2020). Stage de master 2 de CSMI au sein de Vivialys. L'objectif du stage est de développer et de mettre en œuvre de multiples méthodes basées sur l'intelligence artificielle pour améliorer l'expérience client et la stratégie marketing de l'entreprise.

- Elizaveta Logosha (15 Février 2020 -15 Août 2020). Stage de master 2 de statistique au sein de l'Institut Français de la Vigne. Le sujet est la mise en place d'une méthodologie de traitement des données pour lutter contre l'ESCA et le *Black Dead Arm*. Ce stage est financé par un PEPS 1 de l'AMIES dont je suis la responsable scientifique.

- Théo Mardoc (01 Octobre 2019 - 01 Octobre 2020). Mémoire d'alternance de fin d'étude pour obtenir le DUAS et le master 2 d'actuariat à la MATMUT. Le titre du mémoire est le calcul et la projection d'un score sociétaire.

- Mélanie Piot (08 Mars 2020 – 08 Septembre 2020). Stage de fin d'étude pour l'obtention du titre d'ingénieur (Mélanie doit rédiger son mémoire qui doit avoir un niveau de master 2 pour ensuite commencer une thèse de doctorat). L'objectif du stage est l'exploitation des données cliniques de radiothérapie. Ce stage est financé par un PEPS1 de l'AMIES dont je suis la co-responsable scientifique avec Frédéric Bertrand.

- Nikkora Pra Ankhann (01 Octobre 2019 - 01 Octobre 2020). Mémoire d'alternance de fin d'étude pour obtenir le DUAS et de master 2 d'actuariat à Actuelia. Le titre du mémoire de fin d'étude est l'implémentation du PER au sein d'un assureur.

- Oleksandr Sorochynskyi (01 Octobre 2019 - 01 Octobre 2020). Mémoire d'alternance de fin d'étude pour obtenir le DUAS et le master 2 d'actuariat à Prim'Act. Le titre du mémoire est la quantification de l'impact des caractéristiques d'un assuré sur le marge espéré dans un contexte IARD.

- Michel Venniro (01 Octobre 2019 - 01 Octobre 2020). Mémoire d'alternance de fin d'étude pour obtenir le DUAS et le master 2 d'actuariat aux ACM de Strasbourg. Le titre du mémoire est Zonier MRH France : risque d'inondation.

Plusieurs des étudiants que j'ai encadré ont commencé, ou vont commencer, une thèse à l'issue de leur stage comme Nicolas Jung (2011), Fatima Fahs (2018), Mélanie Piot (2020), Elizaveta Logosha (2020).

1.2.2 Encadrement de doctorants

- Nicolas Jung (1er octobre 2011 – 03 décembre 2014) : thèse de doctorat en biologie des systèmes dirigée par le PU-PH Seiamak Bahram du centre de recherche d'immunologie et d'hématologie, directeur du LabEx Transplantex, également membre senior de l'IUF, intitulée « Modélisation de phénomènes biologiques complexes : application à l'étude de la réponse antigénique de lymphocytes B sains et tumoraux ». Pour la partie « Modélisation de phénomènes biologiques complexes », j'ai assuré l'encadrement de thèse de Nicolas Jung en me coordonnant avec Frédéric Bertrand qui travaillait en parallèle avec le MCU-PH Laurent Vallat sur le même jeu de données mais avec des approches différentes et des questions biologiques parfois différentes. La thèse a été soutenue le 03 décembre 2014 à l'institut d'hématologie de l'hôpital civil. Voici le résumé du sujet de thèse : la biologie des systèmes complexes est le cadre idéal pour l'interdisciplinarité. Dans cette thèse, les modèles et les théories statistiques répondent aux modèles et aux expérimentations biologiques. Nous nous sommes intéressés au cas particulier de la leucémie lymphoïde chronique à cellules B, qui est une forme de cancer des cellules du sang. Nous avons commencé par modéliser le programme génique tumoral sous-jacent à cette maladie et nous l'avons comparé au programme génique d'individus sains. Pour ce faire, nous avons introduit la notion de réseau en cascade. Nous avons ensuite démontré notre capacité à contrôler ce système complexe, en prédisant mathématiquement les effets d'une expérience d'intervention consistant à inhiber l'expression d'un gène. Cette thèse s'achève sur la perspective d'une modulation orientée, c'est-à-dire le choix d'expériences d'intervention permettant de « reprogrammer » le programme génique tumoral vers un état normal. Pour résumer, dans cette thèse, nous nous sommes intéressés au problème de reconstruction de réseau de régulation génique. Les outils utilisés que nous avons utilisés pour résoudre ce problème sont des régressions pénalisées de type *lasso* (autrement dit, avec une norme L_1). La thèse de Nicolas Jung était confidentielle jusqu'en décembre 2019 (dedans y étaient mentionnés de nouveaux résultats biologiques sur le cancer qui étaient encore entrain d'être vérifiés au moment de la soutenance). La thèse a donné lieu à deux principales publications (Vallat *et al.* (2013) et Jung *et al.* (2014)). Nicolas Jung a également contribué à deux études statistiques qui ont donné deux autres articles scientifiques.
  Nicolas Jung nous a quitté, après une année de post-doctorat au sein de l'IRMA (que je développerai par la suite), pour rejoindre le monde bancaire, le Crédit Agricole Alsace Vosges, en mars 206, comme *data scientist*.

- Emmanuelle Claeys (1er septembre 2016 – 12 novembre 2019) : thèse de doctorat en informatique financée par un dispositif comparable à celui du Cifre en collaboration avec l'entreprise AB Tasty, située à Paris, co-encadrée par le Professeur Pierre Gançarski, de l'équipe Science des données du laboratoire I-Cube et moi-même intitulée « Clusterisation incrémentale, multicritères de données hétérogènes pour la personnalisation d'expérience utilisateur ». La thèse a été soutenue le 12 novembre 2019. Voici le résumé du sujet de thèse : dans de nombreux domaines (santé, vente en ligne,...) concevoir *ex nihilo* une solution optimale répondant à un problème défini (trouver un protocole augmentant le taux de guérison, concevoir une page web favorisant l'achat d'un ou plusieurs produits,...) est souvent très difficile voire impossible. Face à cette difficulté, les concepteurs (médecins, *web designers*, ingénieurs de production,...) travaillent souvent de façon incrémentale par des améliorations successives d'une solution existante. Néanmoins, définir les modifications les plus pertinentes restent un problème difficile. Pour tenter d'y répondre, une solution adoptée de plus en plus fréquemment consiste à comparer concrètement différentes alternatives (appelées aussi variations) afin d'en déterminer celle(s) répondant le mieux au problème via un test A/B. L'idée est de mettre en œuvre réellement ces alternatives et de comparer les résultats obtenus, c'est-à-dire les gains respectifs obtenus par chacune des variations. Pour identifier la variation optimale le plus rapidement possible, de nombreuses méthodes de test utilisent une stratégie d'allocation dynamique automatisée. Le principe est d'allouer le plus rapidement possible et automatiquement, les sujets testés à la variation la plus performante, par un apprentissage par renforcement. Parmi les méthodes possibles, il existe en théorie des probabilités les méthodes de bandit manchot. Ces méthodes ont montré leur intérêt en pratique mais également des limites, dont en particulier un temps de latence (c'est-à-dire un délai entre l'arrivée d'un sujet à tester et son allocation) trop important, un déficit d'explicabilité des choix et la non-intégration d'un contexte évolutif décrivant le comportement du sujet avant d'être testé. L'objectif de cette thèse est de proposer une méthode générique d'un test A/B permettant une allocation dynamique en temps réel capable de prendre en compte les caractéristiques des sujets, qu'elles soient temporelles ou non, et interprétable *a posteriori*.
  Depuis le 09 décembre 2019, Emmanuelle Claeys est post-doctorante à l'UTT et chez *Your Data Consulting*, *start-up* parisienne qui finance environs 75% du projet. L'autre partie du financement a été obtenu par un PEPS2 de l'AMIES. À l'heure où j'écris ce manuscrit, Emmanuelle Claeys a obtenu un poste de MCF à l'Université de Toulouse en 27ème section.

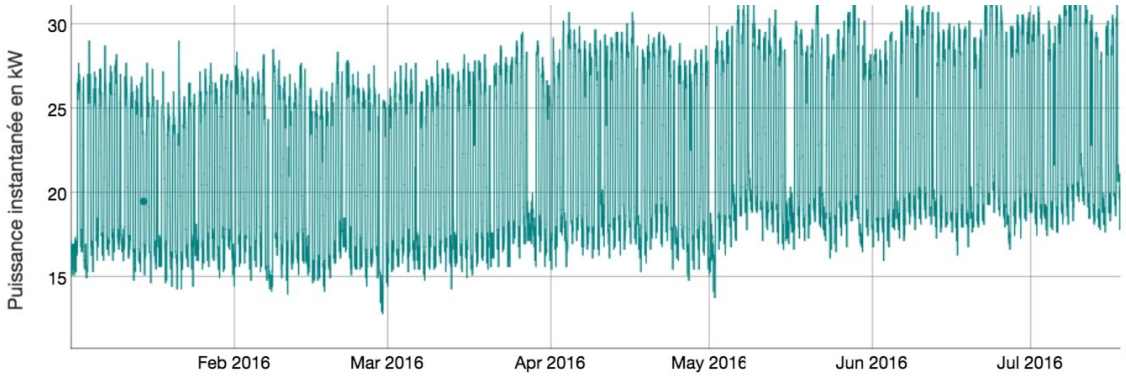


Figure 1.1 : *Exemple de courbe de charge.*

- Fatima Fahs (02 mai 2019 – mai 2022) : thèse de doctorat en mathématiques appliquées financée par un dispositif Cifre en collaboration avec l'entreprise Électricité de Strasbourg (abrégée en ES), co-encadrée par Frédéric Bertrand, Professeur à l'Université Technologique de Troyes et moi-même intitulée « Analyse des courbes de charge d'électricité et prédiction à court terme dans les secteurs résidentiel et tertiaire ».

  Le sujet de thèse a pour objectif principal de déployer des modèles statistiques ou à base de techniques d'intelligence artificielle qui permettraient de déterminer client par client (la majorité des clients de l'ES sont des particuliers) la consommation électrique le jour j à court terme ($j+1$ à $j+3$) à partir des historiques de consommation de chacun d'entre eux, des données météorologiques et d'autres paramètres qualitatifs pour les secteurs résidentiel et tertiaire. Pour cela, nous nous concentrons sur l'analyse des courbes de charge (voir Figure 1.1).

  Les données étudiées sont ici des courbes journalières de charge des clients de l'ES, avec un pas de 30 minutes. Les saisonnalités observées sur les courbes de charge (voir Figure 1.2) sont liées à l'activité des clients de l'ES :

  1. saisonnalité annuelle pour les changements de température (clients thermo-sensibles),
  2. saisonnalité hebdomadaire (différence de consommation entre la semaine et le *week-end*),
  3. saisonnalité journalière (consommation plus faible la nuit).

  À ces courbes de charge sont traditionnellement ajoutées des variables exogènes comme par exemple la météorologie locale, et plus particulièrement la température extérieure, qui aura un impact chez les clients disposant de chauffage électrique individuel (thermo-sensibles, voir Figure 1.3).

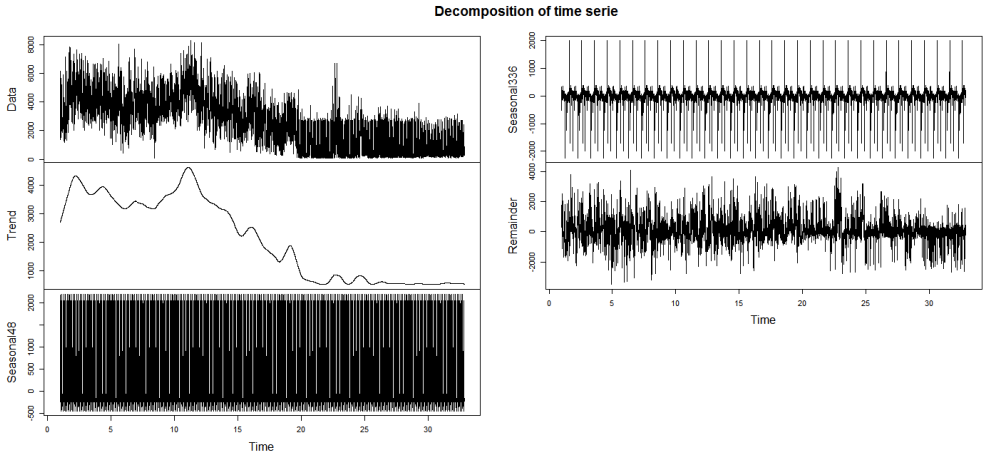


Figure 1.2 : *La décomposition d'une courbe de charge d'un client montre les deux saisonnalités hebdomadaire (seasonal 336) et journalière (seasonal 48).*

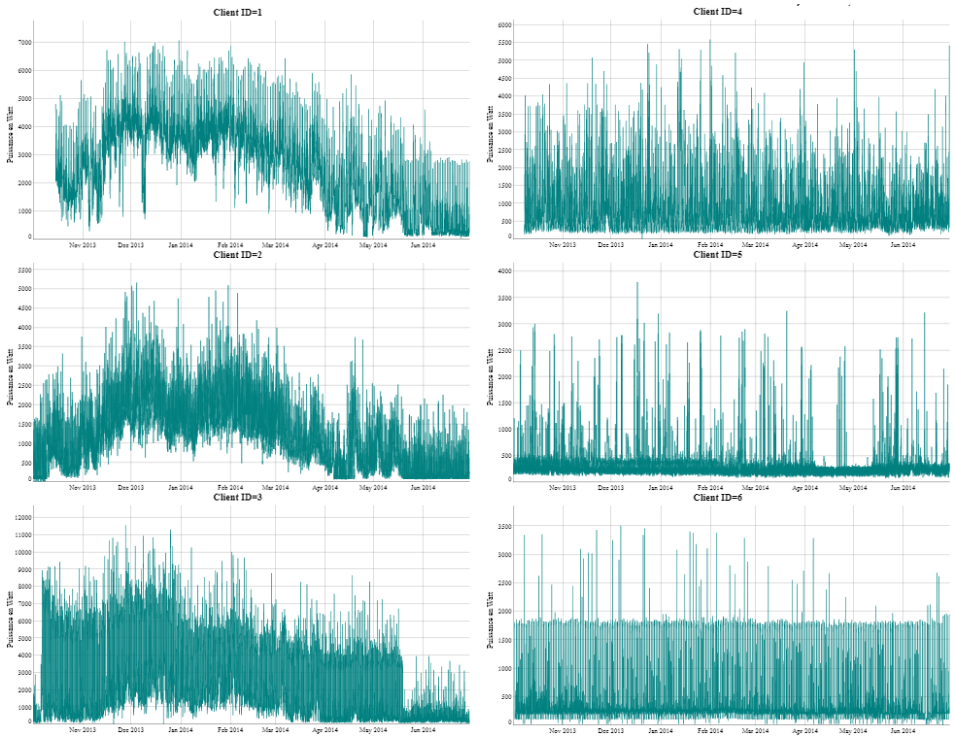


Figure 1.3 : *Client thermo-sensible (à gauche) et non thermo-sensible (à droite).*

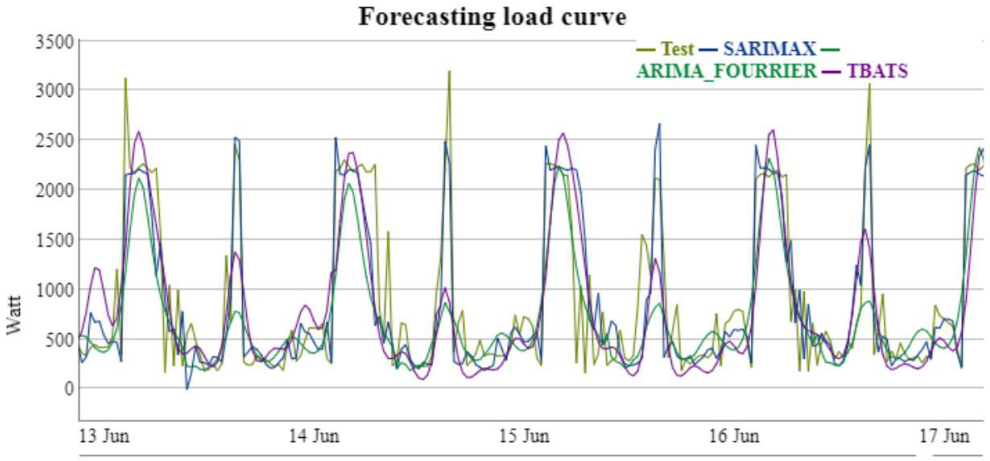


Figure 1.4 : *Résultats préliminaires de prévision.*

Cette approche, qui avait été validée à l'occasion de travaux préliminaires réalisés pendant le stage de six mois de Fatima Fahs pour valider son semestre 4 du master CSMI, doit permettre de procurer à chaque client une prévision à court terme ($j+1$ à $j+3$) de sa propre consommation (voir Figures 1.4 et 1.5), de détecter et de lui communiquer d'éventuelles anomalies telles que des écarts anormaux de la consommation mesurée par rapport à la consommation prévue sur la base des historiques de consommation du client et des conditions météorologiques.

Afin de présenter aux clients de l'ES des analyses de leur consommation journalière dans un format pratique et une interface homme-machine qui leur sera accessible depuis leur téléphone mobile, il leur sera proposé non pas une analyse de leur courbe journalière de charge (Watt en fonction du temps), mais de leur consommation journalière d'énergie (kWh) (voir Figure 1.6) en fonction de l'heure pour chaque jour de l'année.

L'approche étudiée dans le cadre du stage de master CSMI de deuxième année a permis de valider le principe de remontée des anomalies au client basées sur une analyse des laps de temps pendant lesquels la consommation d'énergie se situe en dehors d'intervalles de confiance de type BCa (*bias corrected and accelerated*) à 95% de confiance (voir Figure 1.7).

Avec de tels modèles il est également prévu de développer et de tester des algorithmes de détection d'anomalies dans les courbes de charge des clients afin de développer des services à valeur ajoutée personnalisés à destination de chaque client. Nous pouvons par exemple citer la détection « passive » de pannes sur des équipements du client par recherche

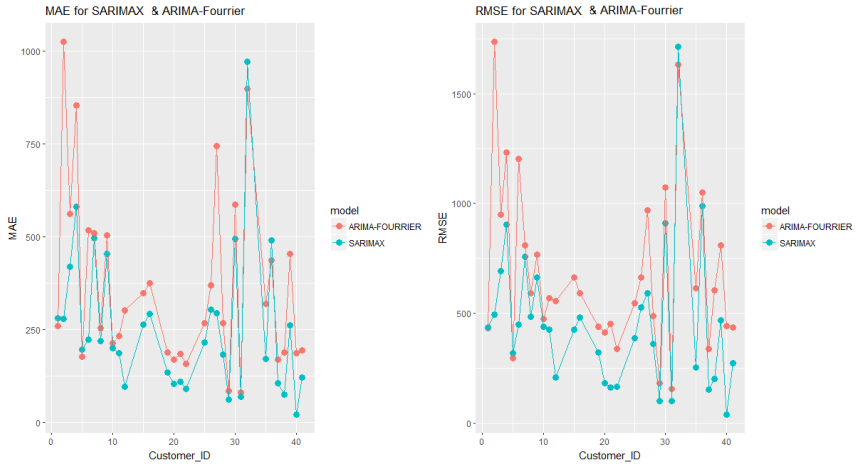


Figure 1.5 : *Les mesures d'erreur de prévision de deux modèles SARIMAX et ARIMA-Fourrier pour 50 clients.*

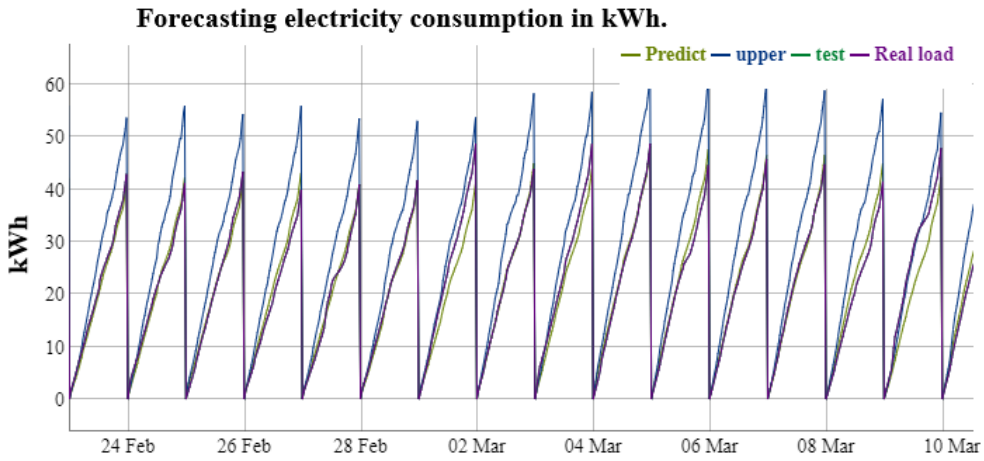


Figure 1.6 : *Énergie journalière consommée en fonction du temps.*

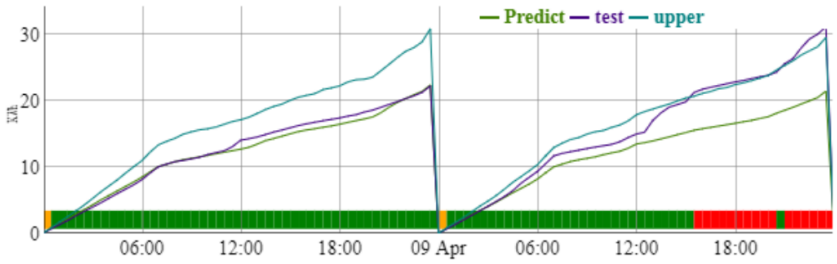


Figure 1.7 : *En rouge les comportements sortant de l'IC et donnant lieu à une alerte.*

de sous-consommation ou de sur-consommation (par exemple au niveau des chambres froides, des systèmes de chauffage-ventilation, des chauffe-eaux ou des pompes à chaleur dans le secteur tertiaire, artisanal ou dans les collectivités locales).

Les principaux verrous scientifiques sont le développement d'algorithmes de prévisions individuelles, leurs qualifications sur des segments de clients très disparates, la sélection de modèles adaptés par un sous-ensemble d'usagers (tertiaires, artisanat, client particulier, ...). L'optimisation de la performance de ces calculs sera également un point clé du travail envisagé.

- Mélanie Piot (15 septembre 2020 – septembre 2023) : thèse de doctorat en mathématiques appliquées soutenue par un co-financement région Grand Est et Université Technologique de Troyes, co-encadrée par Frédéric Bertrand, Professeur à l'Université Technologique de Troyes et moi-même intitulée « Exploitation des données cliniques en radiothérapie. Approches multicentriques et causales ». Voici le résumé du sujet de thèse : la radiothérapie externe est une des techniques utilisées au Centre Paul Strauss de Strasbourg pour la prise en charge des patients atteints d'un cancer. Elle consiste à irradier la tumeur au moyen d'un faisceau de rayons X de quelques méga volts, ce qui est mis en œuvre à l'aide de données dosimétriques en suivant un *Treatment Planning Systems*. L'évolution de la radiothérapie tend vers une personnalisation accrue des traitements, et donc une prise en compte massive des données disponibles. Il faudra pour cela enrichir les données dosimétriques par des données cliniques capturées lors des consultations médicales durant le traitement et au-delà.

  Le département de radiothérapie du Centre Paul Strauss (Institut Régional du Cancer) s'est engagé sur cette voie depuis trois ans, et a mis en place une méthode de saisie des données cliniques au moyen d'un *Oncology Information System*. Environ 10 000 données cliniques structurées sont produites par mois et une base de données de plus de 1 600 000 données cliniques sont disponibles et exploitables à Strasbourg. Un déploiement multicentrique est en cours sur trois centres français. Cette base de données permet actuellement de répondre à des questions médicales précises. Au moyen d'un modèle prédictif simple, elle permet également d'identifier les patients qui présentent une dynamique singulière d'apparition d'une complication lors de son traitement. La prochaine étape que nous envisageons de franchir est la recherche de liens multiples entre les effets secondaires en utilisant et en développant des outils spécifiques d'extraction de connaissance. C'est ce travail qui sera mené dans un contexte d'interdisciplinarité : deux mathématiciens, deux physiciens médicaux et deux radio-

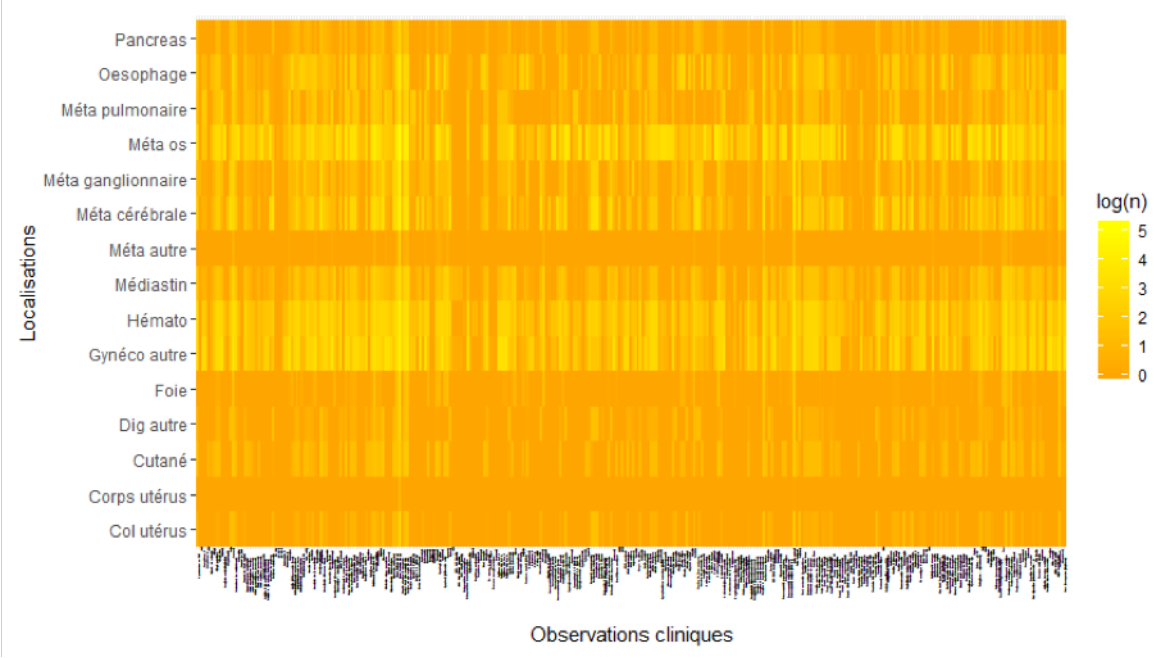


Figure 1.8 : *Cartographie des données disponibles.*

thérapeutes. Ces derniers enrichiront l'analyse statistique par les relations entre les paramètres qui sont déjà connues dans la littérature médicale. La cartographie (voir Figure 1.8) donne le nombre de données disponibles par localisation anatomique et par complication clinique. Elle illustre la richesse de la base de données qui a été constituée par l'équipe de Christophe Mazzara et son potentiel prédictif.

Ce travail permettra de construire les premiers modèles causaux qui seront d'un intérêt premier pour la compréhension des phénomènes. Un coencadrement pour les aspects physique médicale et radiothérapie sera assuré par Christophe Mazzara, responsable du service physique médicale et radioprotection, membre du laboratoire des sciences de l'ingénieur, de l'informatique et de l'imagerie (ICube), UMR 7357, de l'Université de Strasbourg. Christophe Mazzara se fera accompagné par un physicien médical (Philippe Meyer) et par deux radiothérapeutes (Jean-Baptiste Clavier et Sébastien Guihard) comme mentionné auparavant.

Des premiers résultats ont été obtenus à l'aide des réseaux bayésiens lors du travail de stage de Mélanie Piot. Ils ont été accueillis très positivement par les médecins.

1.2.3 Encadrement de post-doctorats

- Nicolas Jung (décembre 2014 – février 2016) : post-doctorat

  L'objectif du sujet de recherche, proposé à Nicolas Jung pour son travail post-doctoral, se concentre autour du thème suivant : quand une cellule reçoit une stimulation de l'environnement, un programme génique est activé. Plusieurs centaines de gènes sont exprimés et traduits en protéines qui apportent alors une réponse adaptée au stimulus.

  La réponse génique complexe et dynamique peut être modélisée statistiquement par un réseau dans lequel les nœuds correspondent aux gènes et les liens correspondent à leurs interactions. Depuis l'introduction de technologies à haut débit qui permettent de mesurer simultanément l'expression de milliers de gènes, beaucoup de méthodes statistiques ont été proposées pour l'inférence de ces réseaux de régulation. Ces méthodes peuvent être regroupées en trois catégories principales :

  1. les méthodes dites d'interactions,
  2. les méthodes dites d'optimisation et
  3. les méthodes basées sur des équations différentielles ou des régressions.

  Dans ces différentes approches, seule l'expression génique est prise en considération pour l'inférence de ces modèles. Nicolas Jung a commencé ses recherches post-doctorales par l'amélioration des résultats qu'il avait obtenus dans sa thèse sur la sélection sûre de prédicteurs dans les réseaux de gène puis il a continué ses recherches dans cette voie cherchant à appliquer des techniques d'inférence bayésienne de type *Approximate Bayesian Computation* (abrégé en *ABC*) à l'inférence des réseaux de gènes.

  Une des difficultés à laquelle Nicolas Jung a été confronté est le coût algorithmique de ces approches de type *ABC*. Avec Nicolas Jung, nous avons réfléchi à la diminution de ces coûts en créant un nouveau modèle d'inférence bayésienne pour les réseaux biologiques, voir Figure 1.9. Celle-ci devait être implémentée avec un langage très rapide comme le langage C++. C'est pourquoi j'ai déposé une demande de financement d'un ingénieur d'étude auprès du LabEx IRMIA (`http://labex-irmia.u-strasbg.fr`), qui a été acceptée. C'est ainsi que Khadija Musayeva a été recrutée sur ce poste. Ces travaux ont donné naissance au *package* `networkABC` (Bertrand et Maumy-Bertrand (2020b)) pour le langage `R`. Ce travail a été accepté à la conférence useR ! 2020 qui aurait du avoir lieu à Saint-Louis (Missouris) en

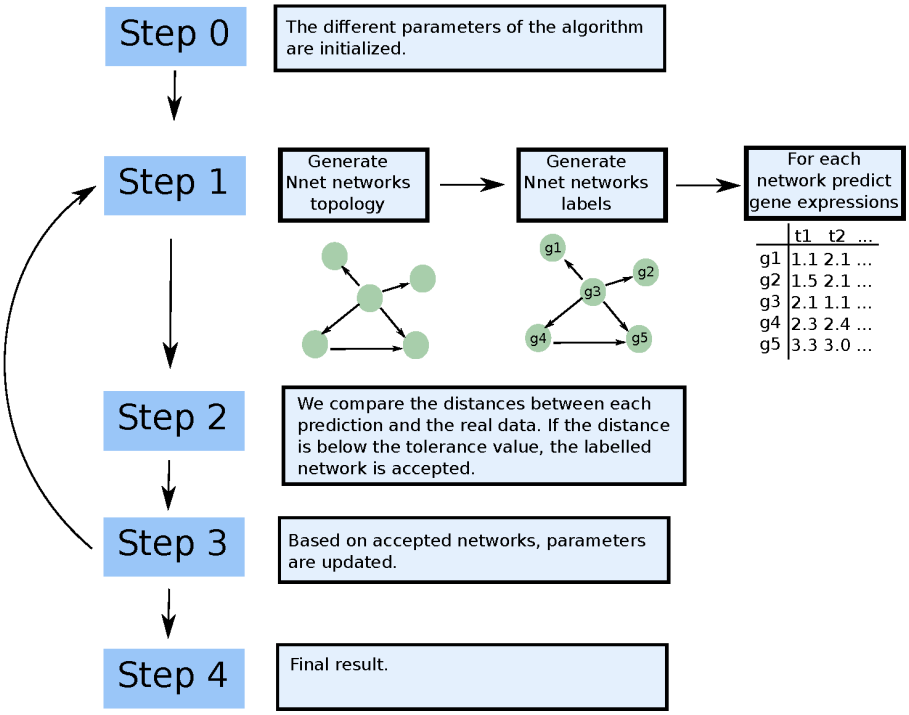


Figure 1.9 : *Algorithme ABC pour l'inférence de réseau biologique.*

juillet 2020 (Bertrand et Maumy-Bertrand (2020c)) mais celle-ci est depuis devenue virtuelle à cause de la crise sanitaire liée au Covid-19.

Notre objectif ultime était de réaliser une inférence conjointe du réseau de régulation génique et protéique en réponse à une stimulation cellulaire dynamique car nous disposions d'un jeu de données d'expression temporelle de protéines (Perrot *et al.* (2011)), mesurée conjointement à l'expression de gènes (Vallat *et al.* (2007)) dans notre modèle cellulaire, mais qui n'avait pas été exploité dans le cadre d'une inférence statistique. Celle-ci a été réalisée mais bien après la fin du travail post-doctoral de Nicolas Jung (Bertrand et Maumy-Bertrand (2020e)). Ce travail a aussi été accepté à la conférence useR ! 2020 (Bertrand et Maumy-Bertrand (2020a)) et aurait aussi dû y être présenté.

- Emmanuelle Claeys (décembre 2019 – décembre 2021) : post-doctorat. Ce projet s'inscrit dans le cadre d'une collaboration de recherche entre la société *Your Data Consulting*, jeune entreprise innovante basée à Paris, et deux enseignants-chercheurs à savoir Frédéric Bertrand et moi-même, sur la thématique de la science des données et plus particulièrement de l'intelligence artificielle. *Your Data Consulting* est propriétaire de la plate-forme *SaaS LiveJourney* qui est présentée à l'adresse suivante : `https://www.livejourney.com`
  Cette plate-forme permet aux entreprises de vente par correspondance ou aux entreprises de production ou celles de livraison de visualiser et d'analyser de façon dynamique les parcours de leurs clients, ou de leurs produits ou de leurs colis. De plus, elle permet aux entreprises d'anticiper des événements qui pourraient ralentir le bon déroulement du processus et de détecter les goulots d'étranglement.

  Plusieurs thématiques ont été sélectionnées pour faire l'objet d'un travail de recherche post-doctoral et être particularisées au contexte du *process mining*. Les voici :

  - *root cause analysis*,
  - prédiction des processus,
  - *clustering* de séries temporelles,
  - causalité et intelligence artificielle.

Un exemple de *process* est représenté à la Figure 1.10.

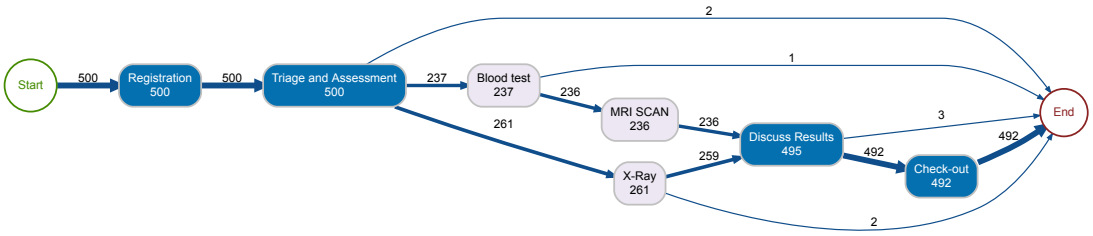


Figure 1.10 : *Visualisation d'un exemple de process.*

1.2.4 Encadrement d'ingénieurs d'étude ou de recherche

Dans le cadre du projet pluridisciplinaire GenPred, `http://www.math.unistra.fr/genpred/`.

- mai 2015 – octobre 2015 : co-encadrement à 50% d'une ingénieure d'étude, Khadija Musayeva. Ces six mois (que j'ai obtenus pendant le post-doctorat de Nicolas Jung afin de renforcer l'équipe GenPred) ont permis la réalisation d'un code écrit en langage C pour l'inférence de réseaux avec la méthode ABC. En deux mots, les méthodes de type ABC sont de plus en plus utilisées, et ce, dans des domaines les plus divers. Toutefois, le domaine de la biologie apparaît comme étant un terrain d'étude particulièrement adapté pour les méthodes ABC. Notre recherche était avant tout portée par un impératif biologique : celui de trouver des cibles dans le réseau de gènes pour moduler le comportement de la cellule. L'idée sous-jacente est qu'en diminuant l'expression de certains gènes, nous pourrions rendre le cancer moins nocif. Pour répondre à cet objectif, la méthode développée dans la thèse de Nicolas Jung trouve une limite. En effet, s'il est possible des caractéristiques globales du réseau de gènes, la méthode ne permet pas de déterminer avec quelle assurance les liens inférés sont forts. Autrement dit, nous avons de l'assurance pour déterminer les propriétés globales du réseau, mais nous manquons d'information sur les propriétés locales. La méthode ABC permet de répondre à cette question. Cependant, l'utilisation d'une telle méthode pour l'inférence de gènes pose des défis à la fois informatiques et mathématiques. Pour note, seul un article était paru dans la littérature pour traiter ce problème (Rau *et al.* (2012)). Les défis informatiques sont liés à la nature même des algorithmes ABC. En effet, ces derniers sont gourmands en temps de calcul. Il est donc nécessaire d'avoir un code écrit dans un langage rapide et de façon optimisée. C'était là qu'intervient le rôle de l'ingénieure d'étude. Khadija Musayeva a poursuivi son cursus universitaire par une thèse en apprentissage statistique au LORIA (Nancy) dirigée par Yann Gueurmeur et qui a été soutenue en octobre

2019. J'ai été membre du jury de thèse de Khadija Musayeva.

- octobre 2014 – mai 2016 : co-encadrement à 50% d'un ingénieur de recherche, Marius Kwemou avec Frédéric Bertrand. Ce poste d'ingénieur de recherche a été financé par l'ITMO Cancer. Le sujet porte sur l'inférence de réseaux pour modéliser le programme génique. Marius Kwemou a également travaillé sur des sujets connexes en collaboration avec le Professeur Seiamak Bahram ce qui a amené à trois publications au total.

Dans le cadre du projet région CheckWave et du PEPS1 DataFlow.

- décembre 2018 – janvier 2020 : encadrement à 100% d'un ingénieur de recherche, Jean-Baptiste Wahl au sein de CEMOSIS. Il a travaillé sur le PEPS1 DataFlow (Walh *et al.* (2019)) qui est devenu le projet CheckWave, projet en collaboration avec la filiale française de Bürkert.

- avril 2020 – juin 2021 : encadrement à 100% d'un ingénieur de recherche, Yannick Stoll au sein de CEMOSIS. Il travaille sur le projet région CheckWave, projet en collaboration avec la filiale française de Bürkert.

Le projet région CheckWave vise à développer et produire des produits de technologies plus sensibles, qui impliquent le développement d'un nouveau processus de fabrication et de test très précis et répétable, basé notamment sur le traitement et la qualification des données.
Ce projet CheckWave a été labellisé par HYDREOS[1]. Il bénéficie du soutien de la Région Grand Est via son dispositif « Aide aux projets de recherche et de développement et d'innovation des grandes entreprises ». Il est porté par Bürkert en partenariat avec la plateforme CEMOSIS de l'Université de Strasbourg.
Ce projet consiste en la réalisation d'un nouveau banc de test avec un double objectif. Tout d'abord, pour permettre à Bürkert de qualifier les nouveaux développements, mais aussi, pour calibrer les produits. L'objectif ambitieux de cette installation est d'atteindre des niveaux de précision très importants ($< 0{,}1\%$) avec une répétabilité certaine, le tout de façon rapide et automatisée pour une gamme vaste de produits. De plus l'installation sera certifiée COFRAC/ISO17025 ce qui permettra d'ouvrir le laboratoire à d'autres sociétés.
La partie exploitation des données est également un axe important du projet qui nécessitera l'intégration d'une approche type analyse big data. L'objectif étant d'améliorer les processus de développement et de production à la lumière de ces analyses de données.

[1] `https://www.hydreos.fr/news/236/34/Financement-des-projets-HydroScreen-et-CheckWave-labellisés-par-HYDREOS.html`

1.3 Expériences en matière d'enseignement

C'est naturellement que je suis intervenue et que j'interviens encore dans divers composantes (biologie, chimie, neurosciences, psychologie, service de formation continue) ou à l'école doctorale de la vie et santé de l'Université de Strasbourg en plus de l'UFR de mathématique et d'informatique.

Par conséquent j'ai été confronté à des publics variés tous de niveaux master ou plus lorsqu'il s'agit d'intervention dans un module d'école doctorale. Cette diversité a été à la fois très intéressante et stimulante mais aussi très chronophage. Ainsi, par exemple, j'ai dû me former à différents logiciels ou langages informatique en fonction des besoins du public concerné. J'utilise R, Python, Excel et XLStat dans mes enseignements mais j'ai aussi dû utiliser dans le passé les logiciels payants comme SPSS et Statistica.

Depuis 2009, j'ai rédigé douze livres, avec des co-auteurs. Deux d'entre eux ont été réédités, et par conséquent retravaillés, trois fois, pour un public d'étudiants de licence, ou de master ou encore d'école d'ingénieurs : Fredon *et al.* (2009a), Fredon *et al.* (2009b), Fredon *et al.* (2009c), Bertrand *et al.* (2011), Bertrand et Maumy-Bertrand (2011), Bertrand et Maumy-Bertrand (2012), Bertrand *et al.* (2013a), Bertrand *et al.* (2016), Bertrand et Maumy-Bertrand (2018a) 3^{e} édition, Bertrand *et al.* (2018), Meyer *et al.* (2018) 3^{e} édition, Bertrand *et al.* (2019b).

1.4 Expériences en matière d'expertise scientifique

Depuis janvier 2003, je participe, sans interruption, à des projets de recherche en partenariat avec des acteurs du monde industriel. J'ai exercé mon expertise scientifique avec les entreprises suivantes :

- pour l'Observatoire Régional de Tourisme en Bretagne (2003 - - 2006),
- pour Générale des eaux sur plusieurs sujets (2004 - - 2007 et 2007 - - 2008),
- pour Lilly (2005 - - 2006),
- pour Interstat (2006 - - 2010),
- pour Électricité de Strasbourg (depuis 2006),
- pour GlaxoSmithKline (2007 - - 2008),
- pour Merck (depuis 2014),
- pour *Your Data Consulting* (depuis 2016),

- pour Bürkert (depuis 2018),
- pour Le Sportif (depuis 2019),
- pour l'Institut Français de la Vigne et du Vin (depuis 2019),
- pour l'Institut de Cancérologie de Strasbourg (depuis 2019).

1.5 Expériences en matière d'animation de la communauté scientifique

- Depuis 2015, j'organise les journées d'étude en statistique (JES) de la Société Française de Statistique. Ces journées, qui ont lieu tous les deux ans (année paire), ont pour objectif de se consacrer pendant une semaine à l'approfondissement d'un thème bien défini, dans un lieu favorisant rencontres et discussions. Je m'occupe avec Frédéric Bertrand principalement de la mise en forme et de la relecture approfondie du livre composé des exposés des orateurs et puis ensuite nous confions le volume aux éditions Technip, (Bertrand *et al.* (2017), Maumy-Bertrand *et al.* (2018), Bertrand *et al.* (2019a)).
- Depuis septembre 2017, je suis correspondante pour la région Grand-Est pour l'AMIES (agence pour les mathématiques en interaction avec l'entreprise et la société). Mon rôle consiste à mettre en relation les entreprises qui pourraient avoir besoin au sein de leur département de recherche et de développement des mathématiques et un ou des chercheurs des laboratoires se trouvant dans le Grand Est et plus particulièrement à Strasbourg.
- Depuis janvier 2018, je co-organise (Emmanuelle Crépeau en décembre 2018, Jérôme Lelong et Nicolas Seguin en octobre 2019 et en octobre 2020), chaque année, le forum emploi maths, manifestation nationale soutenue par l'AMIES et les sociétés savantes de mathématiques qui réunit plusieurs acteurs : les étudiants de la licence jusqu'au doctorat, les jeunes diplômés, les formations diplômantes en mathématiques et bien sûr les entreprises.
- Du 1er septembre 2018 au 30 novembre 2018, j'ai été membre du comité d'organisation de la Semaine d'Etude Maths-Entreprises qui a eu lieu à Strasbourg. `https://seme2018.cemosis.fr`.
- De mai 2018 à septembre 2019, j'ai été secrétaire du comité d'organisation du 6ème Colloque Francophone International sur l'Enseignement de la Statistique (CFIES), manifestation qui a eu lieu à Strasbourg du 25 au 27

septembre 2019. L'objectif du colloque est de rassembler durant trois jours les enseignants et les chercheurs des disciplines concernées : la statistique évidemment, mais aussi les mathématiques, les sciences de l'éducation, l'ingénierie des connaissances, la didactique de la statistique, l'informatique, sans oublier les disciplines utilisatrices de l'outil statistique, tant du côté des sciences exactes que de celui des sciences humaines ou de la santé et enfin les entreprises qui ont à traiter de l'information en ayant recours aux techniques statistiques. `https://cfies2019.sciencesconf.org`.

1.6 D'autres expériences

1.6.1 Au niveau national

- J'ai obtenu la prime d'encadrement doctoral et de recherche (PEDR) en 2012. Elle a été renouvelée en 2016. À l'heure où j'écris ce manuscrit, je l'ai demandée à nouveau et je suis donc en attente des résultats.

- J'ai été membre élu du conseil national des universités (CNU) de la 26ème section (Mathématiques appliquées et applications des mathématiques) de 2011 à 2015, puis membre réélu en 2015. Pendant quatre années (de novembre 2015 à novembre 2019), j'ai été membre du bureau comme assesseur de rang B.

- J'ai été, à partir de janvier 2015, membre du comité d'organisation des 50 ans de l'Institut de Recherche en Mathématique Avancée (IRMA) de l'Université de Strasbourg. Cette manifestation a eu lieu du 07 au 09 Janvier 2016. `https://www.unistra.fr/index.php?id=19773&tx_ttnews%5Btt_news%5D=11434&cHash=301775a1136070bdc42a4c9a593bd8fc`

- J'ai encadré, à partir de 2017 et ce pendant trois années, chaque année un(e) étudiant(e) dans le dispositif EAP (Étudiant Apprenti Professeur). Ce contrat d'apprentissage permet à l'étudiant dès sa deuxième année de licence de découvrir le monde de l'enseignement. Il vise à permettre à davantage d'étudiants d'origine modeste de s'orienter vers les métiers de l'enseignement en leur proposant un parcours professionnalisant et en les accompagnant financièrement. Mon rôle consistait à accompagner l'étudiant(e) tout au long de l'année scolaire en échangeant régulièrement avec lui(elle) sur les nouvelles notions de mathématiques à introduire aux lycéens, la difficulté à enseigner cette discipline devant un public moins sensible aux mathématiques. lorsqu'il s'agissait de lycéens non scientifiques.

- J'ai participé, pendant l'année 2017-2018, avec un groupe d'élèves de terminale du Lycée Marguerite Yourcenar situé à Ernstein au Challenge de statistique de « Graines de Sondeur ». Ce challenge est ouvert à tous les lycéens de la voie générale, technologique ou professionnelle des académies participantes. Cette année-là les académies participantes étaient Bordeaux, Dijon, Lyon et Strasbourg. Lors du 10ème colloque francophone sur les sondages qui se tenait à Lyon du 24 au 26 octobre 2018, une « finale nationale » rassemblant les équipes lauréates du challenge a été organisée. Cette finale a pris la forme d'une session du Colloque, où chaque équipe a présenté ses travaux pendant une quinzaine de minutes. Ces présentations ont été suivies d'une délibération et de la remise de prix nationaux par les membres de l'équipe d'organisation du challenge lors d'une session plénière. Notre équipe a obtenu le premier prix national ex æquo avec l'équipe du lycée Albert Schweitzer de Mulhouse. Pour plus de détails, je renvoie au rapport des activités de l'IREM : `https://mathinfo.unistra.fr/websites/mathinfo/irem/Secretariat/Rapport_activite_IREM/ra-irem_2017-18.pdf`.

- Enfin, depuis 2017, je participe à MATh.en.JEANS qui, depuis 1989, vise à faire vivre les mathématiques par les jeunes, selon les principes de la recherche mathématique. Elle permet aux jeunes de rencontrer des chercheurs et de pratiquer en milieu scolaire une authentique démarche scientifique, avec ses dimensions aussi bien théoriques qu'appliquées et si possible en prise avec des thèmes de recherche actuels. Dans ce cadre-là, pendant l'année scolaire 2018-2019, j'ai encadré un groupe de lycéens du lycée Docteur Eugène Koeberlé de Sélestat. D'ailleurs j'ai présenté le fruit issu de notre travail au colloque du CFIES qui a eu lieu à Strasbourg en septembre 2019.

1.6.2 Au niveau international

J'ai été élue, pour la période 2020 à 2024, membre du *board of directors* de l'*IASC*, *European section*.

Chapitre 2

Modélisation statistique en biologie

2.1 Introduction

Mes premières collaborations en biologie m'ont permis de me rendre compte des besoins des utilisateurs de la statistique, d'être confronté à des jeux de données réels et de proposer des développements théoriques ou méthodologiques issus de problématiques réelles. Ces collaborations, le plus souvent sur plusieurs années, m'ont également formé à la gestion de projet. Cela m'a d'ailleurs beaucoup aidé plus tard lorsque je suis devenue chargée de missions pour la région Grand-Est pour l'agence pour les mathématiques en interaction avec l'entreprise et la société, abrégé en AMIES, (`https://www.agence-maths-entreprises.fr/a/`) et responsable des relations maths et entreprise pour le centre de modélisation et de simulation de Strasbourg (plateforme CEMOSIS, `http://www.cemosis.fr`).

En effet, j'ai participé à l'analyse des données d'un observatoire national pendant près de dix ans et à l'encadrement des étudiants qui ont mis en œuvre ce traitement statistique. J'ai procédé moi-même au dépouillement des résultats de six thèses pour lesquelles des techniques statistiques sophistiquées étaient requises afin de tirer le meilleur parti des données collectées. Il s'agit des thèses d'Alice Couégnas (Couégnas (2011)), d'Anne Rolland (Rolland *et al.* (2009); Bertrand et Maumy (2010)), d'Armand Jacobs (Jacobs *et al.* (2008)), de Marie Bourjade (Bourjade *et al.* (2009, 2014)), de Perrine Bellusso (Bellusso *et al.* (2014)) et de Jad Hamaoui (travail en cours).
J'ai également encadré deux ingénieures, une d'étude, Khadija Musayeva, pendandt six mois, et l'autre de recherche, Céline Caldini-Queiros, également six mois sur le projet COSINUS (en lien avec l'addictologie), en leur indiquant des choix méthodologiques d'analyse pour mener à bien leurs études.

2.2 Approches factorielles, modèles linéaires généralisés

2.2.1 Phytopathologie – Observatoire National des Maladies du Bois de la Vigne

L'observatoire a duré de 2003 à 2008. Un article de Jacques Grosman et de Bruno Doublet (Grosman et Doublet (2012)) fait la synthèse de ces cinq années d'observations. La création de l'observatoire était liée à un problème ayant un fort impact pour la communauté viticole, comme le présente le résumé de Bertrand *et al.* (2008) :

> L'objectif de l'Observatoire National des Maladies du Bois de la Vigne est de dresser un état des lieux de la répartition, de la fréquence et de l'intensité de l'expression des symptômes foliaires des maladies du bois, pour répondre à la question de leur progression dans le vignoble français. En effet, suite à une interdiction de l'utilisation de l'arsénite de soude, les viticulteurs ne disposent plus d'aucune méthode de lutte chimique curative homologuée contre les maladies du bois de la vigne. Cet observatoire collecte, chaque année, depuis 2003, un ensemble de données cohérentes. Le jeu de données est complexe : il comporte des variables qualitatives et quantitatives qui évoluent au cours du temps. La problématique de l'étude est de dégager les grandes tendances en matière d'épidémiologie végétale afin de déterminer quelles sont les mesures prophylactiques à mettre en œuvre collectivement et à grande échelle.

Si la DRAF-SRPV Alsace (direction régionale de l'agriculture et de la forêt et services régionaux de la protection des végétaux), chargée de l'étude des résultats de l'observatoire par son Ministère de tutelle, nous a sollicité, c'est avant tout pour avoir un fort appui méthodologique, concernant les problématiques statistiques, dans le traitement de ces données. En effet, devant la complexité de l'analyse à mener, un avis d'expert était nécessaire pour obtenir des résultats fiables. Pour les aspects relevant de la phytopathologie, une collaboration avec l'UMR Santé Végétale du centre Inra de Bordeaux (Serge Savary) ainsi que la DRAF-SRPV Rhône Alpes (Jacques Grosman) a été mise en place afin d'essayer de tirer le meilleur parti des données en cours de collecte.

Notre contribution en tant que chercheur en statistique à ce travail a été de proposer une méthodologie pertinente et d'encadrer successivement plusieurs étudiants du master de biostatistique de deuxième année sur ce sujet. En effet, l'ob-

servatoire a collecté des données pendant plusieurs années successives et une mise à jour annuelle des résultats était donc nécessaire.

Le travail statistique s'est articulé en trois points.

1. Nous avons mis en évidence des relations entre les différentes variables de l'étude (ces variables étant de nature qualitative et quantitative), puis nous avons utilisé l'analyse des correspondances multiples, l'analyse en composantes principales et l'analyse factorielle de données mixtes. Ces premiers résultats ont été suffisamment pertinents pour être publiés, non seulement dans une revue de vulgarisation scientifique à l'attention de la communauté viticole, (Kobes *et al.* (2007)), mais aussi dans la revue de société américaine de phytopathologie spécialisée dans le domaine (Fussler *et al.* (2008)). Je tiens à souligner que l'expertise de Serge Savary a beaucoup contribué à l'interprétation des résultats statistiques obtenus par ces différentes méthodes.

2. Ensuite, pour tenir compte des différences de nature entre les variables nous avons utilisé l'analyse factorielle des données mixtes, puis pour intégrer le facteur temps, nous avons employé des méthodes d'analyse factorielle de *K-tables CA*. Ces approches par tableaux multiples ont été réalisées dans un second temps puisqu'elles n'ont été possibles qu'à partir du moment où la durée d'observations a été suffisante. Nous avons soumis notre travail puis nous avons été sélectionnés pour présenter cette analyse complétée à une session spéciale d'études de cas organisée lors des journées de la statistique en 2007 à Angers, ce qui a amené à la publication Bertrand *et al.* (2007).

3. Enfin, afin de préciser les relations décelées, notre choix s'est porté sur des modèles de régressions logistiques binaires et ordinales. Nous avons utilisé des techniques *bootstrap* pour construire des régions de confiance autour de leurs paramètres. Bien que ces études aient utilisé des outils statistiques existants, elles ont fait l'objet de choix méthodologiques qu'il nous a semblé pertinent de publier dans Bertrand *et al.* (2008).

Enfin, c'est suite à une sollicitation locale de Philippe Kuntzmann, ingénieur agronome et œnologue à l'Institut Français de la Vigne et du Vin (IFV), Pôle Alsace, qu'une nouvelle collaboration sur cette thématique des maladies du bois de la vigne a eu lieu, et ce, jusqu'au départ de l'IFV de Philippe Kuntzman en juin 2014. Un nouveau jeu de données, constitué par l'IFV, a été analysé avec des techniques d'analyse des correspondances multiples pour la partie exploratoire mais également avec de la régression des moindres carrés partiels pour la partie modélisation car malheureusement cette base de données présentait des valeurs

manquantes et des variables présentant des problèmes de colinéarité. L'analyse statistique a été mise en œuvre pendant le stage de fin d'étude (six mois) d'une étudiante de master de biostatistique de deuxième année (Julie Barbe) en 2011 et a donné lieu à une publication internationale (Kuntzmann *et al.* (2013)).

En juin 2018, nous avons été contacté par Solène Malblanc, elle aussi ingénieure viticole à l'IFV, suite à sa lecture de l'article Bertrand *et al.* (2008). À l'époque Solène Malblanc travaillait sur les données des maladies du bois de la vigne, récoltées par l'observatoire alsacien et son étude faisait suite à celle réalisée par Philippe Kunztmann et publiée en 2013 Kuntzmann *et al.* (2013). Les nouveautés de l'étude à laquelle Solène Malblanc s'intéressait étaient :

1. le nombre d'années de suivi (de 2003 à 2018) plus grand et
2. les enquêtes auprès des viticulteurs qui prenaient en compte l'environnement des parcelles et les pratiques culturales (avec les changements éventuels de pratiques depuis 2003)

ce qui lui donnait un grand nombre de variables potentiellement explicatives. À l'époque Solène Malblanc avait déjà réalisé son mémoire de fin d'étude de sa formation d'ingénieur sur cette problématique, en se concentrant sur les variables qui ne varient pas dans le temps (environnement de la parcelle : cépage, type de sol, ...) et en utilisant une expression de la maladie moyennée au niveau de la parcelle. Elle devait à l'époque traiter les données en incluant une notion de temporalité pour faire le lien avec les changements de pratiques et leur impact. Nous lui avions conseillé à l'époque quelques modèles spatio-temporels qui pouvaient répondre aux questions posées. De ce travail une présentation a eu lieu à la fin du mois d'octobre 2018 à Dijon (Abidon et Malblanc (2018)).
Solène Malblanc est revenue avec plusieurs interrogations qui ont été soulevées par ses collègues lors de cette manifestation . À partir de ce moment-là, nous lui avons proposé de poursuivre notre collaboration, puisque Solène Malblanc n'ayant pas suivi au cours de son cursus d'ingénieure viticole une formation approfondie en statistique, elle ne souhait pas faire les nouveaux développements statistiques. Nous avons donc réfléchi ensemble comment nous pouvions collaborer de façon productive et la meilleure solution était que Solène Malblanc se fasse assister par une étudiante de master de statistique de première année. Ainsi du 02 mai au 31 juillet 2019, Elizaveta Logosha a accompagné Solène Malblanc dans cette nouvelle analyse en complétant le travail existant avec des modèles explicatifs utilisant la pénalisation de type *lasso* afin de sélectionner les variables les plus pertinentes.

En parallèle de cela, Solène Malblanc a en charge le projet Euréka. Ce dernier a pour vocation de proposer de nouvelles voix de lutte contre les dépérissements

en étant un « connecteur de connaissance », c'est-à-dire en faisant le lien entre la recherche scientifique (celle que nous menons avec l'IFV et d'autres collègues dans d'autres disciplines) et les viticulteurs. L'ensemble des connaissances et des hypothèses de travail seront appliquées, testées et expliquées dans une parcelle conçue comme un atelier d'innovations. Le programme de recherche se concentre sur les maladies du bois de la vigne. L'approche retenue implique tous les acteurs de la filière et repose sur une vision globale du processus de culture « Plante / Environnement / Pratiques Viticoles ». À partir d'une synthèse précise de toutes les pratiques utilisées au cours des siècles, les chercheurs étudieront des approches complémentaires en fonction de l'état d'avancement des maladies du bois de la vigne, à savoir :

- une approche curative innovante via des techniques d'endothérapie végétale bio-chimique applicable à différents pathogènes de façon ciblée et évitant une dispersion aérienne peu respectueuse de l'environnement,

- une approche préventive simple qui repense l'architecture du pied de vigne à l'instar de l'utilisation de porte-greffe pour lutter contre le phylloxéra,

- une étude statistique de l'impact du greffage et du recépage qui permet de renforcer les approches curatives et préventives.

- Il sera également sûrement possible d'étudier statistiquement le lien entre les pratiques culturales et le développement des maladies du dépérissement de la vigne.

Pour conclure, cette collaboration n'est pas encore achevée à ce jour. D'ailleurs, Solène Malblanc a sollicité à nouveau Elizaveta Logosha pour continuer ses travaux réalisés pendant son stage de master de statistique de première année. Cette dernière a accepté et est actuellement en stage de master de statistique de deuxième année à l'IFV situé à Colmar et ce depuis le 1er mars 2020 pour une durée de six mois.
Elizaveta Logosha s'intéresse plus particulièrement aux questions spatiales et temporelles suivantes : la localisation des ceps malades ou morts et son évolution temporelle au sein d'une parcelle est-elle aléatoire ? Pour rapporter des éléments de réponse, nous avons dû tenir compte de la structure particulière de l'échantillonnage qui a été réalisé au sein d'une parcelle (Figure 2.1) ainsi que des données d'observations longitudinales qui ont été collectées (Figure 2.2). Une gestion spécifique des valeurs manquantes a dû être mise en place afin d'éviter une « contamination » croissante au fil du temps des résultats d'incidence groupés par parcelle. Ce phénomène est illustré est à la Figure 2.3.

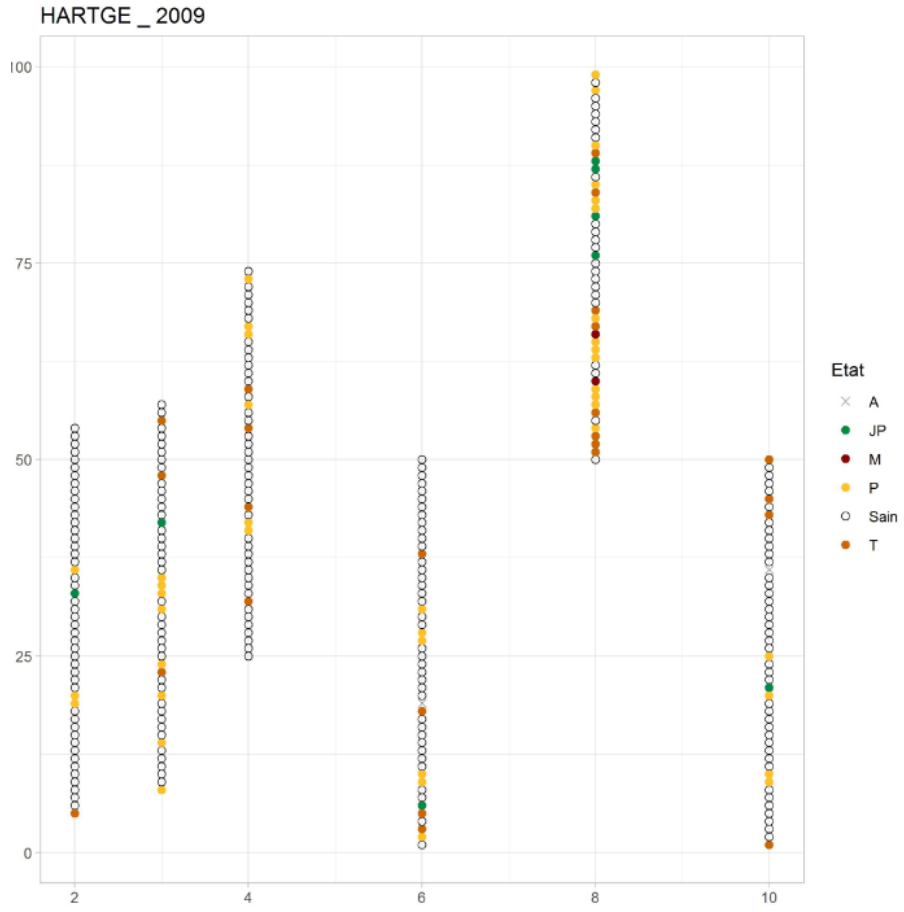


Figure 2.1 : *Localisation des ceps échantillonnés au sein d'une parcelle*

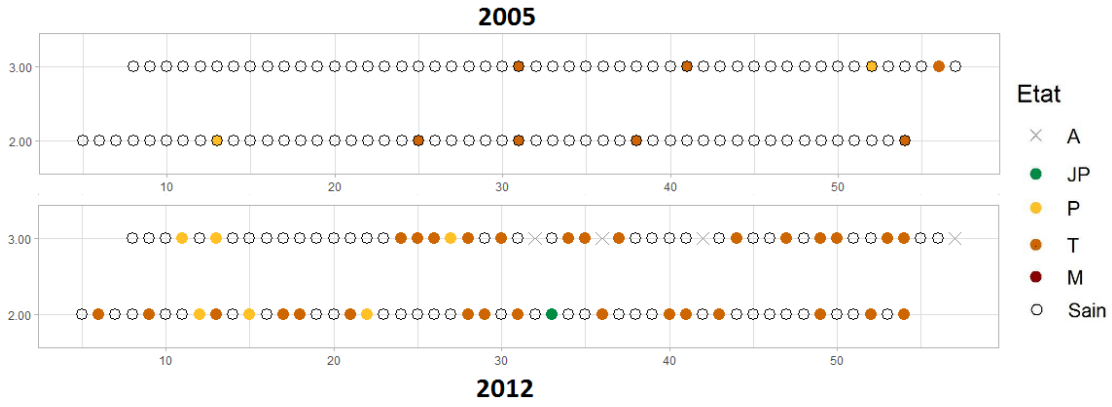


Figure 2.2 : *État des ceps, en 2005 et 2012, échantillonnés au sein d'une même parcelle*

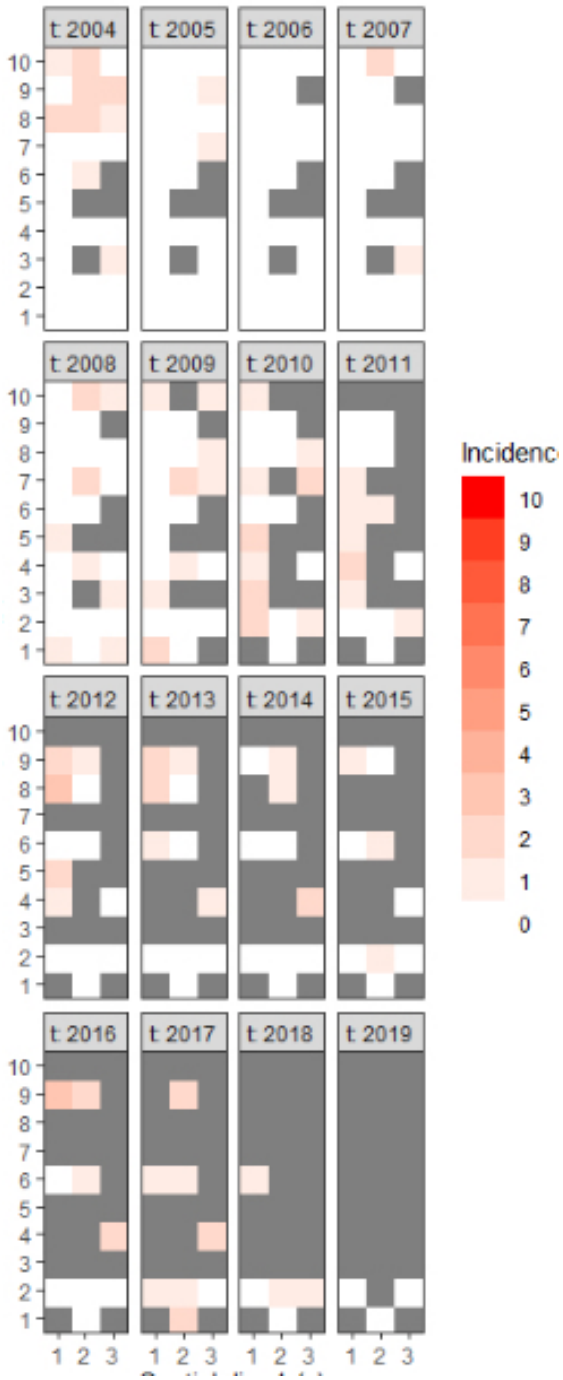


Figure 2.3 : *Contamination par des valeurs manquantes des valeurs d'incidence groupées par parcelle*

2.2.2 Écologie – Centre d'Écologie Fonctionnelle et Évolutive de Montpellier

Cette étude (Grandgeorge *et al.* (2008)) réalisée avec Marine Grangeorge, (actuellement maître de conférences à l'université de Rennes I et habilitée), lors de son mémoire de master 2 d'éthologie et d'écophysiologie, porte sur l'utilisation de techniques factorielles pour étudier la dynamique des communautés d'oiseaux de mer britanniques et irlandais au XX^{e} siècle. Les résultats obtenus contrastent avec les nombreux échecs de reproduction enregistrés au cours des premières années du XXI^{e} siècle, qui indiquent que certaines espèces de la communauté sont aujourd'hui gravement perturbées. Les outils statistiques qui ont permis d'obtenir ces résultats sont encore une fois issue des techniques d'analyse exploratoire comme l'analyse en composantes principales.

2.3 Techniques multitableaux

Ce travail s'inscrit dans une collaboration à une thèse de l'UMR INRA 42 (CARRTEL), Station d'Hydrobiologie Lacustre. Les micro-organismes, en particulier les espèces de phytoplancton, peuvent être considérés comme des indicateurs des changements locaux et plus globaux dans les écosystèmes aquatiques et peuvent donc constituer un excellent biomarqueur de la qualité de l'eau. évaluer l'influence des variables d'environnement biologiques, chimiques et physiques sur la régulation du phytoplancton est une étape clef pour parvenir à comprendre la structure et la dynamique des populations, leur diversité et leur succession afin de proposer, si nécessaire et si possible, une intervention humaine avant que toute prolifération excessive d'algues puisse se produire. Ces questions sont d'un intérêt premier pour les scientifiques et les gestionnaires d'eau afin de permettre aux réservoirs d'eau de grande taille, lacs et étangs, d'atteindre le « bon état écologique » recommandé par la directive-cadre sur l'eau (DCE) d'ici 2015.

Ce projet se concentrait sur l'étude du réservoir Marne (bassin de la Seine), l'un des plus grands réservoirs en Europe occidentale. En 2006, c'est-à-dire la première année du projet, le réservoir a été échantillonné une fois par mois en mars et avril, puis une fois toutes les deux semaines entre mai et septembre. Pour évaluer l'hétérogénéité spatiale, six stations et différentes profondeurs pour chaque station ont été étudiée.

Malheureusement, comme indiqué dans Rolland *et al.* (2009), ces jeux de données posent des problèmes spécifiques dus à leur échantillonnage spatio-temporel. La dynamique et la diversité du phytoplancton sont particulièrement diffi-

ciles à analyser, en particulier lorsque

1. la granularité de l'analyse se situe au niveau de l'espèce,
2. la diversité des espèces est élevée,
3. l'étude couvre plusieurs saisons et
4. l'échantillonnage a été réalisé dans de nombreuses stations de l'écosystème.

L'analyse triadique partielle (introduite en écologie par Thioulouse et Chessel en 1987, Thioulouse et Chessel (1987)), est une méthode d'analyse multitableaux qui est un outil statistique adapté pour obtenir une représentation claire d'une série chronologique, une pour chaque date de prélèvement, de matrices observées (abondances des espèces à chacune des stations). Elle permet l'analyse en composantes principales simultanée de plusieurs matrices et permet de trouver une structure spatiale commune à chacune de ces matrices et d'étudier la stabilité temporelle de celle-ci. L'analyse triadique partielle commence par la recherche d'un tableau moyen appelé **compromis**. Le tableau compromis est ensuite analysé et sa reproductibilité pour chacune des tables initiales est finalement étudiée.

Cette technique a été exploitée avec succès dans Rolland *et al.* (2009) pour décrypter l'organisation spatio-temporelle des assemblages d'espèce de phytoplancton. La méthodologie employée pour l'analyse statistique a été publié comme une étude de cas dans Bertrand et Maumy (2010).

2.4 Modèles mixtes

2.4.1 Modèles linéaires généralisés en éthologie

Éthologie – Armand Jacobs

J'ai participé à l'analyse de données d'éthologie animale sur le comportement de lémuriens lors d'une collaboration avec deux membres (Armand Jacobs et Odile Petit) de l'équipe d'Éthologie des Primates du Département Écologie, Physiologie et Éthologie, de l'Institut Pluridisciplinaire Hubert Curien (IPHC) rattaché au CNRS et à l'Université Louis Pasteur à l'époque.
Voici le sujet de l'article (Jacobs *et al.* (2008)) : des études sur le *leadership* lors de mouvements de groupe dans plusieurs espèces de lémuriens ont montré que les femelles étaient responsables des choix de voyage concernant l'heure et la direction. Il est intéressant de noter que dans ces espèces, les femelles sont plus

nombreuses que les mâles. Nous avons étudié l'influence de l'organisation sociale sur les processus de direction en étudiant une espèce de lémurien dont l'organisation sociale est caractérisée par l'absence de domination féminine : le lémurien brun. L'étude a été menée sur un groupe de 11 individus en semi-liberté et l'analyse effectuée sur 69 mouvements de groupe a montré que tous les individus pouvaient initier un mouvement de groupe. Dans 34 cas, le groupe entier s'est déplacé. Il n'y a pas eu de différence significative dans le nombre de tentatives de démarrage ou dans le nombre de membres du groupe impliqués d'un initiateur à l'autre. En outre, le sexe ou l'âge de l'initiateur n'a eu aucun effet sur le nombre de personnes qui le suivaient ou sur la rapidité du processus d'adhésion. Par conséquent, le leadership observé est largement diffusé à tous les membres du groupe. Ces résultats soutiennent l'hypothèse d'une influence de l'organisation sociale sur les processus de décision mais restent à étudier dans un contexte écologique plus pertinent.
La nature des données a requis l'utilisation de modèles mixtes linéaires et généralisés combinés à des tests de permutation. L'utilisation des tests de permutation provenait de l'absence d'hypothèse de normalité et également la faible taille de l'échantillon constitués par les lémuriens.

Éthologie – Marie Bourjade

Toujours au sein de l'équipe d'Éthologie des Primates, Département Écologie, Physiologie et Éthologie de l'IPHC rattaché au CNRS et à Université Louis Pasteur, j'ai participé à mise en place de l'analyse statistique des résultats obtenus pendant la thèse de Marie Bourjade, encadrée par Odile Petit et Bernard Thierry. Le sujet de la thèse de Marie Bourjade porte sur le comportement de certaines espèces de chevaux.
Voici plus en détails le sujet de l'article que nous avons rédigé (Bourjade *et al.* (2009)) : nous avons abordé les processus décisionnels dans les mouvements collectifs de deux groupes de chevaux Przewalski vivant dans une population semi libre. Nous avons cherché à savoir si différents modèles de mouvements collectifs sont liés à certains contextes écologiques (utilisation de l'habitat et activité du groupe) et analysé les éventuels processus décisionnels impliqués. Nous avons trouvé deux schémas distincts : les mouvements de "combat unique" et de "combat multiple" se sont produits dans les deux groupes d'étude. Les mouvements ont été définis par l'occurrence d'arrêts collectifs entre les combats et se distinguent par leur durée, la distance parcourue et le contexte écologique. Pour les deux types de mouvements, nous avons constaté qu'une période préliminaire impliquant plusieurs chevaux avait eu lieu avant le départ. Dans les mouvements à un seul arrêt, tous les membres du groupe ont rapidement rejoint le premier cheval en mouvement, indépendamment de la période préliminaire.

En revanche, dans les mouvements à sorties multiples, le processus d'adhésion a été plus long, notamment lorsque le nombre de décideurs et leur comportement avant le départ ont augmenté. Les mouvements de combats multiples étaient plus souvent utilisés par les chevaux pour changer d'habitat et d'activité. Cette observation montre que les chevaux ont besoin de plus de temps pour résoudre les conflits de motivation avant ces départs. Nous avons conclu que la prise de décision chez les chevaux de Przewalski est basée sur un processus consensuel partagé, piloté par des déterminants écologiques. La nature des données a requis l'utilisation de tests non-paramétriques et d'approches spatiales puisque comme le mentionne le résumé de l'article les chevaux sont en mouvement dans un environnement semi-libre.

Marie Bourjade a enchaîné ensuite quatre contrats de post-doctorat avant d'obtenir un poste de maître de conférences en psychologie du développement à l'Université de Toulouse en 2014. Nous sommes toujours restées en contact et en 2013, pendant son post-doctorat à l'Université d'Aix-Marseille, Marie Bourjade m'a proposé de participer à une nouvelle étude mais cette fois-ci sur le comportement de singes et avec de nouvelles collaboratrices et j'ai accepté.
La nature des données a requis l'utilisation de modèles mixtes généralisés. Les résultats de ces recherches ont été publiés dans l'article Bourjade *et al.* (2014).

2.4.2 Modèles linéaires généralisés en microbiologie

Ce travail, qui a été réalisé avec deux membres (Marisa Hohnadel et Renaud Chollet) de l'équipe de l'entreprise Merck, a abouti à une publication (Hohnadel *et al.* (2018)). Depuis près d'un siècle, les méthodes microbiologiques conventionnelles sont la norme pour détecter et identifier les agents pathogènes dans les aliments. Néanmoins, la sécurité microbiologique des aliments s'est améliorée et diverses méthodes rapides ont été développées pour surmonter les limites des méthodes conventionnelles. Les méthodes alternatives devraient permettre de détecter un faible nombre de cellules, puisque la présence dans les aliments d'un organisme pathogène, même d'une seule cellule, peut être infectieuse. En ce qui concerne les faibles niveaux de population, la performance d'une méthode de détection est évaluée en produisant des dilutions en série d'une suspension bactérienne pure pour inoculer des matrices alimentaires représentatives avec des cellules bactériennes fortement diluées (moins de 10 UFC/ml). L'exactitude des données obtenues par les techniques de dilution multiples n'est pas certaine et n'exclut pas certaines colonies issues d'amas de cellules. Les techniques de micromanipulation pour capturer et isoler des cellules individuelles à partir d'échantillons environnementaux ont été introduites il y a plus de 40 ans. La principale limite de la technique de micromanipulation actuelle reste le

faible taux de récupération pour la croissance d'une seule cellule dans le milieu de culture. Dans cette étude, nous décrivons une nouvelle méthode d'isolement cellulaire et démontrons qu'elle peut être utilisée avec succès pour cultiver différents types de micro-organismes à partir de cellules individuelles sélectionnées. Des tests avec des organismes Gram-positifs et Gram-négatifs, dont des cocci, des bâtonnets, des aérobies, des anaérobies, des levures et des moisissures, ont montré des taux de récupération de la croissance de 60 à 100% après micromanipulation. Nous soulignons également l'utilisation de notre méthode pour évaluer et remettre en question les limites de détection des méthodes de détection standard dans les échantillons d'aliments contaminés par une seule cellule de Salmonella enterica.
D'un point de vue statistique, ce sont des modèles de régression logistique ordinale qui ont été utilisés.

2.5 Confiance et inférence de réseaux biologiques

Des travaux précédents comme Vallat *et al.* (2013) et la thèse de Nicolas Jung, m'ont amené à m'intéresser à l'inférence de réseaux en biologie et en particulier au problème de la confiance à accorder aux résultats de telles inférences. Dans le cadre des recherches post-doctorales de Nicolas Jung nous nous sommes concentrés à construire des techniques d'inférence de réseaux de gènes pour lesquelles il est possible d'évaluer la confiance à accorder aux liens qui ont été révélés.

L'élucidation du réseau de régulation des gènes est une étape importante pour la compréhension de la physiologie cellulaire normale ou pathologique. Le *reverse-engineering* (décodage) consiste à utiliser l'expression des gènes au fil du temps, ou dans différentes conditions expérimentales, pour découvrir la structure du réseau de gènes dans un processus cellulaire ciblé. Le fait que les données sur l'expression des gènes sont généralement bruitées, fortement corrélées et de grande dimension explique la nécessité de recourir à des méthodes statistiques spécifiques pour le décodage du réseau de gènes. Parmi les méthodes connues, les algorithmes de calcul approximatif bayésien (ABC) n'avaient pas encore été pleinement appliquées à ce problème, en particulier en raison de la charge de calcul nécessaire à leur utilisation qui les limitaient à un petit nombre de gènes. Dans ce travail, nous avons développé une nouvelle approche ABC multi-niveaux, nommée `networkABC`, implémentée en `C++` avec une interface `R` (Bertrand et Maumy-Bertrand (2020b,c)), qui a un coût de calcul moins élevé. Au premier niveau, la méthode capture les propriétés globales du réseau, telles que la liberté d'échelle et les coefficients de regroupement, tandis que le second niveau vise à capturer

les propriétés locales, y compris la probabilité que chaque couple de gènes soit lié, ce qui est particulièrement intéressant pour notre problématique d'inférence avec confiance car cela nous permet d'avoir une idée de la confiance à accorder aux liens qui sont découverts.

Nous nous sommes partiellement appuyés sur des travaux de Di Camillo *et al.* (2009) pour créer notre algorithme de génération de réseaux qui est nécessaire à l'application de l'algorithme ABC.
Voici quelques détails sur l'algorithme (voir aussi la Figure 1.9).
Appelons V l'ensemble des nœuds à connecter dans le graphique G à l'itération actuelle t et H l'ensemble des nœuds à connecter à l'itération $t+1$. V est initialisé comme $V = \{1,...,N\}$, c'est-à-dire avec tous les N nœuds en G alors que H est initialisé comme l'ensemble vide $H = \emptyset$. Les nœuds sont ensuite reliés les uns aux autres par une procédure itérative, qui comprend trois étapes principales, expliquées en détail ci-dessous.

1. Trois modules candidats sont générés. La structure est échantillonnée à partir d'un ensemble de motifs, avec possibilité de modifications aléatoires. Le nombre de nœuds du module est fixé au hasard. Dans cet algorithme, nous considérons les motifs suivants : rétroaction, *feedforward* et boucles.

2. Un score est attribué à chaque module, et l'un des trois modules est échantillonné avec une probabilité proportionnelle à ce score ; désignons le module échantillonné par M et le nombre de ses nœuds par m.

3. Les nœuds sont échantillonnés à partir de V et liés entre eux dans le graphique G selon la structure de module choisie M ; V est mis à jour en supprimant les m nœuds échantillonnés ; H est mis à jour par l'ajout des nœuds. À la fin de ce processus, V est vide alors que H est composé de nombreux motifs. Pour relier les motifs entre eux, nous devons choisir un nœud dans chaque motif qui est la première position. Cet ensemble de nœuds est alors considéré comme l'ensemble V.

Un exemple de paramétrage de l'algorithme est donné à la Table 2.1 et un exemple de résultats à la Table 2.2 ainsi qu'aux Figures 2.4, 2.5 et 2.6. Au départ de Nicolas Jung, plusieurs problèmes techniques avec l'algorithme et son utilisation sous le logiciel `R` restaient à résoudre. Nous les avons résolus depuis et nous sommes en train d'évaluer l'algorithme ainsi que de rédiger un article.

Table 2.1 : *Exemple de paramétrage de l'algorithme* `networkABC`.

```
abc(data=M,
    clust_coeffs=0.33, #plus d'un coefficient peut être spécifié
    tolerance=3.5, #distance maximale entre les données simulées
                   #et les données réelles pour pouvoir accepter
                   #le réseau
    number_hubs=3, #le nombre de hub
    iterations=10, #le nombre d'itérations
    number_networks=1000000, #le nombre de réseaux simulé à chaque
                             #itération
    hub_probs=NA, #spécification de la probabilité a priori
                  #pour qu'un gène soit un hub
    neighbour_probs=NA, #spécification de la probabilité a priori
                        #pour qu'un couple de gènes soit relié
    is_probs=1) #cette option doit être fixée à 1.
```

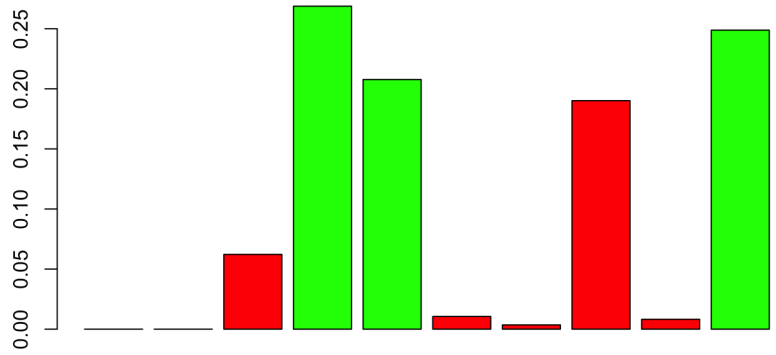


Figure 2.4 : *Visualisation des probabilités d'être un hub.*

Table 2.2 : *Exemple de résultat d'utilisation de l'algorithme* `networkABC`.

```
#> First run of abc to find tolerance
#> ================================
#> Iteration=1
#> Accepted:1000
#> Probabilities of clustering coefficients:
#> 0.325000 0.349000 0.326000
#> Tolerance value
#>       5%
#> 4.523488
#> ================================
#> Beginning main run of abc
#> ================================
#> Iteration=1
#> Accepted:45
#> Probabilities of clustering coefficients:
#> 0.488889 0.311111 0.200000
#> ================================
[...]
#> ================================
#> Iteration=10
#> Accepted:238
#> Probabilities of clustering coefficients:
#> 0.907563 0.079832 0.012605
```

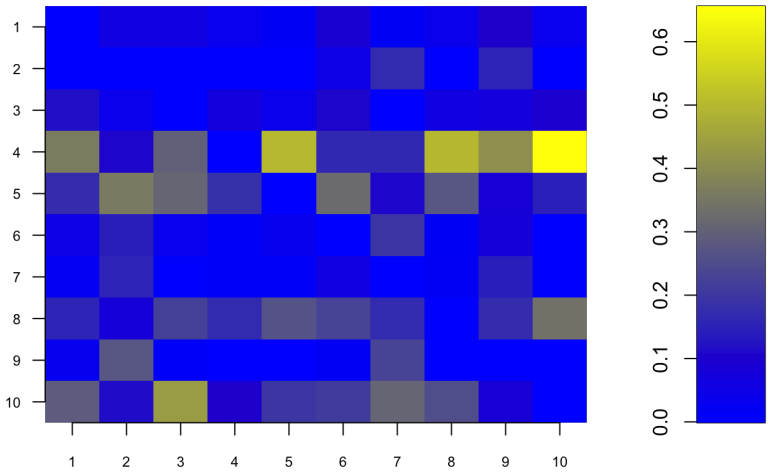


Figure 2.5 : *Visualisation des probabilités de voisinage.*

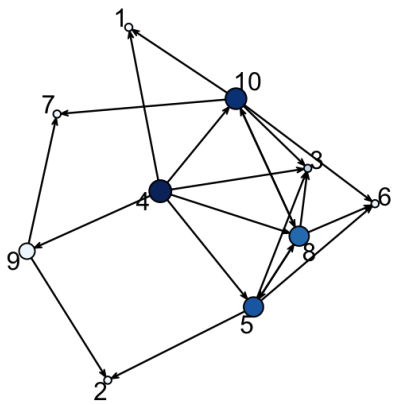


Figure 2.6 : *Visualisation d'un réseau décodé.*

Chapitre 3

Modélisation statistique en médecine

3.1 Introduction

J'ai procédé moi-même au dépouillement des résultats de deux thèses de doctorat de science pour lesquelles des techniques statistiques sophistiquées étaient requises afin de tirer le meilleur parti des données collectées. Il s'agit des thèses de Jérôme Grenèche (Grenèche *et al.* (2011a, 2013)) et de Liliane Tetsi (Tetsi *et al.* (2019)). Je dois préciser qu'après sa thèse de psychologie cognitive , Jérôme Grenèche est devenu d'abord journaliste scientifique et réalisateur de documentaires sur les animaux pour Arte, France5 tandis que Liliane Testi, qui est à la retraite, continue même après sa thèse de faire de la recherche au sein de son laboratoire.
J'ai également encadré trois ingénieurs statisticiens, un de recherche et deux d'étude, (Marius Kwemou, Ismaïl Aouadi et Martin Verniquet) en leur indiquant des choix méthodologiques d'analyse puis en les aidant à mettre en forme leurs résultats afin de pouvoir les diffuser auprès des chercheurs. Tous les trois ont travaillé successivement au Laboratoire d'Immuno Rhumatologie Moléculaire, unité mixte de recherche 1109 entre l'INSERM et l'Université de Strasbourg.

3.2 Premiers pas et petits échantillons

J'ai commencé par effectuer l'analyse des résultats d'études cliniques réalisées par deux chirurgiens du service de chirurgie digestive de l'IRCAD (institut de

recherche contre les cancers de l'appareil digestif). Frédéric Bertrand, à l'époque en thèse, m'a aussi aidé dans l'analyse de ces résultats.
Étant jeune maitre de conférences à l'époque, je n'ai pas particulièrement demandé à être associée aux publications qui concernaient trois études (Pugliese *et al.* (2008); Forgione *et al.* (2009); Rubino (2008)) dont la troisième porte sur la possibilité d'un traitement chirurgical du diabète. Cette dernière a eu un impact considérable dans le monde médical puisqu'un article du Washington Post (Stein (2008)) a été publié et un reportage de CBS (Lesley Stahl (2008)) a été réalisé qui expliquait l'intérêt pour la lutte contre le diabète de la découverte précédente. La difficulté statistique était principalement de travailler avec des échantillons de petite taille. Il faut noter que les échantillons étaient constitués de souris, ce qui nous permet de comprendre pourquoi la taille était réduite. Les deux chirurgiens (Antonello Forgione et Francesco Rubino) faisaient leurs expériences sur ces dernières. De plus, il faut ajouter que l'opération chirurgicale durait trop longtemps et par conséquent rares sont les souris qui survivaient.
Ainsi, très rapidement, je me suis rendue compte que les hypothèses usuelles du modèle linéaire gaussien ne sont pas valides dans un grand nombre de situations expérimentales. En effet, comment tester l'hypothèse de normalité avec des tailles d'échantillon allant d'une dizaine à deux voire dans de rares cas trois dizaines de valeurs ? Il est reconnu que les meilleurs tests de normalité *omnibus* ne commencent à déceler « efficacement » les défauts de normalité que pour des échantillons de taille au moins égale à 30 (Thode (2002)). Ces expériences m'ont fait connaître d'autres approches pour sélectionner ou valider des modèles statistiques comme la validation croisée (Hastie *et al.* (2008)) ou des approches par permutations (Good (2005)). Ces influences se sont manifestées lorsque j'ai moi-même développé de nouveaux outils (voir les chapitres 4, 5 et 6).

3.3 Risques compétitifs en médecine

Marius Kwemou qui était post-doctorant de l'équipe GenPred, spécialité statistique, a travaillé au sein Laboratoire d'Immuno Rhumatologie Moléculaire, INSERM et unité mixte de recherche 1109[1]. D'abord avec Nicolas Jung puis avec Marius Kwemou, Frédéric Bertrand et moi avons en particulier élaboré le modèle statistique qui pouvait expliquer les résultats des recherches coordonnées sur la maladie du greffon contre l'hôte (GvHD). Cette dernière est la complication majeure de la greffe de moelle osseuse provenant d'un donneur étranger (greffe allogénique). Elle peut survenir de façon aiguë, dans les semaines qui

[1]Plateforme GENOMAX, Faculté de Médecine, Fédération Hospitalo-Universitaire OMICARE, Fédération de Médecine Translationnelle de Strasbourg et LabEx TRANSPLANTEX, Faculté de Médecine, Université de Strasbourg

suivent la greffe, ou de façon chronique, dans les mois ou les années qui suivent.

- Les atteintes sont avant tout hépatiques, digestives et cutanées.
- Les manifestations cliniques peuvent parfois être gravissimes avec insuffisance hépatique majeure, désordres hydro-électrolytiques et risque accru de maladies infectieuses sévères : bactériennes, fongiques (aspergillose,...) et virales (cytomégalovirus, herpès, zona, ...).

Le gène A lié à la chaîne du CMH de classe I, MICA, hautement polymorphe, code pour une glycoprotéine induite par le stress et exprimée principalement sur l'épithélium.
Le MICA interagit avec le récepteur d'activation invariant NKG2D, exprimé par les lymphocytes cytotoxiques, et est situé dans le CMH, à côté du HLA-B. Le MICA possède donc les attributs requis d'un antigène de transplantation authentique.
En utilisant un génotypage haute résolution basé sur la séquence des MICA, nous avons analysé rétrospectivement l'effet clinique des mésappariements des MICA dans une cohorte multicentrique de 922 paires de HLA-A, HLA-B, HLA-C, HLA-DRB1 et HLA-DQB1 de donneurs non apparentés appariés par un allèle 10/10.
Parmi les 922 paires, 113 $(12{,}3\%)$ ont été mal appariées dans le cadre du TCM ; les malappariements TCM ont été associés de manière significative à une incidence accrue de GVHD aiguë de grade III-IV (hazard ratio (HR) : 1,83 ; IC de 95%, 1,50-2,23 ; $p < 0{,}001$), de la RGCH chronique (HR, 1,50 ; IC de 95%, 1,45-1,55 ; $p < 0{,}001$) et de la mortalité non due à un accident (HR, 1,35 ; IC de 95%, 1,24-1,46 ; $p < 0{,}001$).
Le risque accru de GVHD se reflète dans un risque de rechute plus faible (HR, 0,50 ; IC de 95%, 0,43-0,59 ; $p < 0{,}001$), ce qui indique un effet possible de la greffe contre la leucémie. En conclusion, lorsque c'est possible, la sélection d'un donneur compatible avec le MICA influence de manière significative les principaux résultats cliniques de l'HCT dans lesquels une réduction marquée de la GVHD est primordiale. Le déséquilibre étroit entre les MICA et les HLA-B rend l'identification d'un donneur compatible avec les MICA facilement réalisable dans la pratique clinique.
La quantité de données à traiter et la complexité du problème posé a nécessité le travail d'un chercheur post-doctorant (Marius Kwemou) puis d'un ingénieur statisticien (Ismaïl Aouadi) que j'ai encadrés sur ce sujet. Les modèles utilisés relèvent de l'analyse de survie et des risques compétitifs. Les résultats de ces recherches ont été publiés dans l'article Carapito *et al.* (2016).
Des recherches similaires, concernant un second paramètre génétique (MICB), ont été réalisées et un deuxième article a été publié (Carapito *et al.* (2020)).

3.4 Courbes néonatales

3.4.1 Contexte

Le poids de naissance reste encore à l'heure actuelle un critère majeur d'évaluation en périnatalité. Depuis le début des années 1960, de nombreuses courbes de poids néonatal en fonction du terme d'accouchement sont publiées de part le monde. Celles de Lubchenco *et al.* (Lubchenco *et al.* (1963)), Leroy et Lefort (Leroy et Lefort (1971)) et plus récemment de l'Association des utilisateurs de dossiers informatisés en pédiatrie, obstétrique et gynécologie (Mamelle *et al.* (1996)) se sont imposées en France, aussi bien pour une utilisation obstétricale que pédiatrique. Cependant, ces courbes peuvent parfois souffrir de biais et/ou d'insuffisances méthodologiques pouvant affecter la précision et la fiabilité des données. En effet, un des problèmes de construction de ces courbes est la mesure de l'âge gestationnel, abrégé en AG [2]. En dehors de l'arrondi à la semaine de l'AG la plus proche ou à la semaine révolue, se pose le choix de la datation, basée sur la date des dernières règles et/ou sur l'échographie. La mesure de l'AG est surestimée lorsqu'elle s'appuie sur la date des dernières règles, et aboutit à un infléchissement de la courbe de poids chez les enfants considérés comme post-termes (Gardosi et Francis (2000)). L'utilisation de l'échographie gomme en partie les disparités de poids fœtal, et a pour conséquence le resserrement des valeurs autour de la moyenne. Un autre problème connu est que certaines courbes sont construites sur des effectifs de population trop faibles ce qui amène à des calculs de percentiles peu fiables. Il reste par ailleurs difficile d'exclure les erreurs de terme comme en témoignent certaines publications donnant des poids[3] de naissance jusqu'à 44, voire 45 semaines d'aménorrhée (abrégées en SA). Plus récemment, Salomon et ses co-auteurs (Salomon *et al.* (2007)) ont publié des courbes de poids de naissance en population française en se basant sur une méthodologie très stricte. Toutefois, il ne peut s'agir au sens strict de courbes dites de normalité car les bébés présentant une anomalie pathologique du poids n'ont pas été exclus de l'échantillonnage, ce qui peut introduire un biais.
Depuis 12 ans, il existe, en France, et en particulier en Bourgogne, d'autres courbes de référence (Rousseau *et al.* (2008)) qui sont construites à partir des mesures du fœtus et non à partir des mesures provenant des échographies. Elles ont aussi intégrées tous les défauts : effectifs suffisants, données récentes avec utilisation massive de l'échographie pour dater la grossesse, exclusion des pathologies maternelles et fœtales pouvant retentir sur le poids à partir d'une base

[2]Nous rappelons que l'âge gestationnel est déterminé par la date des dernières règles, confirmé ou modifié par l'échographie du premier trimestre si nécessaire.

[3]Nous parlons de « poids » mais c'est bien de la masse qu'il s'agit, une quantité en grammes et non un vecteur exprimé en Newton.

de données médicale informatisée périnatale.Ces courbes de référence du poids de naissance en fonction de l'AG ont été construites à partir d'une population de nouveau-nés issus de grossesses uniques et non compliquées de pathologies pouvant influer sur le poids de naissance. L'échantillon a restreint volontairement aux naissances entre 28 et 42 SA afin de tenir compte d'un effectif suffisant dans chaque classe d'âge gestationnel pour calculer des percentiles fiables. L'échantillon ne contient pas les morts in utero et les mort-nés spontanés (j'insiste sur ce point car c'est précisément cette situation-là qui va m'intéresser par la suite) ou après interruption de grossesse, les grossesses multiples et les enfants porteurs d'anomalie chromosomique. Enfin les bébés ayant un poids de naissance incompatible pour leur AG ont été exclus après qu'ils aient été identifiés par un algorithme estimant la probabilité pour un enfant d'avoir un âge gestationnel inapproprié, basé sur la distribution spécifique des poids de naissance selon l'âge gestationnel dans la base de données entière. Aucune autre exclusion n'a été effectuée. Le poids, le terme de naissance et le sexe des enfants ont été pris en compte pour l'exploitation des résultats. Après validation statistique de ces données, la distribution du poids de naissance selon l'AG d'abord par sexes confondus, puis par sexe, correspondant à la population « normale », a été modélisée graphiquement.

3.4.2 Notre problématique

Le Professeur Bernard Foliguet, fœtopathologiste au CHRU de Nancy (centre hospitalier régional universitaire), a collecté pendant plusieurs années des données sur des fœtus morts in utero dans son service médical d'obstétrique et médecine fœtale du CHRU. C'est plus de 600 cas de fœtus morts in utero qui ont été enregistrés à Nancy.
Les variables (masse, taille, périmètre crânien,...) que le Professeur Bernard Foliguet a récoltées sont mesurées au moment de l'expulsion du fœtus. Il est important de noter que nous avons, par fœtus mort, des mesures faites à cet instant donné (juste après l'accouchement) et que ces dernières ne sont pas de type longitudinal, comme nous pourrions le penser. En effet, le caractère longitudinal permettrait peut-être d'anticiper l'expulsion du fœtus par le système biologique de la mère. Il faut noter, et c'est cette remarque-là qui fait la nouveauté de cette étude, que ces grossesses ne présentaient aucune pathologie lors des échographies réalisées pendant chaque grossesse. Le Professeur Bernard Foliguet émet l'hypothèse suivante : dans la courbe de croissance du poids fœtal par exemple, il doit y avoir, à un moment donné, une rupture et c'est cette dernière qu'il faudrait détecter afin de diminuer le nombre de fœtus morts in utero. En France, elle touche 1% des naissances chaque année et à l'heure actuelle, le monde médical n'a aucun argument scientifique pour expliquer une telle situation.

Le Professeur Bernard Foliguet s'est intéressé dans un premier temps à la mise au point d'un modèle qui permettrait de comprendre la croissance fœtale (aussi bien au niveau du poids, que de la taille ou du périmètre crânien) dans ce contexte particulier des fœtus expulsés inopinément par la mère. N'étant pas convaincu par l'utilisation des méthodes polynomiales classiques (méthodes utilisées par les logiciels installés sur les échographes) dans ce cadre-là, le Professeur Bernard Foliguet a donc décidé de prendre contact avec un chercheur en statistique qui serait capable de développer un modèle en adéquation avec ce type de données. En faisant des recherches autour de ses collègues, le nom de Sandie Ferrigno est apparu au Professeur Bernard Foliguet. Il faut préciser que Sandie Ferrigno (docteur en statistique mathématique de l'Université de Montpellier) travaillait en tant qu'ingénieur de recherche au CHU de Montpellier avant de prendre son poste de maître de conférences attaché à l'école nationale supérieure des mines de Nancy (ENSMN) et à l'école européenne d'ingénieurs en génie des matériaux (EEIGM) et également membre de l'IECL (institut Élie Cartan de Lorraine).

Sandie Ferrigno m'a contactée assez rapidement pour me parler de cette problématique puisque nous avions déjà travaillé ensemble dans le domaine de l'estimation non paramétrique. À l'issue de plusieurs discussions entre nous, nous avons décidé de proposer au Professeur Foliguet d'adapter des méthodes non paramétriques d'estimation. Ces dernières seront basées sur les méthodes à noyau ou encore sur les estimations polynomiales locales.
L'idée étant ensuite de pouvoir comparer les modèles que nous allions développer à ceux appliqués aux fœtus sains (et donc viables), nous avons donc cherché à obtenir des données adaptées. Le Professeur Bernard Foliguet nous a orientées vers le Professeur Olivier Morel, chef de pôle à la Maternité Régionale de Nancy. Ce dernier nous a permis de rentrer en contact avec Barbara Heude, chercheuse en épidémiologie et santé publique à l'INSERM et coordinatrice de la cohorte EDEN (Heude *et al.* (2016))que nous allons présenter ci-dessous.

La difficulté de ce sujet de recherche repose sur le fait qu'il est extrêmement rare d'avoir des données réelles sur le poids des bébés nés avant terme. Or le projet EDEN que nous allons présenter ci-dessous a ce type de données.

Enfin, la finalité de nos travaux de recherche que nous menons avec Sandie Ferrigno et le Professeur Bernard Foliguet qui nous aide à interpréter les résultats est de comparer les poids de ces fœtus à problèmes aux poids des fœtus normaux [4] nés avant terme.

[4]Les fœtus normaux sont caractérisés comme des fœtus nés sans problèmes de développement et ne présentant pas de problèmes aux échographies.

3.4.3 L'étude EDEN

L'étude EDEN est la première **E**tude de cohorte généraliste, menée en France sur les **D**éterminants pré et posts natals précoces du développement psychomoteur et de la santé de l'**EN**fant. Son objectif est de mieux établir l'importance des déterminants précoces sur la santé des individus, en particulier en regard des facteurs d'environnement qui l'influencent au cours de l'enfance, puis de la vie adulte. Il s'agit d'une étude épidémiologique longitudinale dont le but dans un premier temps était de suivre une cohorte d'enfants dès la fin du premier trimestre de grossesse jusqu'à l'âge de cinq ans, en prenant en compte un large éventail de renseignements recueillis auprès de la mère, du père, lors des examens de l'enfant et en s'appuyant sur un recueil d'échantillons biologiques. EDEN n'est pas seulement une grande enquête scientifique permettant de repérer un certain nombre de facteurs de risque et de comprendre certains mécanismes. Elle devrait aussi permettre d'identifier les mères à risque et, par là, de leur proposer un suivi plus adapté, afin de réduire les inégalités de santé qui en découlent pour leur enfant. L'étude comporte une multitude d'informations sur 2002 femmes qui ont été recrutées dans les maternités de Nancy et de Poitiers entre 2003 et 2006. De nombreux examens sont réalisés sur la mère et l'enfant avant, au moment et après l'accouchement. En particulier, à l'accouchement des mesures précises des caractéristiques biologiques et du poids du bébé sont réalisées.
L'accès à cette cohorte est pour nous une porte ouverte pour la suite de nos travaux de recherche car pouvoir avoir accès à une telle cohorte nous permet d'une part de comparer les modèles que nous développons aux autres développements statistiques réalisés dans un même contexte et d'autre part, d'étudier ces éventuelles « ruptures » qui apparaitraient soudainement dans le développement des fœtus.

3.4.4 Résumés graphiques et numériques de la base EDEN

Nous avons extrait de cette base les données de 1899 nouveau-nés et avons sélectionné seulement trois variables. Nous connaissons les semaines d'aménorrhée (SA), le poids de naissance et le sexe de l'enfant. Nous avons calculé quelques statistiques descriptives résumant les données pour les deux sexes puis pour les sexes féminins et masculins. Les données ne sont pas exhaustives pour des semaines d'aménorrhée « faibles » (*i.e.* aux alentours de la 20ème SA). Nous ne possédons des informations qu'à partir de la 27ème SA car les fœtus avant la 27ème SA sont rares. Nous avons un nombre suffisant de fœtus pour commencer à calculer des statistiques à partir de la 30ème SA. La Table 3.1 présente les statistiques descriptives de la variable semaine d'aménorrhée et la Table 3.2 donne

celles de la variable masse. Puis nous avons tracé le *barplot* du nombre de fœtus (sexes confondus) en fonction de la SA (voir la Figure3.1) grâce à la Table 3.3 ainsi que le nuage de points du poids du fœtus en fonction de la semaine d'aménorrhée et par sexe (voir la Figure3.3). Nous allons restreindre la période de temps d'étude pour la suite de l'étude par manque de données (<5 données) comme nous pouvons le constater au sein de chacun des sexes (voir la Figure 3.2).

	Sexes confondus	Sexe féminin	Sexe masculin
Effectif	1899	901	998
Minimum	27	28	27
1er quartile	39	39	38
Médiane	39	40	39
Moyenne	39	39	39
3ème quartile	40	40	40
Maximum	42	42	42
Ecart type	1,74	1,65	1,81
Etendue inter-quartile	1	1	2

Table 3.1 : *Statistiques descriptives de la variable semaine d'aménorrhée (en semaines)*

	Sexes confondus	Sexe féminin	Sexe masculin
Effectif	1899	901	998
Minimum	585	1100	585
1er quartile	3000	2950	3040
Médiane	3300	3240	3355
Moyenne	3279	3211	3340
3ème quartile	3620	3520	3680
Maximum	5260	4470	5260
Ecart type	512	475	537
Etendue inter-quartile	620	570	640

Table 3.2 : *Statistiques descriptives de la variable masse (en g)*

SA	Sexes confondus	Sexe féminin	Sexe masculin
27	1	0	1
28	2	1	1
29	1	0	1
30	5	2	3
31	6	4	2
32	7	3	4
33	7	2	5
34	14	5	9
35	18	7	11
36	46	22	24
37	77	35	42
38	276	128	148
39	495	241	254
40	544	273	271
41	363	165	198
42	37	13	24

Table 3.3 : *Nombre des fœtus de 27 à 42 SA*

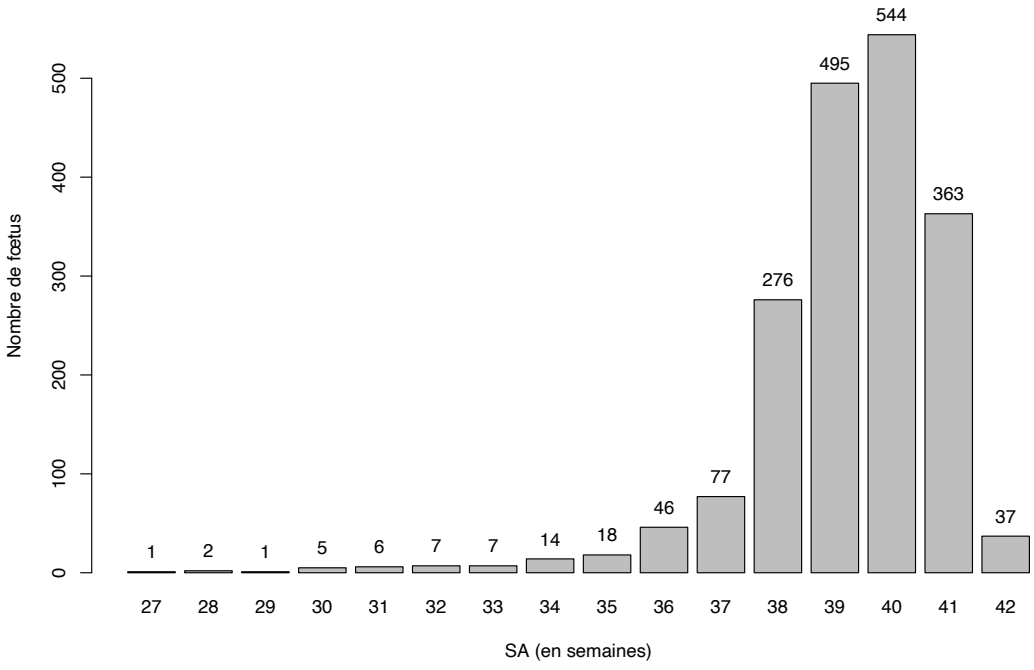


Figure 3.1 : *Répartition des fœtus (sexes confondus) par semaine d'aménorrhée*

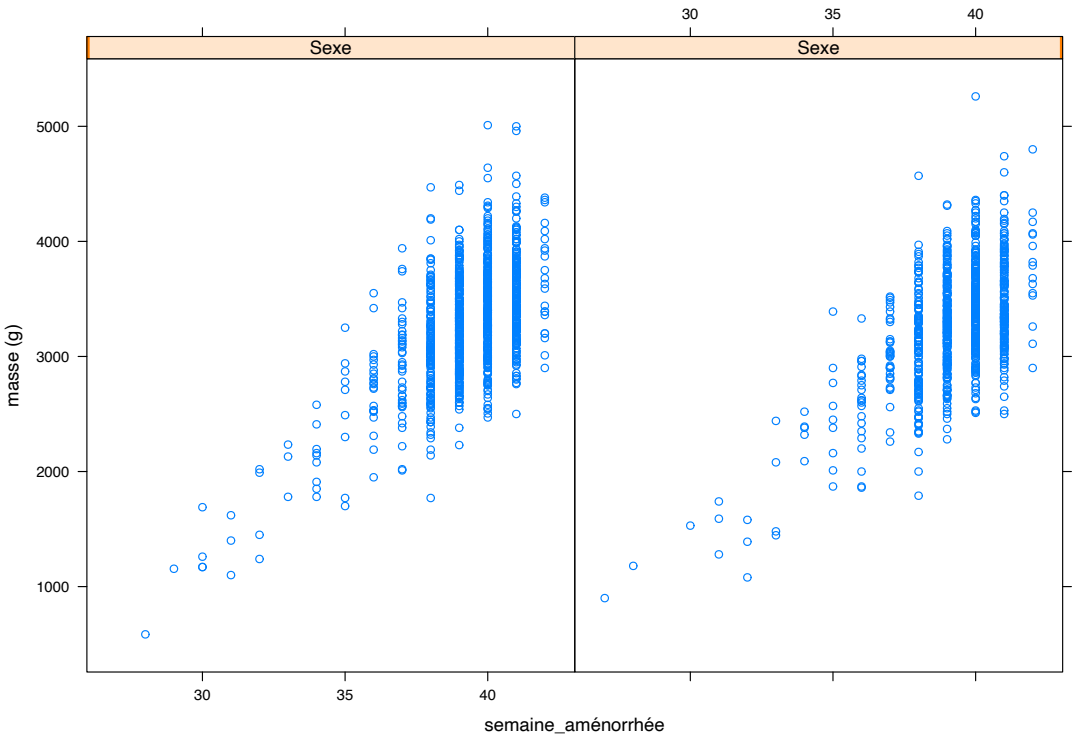


Figure 3.2 : *Nuage de points (par sexe) de la masse en fonction de la semaine d'aménorrhée*

3.4.5 Les méthodes paramétriques pour traiter la problématique

L'estimation empirique des centiles

La première méthode à laquelle nous pensons pour tracer des courbes de croissance néonatales est l'estimation empirique des centiles. Nous partons de l'équation suivante : $C_{100(1-\alpha)}(t) = \mu_t + u_{100(1-\alpha)} * \sigma_t$ où t désigne la semaine d'aménorrhée et $u_{100(1-\alpha)}$ le quantile d'une loi normale centrée réduite à $100(1-\alpha)$. Puis pour tracer les courbes de centiles estimés $\widehat{C}_{1-\alpha}(t)$ à une semaine d'aménorrhée donnée, nous calculerons les estimations empiriques de chacun des deux paramètres μ_t et σ_t à partir des observations recueillies pour la semaine t donnée. Enfin nous les injecterons directement dans la formule ci-dessus pour obtenir l'estimation attendue. Malheureusement cette méthode réclame une distribution normale de la variable étudiée pour chacune des semaines d'aménorrhée. Or comme nous pouvons le constater sur la Figure3.3 pour les SA entre la 30ème et la 35ème, la taille de l'échantillon étant inférieure à 30, il sera très difficile de faire confiance à la conclusion du test de normalité et de penser que la puis-

sance est suffisamment grande. Par absence de normalité de la variable étudiée (par exemple ici la masse du fœtus), nous n'allons pas nous approfondir cette méthode.

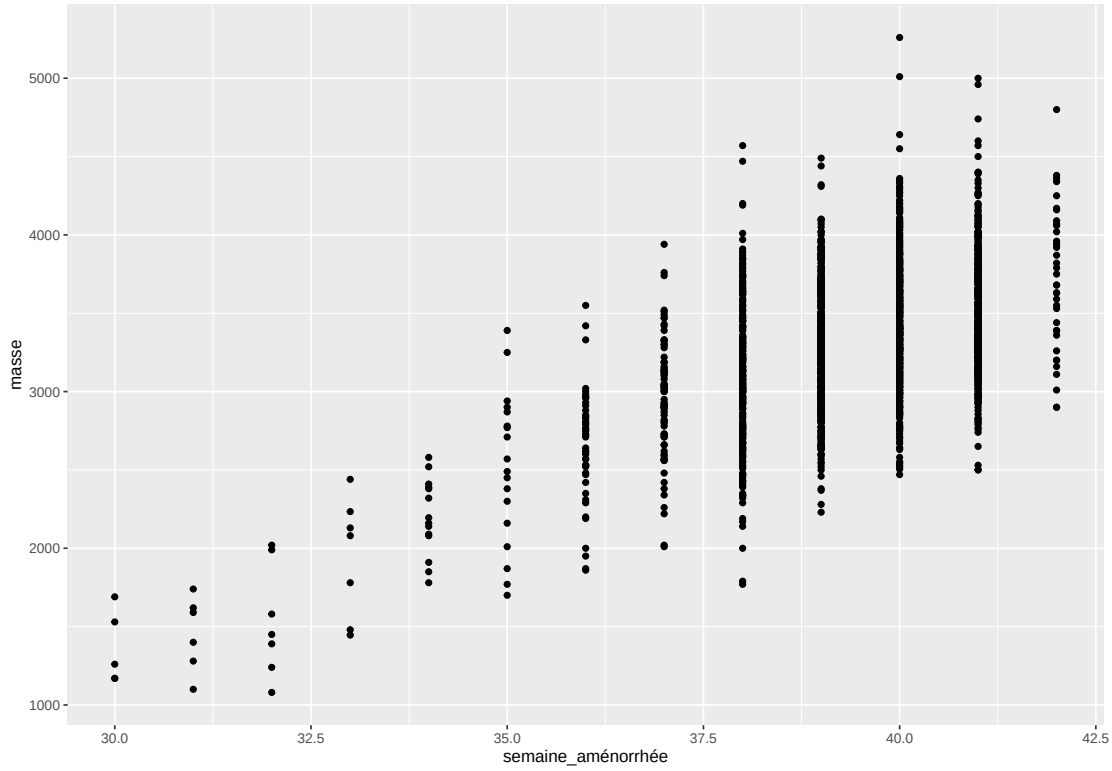


Figure 3.3 : *Nuage de points (sexes confondus) de la masse en fonction de la semaine d'aménorrhée*

Les méthodes paramétriques

Lambda-Mu-Sigma, abrégée en LMS (Cole (1990)) et *Box-Cox Power Exponential*, abrégée en BCPE (Rigby et Stasinopoulos (2004)) sont des méthodes utilisées pour obtenir des courbes de centiles lissées, en particulier pour des données transversales, et pour divers paramètres de croissance chez les bébés. Nous rappelons encore une fois que la base de données du Professeur Bernard Foliguet n'est pas une base de données longitudinales.

Malgré une utilisation aussi diversifiée, les méthodes semblent n'avoir jamais été expliquées explicitement. Des détails mathématiques ont été fournis Rigby et Stasinopoulos (2004) et de Onis (2007) mais ces derniers semblent trop complexes pour les professionnels de la médecine. Même après des justifications, ce n'est pas évident pour la communauté médicale que de telles méthodes si compliquées soient nécessaires si les fœtus ne posaient pas de problèmes parfois...

Les médecins ne comprennent pas quelle partie des méthodes est destinée à l'estimation du centile et quelle partie est destinée au lissage, et parmi le lissage, ce qui est destiné aux estimations des paramètres et ce qui s'applique aux courbes du centile. En raison de plusieurs étapes inexpliquées, l'application n'a jusqu'à présent pas été généralisée. Cela a été fait soit par des experts qui connaissent les subtilités des méthodes, soit par des inexperts qui utilisent un logiciel qui a implémenté ces deux méthodes comme une boîte noire. D'ailleurs c'est exactement ce point-là qui est pointé par le Professeur Bernard Foliguet et c'est pour cette raison qu'il a fait appel à Sandie Ferrigno et à moi afin que nous lui expliquions le côté boîte noire des logiciels commercialisés dans son domaine. Je tente donc de rédiger le plus fidèlement possible ce que nous avons expliqué au Professeur Bernard Foliguet.

La régression polynomiale
La régression polynomiale est l'une des méthodes les plus courantes pour modéliser des données de croissance notamment pendant la période prénatale (Royston et Wright (1998)). Elle est basée sur l'hypothèse qu'à chaque temps (ici le temps est représenté par la semaine d'aménorrhée) la mesure de la variable que nous étudions (ici par exemple la masse du fœtus) suit une loi normale et que la moyenne et l'écart-type varient en fonction du temps au travers d'un polynôme. Les courbes en centiles seront donc toujours estimées à l'aide de la formule précédente à savoir : $C_{100(1-\alpha)}(t) = \mu(t) + \sigma(t) * u_{100(1-\alpha)}$ où α est un nombre réel compris entre 0 et 1, les fonctions $\mu(\cdot)$ et $\sigma(\cdot)$ sont respectivement la moyenne et l'écart-type en fonction du temps (ici ce sera la semaine d'aménorrhée) et $u_{100(1-\alpha)}$ est le quantile de la loi normale centrée et réduite d'ordre $100(1-\alpha)$. Dans cette méthode, contrairement à l'estimation empirique des centiles, ces deux fonctions seront respectivement estimées par des fonction polynomiales d'ordre trois et un. Ces ordres ont été choisis par une méthode de sélection de variables. Ces ordres ont été choisis à l'aide d'une sélection de modèles à l'aide de différents critères (et celui que nous avons retenu est l'AIC). D'ailleurs, Royston et Wright ont également dans leur article (Royston et Wright (1998)) des fonctions polynomiales du même ordre pour la circonférence de la tête du fœtus. Nous injecterons l'ensemble des estimations pour obtenir les centiles estimés (voir la Figure En effet les fonctions polynomiales sont recommandées pour fournir souvent des formes de courbe adéquates pour représenter les relations entre les variables fœtales et l'âge gestationnel (voir la Figure 3.4).

Cette méthode a un inconvénient majeur. Comme pour l'estimation empirique des centiles, la normalité de la distribution de la variable étudiée est nécessaire pour pouvoir l'appliquer. Nous testons la normalité de la variable à chaque SA formellement en utilisant la statistique de Shapiro-Wilk (Royston (1982)), qui est basée sur le coefficient de corrélation au carré des valeurs de l'échantillon or-

donné tracées par rapport à leurs statistiques d'ordre normal attendues. Nous savons qu'il s'agit d'un « bon » test omnibus en l'absence d'asymétrie mais comme nous avons pu le constater pour certaines semaines d'aménorrhée (en particulier la 41ème, pour le sexe masculin) l'asymétrie est présente (voir la Figure 3.5). Il faut donc envisager une autre méthode. Je tiens à mentionner que cette partie de modélisation a été réalisée pendant le stage de Chloé Biabiany pendant l'été 2019 à l'institut Elie Cartan de Lorraine, stage de master de deuxième année du master de statistique de l'Université de Strasbourg rémunéré par un PEPS1 de l'AMIES.

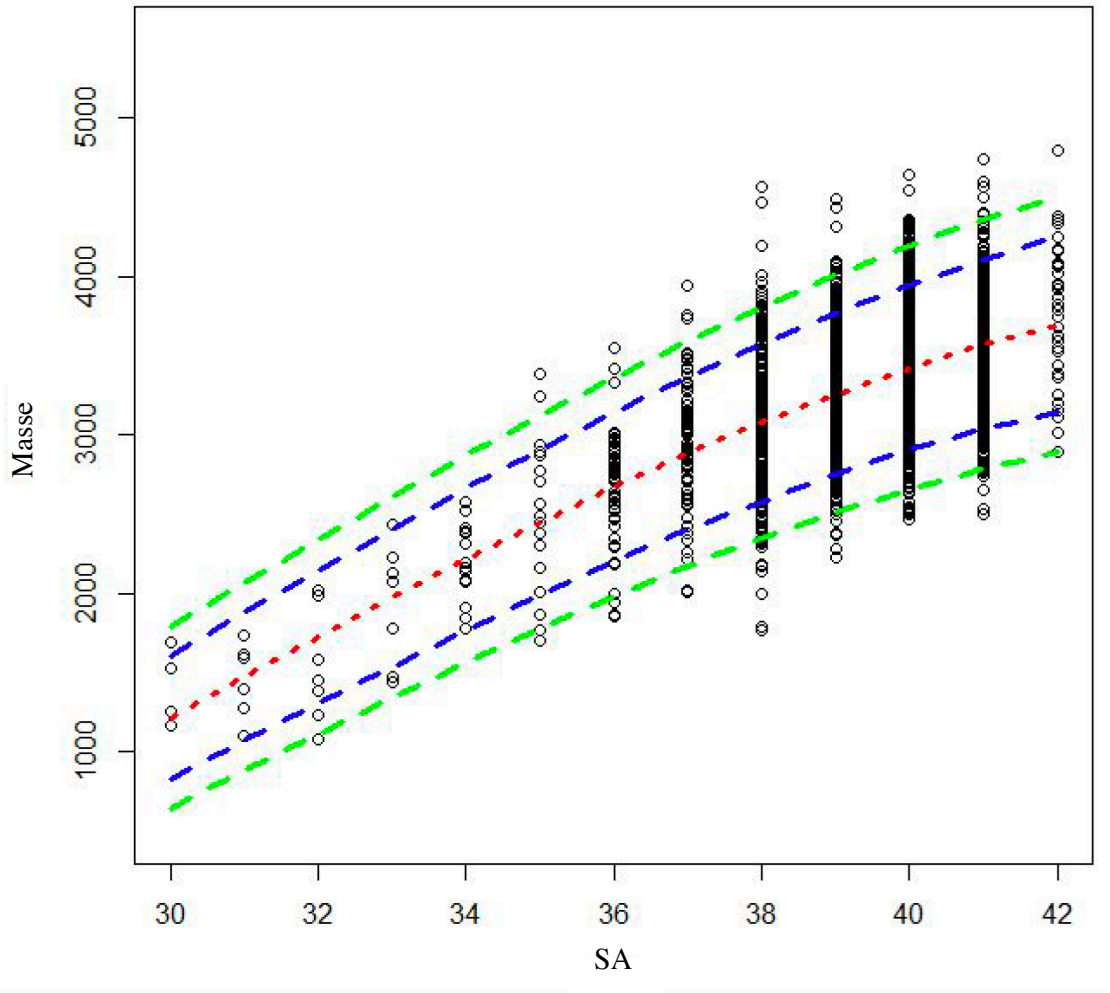


Figure 3.4 : *Régression polynomiale pour construire les courbes 3^e, 10^e, 50^e, 90^e et 97^e centiles.*

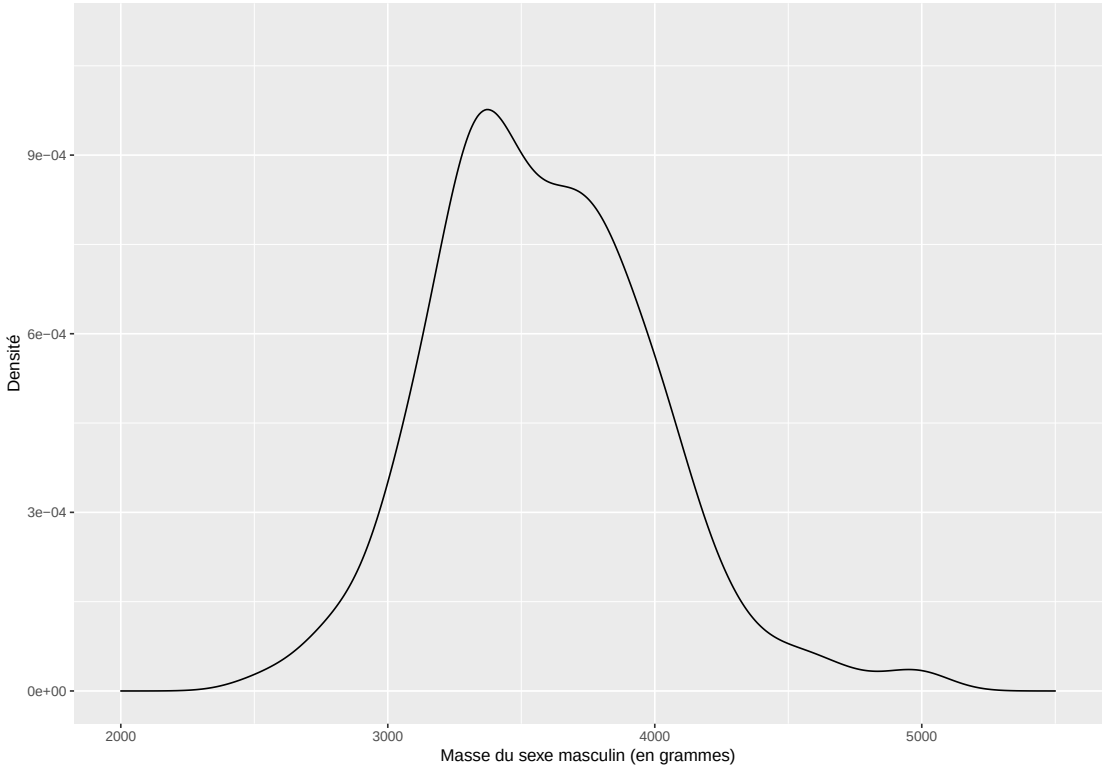


Figure 3.5 : *Densité asymétrique de la masse des fœtus de sexe masculin au cours de la 41e semaine d'aménorrhée*

La méthode LMS

La méthode LMS, (L comme Lambda, M comme Mu et S comme Sigma), est un exemple de modèle construit pour résoudre un problème concret. D'abord en 1988 (Cole (1988)) puis en 1990 (Cole (1990)), Cole a développé cette méthodologie pour résoudre des problèmes spécifiques de la courbe de centiles de croissance, puis ensuite elle a été appliquée à une variété de domaines comme la médecine, la nutrition, etc. La méthode LMS est principalement destinée à corriger l'asymétrie de la distribution d'une variable. Par contre, elle ne sait pas gérer l'aplatisement de la distribution. Il faudra avoir recours à la méthode BCPE (*Box-Cox Power Exponential*).

La première étape de méthode LMS exige que la réponse y soit divisée en K groupes, correspondant à des valeurs ou à des plages de valeurs du temps t. Dans notre cas, le nombre K de groupes est égal au nombre de SA entre la 30ème et la 42ème.

La deuxième étape est que pour chaque groupe i, nous calculons l'estimateur par maximum de vraisemblance du paramètre λ_i. Ce paramètre λ provient de la

transformation de Box-Cox (Box et Cox (1964)) que nous rappelons maintenant :

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & \text{avec } \lambda \neq 0 \\ \ln y & \text{avec } \lambda = 0. \end{cases}$$

La troisième étape est le calcul des deux autres paramètres. Notons par $\widehat{\nu}_{y^\lambda}$ et $\widehat{\sigma}_{y^\lambda}$ la moyenne observée et l'écart-type observé de la variable y^λ. Nous supposons que les deux estimateurs dont sont issues les deux estimations précédentes sont normalement distribués. La médiane associée à la distribution de la variable y^λ est estimée par $\widehat{\nu}_{y^\lambda}$, de sorte qu'une estimation de la médiane de la variable y est donnée par $\widehat{\nu}_{y^\lambda}^{1/\lambda}$. Pour le cas où $\lambda = 0$, une estimation de la médiane est donnée par $\exp(\widehat{\nu}_{y^\lambda})$ tant que $\widehat{\sigma}_{y^\lambda}$ et σ (l'écart-type de la quantité y^λ / y'^λ) coïncident.
L'estimateur du maximum de vraisemblance pour le paramètre λ est la valeur qui minimise le paramètre σ. En outre, la valeur exacte du paramètre λ n'est pas critique puisqu'elle doit être lissée. Cela signifie que le choix des valeurs -1, 0 et $+1$ couvre une plage raisonnable de valeurs et évite les puissances non intégrables, ce qui permet de gagner du temps.
Maintenant que les estimations des paramètres $\widehat{\lambda}_i$, $\widehat{\mu}_i$ et $\widehat{\sigma}_i$ sont calculées à chaque temps t_i, nous allons tracer les courbes lissées, respectivement appelées courbe de puissance et notée $L(t)$, courbe de moyenne et notée $M(t)$ et enfin courbe d'écart-type notée $S(t)$.
Le lissage des trois courbes peut être effectué par n'importe quelle méthode, par exemple les splines cubiques (Silverman (1985)), les méthodes du noyau (Gasser *et al.* (1984)), les polynômes, d'autres fonctions mathématiques spécifiquement adaptées (Jenss et Bayley (1937)) ou simplement en s'ajustant à l'œil.
L'écart-type obtenu à partir de la courbe $S(t)$ peut être ramené à l'unité initiale de la variable y en multipliant par $L(t)\dot{y}_i^{L(t)}$. Toutefois, cela nécessite de lisser les valeurs de $\dot{y}_i$, en plus de celles de $\widehat{\mu}_i$, pour obtenir des valeurs pour $t \neq t_i$. Comme les deux moyennes sont en pratique très similaires, en particulier si $L(t_i)$ est proche de zéro, la méthode est quelque peu simplifiée si $M(t_i)$ est plutôt utilisé que $\dot{y}_i$.
Grâce à cette simplification, les courbes L, M et S peuvent être utilisées pour générer n'importe quelle courbe de centile lisse sur toute la tranche d'âge.
Dans cette méthode, l'équation de la courbe en centiles est donnée par :

$$C_{100(1-\alpha)}(t) = M(t)\left(1 + u_{100(1-\alpha)}L(t)S(t)\right)^{1/L(t)} \quad \text{pour } L(t) \neq 0. \tag{3.1}$$

Lorsque $L(t) = 0$, la forme équivalente de (3.1) est donnée par :

$$C_{100(1-\alpha)}(t) = M(t)\exp\left(u_{100(1-\alpha)}S(t)\right).$$

La méthode LMS utilise l'équation ci-dessous, qui est une forme généralisée de la transformation de Box-Cox (Box et Cox (1964)) pour calculer un score Z

adapté à la distribution des données :

$$Z_{LMS} = \frac{1}{L(t)S(t)} \left[\left(\frac{Y}{M(t)} \right)^{L(t)} - 1 \right] \quad \text{pour } L(t) \neq 0.$$

Lorsque $L(t) = 0$, le score équivalent est égal à : $Z_{LMS} = S(t)^{-1} (Y/M(t) - 1)$. Nous rappelons que les scores Z sont des outils utiles pour évaluer l'ajustement d'un modèle. Ces derniers représentent les mesures exprimées sur une échelle « gaussienne standard », la moyenne et l'écart type étant tous deux ajustés en fonction de la semaine d'aménorrhée. Les mesures relevées sur le fœtus telles que le poids par exemple peuvent avoir une distribution asymétrique avec un mode unique (voir la Figure 3.5). La distribution du score Z_{LMS} avec cette transformation est désormais normale. Ce qui permet de donner un score Z correct pour calculer le centile à condition que l'aplatissement soit déjà égal à zéro.

Maintenant je vais juste rappeler rapidement comment calculer l'estimation des trois paramètre en pratique.
L'écart-type de la mesure étudiée (ici nous prendrons le poids) est calculé pour trois valeurs connues de λ. Les choix les plus simples et les plus économiques sur le plan du calcul sont 1, 0 et -1, correspondant au poids de transformation lui-même (la transformation la plus naturelle), au poids logarithmique et au poids inverse. La moyenne géométrique du poids est également requise, qui est le logarithme inverse (ou l'exponentielle) du log-poids moyen. L'écart-type du poids est divisé par la moyenne géométrique du poids pour donner une forme de coefficient de variation, tandis que l'écart-type inverse du poids est multiplié par la moyenne géométrique ; l'écart-type du log-poids reste inchangé. Les trois écarts types, ou plus précisément les coefficients de variation, se révèlent maintenant très similaires. L'objectif est d'interpoler entre eux pour trouver la valeur minimale du coefficient de variation, qui correspond alors à la meilleure valeur de λ. Appelons respectivement les coefficients de variation ceux obtenus à partir des poids, log-poids et poids inverse s_+, s_0 et s_- . L'estimation du paramètre λ est donnée par :

$$\frac{\ln(s_-/s_+)}{2\ln(s_- s_+/s_0^2)}.$$

Ce processus est répété pour chaque groupe, et les valeurs obtenues de λ sont reportées en fonction de la SA. La courbe lissée L est ensuite tracée à travers tous les points.
Pour trouver les deux autres paramètres, à savoir la moyenne et le coefficient de variation pour chaque SA, le poids est porté à la puissance λ. La moyenne et l'écart-type du poids qui a subi la transformation puissance sont calculés, et cet

écart-type est divisé à la fois par λ et par la moyenne géométrique du poids porté à la puissance λ. C'est le coefficient de variation minimum, et il doit donc être légèrement inférieur à s_-, s_0 et s_+. La moyenne de la variable poids transformée est re-transformée en « poids moyen » en l'élevant à la puissance $1/\lambda$.
Tout comme pour la puissance, la moyenne et le coefficient de variation de chaque SA sont calculés en fonction de la SA, et nous obtenons la courbe lisse pour la moyenne et celle pour le coefficient de variation. Une fois que les courbes L, M et S ont été obtenues, elles peuvent être substituées dans l'équation 3.1 pour des valeurs données de SA afin d'obtenir un ensemble complet de courbes en centiles. Nous présentons ci-dessous les résultats obtenus par cette méthode que nous avons présentés à CMSTATS qui a eu lieu à Londres en décembre 2019 (Ferrigno et Maumy-Bertrand (2019)).

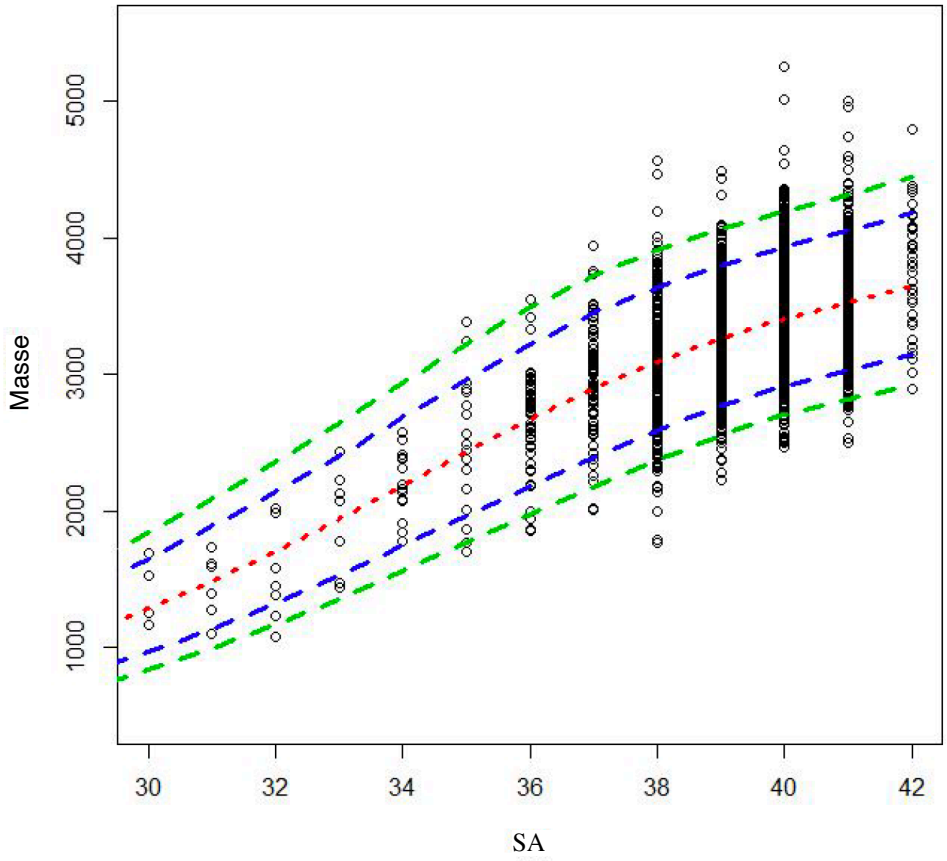


Figure 3.6 : *Méthode LMS pour construire les courbes 3^e, 10^e, 50^e, 90^e et 97^e centiles*

3.4.6 Les méthodes non paramétriques pour traiter la problématique

Le choix d'une méthode non paramétrique s'est imposé à nous suite à plusieurs constatations D'abord, le test de normalité (Royston (1982)) sur la variable poids est parfois significatif au seuil α. En effet, pour la 40ème SA et la 41ème SA le test de Shapiro-Wilk indique que le poids des fœtus de sexe masculin ne suit pas une loi normale pour ces deux SA.
Nous pouvons supposer d'un point de vue biologique que la masse du fœtus, à une SA donnée, dépend fortement de la semaine précédente.
Nous pouvons également supposer que cette dépendance diminue fortement en s'éloignant temporellement de cette SA donnée. Nous venons d'illustrer le principe de la fenêtre d'observations qui est l'élément incontournable des méthodes non paramétriques.
La régression polynomiale locale
Nous cherchons à minimiser, selon le critère des moindres carrés pondérés la fonction suivante :

$$\sum_{i=1}^{n} \frac{1}{h_n} K\left(\frac{X_i - x}{h_n}\right) (Y_i - \mu(X_i))^2$$

où $\mu(X_i)$ représente la moyenne au point X_i et K un noyau symétrique. L'idée qui va permettre d'estimer entre autre la moyenne μ est d'utiliser un développement de Taylor à l'ordre p de la fonction μ au voisinage du point x. De plus, nous ne supposons aucune forme particulière sur le couple de données (X_i,Y_i) pour estimer les paramètres. Le problème se ré-écrit ainsi :

$$\hat{\boldsymbol{\beta}}_h = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} K_{h_n}(X_i - x) \left(Y_i - \sum_{j=0}^{p} \beta_j (X_i - x)^j\right)^2$$

où nous posons $K_{h_n}(z) = \frac{1}{h_n} K\left(\frac{z}{h_n}\right)$ avec K un noyau symétrique et $\beta_j = \frac{\mu^{(j)}(x)}{j!}$.
La résolution de ce problème d'optimisation est plus aisée en notation matricielle. Posons :

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 - x & \cdots & (X_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n - x & \cdots & (X_n - x)^p \end{pmatrix}_{n\times(p+1)}$$

$$\mathbf{W}_x = \mathrm{diag}(K_h(X_1 - x), \ldots, K_h(X_n - x)) \quad \text{et} \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}_{n\times 1}.$$

Le problème se ré-écrit de façon matricielle ainsi :

$$\widehat{\boldsymbol{\beta}}_h = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{W}_x(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}'\mathbf{W}_x\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}_x\mathbf{Y}.$$

L'estimateur de $\mu^{(j)}$, dérivée d'ordre j de la moyenne conditionnelle, est donc égal à :

$$\widehat{\mu}(x;j,h_n) = j!\widehat{\beta_j} = j!\mathbf{e}_j'(\mathbf{X}'\mathbf{W}_x\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}_x\mathbf{Y} \quad \text{pour } j=0,\ldots,p,$$

où $\mathbf{e}_j \in \mathbb{R}^{p+1}$ indicateur de la $j$ème composante.
Remarques.
Si $p=0$, alors nous retrouvons les estimateurs de Nadaraya-Watson (Nadaraya (1964), Watson (1964)) pour la moyenne et la variance :

$$\widehat{\mu}(x;0,h_n) = \sum_{i=1}^{n} \frac{K_{h_n}(x-X_i)}{\sum_{i=1}^{n} K_{h_n}(x-X_i)} Y_i$$

et

$$\widehat{\sigma^2}(x;0,h_n) = \sum_{i=1}^{n} \frac{K_{h_n}(x-X_i)}{\sum_{i=1}^{n} K(\frac{x-X_i}{h_n})} {Y_i}^2 - \widehat{\mu}^2(x;0,h_n).$$

Si $p=1$, alors il s'agit d'une régression non paramétrique linéaire locale. Posons

$$W_i(x;1,h_n) = \frac{1}{n}\frac{\widehat{s}^2(x;h_n) - \widehat{s}^1(x;h_n)(X_i-x)}{\widehat{s}^2(x;h_n)\widehat{s}^0(x;h_n) - \widehat{s}_1(x;h_n)^2} K_{h_n}(x-X_i),$$

où

$$\widehat{s}^r(x;h_n) = \frac{1}{n}\sum_{i=1}^{n} K_{h_n}(x-X_i)(X_i-x)^r.$$

Les estimateurs pour la moyenne et la variance sont alors égaux à :

$$\widehat{\mu}(x;1,h_n) = \sum_{i=1}^{n} W_i(x;1,h_n) Y_i$$

et

$$\widehat{\sigma}^2(x;1,h_n) = \sum_{i=1}^{n} W_i(x;1,h_n) Y_i^2 - \widehat{\mu}^2(x;1,h_n)$$

Nous calculons d'abord les deux estimations dans les deux cas puis nous les injectons dans l'équation qui estime le centile pour obtenir l'estimation du centile d'abord par l'estimateur de Nadaraya-Watson puis par la régression linéaire locale.

Le choix de la fenêtre optimale
Le choix de la fenêtre est l'élément crucial pour la qualité de l'estimation des différentes fonctions servant à estimer le centile. Il existe plusieurs méthodes pour déterminer la fenêtre optimale. Dans la littérature, nous rencontrons deux types de fenêtre :

1. Les fenêtres constantes. Elles sont fixées indépendamment des données (X_i,Y_i) et des valeurs x du support de la densité marginale de X pour lesquelles nous calculons l'estimateur de la moyenne.

2. Les fenêtres variables. Nous distinguons les fenêtres variables globales, ne dépendant que des observations et les fenêtres variables localement qui dépendent des valeurs de x, pour lesquelles nous calculons l'estimateur.

Une petite fenêtre réduit le biais de l'estimateur, mais augmente sa variance. Une grande fenêtre réduit sa variance mais augmente son biais. Or, le problème est de choisir la fenêtre minimisant, simultanément, biais et variance.
Dans ce but, nous utilisons comme critère à minimiser, l'erreur quadratique moyenne, abrégée en MSE, (dans le cas de fenêtres variables localement) ou l'erreur quadratique moyenne intégrée, abrégée en MISE (dans le cas de fenêtres variables globalement).

La fenêtre optimale locale est approchée par minimisation de l'erreur quadratique moyenne asymptotique. En utilisant les formules de biais et de variance de l'estimateur et en effectuant la minimisation nous obtenons :

$$h_{opt}(x) = C_{j,p}(K) \left(\frac{\sigma^2(x)}{m^{(p+1)}(x)^2 f(x)} \right)^{\frac{1}{2p+3}} n^{-\frac{1}{2p+3}},$$

où

$$C_{j,p}(K) = \left(\frac{(p+1)!^2(2j+1)\nu_0(K_j^*)}{2(p-j+1)\mu_{p+1}^2(K_j^*)} \right)^{\frac{1}{2p+3}}$$

où $\mu_{p+1}(K_j^*) = \int u^{p+1} K_j^*(u)\,\mathrm{d}u$, $\nu_0(K_j^*) = \int (K_j^*(u))^2\,\mathrm{d}u$.
La fenêtre globale peut être obtenue par minimisation de l'erreur quadratique moyenne intégrée asymptotique. En utilisant les formules de biais et de variance de l'estimateur et après minimisation, nous obtenons :

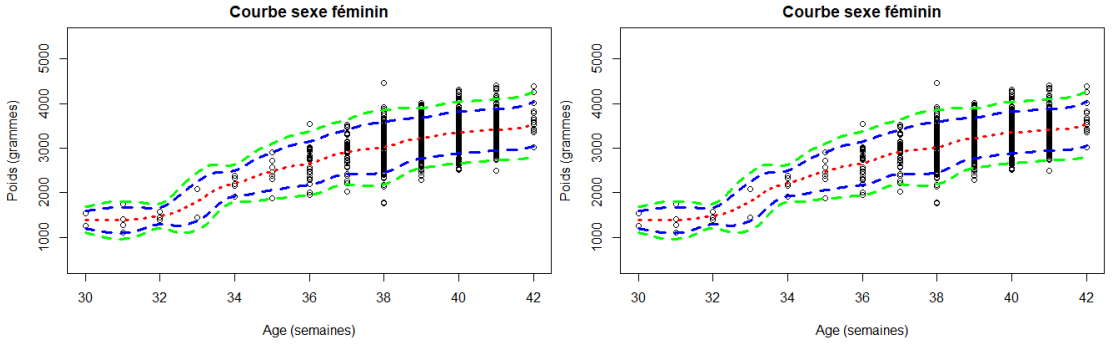
$$h_{opt}(x) = C_{j,p}(K) \left(\frac{\int \sigma^2(x) w(x)/f(x) \mathrm{d}x}{\int m^{(p+1)}(x)^2 \omega(x) \mathrm{d}x} \right)^{\frac{1}{2p+3}} n^{-\frac{1}{2p+3}},$$

où ω est une fonction de poids positive. Nous pouvons choisir par exemple $\omega = f$ et la formule ci-dessous se simplifie.

Nous avons rassemblé ci-dessous les résultats obtenus par les deux méthodes citées précédemment avec différentes fenêtres. Nous avons utilisé le noyau gaussien car dans notre cas il est plus adapté de choisir un tel noyau qui ne renvoie pas 0 en dehors de $[-1;1]$. En effet, si nous ne le choisissons pas, alors nous pouvons nous retrouver dans le cas où le poids associé à une donnée est nul et donc la donnée ne sera pas prise en compte dans l'estimation du centile. Nous avons choisi de donner un poids non nul à toutes les données. Après le choix du noyau gaussien, de la fenêtre optimale, le calcul des poids, nous estimons la moyenne et la variance puis le centile. Il ne reste plus qu'à tracer les courbes de référence.

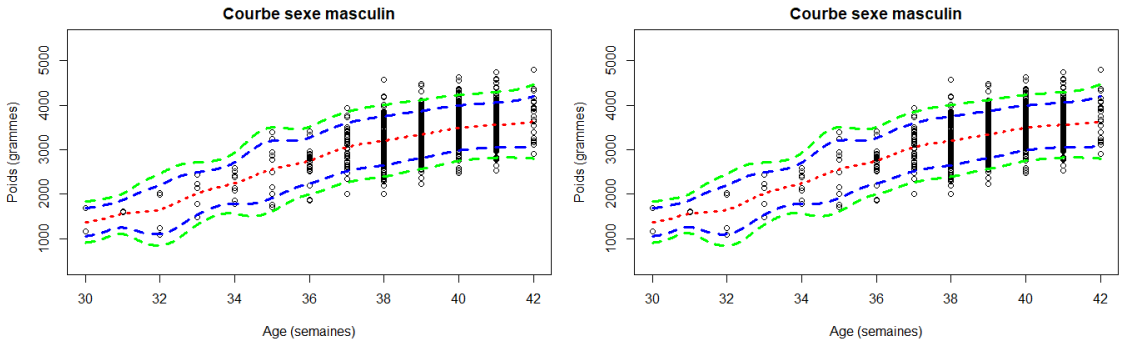
Les résultats pour l'estimateur de Nadaraya-Watson

À gauche, nous avons les résultats avec la méthode *mean square cross validation* et à droite, ceux avec la méthode AIC. Les deux méthodes de choix de fenêtre optimale donnent des résultats quasiment identiques. Nous avons arrondi les fenêtres au centième près.



(a) *Résultat avec h=0,46* (b) *Résultat avec h=0,45*

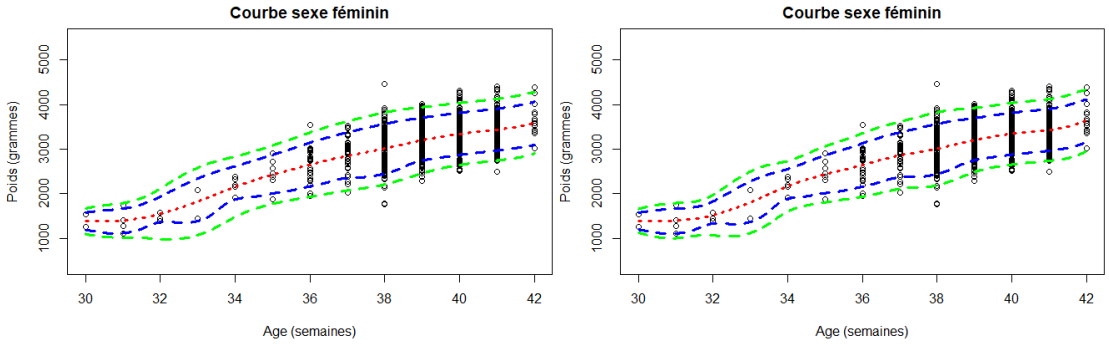
Figure 3.7 : *Comparaison de deux choix de fenêtre pour le sexe féminin*



(a) *Résultat avec h=0,53* (b) *Résultat avec h=0,52*

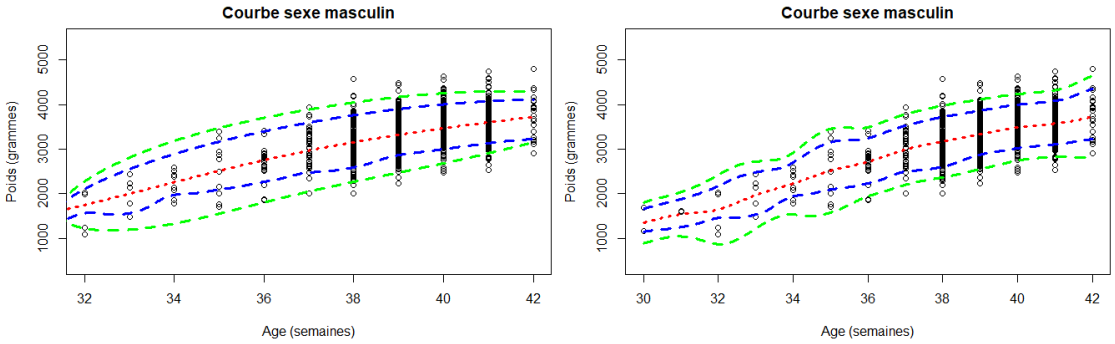
Figure 3.8 : *Comparaison de deux choix de fenêtre pour le sexe masculin*

Les résultats pour la régression linéaire locale
À droite, nous avons les résultats avec la méthode *mean square cross validation* et à gauche, ceux avec la méthode *plug-in*. Les résultats de la méthode *plug-in* donnent de meilleurs résultats que la méthode *mean square cross validation*. En effet, cette dernière donne de mauvais résultats pour les SA où il y a très peu de données. En particulier, elle produit des estimations de variance négatives sûrement en raison d'une mauvaise estimation des paramètres du modèle due au manque de données.



(a) *Résultat avec h=0,90* (b) *Résultat avec h=0,69*

Figure 3.9 : *Comparaison de deux choix de fenêtre pour le sexe féminin*



(a) *Résultat avec h=2,20* (b) *Résultat avec h=0,61*

Figure 3.10 : *Comparaison de deux choix de fenêtre pour le sexe masculin*

Tous ces résultats présentés jusqu'ici ont été présentés à la conférence CMSTAT 2019 (Ferrigno et Maumy-Bertrand (2019)) qui a eu lieu en décembre 2019. Un stage d'élève ingénieur de quatre mois est consacré à l'approfondissement des méthodes non paramétriques et au développement d'un *package* R. Un article est en cours de rédaction et sera soumis à *Statisticis in Medicine*.

La régression quantile linéaire locale

Nous commençons par rappeler quelques notations. Les régressions quantiles tentent d'évaluer comment les quantiles conditionnels $q_\alpha(Y|X)$, définis par :

$$q_\alpha(x) = \inf\{y : F(y|x) \geqslant \alpha\}.$$

Un estimateur de la fonction de répartition $F_{Y|X}$ se définit par :

$$\hat{F}_n(y|x) = \sum_{i=1}^{n} \frac{K(\frac{x-X_i}{h_n})}{\sum_{i=1}^{n} K(\frac{x-X_i}{h_n})} 1_{\{Y_i \leqslant y\}}.$$

Nous pouvons interpréter cet estimateur comme la moyenne pondérée des poids de chaque observation Y_i. Dans le cas qui nous intéresse, la masse du fœtus est inférieure ou égale à la masse que nous souhaitons observer.

Remarque : cet estimateur est très proche de l'estimateur de Nadaraya-Watson. Néanmoins, sa forme est plus complexe et l'obtention du centile nécessite d'inverser cet estimateur ce qui, en pratique, est très coûteux en temps de calcul et plus risqué.

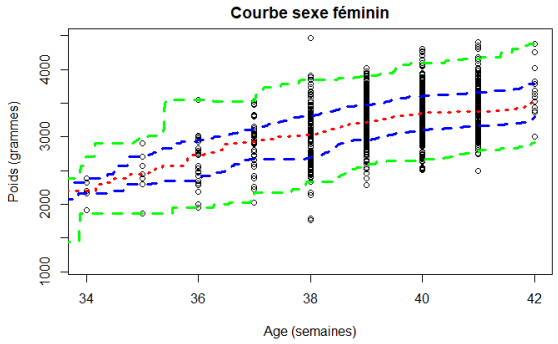
Une caractérisation alternative de quantile conditionnel est obtenue sous forme d'un problème d'optimisation :

$$q_\alpha(x) = \arg\min_a E\left[\rho_\alpha(Y-a)|X=x\right]$$

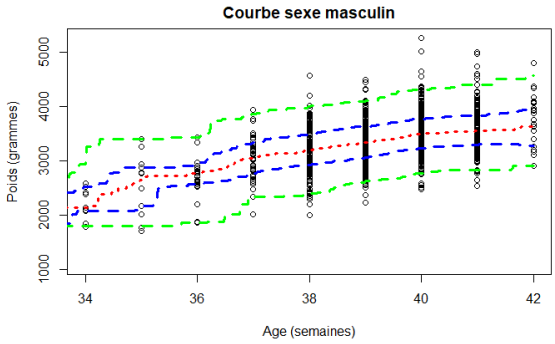
où la fonction ρ_α est définie par : $\rho_\alpha(x) = \alpha x * 1_{[0,+\infty[}(x) - (1-\alpha)x1_{]-\infty;0)}(x)$.

Les résultats

Nous avons utilisé la fonction ***optimize*** du *package* ***stats***. Cette fonction utilise une combinaison de recherche par section d'or et d'interpolation parabolique successive, et a été conçue pour être utilisée avec des fonctions continues. Nous n'avons qu'à fournir la grille de calcul de x. Le choix de la fenêtre est encore une fois le critère le plus important. Nous avons gardé la fenêtre obtenue par la méthode de validation croisée de l'estimateur de Nadaraya-Watson.



(a) *Résultat pour les fœtus de sexe féminin avec h=0,46*



(b) *Résultat pour les fœtus de sexe masculin avec h=0,53*

Limites

Cette régression quantile a quelques limites. Elle est très dépendante des données voisines du point où l'estimateur est calculé. En particulier, aux extrémités, la méthode devient beaucoup moins précise qu'à l'intérieur des données. De plus la base de données EDEN comporte peu de fœtus entre la 27ème SA et la 34ème SA, ce qui accentue le phénomène. C'est pour cette raison que nous avons commencé à la 34ème SA.

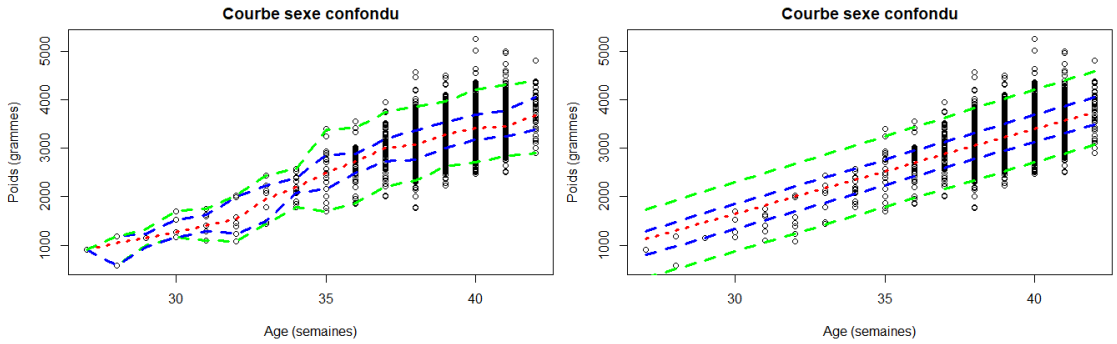
La régression quantile par splines cubiques

La régression par splines est un modèle non linéaire qui est aussi adapté au problème étudié car il sépare l'espace des données en partitions et qui effectue une régression sur chacune des partitions. Soient $\xi_1, \ldots, \xi_M$, M nœuds qui coupent X et qui permettent ainsi de faire une partition des valeurs prises par X. Soient $B_1, \ldots, B_{M+4}$ la base de splines cubiques associée. L'estimateur du quantile se définit par :

$$q_{100(1-\alpha)}(x) = \min_{\theta} \sum_{i=1}^{n} \rho_\alpha(Y_i - \sum_{j=1}^{M+4} \theta_j B_j(X_i))^2.$$

Les résultats

Nous avons utilisé la fonction `qsreg` du *package* `fields` (Douglas Nychka *et al.* (2017)). Cette fonction propose une approche assez sophistiquée pour effectuer le calcul des quantiles. En effet, elle utilise un algorithme itératif pour calculer un estimateur à base de splines cubiques puis lisse cet estimateur. Dans cette fonction, nous ne choisissons que le nombre d'itérations de l'algorithme. Nous avons illustré deux cas extrêmes pour le nombre d'itérations.



(a) *Résultat pour 1 itération* (b) *Résultat pour 80 itérations*

Figure 3.12 : *Comparaison entre deux nombres d'itérations*

Pour déterminer le bon nombre d'itérations, nous avons procédé à tâtons en considérant deux critères qualitatifs à savoir la croissance de chaque centile et le fait que les courbes ne doivent pas se croiser.

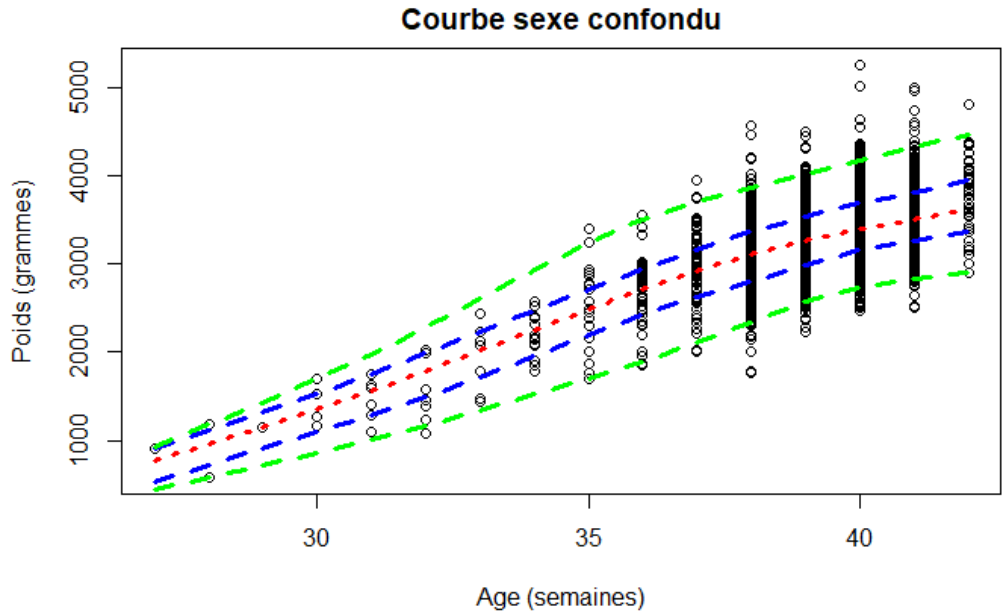


Figure 3.13 : *Résultat pour 25 itérations*

3.5 Modèles mixtes

3.5.1 Mesures doublement répétées et splines cubiques

L'originalité statistique des données étudiées lors de cette collaboration à une thèse avec le Laboratoire de Psychologie des Cognitions (EA 4440-UdS), Strasbourg, France provient de la présence de données doublement répétées sur les sujets, mesures répétées lors de sessions elles-mêmes répétées. Le nombre de répétitions au sein d'une session de 32h pouvant être élevée, au maximum toutes les heures après la nuit de repos soit 25 tests pour l'étude de la puissance des ondes cérébrales par électro-encéphalogramme, des modèles mixtes non paramétriques basés sur des splines cubiques ont été utilisés, voir Grenèche *et al.* (2011b).
En effet, cette étude de suivi a été menée avec des tâches à mémoire et répétée chez 12 patients atteints du SAHOS et 10 témoins sains ayant subi trois séances de 32 heures, la première avant la PPC (T0), la deuxième (T3) et la troisième (T6) respectivement après trois et six mois de traitement pour les patients souffrant d'apnée du sommeil et de syndrome d'hypopnée obstructive (SAHOS). Chaque session comprenait une nuit de sommeil suivie de 24 heures d'éveil prolongé pendant lesquelles les deux groupes effectuaient des tâches sollicitant leur

mémoire à court terme (STM), y compris des tâches numériques (DS) et des tâches Sternberg.
Peu d'études ont examiné l'impact de la thérapie par pression positive continue (CPAP) sur la mémoire à court terme (STM) par rapport à l'état de veille prolongé chez les patients atteints d'apnée du sommeil et de syndrome d'hypopnée obstructive (SAHOS). Nous avons cherché à savoir si le traitement CPAP pouvait inverser la dégradation de la STM dans un paradigme de veille continue de 24 heures. La quantité de données recueillie et exploitée pendant ce travail de thèse était d'une ampleur très conséquente et a donné lieu à trois publications : Grenèche *et al.* (2011b), Grenèche *et al.* (2011a) et Grenèche *et al.* (2013).

3.5.2 Approches permutationnelles et petits échantillons

L'ischémie reperfusion (abrégées en IR) des membres inférieurs est fréquente et associée à une morbidité et une mortalité importantes. C'est un épisode délétère, incontournable de la transplantation d'organes.
Les inhibiteurs de la phosphodiestérase 5 ont démontré des effets antioxydants et bénéfiques dans plusieurs organes soumis à l'ischémies reperfusion, mais leurs effets sur les fonctions mitochondriales des muscles après l'ischémie reperfusion des membres inférieurs sont inconnus. Une ischémies reperfusion unilatérale des membres inférieurs (garrot de 2 heures suivi d'une reperfusion de 2 h) sans ou avec le citrate de sildénafil (1mg/kg ip 30 minutes avant l'ischémie) a été effectuée chez 18 souris.
La capacité d'oxydation maximale (VMax), la contribution relative des complexes de la chaîne respiratoire mitochondriale, la capacité de rétention du calcium (CRC) - un marqueur de l'apoptose - et la production d'espèces réactives à l'oxygène, abrégés en ROS, (pour *reactive oxygen species*) ont été déterminées par respirométrie à haute résolution, spectrofluorométrie et résonance paramagnétique électronique dans les muscles gastrocnémiens des deux membres postérieurs. L'ischémies reperfusion a réduit de manière significative le VMax mitochondrial (de 11,79 $\pm$ 1,74 à 4,65 $\pm$ 1,11 pmol/s*mg poids humide (ww), $p < 0{,}05$, -50,2 $\pm$ 16,3%) et la capacité de rétention du calcium (de 2,33 $\pm$ 0,41 à 0,84 $\pm$ 0,18 μmol/mg poids sec (dw), $p < 0{,}05$; -61,1 $\pm$ 6,8%). Les ROS ont eu tendance à augmenter dans le membre ischémique (+64,3 $\pm$ 31,9%, p = 0,08). Bien que tendant à réduire la production de ROS liée aux IR (-42,4%), le sildénafil n'a pas réussi à réduire les dysfonctionnements mitochondriaux musculaires (-63,3 $\pm$ 9,2%, $p < 0{,}00$ et -55,2 $\pm$ 7,6% $p < 0{,}01$ pour le VMax, et la capacité de rétention du calcium, respectivement).
En conclusion, l'ischémies reperfusion des membres inférieurs a altéré la fonction mitochondriale des muscles squelettiques, mais, malgré une tendance à réduire la production de ROS, le pré-conditionnement pharmacologique avec le

citrate de sildénafil n'a pas montré d'effets protecteurs.
Ces résultats ont été publiés dans Tetsi *et al.* (2019) et Liliane Tetsi a soutenu sa thèse en le 20 décembre 2019.

3.5.3 Mesures doublement répétées et classification sur données manquantes

Une collaboration avec Virginie Doyen, chef de clinique adjoint en immuno-allergologie au CHU Brugmann de l'université libre de Bruxelles et Ahn Poirot du CHU de l'université de Strasbourg a commencé en octobre 2019. Son objectif est d'évaluer la réponse, en fonction du statut asthmatique, après l'exposition à des allergènes. Le plan expérimental consiste en des mesures doublement répétées : chaque sujet se présente à deux visites au cours desquelles deux mesures sont réalisées, l'une avant l'exposition, l'autres six heures après. Plusieurs variables explicatives ont été observées directement sur les patients ou à l'aide de cytométrie en flux. L'analyse de ces données est encore compliquée par la présence de données manquantes, ce qui nécessite un traitement spécifique. Par exemple pour la classification des données, nous avons utilisé l'approche *Variable Selection for Model-Based Clustering of Mixed-Type Data Set with Missing Values* (Marbac et Sedki (2017)) qui permet d'analyser des données continues, catégorielles, entières ou un mélange des types précédents avec une prise en charge des valeurs manquantes ne nécessitant pas de pré-traitement. Certains des résultats sont présentés dans les Figures 3.14, 3.15 et 3.16. Nous allons aussi utilisé l'approche *K-means clustering with build-in missing data imputation* (`https://github.com/pkopper/ClustImpute`) qui combine imputation et algorithme *K-means*.
Une *colored image map* (Weinstein *et al.* (1997), Eisen *et al.* (1998), Lê Cao *et al.* (2011), González *et al.* (2012)) est une visualisation bi-dimensionnelle d'une matrice à valeur réelle avec des lignes et/ou des colonnes réorganisées selon une méthode de classification hiérarchique pour identifier des motifs intéressants. Les dendrogrammes générés à partir du regroupement sont ajoutés au côté gauche et en haut de l'image. Par défaut, la méthode de regroupement utilisée pour les lignes et les colonnes est la méthode de liaison complète et la mesure de distance utilisée est la distance euclidienne. Les résultats sont représentés sur les jeux de données au format *wide* (Figure 3.17) ou *long* (Figure 3.18).
Ce travail est toujours en cours de réalisation mais il a déjà été présenté à deux conférences en 2020 (Doyen *et al.* (2020)), c'est pourquoi le nom des variables n'est pas explicité ici.

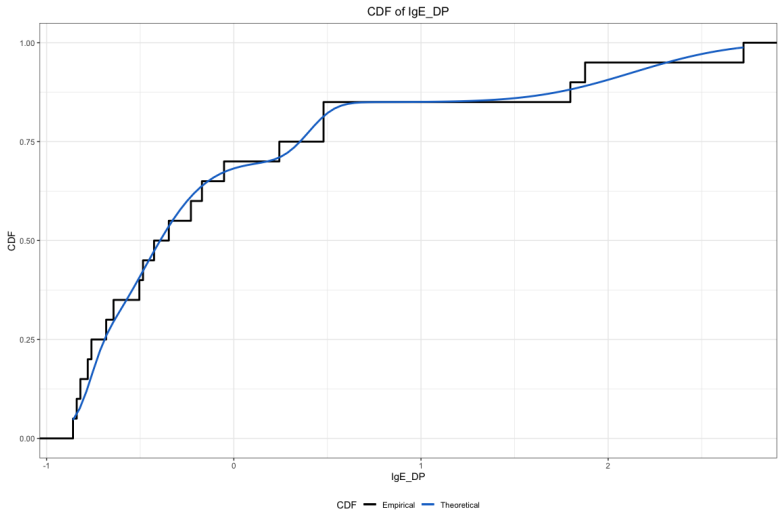


Figure 3.14 : *Vérification de la concordance des distributions empiriques et ajustées*

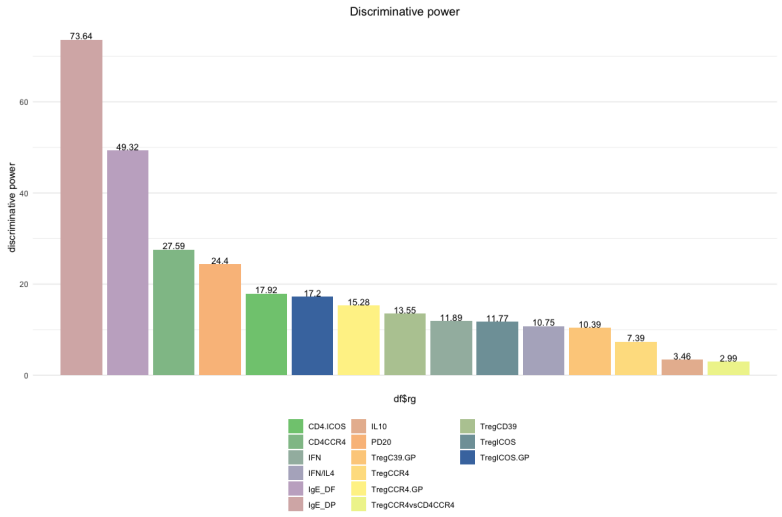


Figure 3.15 : *Pouvoir discriminant des variables*

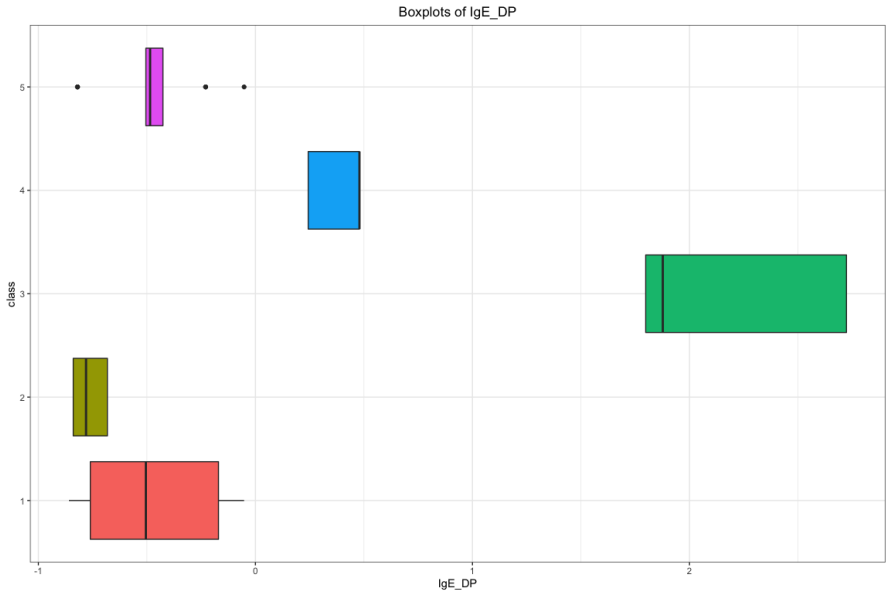


Figure 3.16 : *Variable fortement discriminante et classification obtenue*

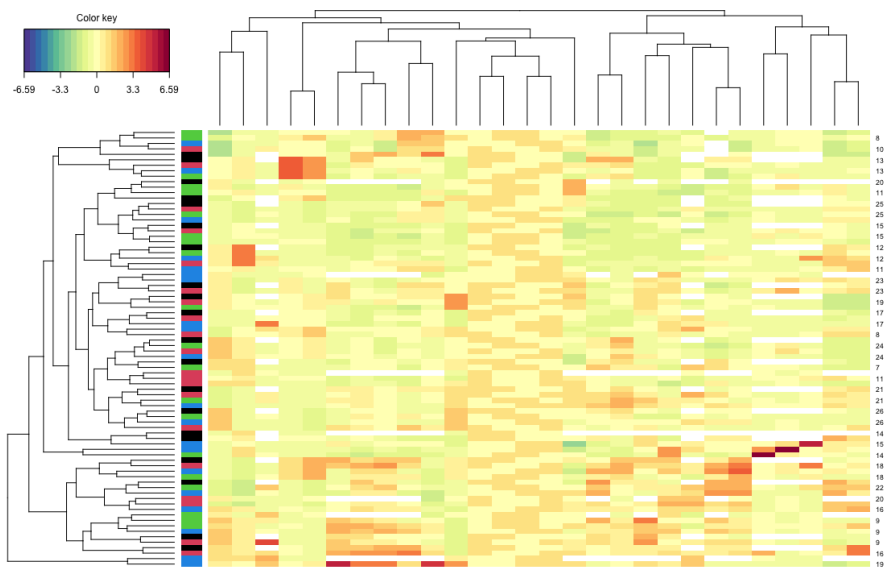


Figure 3.17 : *Colored image maps sur les données au format wide*

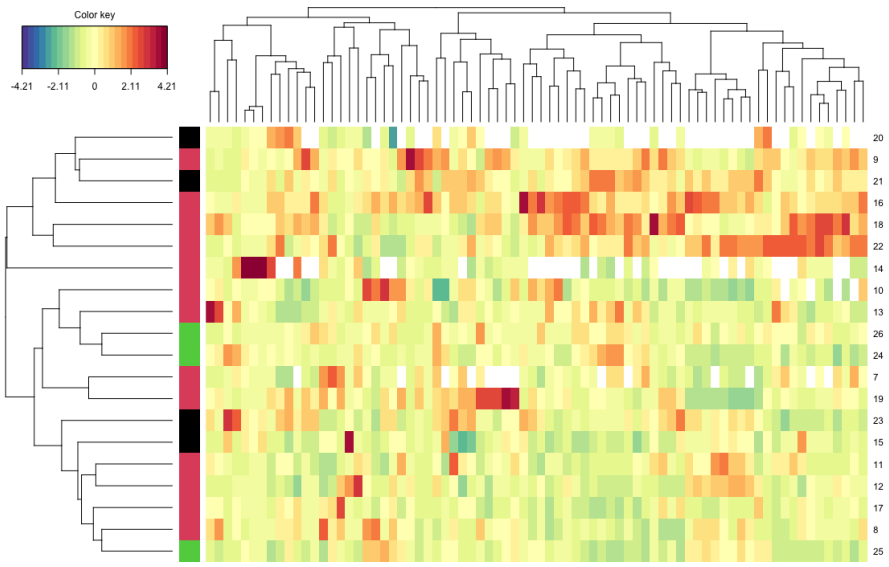


Figure 3.18 : *Colored image maps sur les données au format long*

Chapitre 4

Modélisation statistique dans l'industrie

4.1 Introduction

Je ne développerai pas ici toutes les collaborations que j'ai pu développer avec le monde industriel. J'ai choisi volontairement de ne présenter que celles qui sont actuellement en cours et/ou qui ont donné lieu à un encadrement de thèse ou d'un ingénieur de recherche. Ma première participation dans le monde économique et industriel fut mon intervention dans l'enquête MORGOAT, que j'ai déjà décrite dans le chapitre 1. J'ai découvert grâce à Jean-Claude Deville, que je tiens à remercier sincèrement encore une fois ici, la statistique appliquée. Il m'a expliqué et montré comment des méthodes et des modèles théoriques pouvaient énormément aider à résoudre un problème concret. Puis, j'ai continué, en découvrant une autre branche de la statistique (la statistique bayésienne) et un autre domaine d'application (l'agro-alimentaire) en allant au centre de recherche de Nestlé, situé à Lausanne en Suisse. C'est par l'intermédiaire d'Eric Parent, Professeur à l'ENGREF que j'ai connu Philippe Girard qui m'a accueilli en tant que chercheur post-doctoral pendant six mois (de mars 2004 à août 2004).
Ensuite, j'ai été invitée deux années consécutives (du 1er juin au 31 août 2005 puis la même période en 2006) comme chercheur chez Lilly, Mont-Saint-Guibert en Belgique. D'ailleurs de ce deux visites, un article et une note ont été produits : l'article était plus orienté vers les applications (Boulanger *et al.* (2007)) et la note plus théorique (Bertrand et Maumy (2007)). Puis avec Frédéric Bertrand, nous avons été invités chez GlaxoSmithKline Biologicals pendant quatre mois (du 1er juin au 31 août 2007), toujours en Belgique. Nous avons, entre autres, collaboré avec Walthère Dewé sur la validation de méthodes analytiques ou avec Édouard Ledent sur des problématiques liées à la vaccination et à leurs études cliniques.

4.2 AB Tasty

Cette section concerne le sujet de thèse d'Emmanuelle Claeys que j'ai co-encadrée entre 2016 et 2019 avec Pierre Gançarski, Professeur des Universités à l'IUT d'informatique, Robert Schumann, situé à Illkirch, dans le cadre d'un contrat apparenté à une CIFRE avec la société AB Tasty dont le siège social est situé à Paris. Les sous-sections qui vont suivre sont issues de la thèse d'Emmanuelle Claeys.

4.2.1 Problématique

Dans de nombreux domaines (santé, vente en ligne, production industrielle,...) concevoir *ex nihilo* une solution optimale répondant à un problème défini (trouver un protocole augmentant le taux de guérison, concevoir une page Web favorisant l'achat d'un ou plusieurs produits, définir un processus de fabrication débouchant sur une meilleure qualité des produits,...) est souvent très difficile voire impossible. Face à ce problème, les concepteurs travaillent souvent de façon incrémentale par des améliorations successives d'une solution existante. Néanmoins, définir les modifications les plus pertinentes reste un réel problème. Pour tenter d'y répondre, une solution adoptée de plus en plus fréquemment, consiste à comparer concrètement différentes alternatives, appelées aussi variations, afin de déterminer celle(s) répondant le mieux au problème. L'idée est de mettre en œuvre réellement ces alternatives et de comparer les résultats obtenus, c'est-à-dire les gains respectifs obtenus par chacune des variations. Le principe développé dans les méthodes du test A/B (Fisher (1935); Thompson (1933)) est le suivant : à partir d'une solution existante, généralement nommée **variation A**, une ou plusieurs alternatives, appelées **variations B/C/D/...**, sont construites. Ces dernières sont alors mis en œuvre in vivo. Cependant, afin de garantir l'indépendance entre les réalisations des variations, chaque décision d'affectation d'un item à une variation est irrévocable. De cette contrainte découle le fait qu'il est impossible de savoir quel aurait été le gain obtenu en cas d'affectation de l'item à une autre variation. De plus, les sujets soumis au test n'ont pas conscience d'être testés et ignorent même qu'il existe des variations différentes. À la fin de cette phase dite d'**exploration**, les performances de chaque variation sont alors comparées en fonction de l'objectif visé. Cette comparaison repose sur une ou plusieurs méthodes statistiques permettant de quantifier et/ou qualifier une différence de gain global comme par exemple, le cumul ou la moyenne des récompenses produites par chacune des variations. La meilleure variation peut alors être mise en **exploitation**. Historiquement, en 1925, Fisher (Fisher (1925)) est le premier à suggérer cette idée et en 1935, il propose de définir plus formellement le contexte. Parallèlement, en 1933, Thompson proposa une méthode de test qui complétera les techniques de comparaisons proposées par Fisher.

4.2.2 Stratégie d'allocation et allocation dynamique

La stratégie d'allocation des items aux variations est au cœur du processus de test car c'est elle qui doit permettre de déterminer la meilleure variation à l'issue de la phase d'exploration.
Une première approche du test A/B qualifiée de **statique** consiste à fixer en amont du test la durée de la phase d'exploration et les proportions relatives d'items à affecter à chaque variation. À l'issue de cette phase, les gains respectifs de chacune des variations sont alors comparés et il est ainsi possible de déterminer la meilleure variation. Il s'avère que la durée de cette phase d'exploration est, dans les faits, très difficile à fixer car elle dépend fortement de l'objectif du test, des variations testées et des items eux-mêmes. Or cette phase peut représenter un coût très élevé. En effet, le fait de proposer trop longtemps une variation sous-optimale alors que la variation optimale (*i.e.*, la meilleure variation parmi les variations testées) aurait pu être choisie plus tôt, débouche sur un manque à gagner pouvant être fort : à chaque fois que la méthode affecte une variation de façon sous-optimale, celle-ci produit un gain inférieur à celui qu'aurait produit la version optimale. Or cette dernière est malheureusement inconnue a priori, l'objectif de test étant par définition, de la déterminer.
Pour répondre à ce problème, le concept d'**allocation dynamique** a vu le jour. Il s'agit de permettre de modifier automatiquement les différents ratios lors de la phase d'exploration et ainsi de basculer plus vite vers la solution optimale. L'idée est de maintenir et de mettre à jour les estimations de gain de chacune des variations et d'allouer les items en fonction de celles-ci. Il s'agit à la fois de privilégier la variation la plus prometteuse tout en affinant les estimations des espérances des autres variations en continuant à allouer des items à des variations potentiellement sous-optimales.

4.2.3 Allocation dynamique et modèle de bandits

Pour mettre en place une allocation dynamique efficace, les concepteurs des méthodes des tests A/B se sont orientés vers des stratégies d'allocation basées sur des **modèles de bandits manchots**.
Cette notion de bandit a été introduite par Lai et Robbins (Lai et Robbins (1985)) sous le nom de **bandit multi-bras**. Ce dernier se définit comme un problème dans lequel un ensemble limité de ressources doit être réparti entre des choix (des alternatives) d'une manière qui maximise le gain. Les propriétés de chaque choix ne sont que partiellement connues au moment de la répartition, et ne peuvent être mieux comprises qu'au fil du temps ou en allouant des ressources aux différents choix (Gittins et Jones (1974)). Intuitivement, par analogie avec les machines à sous d'un casino, il s'agit de choisir, pour un joueur, sur une ma-

chine à plusieurs bras, celui qui présente, pour lui, la meilleure espérance de gain. Pour cela, chaque fois qu'il a joué un bras et éventuellement récolté un gain, le joueur met à jour ses estimations de gain sur chaque bras de la machine. Son choix suivant tiendra compte de ces nouvelles estimations. Le but du joueur est de trouver le plus rapidement possible le meilleur bras, appelé **bras optimal** pour le jouer au maximum ou pour minimiser le regret. De façon évidente, ce cadre théorique répond parfaitement au problème de l'allocation dynamique : les bras sont les variations, les parties sont les items et les espérances de gains sont calculées à partir des récompenses réellement obtenues. Les méthodes basées sur les modèles de bandits manchots telles que Ucb (Auer *et al.* (2002)) ou Thompson Sampling (Thompson (1933)) ont rapidement prouvé leur efficacité dans plusieurs domaines.
Néanmoins, elles ne permettaient pas jusqu'à récemment de tenir compte des items eux-mêmes. En effet, ces approches sont **non informées** et ne prennent pas en compte les informations potentiellement disponibles sur les items eux-mêmes. Or, le contexte dans lequel le patient, le visiteur ou l'objet à produire évolue ou se développe est apparu comme indispensable à prendre en compte pour améliorer et réduire fortement la phase d'exploration.
Ainsi, des **méthodes informées** ont vu le jour. Basées sur des **bandits contextuels**, elles permettent d'intégrer les **caractéristiques** des items directement dans le mécanisme d'affectation dynamique des variations. Les méthodes telles que KernelUCB (Valko *et al.* (2013)) et LinUCB (Chu *et al.* (2011)) ont montré l'intérêt mais aussi les limites de ces approches, dont en particulier

- un temps de latence, c'est-à-dire le délai nécessaire à l'algorithme pour allouer un item à une variation, trop important,

- et un déficit d'explicabilité des choix faits par l'algorithme

qui limitent leur utilisation concrète. Des méthodes plus récentes proposent donc d'utiliser des bandits non contextuels indépendants sur des groupes pré-définis. Le premier objectif des travaux de recherche d'Emmanuelle Claeys a été de proposer une méthode d'allocation dynamique informée basée sur une explicitation préalable des groupes autorisant une forte explicabilité des allocations effectuées. Ces travaux ont été présentés oralement à la conférence internationale useR! du 09 au 12 juillet 2019, à Toulouse, en France (Claeys *et al.*, 2019). Un *package* de R qui contiendra l'essentiel de l'algorithme développé dans la thèse d'Emmanuelle Claeys est en train d'être finalisé à l'heure où j'écris ce manuscrit.

4.2.4 Stationnarité, temporalité et test A/B

Par ailleurs, il apparaît de plus en plus indispensable d'étudier l'influence du temps sur les tests, et ce, à différents niveaux. En effet, les méthodes actuelles sont souvent amenées à faire des hypothèses de stationnarité sur les distributions réelles des gains associées aux différentes variations (par exemple, la variation optimale est toujours la même), sur les caractéristiques globales de l'ensemble des items (par exemple, la proportion de femmes et d'hommes est invariant) ou encore sur les caractéristiques des items eux-mêmes. Malheureusement, ces hypothèses s'avèrent de plus en plus contraignantes et limitent leur applicabilité sur des tests A/B. Il est donc indispensable à la fois d'en étudier l'impact sur les tests et proposer de nouvelles méthodes autorisant un relâchement de telles contraintes de stationnarité.

Parallèlement, les informations temporelles liées aux comportements des items avant et pendant le test ne sont actuellement pas prises en compte lors de l'allocation d'un item à une variation. Or, l'utilisation d'informations temporelles comme par exemple, la liste des sites visités par un nouveau visiteur avant d'arriver sur le site hébergeant le test ou celle des pages parcourues et les options cliquées sur celui-ci avant d'arriver à la page testée devrait, à notre avis, améliorer fortement les mécanismes d'allocation.

Si l'étude de nouvelles méthodes des tests A/B permettant de relâcher les contraintes de stationnarité n'a pas pu malheureusement être pas abordée concrètement dans cette thèse, un deuxième objectif de celle-ci a été d'étudier et de proposer une méthode d'allocation permettant de prendre en compte des informations temporelles sur les items.

4.2.5 Nos contributions

L'objectif global des travaux de nos recherches est de proposer une méthode générique d'un test A/B permettant une allocation dynamique en temps réel capable de prendre en compte les caractéristiques des items lorsque les caractéristiques sont définis comme un vecteur de taille finie et strictement identique pour tous les items qu'ils soient temporels ou non.

La première contribution porte sur la prise en compte des caractéristiques d'items lors de la phase d'exploitation. L'idée originale est de postuler qu'il existe dans la population des items des sous-groupes présentant chacun une sensibilité différente aux variations et donc que l'allocation dynamique doit se faire différemment dans chacun de ces sous-groupes. La méthode Ctree-Ucb, que nous avons proposée dans Claeys *et al.* (2017), est constituée de deux phases. La première,

dite *d'explicitation des groupes*, effectuée en pré-traitement (hors ligne) consiste à créer un *modèle de classement* des items en groupes distincts. Ce modèle est appris à partir des caractéristiques d'items ayant été affectés à la variation originale avant le test et de leurs gains obtenus sur cette variation originale. La deuxième phase, correspondant à la phase d'exploration elle-même, consiste à associer à chacun des groupes un bandit non contextuel chargé de l'allocation dynamique des nouveaux items qui auront été classés par le modèle dans ce groupe (Claeys *et al.*, 2018). Ainsi, un nouvel item sera classé dans un groupe en fonction d'un modèle prédéfini et de ses propres caractéristiques, puis affecté à une variation par le bandit associé à ce groupe. Les expériences que nous avons menées ont montré que cette méthode, en plus de déboucher sur une phase d'exploration raccourcie par rapport aux autres méthodes, tout en facilitant l'explication a posteriori des choix, permettait une allocation dynamique en temps réel.

La deuxième contribution est une extension de la méthode précédente aux caractéristiques temporelles Claeys *et al.* (2020a). Il s'agit de mieux préciser les modèles de classement en intégrant la possibilité de prise en compte de données temporelles. Les méthodes DBA-Ctree-Ucb et DBA-LinUcb permettent ainsi d'intégrer à la fois des caractéristiques temporelles historisées (c'est-à-dire liées aux comportements des items avant le test) et dynamiques (c'est-à-dire par exemple, saisie à la volée lors de la navigation d'un visiteur sur le site de test). Nos expériences ont montré que détecter un contexte évolutif et adapter le test en fonction de celui-ci offre un intérêt pour la problématique générale de l'A/B test, et plus particulièrement pour la mise en forme de pages WEB d'e-commerce.
Les contributions sont de différentes natures. Elles se veulent à la fois :

- génériques en proposant des solutions applicables à tout domaine,
- théoriques en validant mathématiquement les propositions faites et
- applicatives en les mettant concrètement en œuvre.

En effet, la thèse d'Emmanuelle Claeys (Claeys, 2019b))a été réalisée en partenariat avec la société AB Tasty. Comme nous l'avons déjà mentionné, notre contribution repose à la fois sur des garanties théoriques par des propriétés statistiques, mais également par des résultats empiriques issus de jeux de données réelles publics ou provenant pour la majorité de cette société. Ainsi, afin de valider cette approche, des expérimentations ont été menées à la fois sur des données dites de *benchmarks* et sur des données issues de tests réels menés par AB Tasty. Actuellement, la société AB Tasty utilise nos développements théoriques et nos résultats empiriques pour développer de nouveaux produits que le service marketing souhaite proposer à leurs clients.

4.3 Électricité de Strasbourg

Électricité de Strasbourg, connue au niveau de la région Grand Est sous l'abréviation ÉS, développe pour le compte de sa filiale de commercialisation d'énergie (ÉS Energies) des équipements (indépendants de l'installation de comptage) et des services de suivi en temps réel de la consommation d'électricité. Ces services sont actuellement en test chez une communauté d'une centaine de clients « grand-public » et professionnels qui est animée par l'équipe de développement de services digitaux d'Électricité de Strasbourg. Dans le cadre des retours d'expérience remontés par cette communauté est apparu le besoin de services d'analyse et de prévision de consommation individualisés. Pour répondre à cette demande, il a été décidé de tester le développement de modèles de prévision à l'échelle individuelle afin de proposer toute une gamme de diagnostics basés sur la comparaison entre la consommation réelle d'électricité avec un modèle reflétant l'historique de consommation propre à chacun des clients.

De nombreux travaux de recherche ont été menés pour permettre aux différents acteurs du secteur commercial ou du secteur marketing de réaliser des prévisions fiables de la demande d'électricité. Ces travaux ont produit des modèles à haute performance pour des agrégats de demande totalisant à minima plusieurs milliers de clients, le plus souvent à l'échelle d'un pays ou d'une région. Ces modèles ont été testés par les équipes du département R&D d'Électricité de Strasbourg et ne sont malheureusement pas adaptés pour la prévision de consommation à l'échelle individuelle. Suite à ce constat, l'ÉS, et en particulier Daniel Wagner (Directeur Digital, Informatique et Infrastructures), m'a contacté afin de mettre en place un sujet de thèse sur cette dernière problématique.

Le projet de doctorat a pour principal objectif d'investiguer et de déployer des modèles statistiques ou à base de techniques d'intelligence artificielle qui permettraient de déterminer, client par client, la consommation électrique à court terme ($j+1$ à $j+3$) à partir des historiques de consommation d'électricité de chacun d'entre eux, des données météorologiques et d'autres paramètres qualitatifs.
Cette approche, que nous avons validée à l'occasion de travaux préliminaires réalisés à l'occasion d'un stage de deuxième année du Master CSMI de l'Université de Strasbourg, doit permettre de procurer à chaque client une prévision à court terme de sa propre consommation, de détecter et de lui communiquer d'éventuelles anomalies telles que des écarts anormaux de la consommation mesurée par rapport à la consommation prévue sur la base des historiques de consommation d'électricité et des conditions météorologiques.

Avec de tels modèles il est également prévu de développer et de tester des algorithmes de détection d'anomalies dans les courbes de charges individuelles des clients des secteurs résidentiel et tertiaire afin de développer des services à valeur ajoutée à leur destination. On peut par exemple citer la détection « passive » de pannes sur des équipements de l'utilisateur final par recherche de sous-consommation ou de sur-consommation (par exemple au niveau de chambres froides, de systèmes de chauffage-ventilation, de chauffe-eau ou de pompes dans le secteur tertiaire, artisanal ou chez les collectivités locales).

Plus précisément, les travaux envisagés dans le cadre du doctorat porteront entre autres à développer et à tester à grande échelle (plusieurs dizaines de milliers de sites de consommation) des modèles adaptés à chaque client de sa propre prévision de consommation d'électricité et de comparer leurs performances respectives en termes de qualité de prévision. Différentes approches seront étudiées notamment celles à base de *deep learning* qui seront comparées aux méthodes plus classiques telles que les méthodes à base de séries auto-régressives. Par ailleurs, compte tenu du volume de données et de l'objectif final recherché, qui est de pouvoir proposer à l'utilisateur final des recalculs en temps réel des paramètres du modèle lui étant attaché, il est prévu d'adapter et de tester certaines implémentations sur les infrastructures de calcul à haute performance dont dispose l'Université de Strasbourg ainsi que sur les fermes de *GPGPU* dont l'Électricité de Strasbourg fera prochainement l'acquisition dans le cadre de ses investissements dans le *big data*.

Durant son stage de deuxième année de Master CSMI effectué au sein de l'Électricité de Strasbourg sur le site du centre ville de Strasbourg, Fatima Fahs a appliqué un modèle *SARIMAX* à la base de données de consommation d'électricité des particuliers, relevée toutes les demi-heures. Ce dernier a donné des résultats très satisfaisants en terme de prédiction parmi plusieurs modèles issus des séries temporelles et du *machine learning* testés au préalable. En revanche, le temps d'estimation des paramètres de ce modèle dans la phase d'entrainement restait assez couteux (une demi-heure sur une machine bureautique) et non négligeable. Ainsi pendant la première année de la thèse de Fatima Fahs, nous nous sommes concentrés sur la recherche de méthodes alternatives de prévision et de leurs éventuelles possibilités d'applications dans le domaine de l'électricité.

Nous avons donc appliqué des modèles non paramétriques plus rapides que ceux testés durant le stage de deuxième année de Master. La prévision des processus à valeurs fonctionnelles est l'une des approches non paramétriques qui permet de décomposer des séries chronologiques fonctionnelles en processus continus à valeurs fonctionnelles et ainsi de pouvoir prédire le futur processus à partir des processus passés.

Nous avons testé un modèle de prévision non paramétrique *KWF* (*Kernel Wavellite Function*) sur des données de consommation pour des ménages thermo-sensibles et d'autres moins thermo-sensibles. Les premiers résultats obtenus montrent que le modèle *KWF* est plus performant que le modèle *SARIMAX* par rapport à la précision de la prévision et le temps d'entrainement du modèle (deux minutes) chez certains profils de clients en particulier chez les clients non thermo-sensibles.

Nous avons également travaillé sur la classification intra journalière des courbes de charge dans le but de chercher des jours ayant des comportements énergétiques similaires pour pouvoir améliorer la qualité de la prévision du modèle. En effet, nous avions mis en évidence une incapacité du modèle à distinguer les jours de la semaine pour une application de la méthode *KWF* (figures 4.1 et 4.2). La figure 4.3 représente les distributions de l'erreur de prévision sur les jours de la semaine pour le modèle *KWF* sans groupes et *KWF* avec groupes de jours.

Le modèle a aussi été adapté pour fonctionner à la fois chez les clients particuliers et les entreprises. Des techniques de calcul parallèle ont été utilisées afin de répondre aux besoins de prévisions quotidiens de l'entreprise.

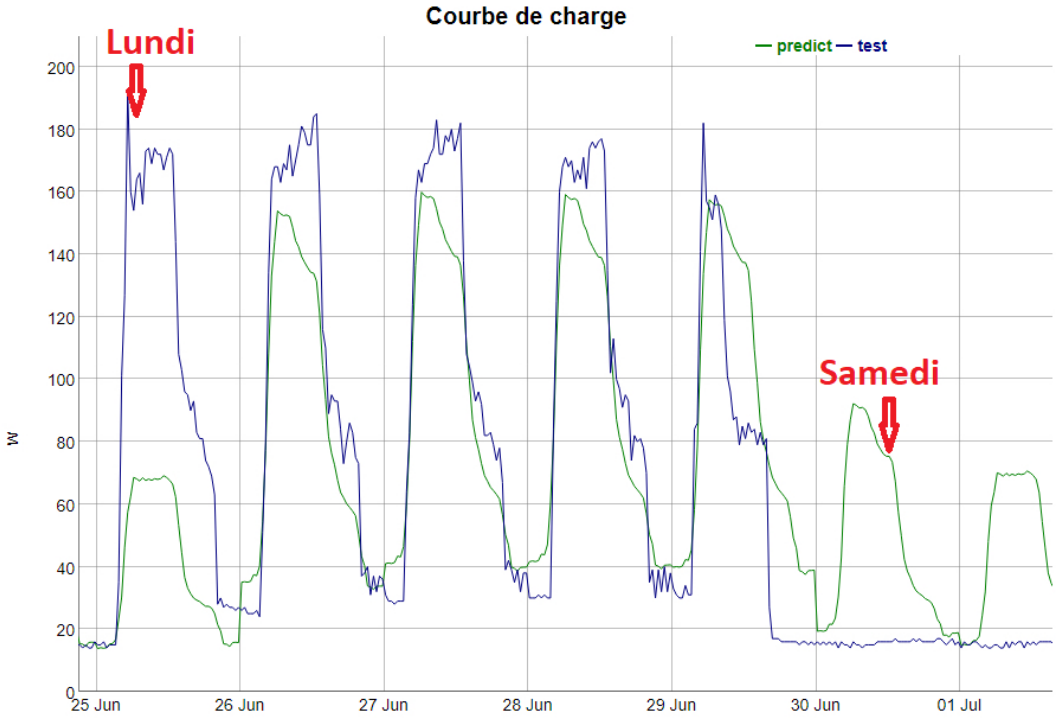


Figure 4.1 : *Incapacité à distinguer les jours de la semaine pour une application de la méthode KWF pour une entreprise B.*

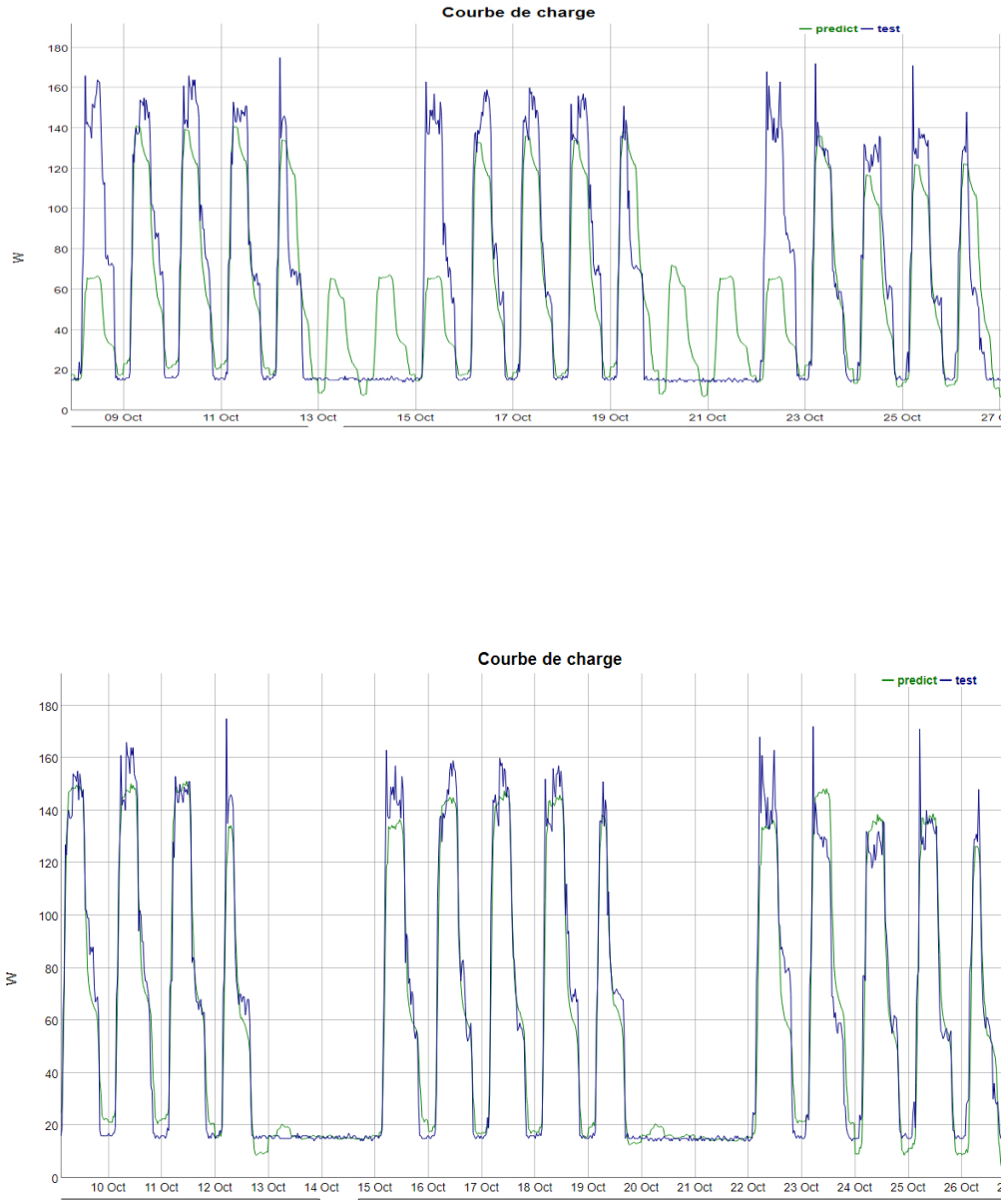


Figure 4.2 : *Pour un même client (entreprise B), prévisions par le modèle KWF de base (haut) et par le modèle KWF avec groupes de jours (bas).*

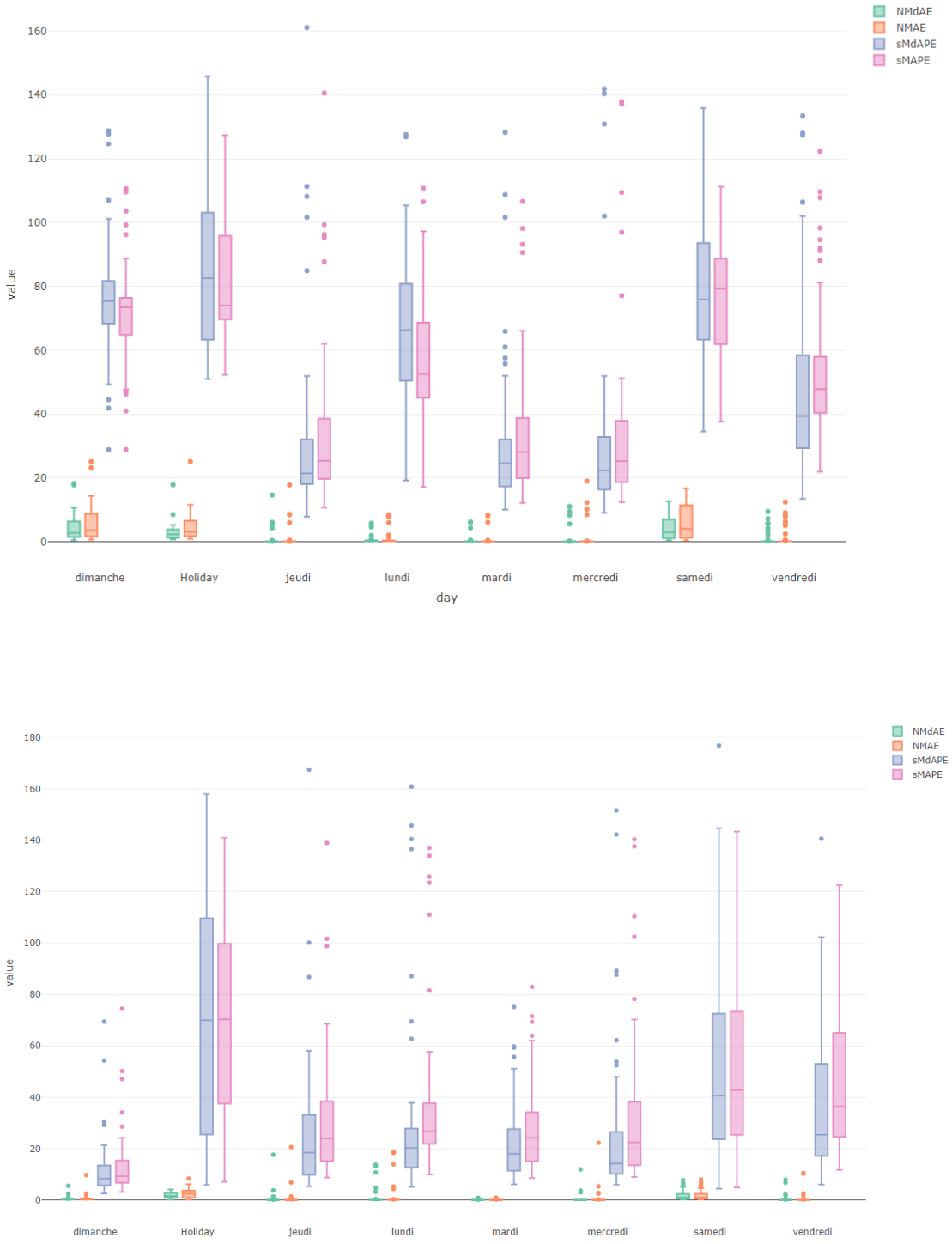


Figure 4.3 : *Distributions de l'erreur de prévision sur les jours de la semaine pour le modèle KWF sans groupes (haut) et KWF avec groupes de jours (bas).*

La méthode de Steinley et Brusco (2008), appliquée à une transformée par ondelettes des données journalières à regrouper, a permis de mettre en évidence plusieurs caractéristiques dont les deux suivantes :

- les échelles significatives pour révéler la structure des groupes sont indépendantes du nombre de groupes utilisés dans l'algorithme de sélection des *features*.

- Les échelles significatives sélectionnées par l'algorithme sont très différentes d'un client à un autre. Chez certains clients les échelles significatives sont celles associées aux moyennes fréquences (figure 4.4). Chez d'autres, ce sont celles associées aux basses fréquences (qui représentent des fréquences qui varient lentement) (figure 4.5) ou celles qui capturent l'activité à très haute fréquence (qui représentent généralement du bruit) (figure 4.6).

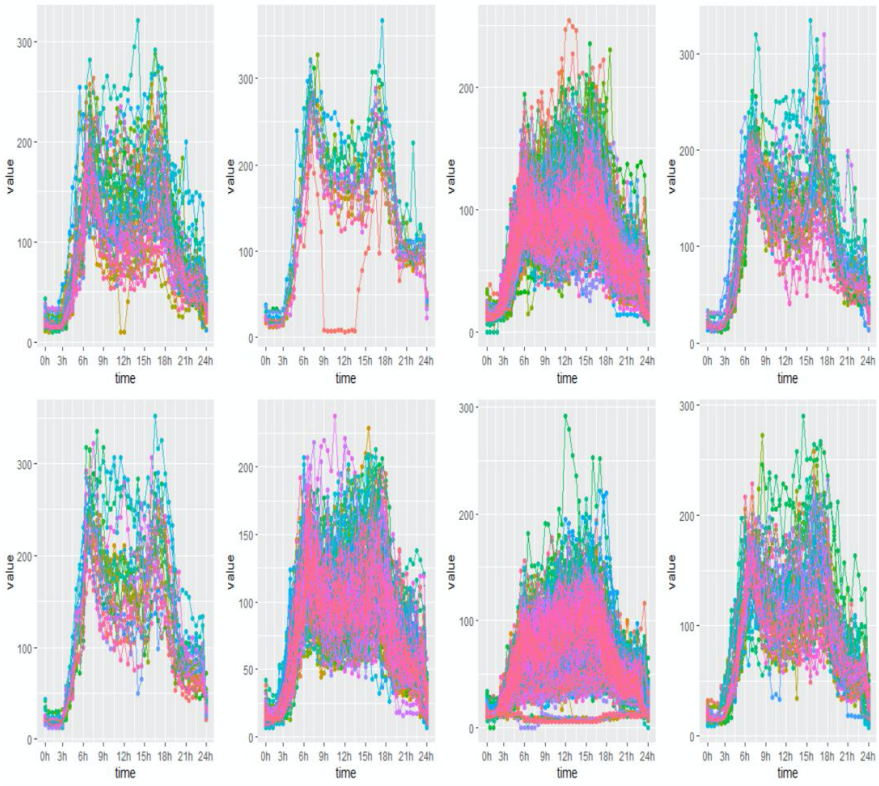


Figure 4.4 : *Classification intra-journalière de la courbe de charge d'une entreprise C selon les moyennes fréquences (l'échelle $j = 3$).*

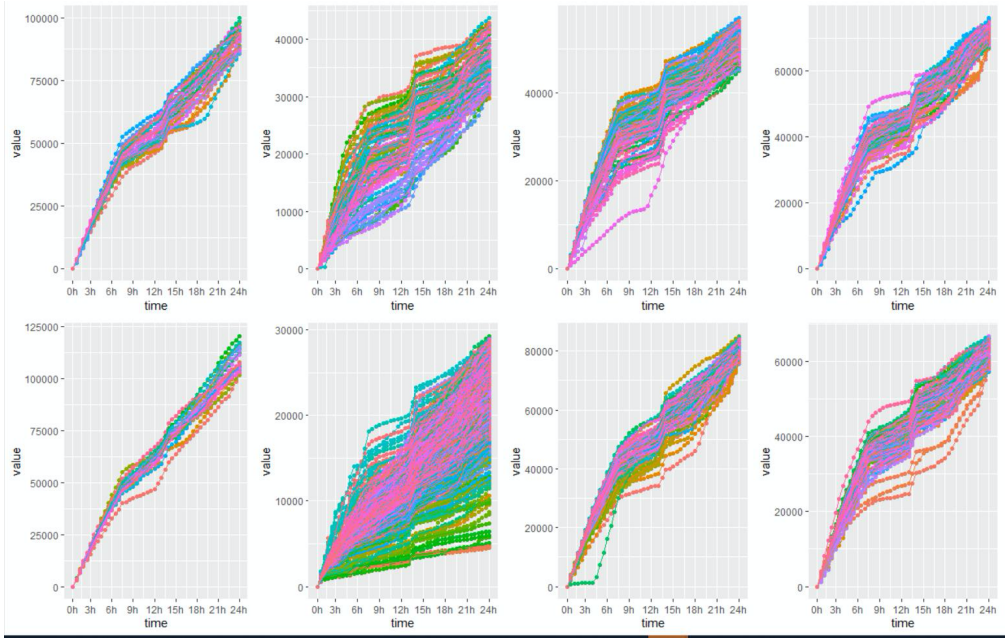


Figure 4.5 : *Classification intra-journalière de la courbe de charge d'une entreprise B selon les basses fréquences* (*l'échelle* $j = 6$).

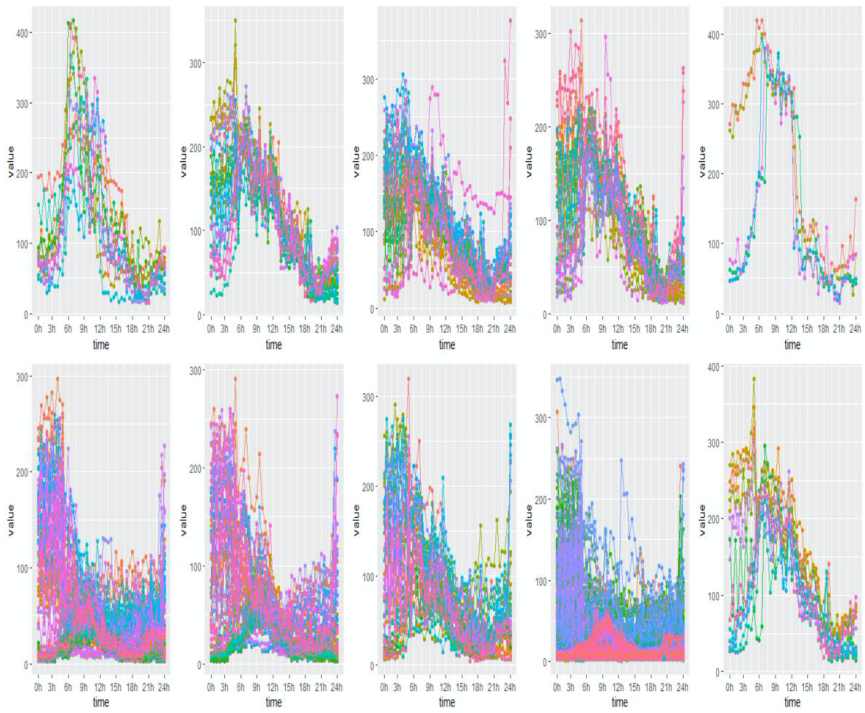


Figure 4.6 : *Classification intra-journalière de la courbe de charge d'une entreprise A selon les hautes fréquences* (*l'échelle* $j = 1$).

4.4 Bürkert

Bürkert, leader allemand de la gestion des systèmes de régulation de fluides, a développé un système de débitmètre qui leur est propre : le système *FLOWave*. En effet, pour les systèmes de mesure du débit, les facteurs cruciaux sont le fonctionnement sans défaillance et sans erreur de mesure même dans les conditions les plus difficiles. La mesure exacte du débit des fluides de nettoyage et de stérilisation en place est également déterminante. Par le passé, il était nécessaire d'utiliser des procédés (relativement coûteux) avec des propriétés aseptiques peu optimales afin de satisfaire aux exigences strictes de précision de mesure. Le nouveau débitmètre *FLOWave* a révolutionné les habitudes de travail. En effet, pour la mesure du débit, les experts en fluidique de Bürkert ont mis au point une technologie basée sur les ondes acoustiques de surface. Cette technologie mesure la propagation des ondes de surface sur le tube de mesure et dans le fluide. L'effet est comparable à la propagation des ondes dans les activités sismiques. Les clients ont rapidement été convaincu par la polyvalence illimitée et les avantages du FLOWave. La surveillance fiable et reproductible des processus NEP et SEP a permis d'identifier les potentiels d'optimisation dans un délai très court. L'efficacité du nettoyage et de la stérilisation, mais aussi l'efficacité globale du système ont été renforcées grâce à la mesure exacte du débit et de la température, ainsi qu'à sa conception simplifiée. Le débitmètre *FLOWave* est donc non seulement un système de mesure mais aussi un outil d'assurance qualité et d'optimisation.

Ce projet consiste en la réalisation d'un nouveau banc de test avec un double objectif : permettre à Bürkert de qualifier les nouveaux développements pour le débitmètre *FLOWave* et calibrer les produits. Il est porté par Bürkert en partenariat avec la plateforme Cemosis (IRMA, Université de Strasbourg et CNRS). Il a tout d'abord été soutenu, en décembre 2018, par un PEPS1 de l'AMIES d'un montant de 25 k€ puis est actuellement soutenu par la région Grand Est, de mars 2019 à mars 2021 pour un montant de 70k€, dans le cadre du pôle de compétitivité Hydreos. Là encore, j'ai encore pu constater l'efficacité du dispositif PEPS1 de l'AMIES qui a permis un démarrage rapide du projet et la rédaction du dossier de demande de financement auprès de la région Grand Est. L'objectif ambitieux de cette installation est d'atteindre des niveaux de précision très importants (< 0.1 %) avec une répétabilité certaine, le tout de façon rapide et automatisée pour une gamme vaste de produits.

La partie exploitation des données est également un axe important du projet qui pourra nécessiter l'intégration d'une approche type analyse *big data*. L'objectif étant d'améliorer les processus de développement et de production à la lumière

de ces analyses de données. De plus l'installation sera certifiée COFRAC/ISO17025 ce qui permettra d'ouvrir le laboratoire à d'autres sociétés.

Plus précisément, un grand nombre de données sont collectées durant le processus de fabrication et de validation de chaque nouveau débitmètre. Des mesures géométriques très précises sont réalisées après l'usinage du support des capteurs. Des mesures d'impédance sont ensuite réalisées à plusieurs instants du processus d'assemblage. Enfin, chaque débitmètre est étalonné et sa précision est validée sur un banc de test. Les objectifs à court terme du projet sont de réaliser une analyse statistique exploratoire des données collectées à l'aide du système de banc de test déjà existant. Cette analyse permettra, dans un premier temps, de déterminer si les mesures réalisées actuellement suffisent pour mettre en place un système qui permettrait d'alerter si la fabrication des pièces est potentiellement non-conforme à l'issue du processus de fabrication ou si, au contraire il faut compléter ces mesures par d'autres types de mesures. Ce point essentiel pourrait alors aboutir à la conception d'un nouveau banc de test qui permettrait de certifier une meilleure précision des débitmètres.

D'un point de vue statistique, le banc de test est utilisé à deux reprises pour réaliser des observations longitudinales multivariées des pièces produites par Bürkert. En effet, le banc de test est utilisé de manière répétée pour réaliser des mesures intermédiaires lors du processus de fabrication d'une même pièce. Nous tirerons les conclusions nécessaires et utiles de ce suivi longitudinal pour estimer les variabilités inter et intra pièces, ce qui nous sera essentiel pour déterminer des couples capteurs/valeurs permettant de déclencher des alertes prédictives de problèmes de conformités de certaines pièces en cours de fabrication. Comme le banc de test est monté à la main par les équipes de Bürkert, ceci induit une source de variabilité qu'il faudra également chercher à estimer.

De premiers résultats ont été exposés au congrès SMAI en mai 2019 à Guidel Plages (Morbihan) par Jean-Baptiste Wahl qui était, à l'époque, ingénieur de recherche, spécialité science des données, de la plateforme Cemosis (Wahl *et al.*, 2019). Jean-Baptiste Wahl nous a quitté en janvier 2020 pour rejoindre le groupe GE Healthcare. Depuis le 1er avril 2020, Yannick Stoll, également ingénieur de recherche en science des données de la plateforme Cemosis a repris le projet sous ma direction. Ce projet se terminera en mars prochain.

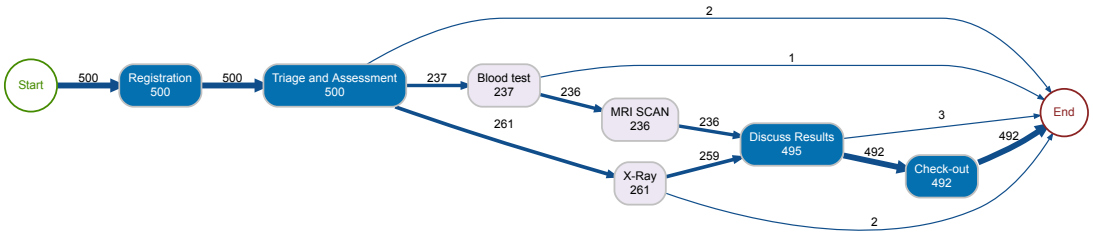


Figure 4.7 : *Visualisation d'un exemple de process.*

4.5 Your Data Consulting

Ce projet s'inscrit dans le cadre d'une collaboration de recherche entre la société *Your Data Consulting*, jeune entreprise innovante, et deux enseignants-chercheurs à savoir Frédéric Bertrand et moi-même, sur la thématique de la science des données et plus particulièrement de l'intelligence artificielle.
Your Data Consulting est propriétaire de la plateforme *SaaS LiveJourney* qui est présentée à l'adresse suivante : `https://www.livejourney.com` Cette plateforme permet aux entreprises de vente par correspondance ou aux entreprises de production ou celles de livraison de visualiser et d'analyser de façon dynamique les parcours de leurs clients, ou de leurs produits ou de leurs colis. De plus, elle permet aux entreprises d'anticiper des événements qui pourraient ralentir le bon déroulement du processus et de détecter les goulots d'étranglement. Un exemple de *process* est représenté à la Figure 4.7. Ce projet a fait l'objet de plusieurs soutiens de l'AMIES (PEPS1-IA 8k€ et PEPS2 50k€) et de plusieurs contrats de collaboration industriels. Voici plusieurs thématiques qui ont été sélectionnées pour être particularisées au contexte du *process mining*. lors de ces travaux

- *root cause analysis*,
- prédiction des processus,
- *clustering* de séries temporelles,
- causalité et intelligence artificielle.

D'une part, depuis le 09 décembre 2019, Emmanuelle Claeys travaille sur ce projet en tant que chercheuse post-doctorante. À l'heure où j'écris ce manuscrit, Emmanuelle Claeys a obtenu un poste de MCF à l'Université de Toulouse en 27ème section. D'autre part, Yoann Valero travaille aussi à ce projet en tant qu'ingénieur d'études en attendant une réponse à une demande de financement de thèse CIFRE qui a été déposée au début du mois de mars 2020.

Chapitre 5

Contributions à la régression pénalisée

5.1 Ajustement multidimensionnel

5.1.1 Contexte

Les tableaux de données multidimensionnels apparaissent naturellement dans de nombreux domaines scientifiques. Par exemple nous pouvons citer la biologie avec l'étude des expressions des gènes (Cheung (2012), Golub *et al.* (1999)), la géographie avec l'analyse des données spatiales des séismes (van der Hilst *et al.* (2007)), l'étude des marchés financiers et en particulier la constitution et l'évaluation de portefeuilles (Jagannathan et Ma (2003)).
La complexité des jeux de données réels est, la plupart du temps, telle qu'il n'est pas possible de réduire l'étude des variables à celle d'une seule. Ainsi une analyse multivariée globale des données (Mardia *et al.* (1979)) est généralement nécessaire, voir les chapitres 3, 6 pour des exemples d'application de certaines de ces techniques.

5.1.2 Ajustement multidimensionnel déterministe

La méthode d'ajustement multidimensionnel (*Multidimensional fitting* que j'ai introduite à l'époque avec Nicolas Wicker, Professeur des Universités à Lille 1 et Claude Berge (Berge *et al.* (2010)) est une méthode d'analyse multivariée de données récemment mise au point et basée sur un ajustement pénalisé de matrices de distance.

Imaginons que nous disposions de deux matrices observées $\mathbf{X}$ et D : la première, $\mathbf{X}$, contient les coordonnées des individus et la seconde, D, des distances entre ces individus. À partir de la matrice X, il est également possible de calculer des distances, D_X, entre les individus. L'originalité de l'ajustement multidimensionnel est de proposer des vecteurs de modification des coordonnées des individus, donc la matrice $\mathbf{X}$, afin de faire se rapprocher les distances, D_X, calculées sur ces coordonnées modifiées des distances initialement présentes dans la matrice D.

Ces travaux ont été repris et appliqués à des jeux de données réels dans Alawieh *et al.* (2017).

5.1.3 Ajustement multidimensionnel stochastique

J'ai étendu cette méthode avec Hiba Alawieh (qui était à l'époque en thèse en co-direction, Nicolas Wicker et Baydaa Al Ayoubi, Professeur à l'Université libanaise de Beyrut) dans Alawieh *et al.* (2018), du cas d'un modèle déterministe pour les vecteurs de déplacement à celui d'un modèle stochastique. En effet, il était nécessaire d'ajouter la possibilité de modéliser des effets aléatoires qui pourraient se produire durant le processus de modification et avoir un effet sur le résultat de celui-ci. Nous avons introduit un modèle stochastique pour l'ajustement multidimensionnel afin de trouver des vecteurs de déplacement optimaux dans un contexte bruité. La fonction objectif n'étant plus déterministe, il existe de multiples critères permettant d'évaluer la pertinence des vecteurs de modification du *Multidimensional fitting* et nous avons proposé deux approches permettant de trouver une solution dans le cas de l'écart quadratique moyen. La première approche conduit à un problème d'optimisation déterministe et repose sur un cadre gaussien. Le cas indépendant et identiquement distribué est présenté dans Alawieh *et al.* (2018). Même si elle n'est pas présentée dans Alawieh *et al.* (2018) pour des raisons de concision, nous avons aussi envisagé une approche gaussienne tenant compte de corrélations éventuelles entre les variables qui décrivent les individus dans la matrice $\mathbf{X}$ et qui repose sur des développements mathématiques plus raffinés impliquant l'utilisation de lois de Wishart non centrales (Anderson (1946)) et le calcul de certains de leurs moments (Letac et Massam (2004)). La deuxième approche est une optimisation stochastique qui repose sur l'utilisation d'un algorithme de Metropolis-Hastings (Metropolis *et al.* (1953)). Elle convient à des contextes non gaussiens ou de corrélation entre les variables explicatives mais est plus longue à mettre en œuvre. L'article Alawieh *et al.* (2018) contient cette extension aléatoire du *Multidimensional fitting*, ainsi qu'une application dans le domaine de la sensométrie.

5.2 Sélection de variables avec confiance

Avec l'émergence de technologies à haut débit, il est possible de mesurer énormément de variables à un coût relativement faible. De telles situations existent dans de nombreux domaines, des sciences expérimentales aux sciences humaines, et la sélection de variables peut être très utile pour répondre aux défis spécifiques à chacune d'elles. La sélection de variables peut permettre de connaître, parmi toutes les variables relevées, celles qui sont essentielles à la compréhension du phénomène ou pas.
J'ai du utiliser des techniques de sélection de variables pendant que j'encadrais la thèse de Nicolas Jung qui portait sur l'inférence de réseaux de gènes. Une des raisons initiales qui a motivé le travail de la création de l'algorithme *SelectBoost* est le problème exposé au chapitre 2 dans le paragraphe 2.5 sur l'inférence des réseaux biologiques et plus particulièrement la détermination de cibles pour une intervention dirigée dans les programmes géniques des cellules cancéreuses.
Une intervention dirigée dans un réseau de gènes en cascade consiste à agir sur certains gènes en amont de la cascade dans le but d'obtenir l'effet souhaité sur d'autres gènes situés plus en aval. Notre problème est non seulement de trouver ces gènes sur lesquels agir mais aussi de choisir parmi eux, ceux pour en lesquels nous avons la plus grande confiance pour obtenir la modification souhaitée.
De nombreuses méthodes ont été proposées pour traiter ce problème, le *lasso* et d'autres régressions pénalisées en étant des cas particuliers. Ces méthodes échouent dans certains cas et la corrélation linéaire entre les variables explicatives est la plus courante des situations pour laquelle cela se produit. Or, celle-ci apparaît naturellement dans les grands ensembles de données. Dans Jung *et al.* (2015), nous avons présenté un algorithme capable d'améliorer la précision de toute méthode de sélection de variables.
J'ai alors proposé une extension de cet algorithme pour le cas des modèles linéaires généralisés. Avec l'aide de Frédéric Bertrand, nous avons également repris et repensé intégralement le code qui avait été développé initialement afin d'améliorer ses performances et de proposer un *package* `R` le contenant, Bertrand *et al.* (2020). Avec l'aide d'Ismaïl Aouadi, ingénieur d'étude, nous avons appliqué ce nouvel algorithme, appelé *SelectBoost*, à différents jeux de données dont des problèmes de régression logistique binaire pénalisée (typiquement du *lasso*). Puis, nous avons récemment écrit une révision majeure de la première prépublication, Aouadi *et al.* (2018). Il s'agit d'une collaboration avec le laboratoire d'Immuno Rhumatologie Moléculaire, Unité Inserm 1109, LabEx TRANSPLANTEX, Centre de Recherche d'Immunologie et d'Hématologie, Fédération de Médecine Translationnelle de Strasbourg, le Laboratoire International Associé INSERM, Strasbourg (France) - Nagano (Japon) et l'ICD, ROSAS, M2S, de l'Université de Technologie de Troyes.

5.3 SelectBoost: a general algorithm to enhance the performance of variable selection methods

Motivation: with the growth of big data, variable selection has become one of the critical challenges in statistics. Although many methods have been proposed in the literature their performance in terms of recall (sensitivity) and precision (PPV) is limited in a context where the number of variables by far exceeds the number of observations or in a highly correlated setting.
Results: in this article, we propose a general algorithm which improves the precision of any existing variable selection method. This algorithm is based on highly intensive simulations and takes into account the correlation structure of the data. Our algorithm can either produce a confidence index for variable selection or be used in an experimental design planning perspective. We demonstrate the performance of our algorithm on both simulated and real data. We then apply it in two different ways to improve biological network reverse-engineering.
Availability: code is available as the `SelectBoost` package on the CRAN, `https://cran.r-project.org/package=SelectBoost`.
Some network reverse-engineering functionalities are available in the `Patterns` CRAN package, `https://cran.r-project.org/package=Patterns`
online.

5.3.1 Introduction

Technological innovations make it possible to measure large amounts of data in a single observation. As a consequence, problems in which the number P of variables is larger than the number N of observations have become common. As reviewed by Fan and Li (Fan et Li, 2006), such situations arise in many fields from fundamental sciences to social science, and variable selection is required to tackle these issues. For example, in biology/medicine, thousands of messenger RNA (mRNA) expressions (Lipshutz *et al.*, 1999) may be potential predictors of some disease. Moreover, in such studies, the correlation between variables is often very strong (Segal *et al.*, 2003), and variable selection methods often fail to make the distinction between the informative variables and those which are not. Similarly, inference of gene regulatory networks (GRNs) from perturbation data can enhance the insights of a biological system, (Morgan *et al.*, 2019). In this paper, we propose a general algorithm that enhances model selection in correlated variables.

First, we will assume a statistical model with a response variable $\mathbf{y} = (y_1, ..., y_N)'$ (with the symbol "'" as the transposed), a variable matrix of size $N \times P$, $\mathbf{X} =$

$(\mathbf{x}_{1.}, ..., \mathbf{x}_{P.})$ and a vector of parameters $\boldsymbol{\beta} = (\beta_1, ..., \beta_P)'$. Then, we will assume that the vector of parameters $\boldsymbol{\beta} = (\beta_1, ..., \beta_P)'$ is sparse. In other words, we will assume that $\beta_i = 0$ except for a quite small proportion of elements of the vector. We note $\mathcal{S}$ as the set of indices for which $\beta_i \neq 0$ and $q < \infty$ is the cardinality of this set $\mathcal{S}$. Without any loss of generality, we will assume that $\beta_p \neq 0$ if and only if $p \leqslant q$.

When dealing with a problem of variable selection, one of the goals is the estimation of the support, in which you want $\mathbb{P}(\mathcal{S} = \hat{\mathcal{S}})$ to be close to one, with $\hat{\mathcal{S}} = \{k : \hat{\beta}_k \neq 0\}$. Here, our interest is mainly as follows, *i.e.* in identifying the correct support $\mathcal{S}$. This kind of issue arises in many fields, for example in biology, where it is of greatest interest to discover which specific molecules are involved in a disease (Fan et Li, 2006).

There is a vast literature dealing with the problem of variable selection in both statistical and machine learning areas (Fan et Li, 2006; Fan et Lv, 2010). The main variable selection methods can be gathered in the common framework of penalized likelihood. The estimate $\hat{\boldsymbol{\beta}}$ is then given by:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^P}{\arg\min} \left[-\ell_N(\boldsymbol{\beta}) + \sum_{p=1}^{P} \underset{\lambda}{\Omega}(\beta_p) \right], \tag{5.1}$$

where $\ell_N(.)$ is the log-likelihood function, $\Omega_\lambda(.)$ is a penalty function and $\lambda \in \mathbb{R}$ is the regularization parameter. As the goal is to obtain a sparse estimation of the vector of parameters $\boldsymbol{\beta}$, a natural choice for the penalty function is to use the so-called $\mathcal{L}_0$ norm ($\|.\|_0$) which corresponds to the number of non-vanishing elements of a vector:

$$\begin{array}{rcccl} \Omega_\lambda & : & \mathbb{R} & \mapsto & \{0, \lambda\} \\ & & x & \mapsto & \left\{ \begin{array}{ll} \Omega_\lambda(x) = \lambda & \text{if } x \neq 0 \\ \Omega_\lambda(x) = 0 & \text{else} \end{array} \right. \end{array} \tag{5.2}$$

which induces $\sum_{p=1}^{P} \Omega_\lambda(\beta_p) = \lambda \|\boldsymbol{\beta}\|_0$. For example, when $\lambda = 1$, we get the Akaike Information Criterion (AIC) (Akaike et Akaïke, 1974) and when $\lambda = \frac{\log(N)}{2}$ we get the Bayesian Information Criterion (BIC) (Schwarz, 1978).

Many different penalties can be found in the literature. Solving this problem with $\|.\|_0$ as part of the penalty is an NP-hard problem (Natarajan, 1995; Fan et Lv, 2010). It cannot be used in practice when P becomes large, even when it is employed with some search strategy like forward regression, stepwise regression (Hocking, 1976), genetic algorithms (Koza *et al.*, 1999). Donoho and

Elad (Donoho et Elad, 2003) showed that relaxing $\|.\|_0$ to norm $\|.\|_1$ ends, under some assumptions, to the same estimation. This result encourages the use of a wide range of penalties based on different norms. For example, the case where $\Omega_\lambda(\beta_p) = \lambda|\beta_p|$ is the lasso estimator (Tibshirani, 1996) (or equivalently Basis Pursuit Denoising (Chen *et al.*, 2001)) whereas $\Omega_\lambda(\beta_p) = \lambda\beta_p^2$ leads to the Ridge estimator (Hoerl et Kennard, 1970). Nevertheless, the penalty term induces variable selection only if:

$$\min_{x \geqslant 0} \left(\frac{\mathrm{d}\,\Omega_\lambda(x)}{\mathrm{d}x} + x \right) > 0. \tag{5.3}$$

Equation (5.3) explains why the lasso regression allows for variable selection, while the Ridge regression does not. The lasso regression is, however, known to lead to a biased estimate (Zou, 2006). The SCAD (Smoothly Clipped Absolute Deviation) (Fan, 1997), MCP (Minimax Concave Penalty) (Zhang, 2010) or adaptive lasso (Zou, 2006) penalties all address this problem. The popularity of such variable selection methods is linked to fast algorithms like LARS (Least-Angle Regression Selection) (Efron *et al.*, 2004), coordinate descent or PLUS (Penalized Linear Unbiased Selection) (Zhang, 2010).

Nevertheless, the goal of identifying the correct support of the regression is complicated and the reason why variable selection methods fail to select the set of non-zero variables $\mathcal{S}$ can be summarized in two words: linear correlation. Choosing the lasso regression as a special case, Zhao and Yu (2006) stated that if an irrelevant predictor is highly correlated with the predictors in the true model, lasso may not be able to distinguish it from the true predictors with any amount of data and any amount of regularization. Zhao and Yu (2006) (and simultaneously Zou (2006)) found an almost necessary and sufficient condition for lasso sign consistency (*i.e.* selecting the non-zero variables with the correct sign). This condition is known as "irrepresentable condition":

$$\left| \mathbf{X}'_{\backslash\mathcal{S}} \mathbf{X}_\mathcal{S} \left(\mathbf{X}'_\mathcal{S} \mathbf{X}_\mathcal{S} \right)^{-1} \mathrm{sgn}(\boldsymbol{\beta}_\mathcal{S}) \right| < \mathbf{1}, \tag{5.4}$$

where $\mathbf{X}_\mathcal{S} = (x_{ij})_{i,j \in \mathcal{S}^2}$, $\mathbf{X}_{\backslash\mathcal{S}} = (x_{ij})_{i,j \neq \mathcal{S}^2}$, $\boldsymbol{\beta}_\mathcal{S} = (\beta_p)_{p \in \mathcal{S}}$. In other words, when $\mathrm{sgn}(\boldsymbol{\beta}_\mathcal{S}) = 1$, this can be seen as the regression of each variable which is not in $\mathcal{S}$ over the variables which are in $\mathcal{S}$. As all variables in the matrix $\mathbf{X}$ are centered, the absolute sum of the regression parameters should be smaller than 1 to satisfy this "irrepresentable condition".

Facing this issue, existing variable selection methods can be split into two categories:

- those which are "regularized" and try to give similar coefficients to correlated variables (*e.g.* elastic net (Zou et Hastie, 2005)),

- those which are not "regularized" and pick up one variable among a set of correlated variables (*e.g.* the lasso (Tibshirani, 1996)).

The former group can further be split into methods in which groups of correlation are known, such as the group lasso (Yuan et Lin, 2006; Friedman *et al.*, 2010a) and those in which groups are not known as in the elastic net (Zou et Hastie, 2005). The latter combines the $\mathcal{L}_1$ and the $\mathcal{L}_2$ norm and takes advantage of both. Non-regularized methods will select some co-variables among a group of correlated variables while regularized methods will select all variables in the same group with similar coefficients.

The main idea of our algorithm is to consider that any observed value of a group of linearly correlated variables of the $\mathbf{X}$ matrix is the independent realization of a given random function. This common random function is then used to perturb the observed values of the relevant correlated variables. Strictly speaking, the use of noise to determine the informative variables is not a new idea. For example, it has been shown that adding random pseudo-variables decreases over-fitting (Wu *et al.*, 2007). In the case where $P > N$ the pseudo-variables are generated either with a standard normal distribution $\mathcal{N}(0, 1)$ or by using permutations on the matrix $\mathbf{X}$ (Wu *et al.*, 2007). Another approach consists of adding noise to the response variable and leads to similar results (Luo *et al.*, 2006). The rationale of this method is based on the work of Cook and Stefanski (Cook et Stefanski, 1994), which introduces the simulation-based algorithm SIMEX (Cook et Stefanski, 1994). Adding noise to the matrix $\mathbf{X}$ has already been used in the context of microarrays (Chen *et al.*, 2007). Simsel (Eklund et Zwanzig, 2012) is an algorithm that both adds noise to variables and uses random pseudo-variables. One new and inspiring approach is stability selection (Meinshausen et Bühlmann, 2010) in which the variable selection method is applied on sub-samples, and informative variables are defined as variables which have a high probability of being selected. Bootstrapping has been applied to the lasso on both the response variable and the matrix $\mathbf{X}$ with better results in the former case (Bach, 2008). A random lasso, in which variables are weighted with random weights, has also been introduced (Wang *et al.*, 2011).

In this article, following the idea of using simulation to enhance the variable selection methods, we propose the SelectBoost algorithm. Unlike other algorithms reviewed above, it takes into account the correlation structure of the data. Furthermore, our algorithm is motivated by the fact that in the case of

non-regularized variable selection methods, if a group contains variables that are highly correlated together, one of them will be chosen with precision.

5.3.2 Methods

The SelectBoost algorithm has been designed in a general framework in order to avoid to select non-predictive correlated features. The main goal is to improve the PPV, *i.e.* the proportion of selected variables which truly belong to $\mathcal{S}$.

Generate new perturbed design matrix

As we assume that the variables are centered and that $\|\mathbf{x}_{p\cdot}\|^2 = 1$ for $p = 1, ..., P$, we know that $\mathbf{x}_{p\cdot} \in \mathcal{S}^{N-2}$. Indeed, the normalization puts the variables on the unit sphere $\mathcal{S}^{N-1}$. The process of centering can be seen as a projection on the hyperplane $\mathcal{H}^{N-1}$ with the unit vector as normal vector. Moreover, the inter-subsection between $\mathcal{H}^{N-1}$ and $\mathcal{S}^{N-1}$ is $\mathcal{S}^{N-2}$. We further define the following isomorphism:

$$\begin{array}{rccl} \phi & : & \mathcal{H}^{N-1} & \rightarrow & \mathbb{R}^{N-1} \\ & & \mathbf{h}_n & \mapsto & \phi(\mathbf{h}_n) = \mathbf{f}_n \quad n = 1, ..., N-1, \end{array} \tag{5.5}$$

where $\{\mathbf{h}_n\}_{n=1,...,N-1}$ is an orthogonal base of $\mathcal{H}^{N-1}$ and $\{\mathbf{f}_n\}_{n=1,...,N-1}$ is the canonical base of $\mathbb{R}^{N-1}$. We define:

$$\mathbf{h}_n = \frac{\sum_{i=1}^{n} \mathbf{e}_i - n\mathbf{e}_{n+1}}{\|\sum_{i=1}^{n} \mathbf{e}_i - n\mathbf{e}_{n+1}\|},$$

with $\{\mathbf{e}_n\}_{n=1,...,N}$ the canonical base of $\mathbb{R}^N$. Note that $\phi(\mathcal{S}^{N-2}) = \mathcal{S}^{N-2}$, and that is why we can work in $\mathbb{R}^{N-1}$ and then return in $\mathbb{R}^N$.

Here, we make the assumption that a group of correlated variables are independent realizations of the same multivariate Gaussian distribution. As the variables are normalized with respect to the $\mathcal{L}_2$ norm, we will use the von Mises-Fisher distribution (Sra, 2012) in $\mathbb{R}^{N-1}$ thanks to the isomorphism ϕ in order to generate new perturbed design matrix. The probability density function of the von Mises-Fisher distribution for the random P-dimensional unit vector $\mathbf{x}$ is given by:

$$f_P(\mathbf{x}; \boldsymbol{\mu}, \kappa) = \widetilde{K}_P(\kappa) \exp\left(\kappa \boldsymbol{\mu}' \mathbf{x}\right),$$

where $\kappa \geqslant 0$, $\boldsymbol{\mu} = (\mu_1, ..., \mu_P)'$, $\|\boldsymbol{\mu}\|_2 = 1$, and the normalization constant $\widetilde{K}_P(\kappa)$ is equal to:

$$\widetilde{K}_P(\kappa) = \frac{\kappa^{P/2-1}}{(2\pi)^{P/2} I_{P/2-1}(\kappa)},$$

where I_v denotes the modified Bessel function of the first kind and order v (Abramowitz et Stegun, 1972).

The multivariate Gaussian distribution assumption is not restrictive. As long as the group of correlated variables are independent realizations of the same distribution, the SelectBoost algorithm can be applied. Directly to assess the stability of the selected variables with perturbated datasets with an increasing noise level, which is still the idea behind the SelectBoost algorithm. After replacing the von Mises-Fisher with a more relevant one.

The SelectBoost algorithm

To use the SelectBoost algorithm, we need a grouping method gr_{c_0} depending on a user-provided constant $0 \leqslant c_0 \leqslant 1$. This constant determines the strength of the grouping effect. The grouping method maps each variable index $1, ..., P$ to an element of $\mathcal{P}(\{1, ..., P\})$ (with $\mathcal{P}(S)$ the powerset of the set S, *i.e.* the set which contains all the subsets of S). Concretely, $gr_{c_0}(p)$ is the set of all variables which are considered to be linked to the variable $\mathbf{x}_p$ and $\mathbf{X}_{gr_{c_0}(p)}$ is the submatrix of $\mathbf{X}$ containing the columns which indices are in $gr_{c_0}(p)$. We impose the following constraints to the grouping function:

$$\forall p \in \{1, ...P\} : gr_1(p) = \{p\} \text{ and } gr_0(p) = \{1, ...P\}. \tag{5.6}$$

Furthermore, we need to have a selection method:

$$select \ : \ \mathbb{R}^{N \times P} \times \mathbb{R}^N \ \rightarrow \ \{0, 1\}^P$$

which maps the design matrix $\mathbf{X}$ and the response variable $\mathbf{y}$ to a 0-1 vector of length P with 1 at position p if the method selects the variable p and 0 otherwise.

We then use the von Mises-Fisher law to generate replacement of the original variables by some simulations (see Algorithm 1) to create B new design matrices $\mathbf{X}^{(1)}, ..., \mathbf{X}^{(B)}$. The SelectBoost algorithm then applies the variable selection method $select$ to each of these matrices and returns a vector of length P with the frequency of apparition of each variable. The frequency of apparition of variable $\mathbf{x}_{p.}$, noted ζ_p is assumed to be an estimator of the probability $\mathbb{P}(\mathbf{x}_{p.} \in \mathcal{S})$ for this variable to be in $\mathcal{S}$. The choice of c_0 is crucial. On the one hand, when this constant is too large, the model is not perturbed enough. On the other hand, when this constant is too small, variables are chosen at random.

The SelectBoost algorithm returns the vector $\boldsymbol{\zeta} = (\zeta_1, ..., \zeta_P)'$. Each of these values has to be compared to a threshold ζ_{min} to determine which variables are selected: we choose to select a variable p if $\zeta_p \geqslant \zeta_{\text{min}}$. The simulation study

showed that the choice of the threshold is critical and the algorithm can be improved if we enforce that the ζ_p values - as functions of c_0 - are non-increasing,

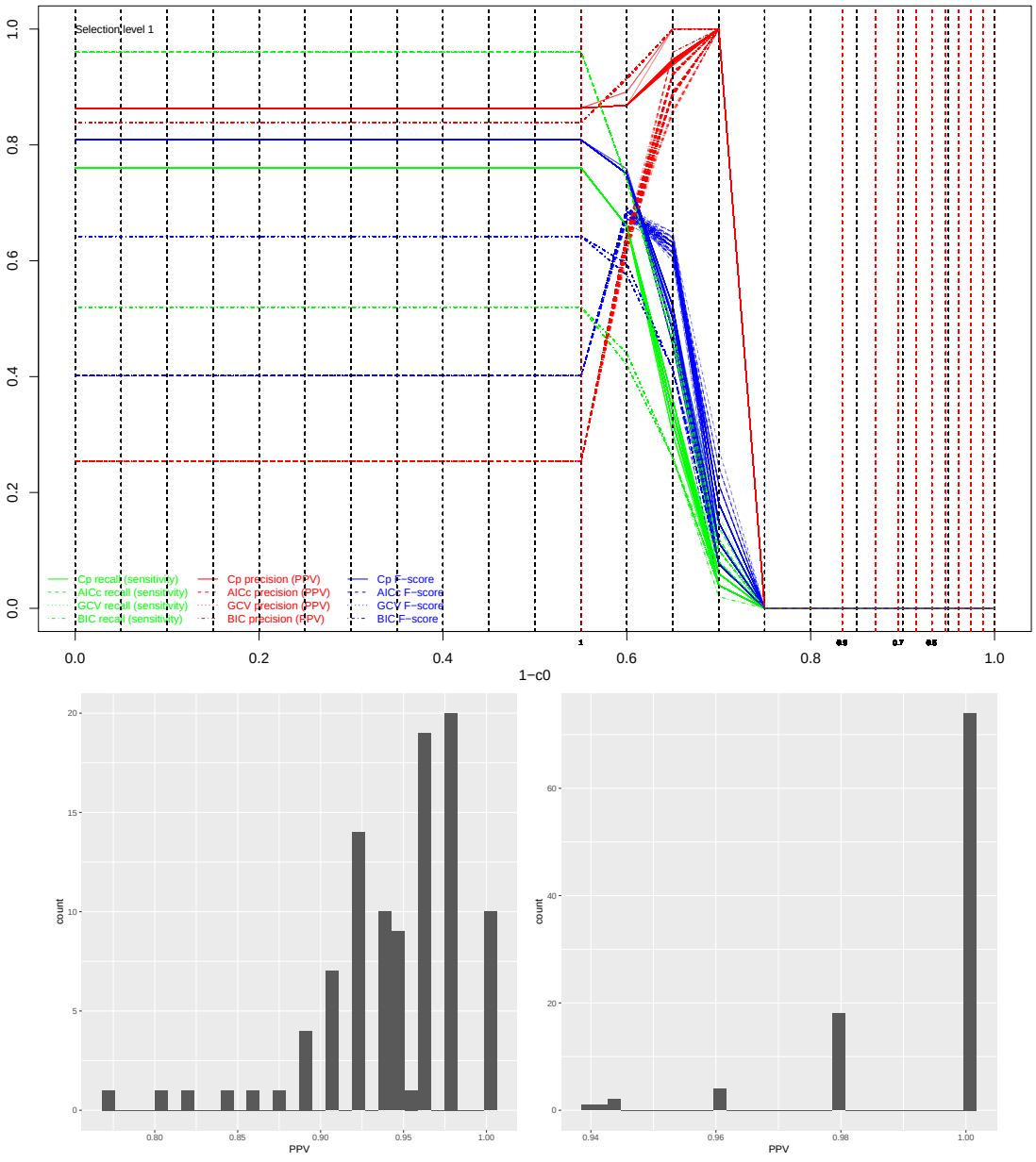


Figure 5.1 : *Top: evolution of the recall, PPV and F-score as a function of $1 - c_0$ for LASSO-based SelectBoost for Type1 simulated data with a non-increasing post-processing step . Bottom: the distribution of the PPV for a 0.25 threshold and $c_0 = \text{mean}(q_{90}, q_{100})$ for SPLS-based SelectBoost and Type1 data.*

Table 5.1 : *Summary of the types of datasets used to benchmark the SelectBoost algorithm*

Name	Data	Individuals	Variables
Type1	Simulated	100	1000
Type2	Simulated	100	1000
Type3	Simulated	400	203
Type4	Simulated	750	102
Leukemia	Observed	72	3571
Huntington	Observed	69	17717
Melanoma	Observed	28	25268

see Figure 5.1 bottom. This additional requirement makes sense: the more variables the resampling process involves -with smaller c_0-, the less a given variable will be selected.

Choosing the parameters of the algorithm

We first have to choose the grouping function. One of the simplest ways to define a grouping function gr_{c_0} is the following:

$$gr_{c_0}(p) = \Big\{ q \in \{1, ..., P\} \;\Big|\; |< \mathbf{x}_{p.}, \mathbf{x}_{q.} >| \geqslant c_0 \Big\}. \tag{5.7}$$

In other words, the correlation group of the variable p is determined by variables whose correlation with $\mathbf{x}_{p.}$ is at least c_0. In another way, the structure of correlation may further be taken into account using graph community clustering. Let $\boldsymbol{C}$ be the correlation matrix of matrix $\mathbf{X}$. Let define $\check{\boldsymbol{C}}$ as follows:

$$\check{c}_{ij} = \begin{cases} \mid \check{c}_{ij} \mid & \text{if} \quad \mid \check{c}_{ij} \mid > c_0 \text{ and } i \neq j \\ 0 & \text{otherwise.} \end{cases}$$

Then, we apply a community clustering algorithm on the undirected network with weighted adjacency matrix defined by $\check{\boldsymbol{C}}$. Using a graph community clustering algorithm is helpful with large datasets while still clustering similar variables together. For instance, the fast greedy modularity optimization algorithm for finding community structure, Clauset *et al.* (2004), runs in essentially linear time for many real-world networks given that they are sparse and hierarchical.

Once the grouping function is chosen we have to choose parameter c_0 Due to the constraints in equation (5.6) the SelectBoost algorithm results in the initial variable selection method when $c_0 = 1$. As we will show in the next subsection,

the smaller the parameter c_0, the higher the precision of the resulting selected variables. On the other hand, it is obvious that the probability of choosing none of the variables (*i.e.* resulting in the choice of an empty set) increases as the parameter c_0 decreases. In the perspective of experimental planning, the choice of c_0 should result of a compromise between precision and proportion of active identified variables. Hence, the c_0 parameter can be used to introduce a confidence index γ_p related to the variable $\mathbf{x}_{p.}$:

$$\gamma_p = 1 - \min_{\mathbf{x}_{p.} \in \hat{\mathcal{S}}_{c_0}} c_0, \text{ hence } 0 \leqslant \gamma_p \leqslant 1 \tag{5.8}$$

Algorithme 5.1 Pseudo-code for the SelectBoost algorithm

Nécessite $gr_{c_0}, select, B, c_0$
  $\zeta \leftarrow \mathbf{0_P}$
  Pour $b = 1, ..., B$ **Faire**
    $\mathbf{X^{(b)}} \leftarrow \mathbf{X}$
    Pour $p = 1, ..., P$ **Faire**
      $\mathbf{x_{p.}^{(b)}} \leftarrow \phi^{-1}\left(\text{random-vMF}\left(\hat{\mu}(\phi(\mathbf{X_{gr_{c_0}(p)}})), \hat{\kappa}(\phi(\mathbf{X_{gr_{c_0}(p)}}))\right)\right)$
    Fin Pour
    $\zeta \leftarrow \zeta + \mathbf{select}(\mathbf{X^{(b)}}, \mathbf{y})$
  Fin Pour
  $\zeta \leftarrow \zeta/\mathbf{B}$

5.3.3 Numerical studies

We benchmarked the algorithm with a large simulation study with four data generation processes and three real datasets.

1. Simulation with 1000 variables and one linear response. A cluster of 50 variables is linked to the response.

2. Simulation with 1000 variables and one binary response. A cluster of 50 variables is linked to the response.

3. Data are 200 uncorrelated ("unlinked") single nucleotide polymorphisms (SNPs) with simulated genotypes, in which the first 20 of them affect the outcome with three covariates. 400 observations;

4. Data are 100 uncorrelated ("unlinked") single nucleotide polymorphisms (SNPs) with simulated genotypes, in which the first 10 of them affect the outcome with two covariates. 750 observations.

5. The Huntington dataset is a real dataset with 28087 variables observed on 69 individuals. We first applied independent filtering and removed 10370 variables. We applied the SelectBoost algorithm to 17717 variables observed on 69 individuals.

6. The leukemia dataset (Golub *et al.*, 1999) are the preprocessed data of (Dettling, 2004) retrieved from the supplementary materials accompanying (Friedman *et al.*, 2010b).

7. The melanoma dataset is the GSE78220 dataset from Hugo *et al.* (2016).

For Type 1 and 2, the number of variables is 1000, and the number of observations is 100. The data are generated from a cluster simulation (Bastien *et al.*, 2015; Bair *et al.*, 2006). Only 50 first predictors are linked to the response Y and

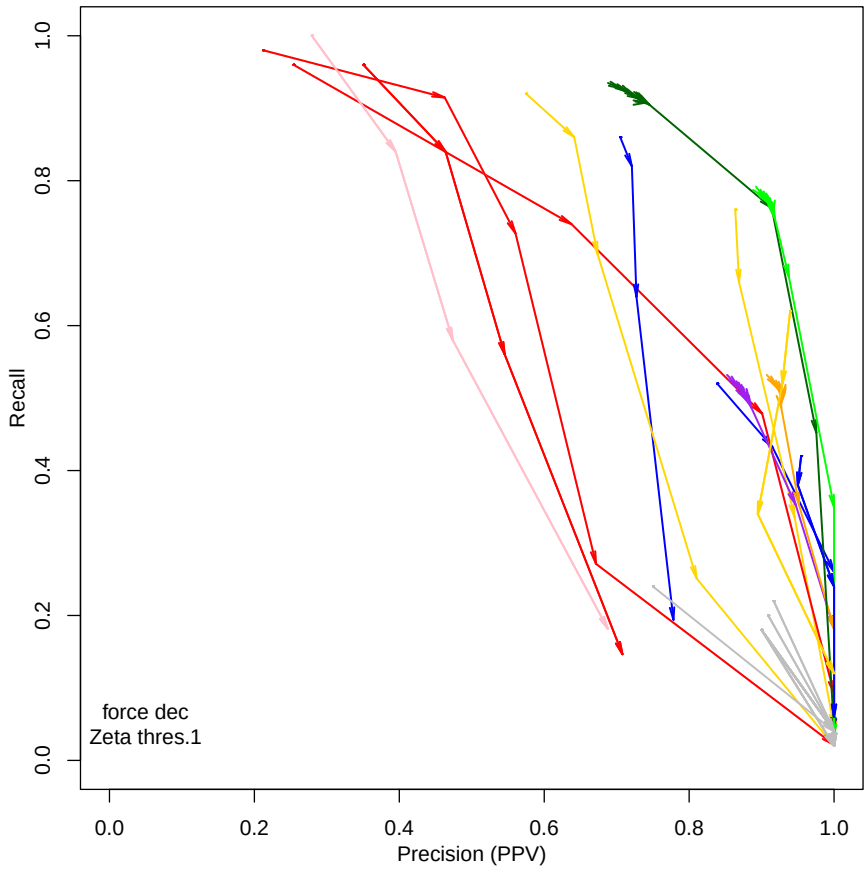


Figure 5.2 : *Recall-precision curve. All models and criteria non-increasing SelectBoost. Type 1 data. Direct grouping. 100 different datasets.* $\zeta_{\min} = 1$

the last 950 variables are randomly generated from a standard normal distribution. For example 3, the response variable is linear but was turned into a binary variable ($+1$ when $Y_i > 0$ and -1 when $Y_i < 0$).

Examples 1 and 3 are linear regression examples whereas 2, 4, 5, 6 and 7 are logistic regression ones, for which we will assume a logistic model with a binary response variable (Peng *et al.*, 2002).

We provide results for twelve different settings based on ten different models.

1. Linear regression (seven types): SPLS (Chun et Keleş (2010), with raw and bootstrap corrected coefficients), LASSO with model choice based on information criteria (AICc, BIC, GCC, Cp), LASSO with model choice based on five fold cross-validation, adaptative LASSO, enet, adapative enet, varbvs linear (Carbonetto et Stephens, 2012; Guan et Stephens, 2011; Zhou *et al.*, 2013).

2. Logistic regression (five types): logistic LASSO (glmnet based) with model choice based on five fold cross-validation, logistic LASSO (glmnet based) with model choice based based on information criteria (AICc, BIC), varbvs binomial, SPLSda (Chun et Keleş, 2010), sgpls (Chun et Keleş, 2010).

The SelectBoost algorithm is based on correlated resampling and hence random. We wanted to assess both the stability and performance of the algorithm. As a consequence, for the four types of simulated data, we focused both on what may be called a repeatability study (a given dataset was analyzed 100 times to estimate the variation only due to the fact that the algorithm is random) and a reproducibility study (100 different datasets were generated and analyzed to estimate the variability due to both data simulation -from the same data generation- and the fact that the algorithm is random).

The repeatability issue raised was raised, for instance by Boulesteix (2014); Magnanensi *et al.* (2017) for PLS models. For those models, random split cross-validation is known to have poor repeatability. We used two types of grouping functions (either determined by variables whose correlation with x_p is at least c_0 -gdirect- or community clustering-based -gcc-).

The cost (memory and time) of the random generation step can be limited thanks to a sparse correlated resampling feature. The remaining cost of the algorithm is $B \times Nc0 \times Time_1$ with B the number of resampling and $Nc0$ the number of c0 values that are investigated and $Time_1$ the time to fit the model once.

To demonstrate the performance of the SelectBoost method, we compared our method with stability selection (Meinshausen et Bühlmann, 2010) and with a naive version of our algorithm, naiveSelectBoost. The naiveSelectBoost algorithm works as follows: estimate β with any variable selection method then if $gr_{c_0}(p)$, as defined in equation (5.7) for example, is not reduced to p, shrink to

0. The naiveSelectBoost algorithm is similar to the SelectBoost algorithm, except that it does not take into account the error which is made choosing at random a variable among a set of correlated variables.

We use four indicators to evaluate the abilities of our method on simulated data. We define:

- recall as the ratio of the number of correctly identified variables (*i.e.* $\hat{\beta}_i \neq 0$ and $\beta_i \neq 0$) over the number of variables that should have been discovered (*i.e.* $\beta_i \neq 0$).

- precision as the ratio of correctly identified variables (*i.e.* $\hat{\beta}_i \neq 0$ and $\beta_i \neq 0$) over the number of identified variables (*i.e.* $\hat{\beta}_i \neq 0$).

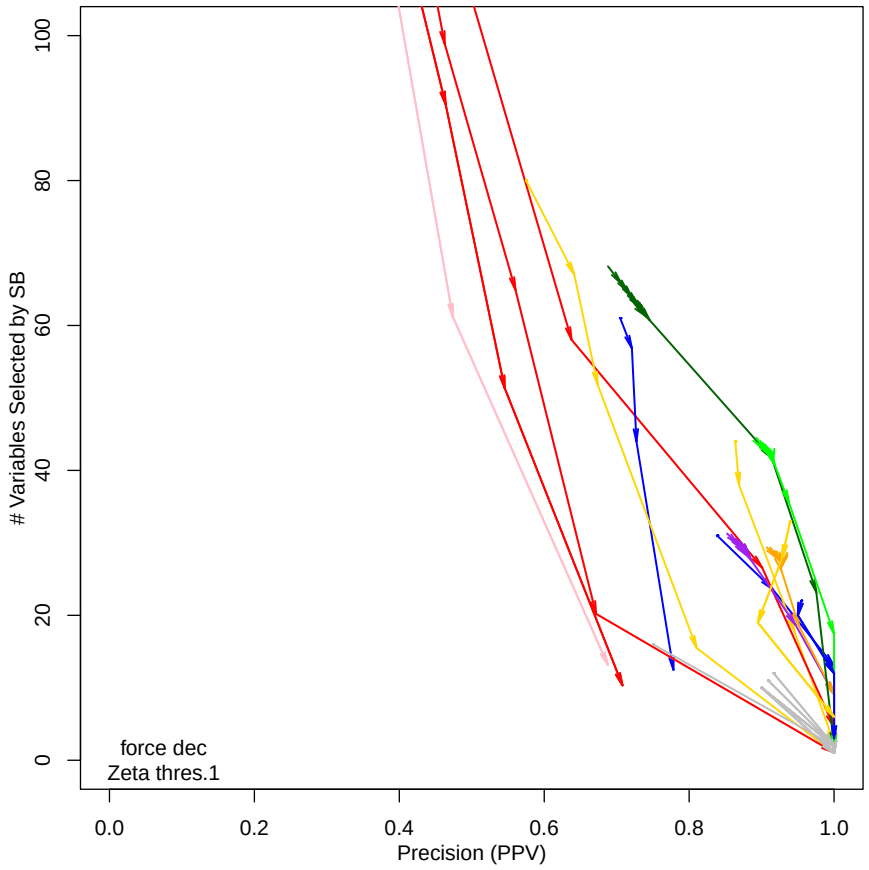


Figure 5.3 : *The average number of identified variables is plotted as a function of the proportion of correctly identified variables for Type1 simulated data and all models.*

- F-score as the following ratio:

$$2 \times \frac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}}.$$

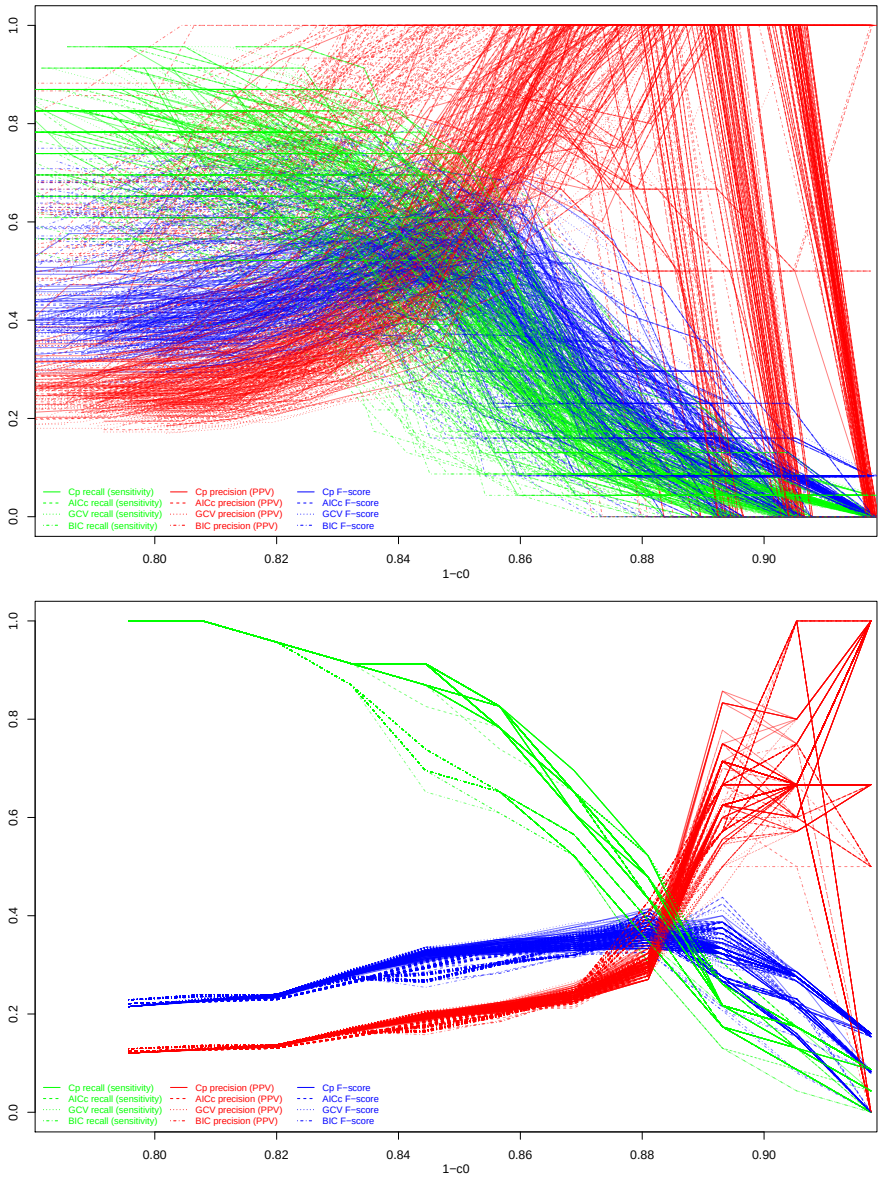


Figure 5.4 : *Top and Bottom: Effect of the SelectBoost algorithm wrt* $1 - c_0$ *for adaptative elastic net with* c_0 *in the range* $[q_{90\%}; q_{100\%}]$ *for 100 different (Middle, reproducibility) or 100 identical (Bottom, repeatability) Type3 simulated data with a non-increasing post-processing step.*

- selection as the average number of identified variables (*i.e.* $\hat{\beta}_i \neq 0$).

Note that our interest is focused on precision, as our goal is to select reliable variables. As stated before, when c_0 is decreasing toward zero, we expect a profit in precision and a decrease in recall. We also compute the F-score, which combines both recall and precision. As an improvement of precision comes with a decrease in the number of identified variables, the best method is the one with the highest precision for a given level of selection.

Results of the numerical studies

We show the evolution of the four criteria (recall, precision, F-score and selection) with regards to the decrease of c_0. When $c_0 = 1$, the SelectBoost algorithm is equivalent to the initial variable selection method. We introduce a post-processing step to enforce that, for a given variable, the proportion of selection is non-increasing. It is the expected behaviour since the correlated resampling is not meant to increase the probability of selection for a variable. Such an increase may happen for small c_0 values when a variable that is not linked with the response is mixed with a variable that is linked to the response. For all the simulations, this post-processing step increases the PPV of the SelectBoost algorithm, see Figure 5.1. As our primary focus is PPV, we recommend the use of this post-processing step. More details can be found in the SI Graphs 7 to 174.

We created precision-recall plots to display the effects of the algorithm on the performance of all the models and criteria used for a given dataset. Identical model fitting criteria share the same colors. The arrows point towards decreasing c_0 values. Direct grouping and community grouping lead to similar results, Figures 5.2 and SI 1, 3 and 4. These Figures also show that the results for a single dataset repeated 100 times are similar to results for 100 different datasets. The Zoom l sequence achieves high PPV, Figure SI 299.

Figure SI 5 displays an example of raw SelectBoost (without the non-increasing post-processing step) for direct grouping and 100 different datasets that should be compared to Figure 5.2. This effect is even stronger smaller values for the ζ_{min} threshold. The non-increasing post-processing step greatly improves the results of the algorithm and leads to monotonic relationships between the recall, the precision and the c_0 value. PPV benefit less from smaller values for the ζ_{min} threshold, Figure SI 6.

All the results of the simulation study showed the good performance and stability of the algorithm, which we then applied once to each of the real datasets. The results of the gdirect and gcc based SelectBoost are similar, the gcc based being a bit more time consuming than the gdirect one. In the following of this article, we reported results and Figures for gdirect based SelectBoost.

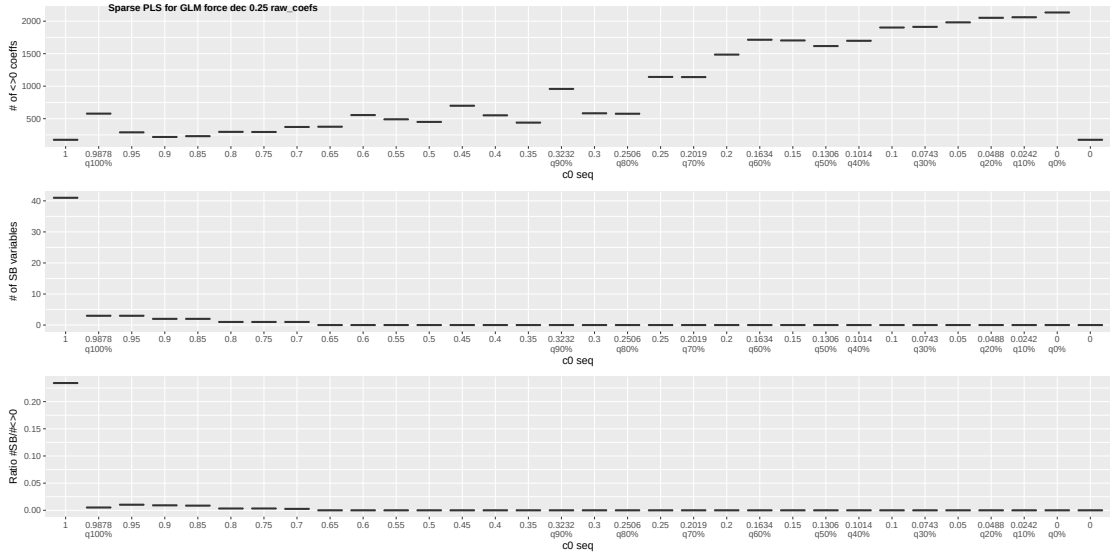


Figure 5.5 : *% of non-zero coefficients wrt to c_0 for SGPLS-based SelectBoost models of the leukemia datasets and threshold= .25.*

According to our simulation studies, Graphs SI 7 to 174, one should choose c_0 between $q_{90\%}$ and $q_{100\%}$, see Figure 5.1 Top. In our simulation studies, we used an 11-steps c_0 sequence, but, according to our results, it could be limited to 6 steps (from mean$(q_{90\%}, q_{100\%})$ to $q_{100\%}$) for the biggest datasets. The value of B should not be lower than 10. $B = 50$ or $B = 100$ will provide more stable results. As a consequence, the minimal time cost of the SelectBoost algorithm will be 60 times the time cost of the regular model fit, which could be afforded in almost every case. The parallel processing support of the SelectBoost package can help to reduce this time. Hence the SelectBoost seems feasible with most of the datasets and even omics datasets as we did in our simulation study with the three real datasets.

Hence to assess the performance of the SelectBoost algorithm, we performed comprehensive numerical studies. As stated before, the SelectBoost algorithm can be applied to any existing variable selection method.

Figure 5.1 Top shows the result for the lasso selection with a penalty parameter chosen using information criteria for Type 1 datasets. In this example, we improve the precision up to 1. Moreover, as shown by Figures 5.1 and 5.4, the proportion of models, for which the precision reaches one, increases with the decrease of c_0. The F-score increases, remains either stable or shows a small decrease indicating that the increase of PPV compensates the loss in recall.

In the previous subsection, we mentioned the possibility of using SelectBoost to obtain a confidence index, corresponding to one minus the lowest c_0 for which

a variable is selected. For each c_0, we plotted the average number of selected variables as a function of the proportion of correctly identified variables (Figure 5.3 and SI 300-304). As expected, the proportion of correctly identified variables increases with the increase of the confidence index and with the decrease of the average number of identified variables. Therefore, the proportion of non-predictive features decreases with the increase of the confidence index.

The SelectBoost algorithm shows its superiority over the naive SelectBoost algorithm. The error made when choosing a variable randomly among a set of correlated variables leads to more incorrect choices of variables. While the intensive simulation of our algorithm allows taking into account this error, the naiveSelectBoost does not.

Finally, we compare the SelectBoost algorithm with stability selection. Stability selection uses a resampling algorithm to determine which of the variables included in the model are robust. In our simulation, stability selection shows performance with high precision but also low recall. Moreover, in contrast to the SelectBoost algorithm, stability selection does not allow to choose a convenient precision-PPV trade-off.

The timings of the algorithm can be found on Figure 5.9 and on SI Figures 175 to 222.

5.3.4 Application to three real datasets

We applied our algorithm to three real datasets. We studied, with respect to the threshold, the number of non-zero variables, the number of variables selected by SelectBoost and their ratio. We found results that were concordant with those of the simulated datasets. Figure 5.5 displays those results for a SGPLS based SelectBoost of the Leukemia dataset with a .25 threshold (tr$= .25$). See SI Figures 247 to 298.

We report the results for the RNA-Seq dataset providing mRNA expressions from Huntington's disease and neurologically normal individuals. This dataset was downloaded from the GEO database under accession number GSE64810 (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64810`). This dataset contains 20 Huntington's Disease cases and 49 neurologically normal controls and includes 28,087 genes as explanatory variables. An independent filtering (Bourgon *et al.*, 2010) preprocessing step was first performed using data-based filtering for replicated high-throughput transcriptome sequencing experiments (Rau *et al.*, 2013). Then we applied the lasso selection method to this reduced dataset (see Figure 5.6 left for the whole path of the solution). We used cross-validation to choose the appropriate level of penalization (*i.e.* the λ parameter in Equation (5.3)).

We then applied our SelectBoost algorithm on the lasso method with penalty parameter chosen by cross-validation. We use a range for the c_0 parameter starting from 1 to 0.7 with steps of 0.05, which corresponds to a confidence index from 0 to 0.3. For each step, the probability of being included in the support $\mathcal{S}$ was calculated with 200 simulations as described in Algorithm 1. We set the threshold of being in the support to 0.95 to avoid numerical instability. We classify the selected variables into three categories: those that are identified for each confidence index from 0 to 0.15 (red), those identified from 0 to 0.25 (orange) and those identified from 0 to 0.3 (green). The last category contains the most reliable variables selected by the SelectBoost algorithm because these variables are identified from low to high confidence index.

With the lasso selection method, 15 variables were selected. Among them, four genes were identified by SelectBoost into the three different categories of confidence index (see Figure 5.6 right): two genes for low confidence (red) (ANXA3 and INTS12), one gene for intermediate confidence (orange) (NUB1) and one gene for high confidence (green) (PUS3).

The interesting point, in these three examples, is that the identified variables are neither the first variables selected by the lasso nor the variables with the highest coefficients (see Figure 5.6 left). This result demonstrates that our algorithm can be advantageous to select variables with high confidence and not just to select variables with the highest coefficients.

Finally, we decided to assess the differential expression of these genes between patients and controls, using the `limma` package (Linear Models for Microarray and RNA-Seq Data) (Ritchie *et al.*, 2015). The four identified genes are significantly down-expressed by neurologically healthy controls confirming the result of a logistic model with these four genes.

5.3.5 Robust reverse-engineering of networks

Sparsity is a well-known feature of most biological networks (Barabási, 2003). An actor can only be regulated by a small number of other actors, whereas it may regulate any number of other actors. Hence, variable selection methods, such as the lasso, ensures that sparsity feature and are often core components of most of the biological network reverse-engineering tools. As a consequence, we propose to apply the SelectBoost algorithm in two different ways in order to improve the biological network reverse-engineering: as a post-processing step after the inference was made or during the inference itself in order to select the most stable predictors for each node in the network. When used as a post-processing step, one can assess for any of the inferred links between the actors of the network its confidence index against correlated resampling of the predictors. When used

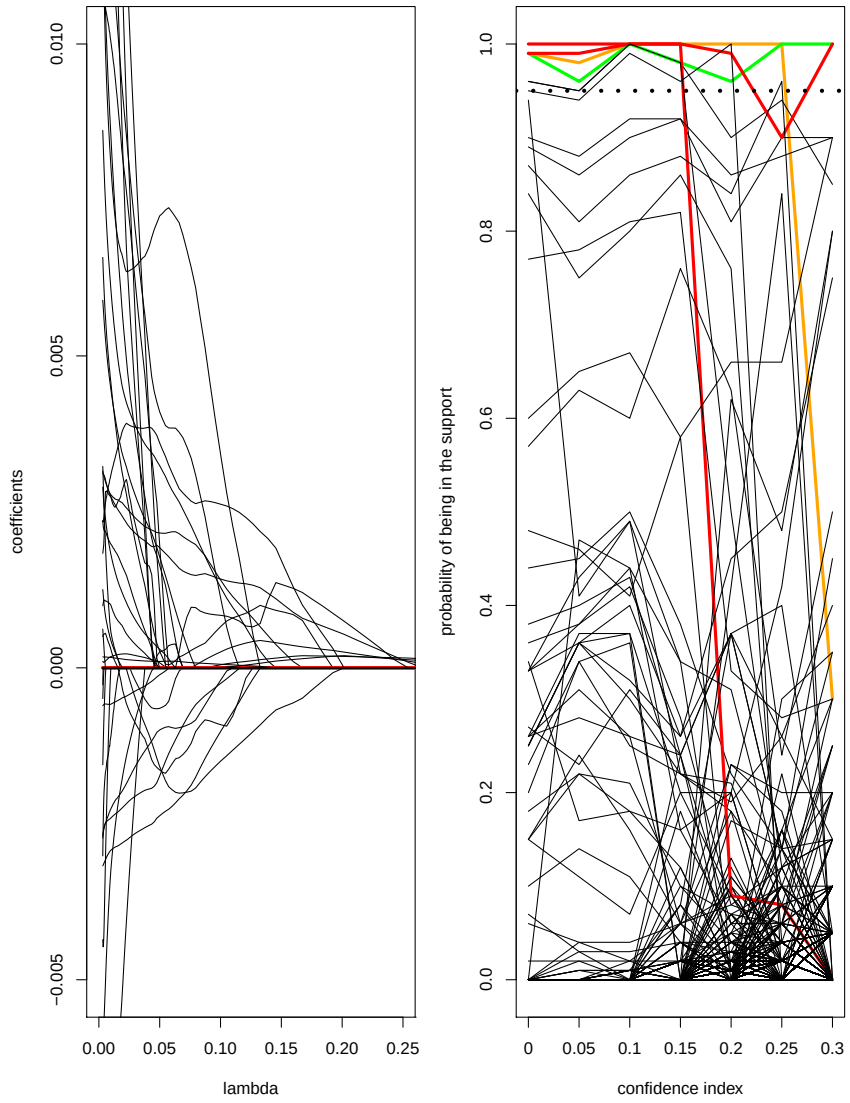


Figure 5.6 : ***Colors:*** *the green is for the most reliable variables selected by the SelectBoost algorithm (confidence index of 0.3; orange is for intermediate confidence (0.25) and red for low confidence (0.15)).* ***Left:*** *evolution of the coefficients in the lasso regression when the regularization parameter λ is varying.* ***Right:*** *evolution of the probability of being in the support of the regression when the confidence index is varying. The dotted line represents the threshold of 0.95.*

during the inference step, one can infer a model that is only built with links with a high enough confidence index. The former is implemented in the SelectBoost package as a new method for the Cascade package. The latter is implemented in the new Patterns CRAN package as a dedicated fitting function and is especially useful when trying to find targets for biological intervention that are strongly related to markers of some diseases through the reverse-engineered network and useful and reliable links.

We benchmarked those two uses of the algorithm with a particular type of biological networks that we have been using for several years: cascade networks (Vallat *et al.*, 2013).

For the post-inference processing, we first fit a model to a cascade network using the Cascade package inference function. Then we compute confidence indices for the inferred links using the SelectBoost algorithm, more details, as well as the code, of the simulations can be found in the vignette of the package "Towards Confidence Estimates in Cascade Networks using the SelectBoost Package", available at `https://fbertran.github.io/SelectBoost/articles/confidence-indices-Cascade-networks.html`. An example of those results is shown in Figure 5.7 with a cascade network for four time points and four groups of 25 actors.

For the use of the SelectBoost algorithm during the fitting step of a cascade network reverse-engineering, we used the Patterns package. Benchmark results were reported as sensitivity, positive predictive value and F-score, shown in Figure 5.8; the code, the simulation details and the remaining results are part of a vignette of the package "Benchmarking the SelectBoost Package for Network Reverse Engineering", that is available at `https://fbertran.github.io/SelectBoost/articles/benchmarking-SelectBoost-networks.html`.

We created an unweighted or a weighted version of the algorithm. The weighted version of the algorithm enables the user to include weights in the model, which means to favour or disfavour some links between the actors, in order, for instance, to take into account biological knowledge.

The results shown in Figure 5.8 of the simulation study are a comparison to a standard set up for stability selection and regular lasso both for an unweighted version of the algorithms and a highly correctly weighted version of the same algorithms.

By highly correctly weighted, we mean that we included influential weights in the model accordingly to the links that existed in the network that was used for data simulation. This network was randomized from one simulation to another. This weighted setting was used to determine if including correct biologi-

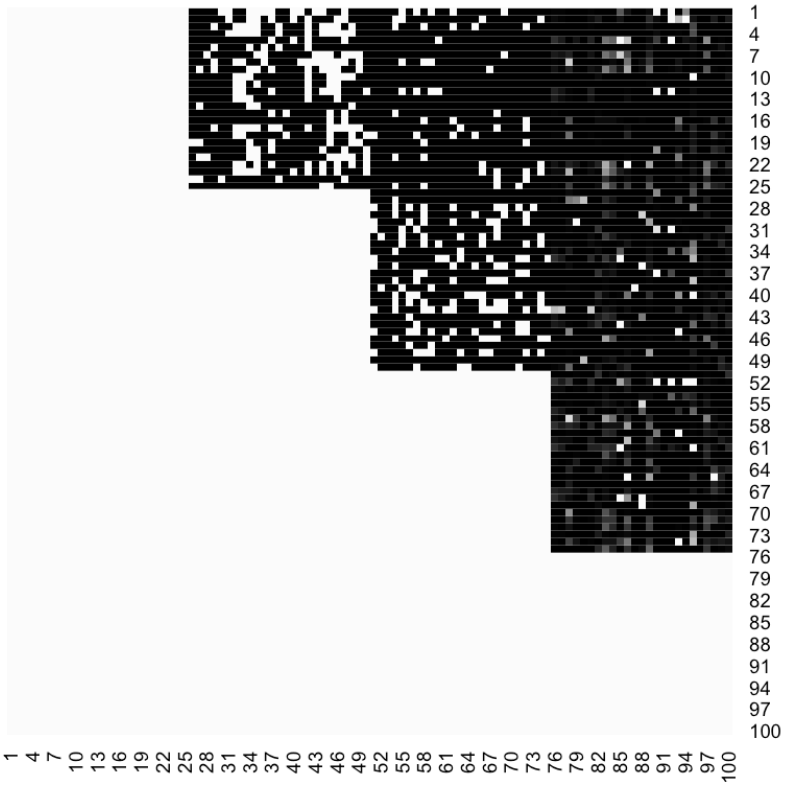


Figure 5.7 : *Post inference analysis of an inferred cascade network. Dark values are tantamount to low confidence. Bright values are tantamount to high confidence. Confidence ranges from 0 (lowest) to 1 (highest). The lower triangular part of the matrix is an area with the highest confidence (1) since we know -and assume so in the model- that for cascade networks those links must be equal to 0.*

cal knowledge would help the reverse-engineering algorithm to retrieve the correct network. If correct biological knowledge is included in the model, all three fitting functions lead to similar and outstanding results for the F-score criterion without even requiring the need to search for an optimal thresholding value as we had to do with the Cascade package.

For each simulated dataset, vertical dots are displayed to show the optimal threshold level that should be used to maximize the F-score. It is computed with respect to the actual values that are unknown for real datasets. Without weights, SelectBoost shrinks the range of optimal values when compared to the lasso or stability selection. With correct weights, none of the methods still requires to

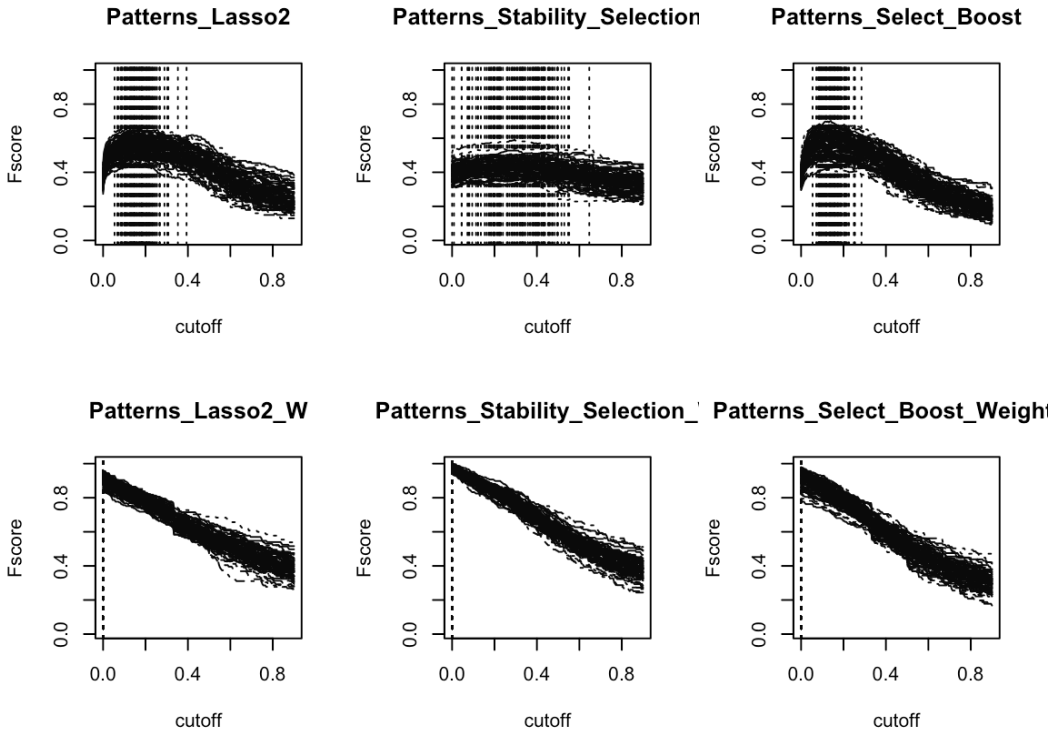


Figure 5.8 : *F-score as a function of the thresholding value: if an inferred coefficient for the network is less than the thresholding value, then it is set to 0. The SelectBoost algorithm is compared to both stability selection and the regular lasso. The upper row displays results for the unweighted version of the algorithms, whereas the lower row displays results for their weighted counterparts.*

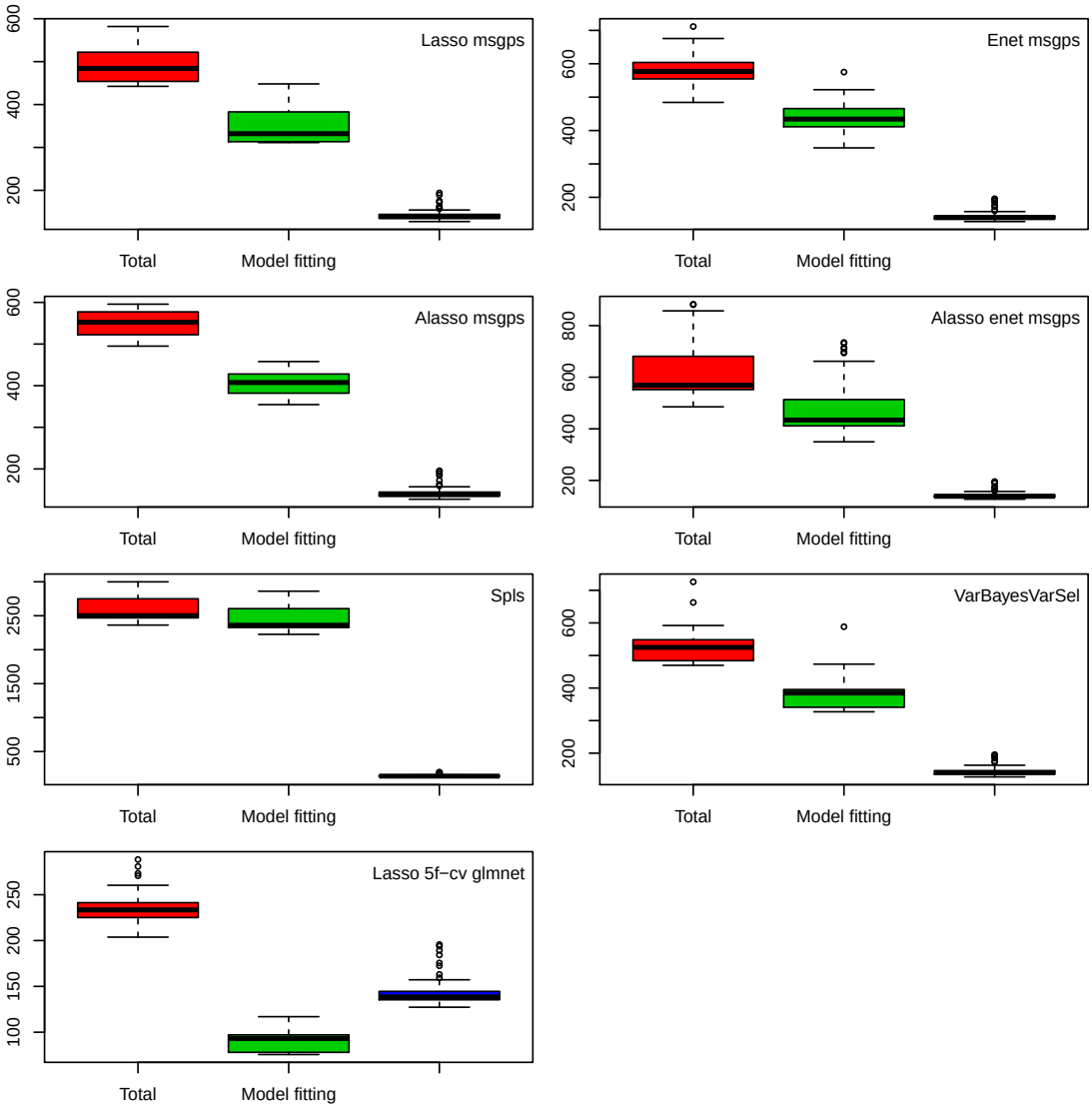


Figure 5.9 : *Timing of SelectBoost for the seven linear regression models and Type 1 datasets.*

use a cut-off value to maximize F-score.

In an unweighted setting, the SelectBoost version of the fitting process shows better performance than stability selection and the lasso as long as the cut-off value is less than 0.4, which is about the double of the optimal thresholding value.

5.3.6 Conclusion

We introduce the SelectBoost algorithm that relies intensive computations to select variables with high precision (PPV). The user of SelectBoost can apply this algorithm to produce a confidence index or choose an appropriate precision-selection trade-off to select variables with high confidence and avoid selecting non-predictive features. The main idea behind our algorithm is to take into account the correlation structure of the data and thus use intensive computation to select reliable variables.

We prove the performance of our algorithm through simulation studies in various settings and recommend the use of c_0 in the range $q_{90\%}$ and $q_{100\%}$ with the non-increasing post-processing step to get the best results. We succeed in improving the PPV, whenever it was possible, of all the twelve selection methods with relative stability on recall and F-score. Il the PPV was already nearing 1, then there is almost no negative effect on the PPV and recall when applying SelectBoost.

Our results open the perspective of a precision-selection trade-off which may be very useful in some situations where many regressions have to be made (*e.g.* network reverse-engineering with one regression made per node of the network). In such a context, our algorithm may even be used in an experimental design approach.

The application to three real datasets allowed us to show that the most reliable variables are not necessarily those with the highest coefficients. The SelectBoost algorithm is a powerful tool that can be used in every situation where reliable and robust variable selection has to be made.

Chapitre 6

Contributions à la régression des moindres carrés partiels

6.1 Introduction

Nous allons dans ce chapitre aborder la régression des moindres carrés partiels classique (Wold *et al.* (1983), Wold *et al.* (1984), Wold *et al.* (2001)) ou parcimonieuse (Lê Cao *et al.* (2009), Chun et Keleş (2010)) mais aussi la problématique plus générale de l'amélioration de la spécificité lors des phases de sélection de variables dans la régressions des moindres carrés partiels parcimonieux (Lê Cao *et al.* (2008), Lê Cao *et al.* (2011)).

6.2 Régression sur données qualitatives

6.2.1 Présentation de la problématique

J'ai commencé à m'intéresser à la régression des moindres carrés partiels (abrégée en régression PLS) lorsqu'il a fallu que je réponde à une question qui m'avait été posée par mon collègue le Professeur Nicolas Meyer, PU-PH à l'Hopital Civil de Strasbourg. Nicolas Meyer avait besoin de modéliser un certain type de données médicales, données qui provenaient de l'une de ces collègues, Dominique Guénot (directeur de recherche 2 du CNRS) : celles issues de l'**allélotypage**, terme que nous allons définir immédiatement.
Un **microsatellite** est une séquence non-codante de l'ADN. L'**allélotypage** consiste à rechercher le statut normal ou altéré d'un ensemble prédéfini de microsatellites, en général dans une cellule cancéreuse. Les données d'allélotypage présentent les spécificités suivantes :

- un ensemble de variables explicatives binaires décrivant l'état global des chromosomes de la cellule ;
- une variable à expliquer, également binaire, du patient ou de la tumeur ;
- le nombre de variables explicatives peut dépasser le nombre de sujets ;
- une éventuelle colinéarité entre les variables explicatives.

La compréhension des mécanismes de cancérogenèse implique également une description multivariée des données. À l'époque où Nicolas Meyer m'avait posé la question, les publications biomédicales traitant de données d'allélotypage ignoraient la structure générée par les variables.

En effet, les analyses statistiques étaient faites uniquement en univarié (voir par exemple Zhu *et al.* (1998)). Les approches multivariées qui ont été tentées étaient des analyses en *clusters* (voir par exemple Weber *et al.* (2007)), analyses qui ne permettaient malheureusement pas de répondre à toutes les questions posées par les biologistes. De plus, nous rappelons que les méthodes exploratoires multivariées ne modélisent pas, ce qui limite leur utilisation pour étudier les relations possibles entre les voies d'altérations et une caractéristique clinique du patient ou de la tumeur cancéreuse.

Si la variable à expliquer avait été une variable quantitative continue, alors l'utilisation de la régression des moindres carrés partiels (voir par exemple Tenenhaus (1998)) aurait été directe. Dans le cas d'une réponse binaire (le cas qui m'intéresse ici), il fallait envisager des dérivées de la PLS des régressions linéaire et logistique comme proposée dans Tenenhaus (1999). Une spécificité supplémentaire des données d'allélotypage est la présence de variables explicatives uniquement qualitatives. Nous avons donc comparé, dans Meyer *et al.* (2010), les performances des variantes PLS des régressions linéaire et logistique sur des variables toutes qualitatives.

6.2.2 Des propriétés intéressantes de la régression PLS

L'analyse des jeux de données qui contiennent un grand nombre de variables est toujours en forte progression dans tous les domaines et plus particulièrement en biologie et en médecine. Toutefois, ces jeux de données présentent souvent deux niveaux de complexité :

- la présence d'une corrélation linéaire significative entre les variables contenues dans les jeux de données,
- le nombre de variables est souvent bien plus élevé que le nombre d'observations,

- la présence de valeurs manquantes dans le jeu de données aussi bien au niveau des variables que des observations.

Il faut alors recourir à des modèles plus raffinés que le modèle linéaire usuel. L'un de ces modèles est justement la régression des moindres carrés partiels (Wold *et al.* (1983), Wold *et al.* (1984), Wold *et al.* (2001)).

6.2.3 Package `plsRglm`

Afin de mettre en œuvre les comparaisons réalisées dans l'article Meyer *et al.* (2010), nous (Frédéric Bertrand et moi) avons créé des fonctions écrites en langage `R`. Il nous a semblé important de le rendre accessible à l'ensemble de la communauté scientifique et surtout médicale sous la forme d'un *package*, nommé `plsRglm`, pour le langage `R`, Bertrand et Maumy-Bertrand (2020d), `https://cran.r-project.org/web/packages/plsRglm/index.html`.

L'objectif du *package* `plsRglm` est de pouvoir traiter soit des jeux de données complets soit des jeux de données présentant des valeurs manquantes (Little et Rubin (2002)). L'originalité du *package* `plsRglm` est qu'il va traiter les donnés manquantes à l'aide de techniques qui n'existaient pas dans les autres *packages*. En effet, il faut noter que les autres *packages* de régression des moindres carrés partiels, existants à ce moment-là, n'autorisaient pas la présence de valeurs manquantes. En effet, l'utilisateur était confronté à deux choix : soit utiliser des *packages* qui supprimaient les lignes ou les colonnes présentant des valeurs manquantes ou utiliser des *packages* qui se terminaient avec une erreur.

Le *package* `plsRglm` implémente la régression des moindres carrés partiels univariée usuelle mais aussi son extension aux modèles de régression des moindres carrés partiels généralisée (PLS due à Bastien *et al.* (2005) et en particulier aux modèles de régression logistique des moindres carrés partiels binaires ou ordinaux).

Pour ces différents modèles, le *package* `plsRglm` propose :

- d'ajuster des modèles de régression PLS univariée ou PLS généralisée (Bastien *et al.* (2005)) à des jeux de données complets ou incomplets,
- d'utiliser des versions pondérées des modèles PLS univariés (Haaland et Howland (1998)) et PLSGLR, une nouveauté,
- de mettre en œuvre, en utilisant différents critères d'évaluation des modèles, des validations croisées de type *k-fold* pouvant être doublement répétées mais aussi *leave-one-out* sur des jeux de données complets ou incomplets,

- d'appliquer des techniques de *bootstrap* (Lazraq *et al.* (2003) et Bastien *et al.* (2005)) pour déterminer des intervalles de confiance pour les prédicteurs d'origine, non seulement dans les cas PLS univariés mais dans le cas PLS-GLR, et ainsi d'évaluer leur significativité.

Avec Frédéric Bertrand, nous avons été les premiers à intégrer à un *package* de méthode PLS, la correction pour le calcul des degrés de liberté proposée par Kraemer et Sugiyama (2011) et implémentée dans le *package* `plsdof`.

Suite à la présentation du *package* à la conférence UseR! 2014, (Bertrand *et al.* (2014c)), le *package* a été intégré à l'offre de modélisation du *package* `caret`, Kuhn. (2018). Pour plus de détails sur les fonctionnalités du package, une prépublication Bertrand et Maumy-Bertrand (2018d) et une vignette détaillée a été rédigée Bertrand *et al.* (2014b). Le *package* est toujours activement maintenu et développé.

6.3 Détermination du nombre de composantes

Une des étapes clefs dans l'utilisation de la régression PLS est la détermination correcte du nombre de composantes du modèle. Il s'agissait d'un problème ouvert et important (Wiklund *et al.* (2007), Kraemer et Sugiyama (2011)). En effet, compte tenu du manque relatif d'hypothèses faites sur le modèle de régression PLS, qui fait de la PLS une approche de type *soft modelling*, (Manne (1987)), il n'est pas possible de développer des tests statistiques reposant sur des lois de probabilité connues pour tester les paramètres du modèle (Wakeling et Morris (1993)). L'approche généralement retenue est alors d'introduire et de comparer des critères numériques à l'aide de campagnes de simulations. Les critères qui ont été les plus mis en avant, pour cette sélection de modèle, sont basés sur la PRESS, introduite par Allen en 1971 (Allen (1971)). Pour être évaluée convenablement, cette statistique nécessite l'utilisation d'un jeu de données test indépendant du jeu de données d'apprentissage. Néanmoins, pour des raisons logistiques, ce jeu de données additionnel n'est que rarement disponible (Efron et Tibshirani, 1993, p. 240).

De ce fait, il est généralement d'usage de recourir à des techniques de validation croisée pour obtenir une estimation de statistiques fonction du PRESS. Or, des difficultés concernant la capacité de la validation croisée à déterminer le pouvoir prédictif des modèles, souvent liées à la grande variabilité des résultats obtenus, ont été mises en avant par (Efron et Tibshirani, 1993, p. 240), Wiklund *et al.* (2007), (Hastie *et al.*, 2009, p. 249), Boulesteix (2014). Lors de la rédaction de l'article Meyer *et al.* (2010) et de la vignette Bertrand *et al.* (2014b) qui impliquaient l'étude, sur des exemples variés, des propriétés de la PLSR et de la PLSGLR,

nous (Frédéric Bertrand et moi) nous sommes rendus compte que le critère du Q^2, pourtant reconnu comme étant le plus performant, posait ces problèmes et d'autres encore : propriétés peu étudiées en présence d'un grand nombre de variables bruitant le signal, comme c'est le cas pour un nombre de plus en plus important de jeux de données biologiques, par exemple génomiques ou protéomiques, ni dans le cas de la PLSGLR, ni en présence de valeurs manquantes. Pour ce dernier point, nous invitons le lecteur à lire la section 6.4 de ce chapitre.

Un sujet de thèse au LabEx IRMIA a été proposé sur ce problème de la détermination du nombre de composantes en régression PLSGLR. La proposition de sujet a été retenue et financée par le LabEx et Frédéric Bertrand a co-encadré cette thèse, le second encadrant étant le PU-PH Nicolas Meyer, pour assurer la majorité du temps l'encadrement mathématique de Jérémy Magnanensi. Je dois souligner ici que j'ai également apporté mon aide dans l'encadrement mais de façon moins intense que Frédéric Bertrand.

Nous avons proposé dans Magnanensi *et al.* (2016a) puis étendu dans Magnanensi *et al.* (2017) un nouveau critère d'arrêt pour déterminer le nombre de composantes en PLSR et en PLSGLR. Il est caractérisé par un grand niveau de stabilité (par rapport au ré-échantillonnage) et de robustesse (par rapport au bruit qui pourrait être présent dans les données). Ce nouveau critère est universel car il est approprié à la fois pour la PLSR et la PLSGLR. Il repose sur l'utilisation de techniques de *bootstrap* non paramétrique (Efron (1979)) et permet de tester l'intérêt de l'ajout de chaque composante supplémentaire au niveau α. La performance et la robustesse de ce critère ont été évaluées par simulations sur des jeux de données à n individus et p variables dans les cas où $n < p$ et $p < n$ avec différents niveaux de bruit, résiduel ou dans les variables explicatives qui constituent la matrice $\mathbf{X}$. La stabilité de ce critère a été évaluée en ré-échantillonnant un jeu de données réel. Un autre point important à souligner est que ce critère donne de meilleures performances que celles existantes dans les cas PLSR et PLSGLR.

Suite à cette étude, nous nous sommes intéressés, dans Magnanensi *et al.* (2016b), aux approches parcimonieuses en régression PLS qui ont récemment attiré beaucoup d'attention dans l'analyse des jeux de données génomiques de grande dimension. En effet, depuis le début des années 2000, des méthodes basées sur la régression par les moindres carrés partiels ont été développées pour effectuer une sélection de variables. La plupart de ces techniques reposent aussi sur le choix d'hyperparamètres, souvent déterminés par des méthodes basées sur la validation croisée, ce qui pose à nouveau d'importants problèmes de stabilité.

Pour surmonter cela, nous avons développé une nouvelle méthode dynamique, basée à nouveau sur le *bootstrap*, pour permettre la sélection des prédicteurs significatifs. Elle est adaptée à la fois à la régression PLS et à son extension aux mo-

dèles linéaires généralisés (GPLS). Elle repose sur l'établissement d'intervalles de confiance *bootstrap*, ce qui permet de tester la significativité des prédicteurs à un niveau de risque α fixé à l'avance, et évite l'utilisation de la validation croisée. Nous avons également développé des versions adaptatives de la régression PLS et de la GPLS parcimonieuse en intégrant le critère d'arrêt que nous avions introduit précédemment dans Magnanensi *et al.* (2017). Enfin, nous avons comparé la fiabilité et la stabilité de la sélection de variables, celles de la détermination des hyperparamètres du modèle, ainsi que leurs capacités prédictives, en utilisant des données simulées pour la PLS et des données réelles issues des expressions géniques de puces à ADN (*microarrays*) pour la classification logistique PLS.

Par rapport aux autres méthodes évaluées, notre nouvelle méthode dynamique présente la propriété de mieux séparer le bruit aléatoire, présent dans la réponse y, des informations pertinentes, conduisant à une meilleure précision et à une amélioration des capacités prédictives, en particulier en présence de niveaux de bruit non négligeables.

D'un point de vue théorique, une de nos objectifs serait d'obtenir une évaluation des degrés de liberté, dans les modèles issus de ces extensions de la régression PLS à la régression généralisée, à la manière de Kraemer et Sugiyama (2011), le cas de la régression PLS logistique ou de la régression PLS de Poisson semblant être les plus simples par lesquels commencer.

6.4 Influence des valeurs manquantes sur la sélection de composantes en PLS

Nous avons proposé à Titin Agustin Nengsih (étudiante indonésienne qui venait avec une bourse du gouvernement indonésien) un sujet de thèse sur l'influence des valeurs manquantes en régression PLS. Nous avions commencé à nous intéresser à la manière dont la PLS gérait les valeurs manquantes et s'il était nécessaire, et si oui dans quels cas, d'utiliser des approches plus sophistiquées. L'arrivée de Titin Agustin Nengsih nous a permis de reprendre ses recherches et de mettre en place une étude par simulation suffisamment conséquente pour aboutir à la rédaction de l'article Nengsih *et al.* (2019).
Les données manquantes (Little et Rubin (2002)) sont connues pour être un sujet de préoccupation pour la recherche appliquée, en particulier dans le domaine médical. Plusieurs méthodes ont été développées pour traiter des données incomplètes. La méthode d'imputation est le processus de substitution des données manquantes avant l'estimation des paramètres du modèle.
La régression PLS est un modèle multivarié pour lequel deux algorithmes (SIMPLS ou NIPALS) peuvent être utilisés pour fournir des estimations des para-

mètres. La régression PLS a été largement utilisée dans le domaine de la recherche en santé en raison de son efficacité pour analyser les relations entre la réponse et plusieurs composantes.
Toutefois, la gestion des valeurs manquantes lors de l'utilisation de la régression PLS fait toujours l'objet d'un débat. L'algorithme NIPALS a la propriété intéressante de pouvoir fournir des estimations à partir de jeux de données incomplets. La sélection du nombre de composantes pour créer un modèle approprié est une étape clef lors de la régression PLS. Plusieurs approches ont été proposées dans la littérature pour déterminer le nombre de composantes à inclure dans un modèle, tels que le critère Q^2, le critère d'information d'Akaike (AIC) ou le critère d'information bayésien (BIC). L'objectif de notre étude de simulations est d'analyser l'impact de la proportion de données manquantes sous l'hypothèse de données manquantes de type MCAR (*Missing Completely At Random*) et de type MAR (*Missing At Random*) sur l'estimation du nombre de composantes d'une régression PLS.
Nous avons comparé les critères de sélection du nombre de composantes d'une régression PLS sur des données incomplètes avec l'algorithme NIPALS (NIPALS-PLSR) et la régression PLS sur un jeu de données imputé en utilisant trois méthodes d'imputation : l'imputation multiple par des équations enchaînées (MICE, *Multivariate Imputation by Chained Equations* van Buuren et Groothuis-Oudshoorn (2011)), l'imputation par les k plus proches voisins (KNNimpute, Kowarik et Templ (2016)) et l'imputation basée sur la décomposition en valeurs singulières (SVDimpute, Perry (2015)). Les critères qui ont été comparés sont Q^2-LOO, Q^2-10-*fold*, AIC, AIC-DoF, BIC et BIC-DoF sur différentes proportions (allant de 5% à 50%) de données manquantes et selon le mécanisme MCAR ou MAR.

1. Les données ont été simulées d'après Li *et al.* (2002b). Le vrai nombre de composantes a été choisi égal à 2, 4 ou 6. Le nombre d'observations n et le nombre de variables p respectent les cinq configurations suivantes :

   - $n = 100$ et $p = 20$,
   - $n = 80$ et $p = 25$,
   - $n = 60$ et $p = 33$,
   - $n = 40$ et $p = 50$,
   - $n = 20$ et $p = 100$.

2. Les données manquantes sont créées sous l'hypothèse d'un mécanisme MCAR ou d'un mécanisme MAR avec un pourcentage de valeurs manquantes allant de 5% à 50% par pas de 5%.

3. Les valeurs manquantes sont imputées en utilisant les méthodes *MICE*, *KNNimpute* et *SVDimpute*.

4. Le nombre de composantes est choisi à l'aide d'une validation croisée *LOO* (Leave One Out) ou 10-*fold* calculée sur les données incomplètes à l'aide des deux méthodes standard et adaptative (qui sélectionne la méthode de prédiction en fonction de la présence de valeurs manquantes dans une ligne du tableau de données, Bertrand et Maumy-Bertrand (2020d)). Pour MICE, le nombre de composantes est le mode des nombres de composantes obtenus par validation croisée pour chacun des m jeux de données imputées où m est égal à 100 x la proportion de valeurs manquantes, White *et al.* (2011).

5. Nous avons aussi fixé à 8 le nombre maximal de composantes pouvant être extraites. Le vrai nombre de composantes est 2, 4 ou 6.

6. Pour chaque combinaison du nombre de vraies composantes, de la proportion de valeurs manquantes, de la configuration ligne-colonne et du mécanisme générateur des valeurs manquantes, 1000 réplicats ont été tirés.

L'étude par simulations a montré que :

- Le Q2-LOO affiche la meilleure performance quelles que soient les méthodes d'imputation. Les performances augmentent lorsque la taille de l'échantillon augmente et diminuent avec une proportion croissante de données manquantes.

- Le nombre de composantes sélectionnées par AIC, AIC-DoF et BIC est presque deux fois plus important que le nombre réel de composants.

- Le nombre réel de composantes d'une régression PLS est difficile à déterminer, en particulier pour un échantillon de petite taille et lorsque la proportion de données manquantes est supérieure à 30%.

- L'exécution de MICE a pris beaucoup de temps. Par exemple, lorsque $n = 100$ et que la proportion de données manquantes = 10%, la durée d'exécution de MICE était environ 11 fois plus lente que celle de NIPALS-PLSR.

Pour plus de détails, nous invitons le lecteur à consulter l'article Nengsih *et al.* (2019).
Les recherches présentées dans ce chapitre sont des extensions de la régression des moindres carrés partiels dans deux nouveaux cas :

- les réponses bornées, et dont le support est connu avant que l'expérience ne soit mise en œuvre,
- les données de survie.

Comme pour les contributions proposées au chapitre 3, je me suis intéressée au problème du choix du nombre de composantes ainsi qu'à celui de la sélection des variables.

6.5 Régression Bêta

6.5.1 Motivation

De nombreuses variables d'intérêt, comme par exemple des résultats expérimentaux, des rendements ou des indicateurs économiques, s'expriment naturellement sous la forme de taux, de proportions ou d'indices dont les valeurs sont nécessairement comprises entre zéro et un ou plus généralement entre deux valeurs non aléatoires et connues à l'avance. La régression Bêta permet de modéliser ces données avec beaucoup plus de souplesse puisque les fonctions de densité issues des lois Bêta peuvent prendre des formes très variées. En effet, l'intérêt pratique de la loi Bêta a été plusieurs fois affirmé par exemple par Johnson *et al.* (1995) : "*Beta distributions are very versatile and a variety of uncertainties can be usefully modelled by them. This flexibility encourages its empirical use in a wide range of applications.*" Plusieurs articles récents se sont intéressés à l'étude de la régression Bêta et à ses propriétés. L'article de Ferrari et Cribari-Neto (2004) mérite d'être mentionné comme introduction à ces modèles et ceux de Kosmidis et Firth (2010), Simas *et al.* (2010) et Grün *et al.* (2012) pour des extensions ou des améliorations des techniques d'estimation de ces modèles.

Toutefois, comme tous les modèles de régression, la régression Bêta ne peut s'appliquer directement lorsque les variables explicatives présentent des problèmes de multicolinéarité ou lorsqu'elles sont plus nombreuses que les unités statistiques observées. Ces situations se rencontrent fréquemment comme le savent tous les praticiens ou les utilisateurs de la statistique. Nous pouvons citer la biologie, la chimie, l'économie, la médecine. Pour contourner cette difficulté, nous avons proposé une extension de la régression PLS pour les modèles de régression Bêta. Celle-ci, ainsi que plusieurs outils, comme la validation croisée et des techniques *bootstrap*, est disponible dans le *package* `plsRbeta` du logiciel libre R. Ce *package* utilise la régression Bêta implémentée dans le *package* `betareg` (Cribari-Neto et Zeileis (2010)) du logiciel libre `R`.

La régression PLS, fruit de l'algorithme NIPALS initialement développée par Wold (1966) et exposée en détails par Tenenhaus (1998), avait été étendue avec succès aux modèles linéaires généralisés par Bastien *et al.* (2005) et aux modèles de Cox par Bastien (2008).

6.5.2 Bootstrap

Nous supposons avoir retenu le nombre m adéquat de composantes d'un modèle de régression Bêta PLS de Y sur $x_1, \ldots, x_j, \ldots, x_p$. Nous proposons l'algorithme suivant pour construire des intervalles de confiance et des tests de significativité pour les prédicteurs x_j, $1 \leqslant j \leqslant p$, à l'aide de techniques *bootstrap*.
Soit $\widehat{F}_{(T|Y)}$ la fonction de répartition empirique étant données la matrice T formée des m composantes PLS et la réponse Y.

Étape 1. Tirer B échantillons de $\widehat{F}_{(T|Y)}$.

Étape 2. Pour tout $b = 1, \ldots, B$, calculer :

$$c^{(b)} = (T^{(b)\prime}T^{(b)})^{-1}T^{(b)\prime}Y^{(b)} \quad \text{et} \quad b^{(b)} = W^* c'^{(b)},$$

où $[T^{(b)}, Y^{(b)}]$ est le b^{e} échantillon *bootstrap*, $c'^{(b)}$ est le vecteur des coefficients des composantes et $b^{(b)}$ est le vecteur des coefficients des p prédicteurs d'origine pour cet échantillon et enfin W^* est la matrice fixe des poids des prédicteurs dans le modèle d'origine comportant m composantes.

Étape 3. Pour chaque j, notons Φ_{b_j} l'approximation de Monte-Carlo de la fonction de répartition de la statistique *bootstrap* de b_j.
Pour chaque b_j, des boîtes à moustaches et des intervalles de confiance peuvent être construits à l'aide des percentiles de Φ_{b_j}. Un intervalle de confiance peut être défini par $I_j(\alpha) =]\Phi_{b_j}^{-1}(\alpha), \Phi_{b_j}^{-1}(1-\alpha)[$ où $\Phi_{b_j}^{-1}(\alpha)$ et $\Phi_{b_j}^{-1}(1-\alpha)$ sont les valeurs obtenues à partir de la fonction de répartition de la statistique *bootstrap* de telle sorte qu'un niveau nominal de confiance de niveau $100(1-2\alpha)\%$ soit atteint. Afin d'améliorer la qualité de l'intervalle de confiance en termes de taux de couverture, c'est-à-dire la capacité de $I_j(\alpha)$ à fournir les taux de couverture attendus, il est possible d'utiliser plusieurs techniques de construction : normale, percentile ou BC_a (Efron et Tibshirani (1993) ou Davison et Hinkley (1997)). Les intervalles ainsi obtenus ne sont pas conçus pour servir à réaliser des comparaisons multiples ou deux à deux et doivent être interprétés séparément.

6.5.3 Choix du nombre de composantes

Un problème crucial pour une utilisation correcte de la régression PLS est la détermination du nombre de composantes.
Le nombre de composantes PLS $\boldsymbol{t}_h$ peut être déterminé en régression PLS classique par validation croisée. Une composante $\boldsymbol{t}_h$ est ajoutée si le *PRESS* (*PRedicted Error Sum of Squares*) de l'étape h est nettement plus petit que le *RESS* (*Residual Sum of Squares*) de l'étape $h-1$. Wold propose dans le logiciel SIMCA (Eriksson *et al.* (2006)) d'introduire la h-ième composante si l'indice de Stone-Geisser

$$Q^2 = 1 - \frac{PRESS_h}{RESS_{h-1}}$$

est au moins égal à $0{,}0975$. Le même type d'approche a été introduit, dans Tenenhaus (1999) et Tenenhaus (2005) pour les extensions de la régression PLS à la régression logistique binaire et à la régression logistique ordinale. Dans le cas de la régression logistique binaire, elle repose sur l'utilisation du χ^2 de Pearson défini par :

$$\chi^2 = \sum_{i=1}^{n} \frac{(Y_i - \pi_i)^2}{\pi_i(1-\pi_i)}$$

où Y_i est la valeur de la variable Y pour l'individu i et π_i la probabilité de l'événement $\{Y = 1\}$ pour un individu ayant les caractéristiques de l'individu i. Le χ^2 de l'étape h peut être calculé par substitution en remplaçant π_i par son estimation à l'aide de la régression logistique sur les composantes $\boldsymbol{t}_1$, ..., $\boldsymbol{t}_h$. Il peut aussi être calculé par validation croisée en estimant π_i sans utiliser l'observation i, ou plus généralement en estimant π_{i_1}, ..., π_{i_k} sans utiliser les observations i_1, ..., i_k. Nous considérons que la composante $\boldsymbol{t}_h$ est significative si le χ^2 calculé à l'étape h par validation croisée est nettement inférieur au χ^2 calculé à l'étape $h-1$ par substitution. En reprenant l'approche de Wold, nous décidons que la composante $\boldsymbol{t}_h$ est significative si l'indice

$$Q^2 = 1 - \frac{\chi^2_{\text{validation croisée, étape } h}}{\chi^2_{\text{substitution, étape } h-1}}$$

est au moins égal à $0{,}0975$.
Plus généralement, une extension au cas des modèles linéaires généralisés à réponse univariée est possible en considérant que les densités f_i des réponses Y_i font parties d'une famille exponentielle uni-dimensionnelle :

$$f(y_i,\theta_i,\phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi/A_i} + c(y_i,\phi/A_i)\right)$$

où θ_i, A_i et ϕ sont des paramètres et les fonctions b et c sont connues. La valeur du χ^2 s'obtient alors à l'aide de la formule

$$\chi^2 = \phi \sum_{i=1}^{n} \frac{(Y_i - \mathbb{E}(Y_i))}{V(Y_i)}.$$

où Y_i est la variable aléatoire réponse Y pour l'individu i, $\mathbb{E}(Y_i)$ l'espérance de Y_i, $V(Y_i)$ sa variance et ϕ un paramètre de dispersion. Dans le contexte de la famille exponentielle uni-dimensionnelle, une expression plus simple du χ^2 ne met en jeu que les des dérivées d'ordre 1 et 2 de la fonction b et les paramètres θ_i et A_i.

$$\chi^2 = \sum_{i=1}^{n} \frac{(Y_i - b'(\theta_i))^2}{b''(\theta_i)/A_i}.$$

Le χ^2 de l'étape h peut être calculé par substitution en remplaçant θ_i par son estimation à l'aide de la régression généralisée sur les composantes $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_h$. Il peut aussi être calculé par validation croisée en estimant π_i sans utiliser l'observation i, ou plus généralement en estimant $\pi_{i_1}, \ldots, \pi_{i_k}$ sans utiliser les observations $i_1, \ldots, i_k$. On considère que la composante $\boldsymbol{t}_h$ est significative si le χ^2 calculé à l'étape h par validation croisée est nettement inférieur au χ^2 calculé à l'étape $h-1$ par substitution. En reprenant l'approche de Wold, nous décidons que la composante $\boldsymbol{t}_h$ est significative si l'indice

$$Q^2 = 1 - \frac{\chi^2_{\text{validation croisée, étape } h}}{\chi^2_{\text{substitution, étape } h-1}}$$

est au moins égal à $0{,}0975$.
Si, dans le cas de la régression PLS originale, le critère du Q^2 est extrêmement efficace Tenenhaus (1998), ses bonnes propriétés disparaissent malheureusement pour les modèles de régression linéaire généralisée PLS. Une étude par simulation s'impose donc afin de déterminer un critère fonctionnel pour choisir le nombre de composantes. Nous avons comparé les critères suivants AIC et BIC, Cribari-Neto et Zeileis (2010), χ^2 de Pearson, R^2 de Pearson et pseudo-R^2, Ferrari et Cribari-Neto (2004), ou critères $Q^2\chi^2$ et $Q^2\chi^2$ cumulé estimés par validation croisée en 5 groupes (5-CV) ou en 10 groupes (10-CV), Bastien *et al.* (2005).

L'algorithme utilisé pour créer les données simulées est une adaptation directe de l'algorithme de Li *et al.*, Li *et al.* (2002a) qui est lui-même une généralisation multivariée de celui de Naes et Martens, Naes et Martens (1985). Ce type de généralisation a déjà été utilisé avec succès dans le cas des modèles de régression logistique PLS, Meyer *et al.* (2010).

De manière générale, les résultats de l'étude par simulations montrent que le $Q^2\chi^2$ (5-CV et 10-CV), déjà connu pour son comportement surprenant en régression logistique PLS (Bastien *et al.* (2005), Meyer *et al.* (2010)), ne se comporte guère mieux pour les modèles de régression Bêta PLS. La maximisation des critères du R^2 ou du pseudo-R^2, s'avère également inefficace. Les critères AIC et BIC retiennent systématiquement quelques composantes de trop. Cette tendance est également connue dans le cas de la régression PLS traditionnelle (Kraemer et Sugiyama (2011)) comme dans celui de la régression logistique PLS (Meyer *et al.* (2010)).

6.5.4 Exemples d'application

Médecine

Les tumeurs cancéreuses représentent l'une des trois principales causes de mortalité dans le monde occidental. La compréhension des mécanismes des pathologies cancéreuses repose actuellement sur l'étude des relations mutuelles des anomalies génétiques acquises, apparaissant dans les tissus au cours du processus de la cancérisation. Ces anomalies sont fréquemment analysées par allélotypages, permettant de déterminer pour un nombre plus ou moins important de sites géniques, la présence ou non d'une modification du nombre de copies de chaque gène. La description multivariée de ces anomalies est informative sur le processus de cancérogénèse. Par ailleurs, l'ensemble de ces sites géniques porteur ou non d'anomalie peut être utilisé pour tenter de prédire certaines caractéristiques cliniques ou biologiques de la tumeur telles que le taux de cellules tumorales sur la biopsie d'une lésion. La modélisation dans un modèle statistique de taux, variable dont l'espace de variation est contenu dans l'intervalle fermé $[0; 1]$ comme variable prédite suggère l'utilisation d'une régression Bêta. Par ailleurs, les données d'allélotypage sont caractérisées par une fréquente colinéarité et par une proportion importante de données manquantes. De plus la matrice des données a souvent des dimensions $(i; j)$ telles que $j > i$, ce qui rend la matrice non-inversible, posant des difficultés dans l'ajustement d'un modèle de régression. La régression Bêta de type PLS que nous avons développée est donc particulièrement adaptée pour traiter les données d'allélotypage dans le contexte particulier de la prédiction d'une variable de type taux.
L'exemple, présenté dans Bertrand *et al.* (2013b), est celui de données d'allélotypage obtenues sur une série de 93 patients atteints de différents types de cancer du poumon et comportant $23{,}2\%$ de valeurs manquantes. La variable prédite est le taux de cellularité tumorale du prélèvement peropératoire de la tumeur. Les variables explicatives sont composées de 56 variables binaires indicatrices de la présence d'une anomalie sur chacun des 56 microsatellites et de trois variables

cliniques.
La sélection de variables est, dans cet exemple, très importante car elle permet de définir un sous-ensemble de prédicteurs, c'est-à-dire de sites géniques, capable de prédire le taux de cellules tumorales. En effet, les pathologies cancéreuses sont des pathologies génétiques acquises et certaines de ces anomalies sont la cause et d'autre la conséquence de la pathologie tumorale. Par ailleurs, l'information contenue dans les différents sites géniques microsatellites est potentiellement redondante. La sélection de variable, séparant les variables jouant probablement un rôle moteur dans le développement tumoral des variables ne faisant que traduire un bruit de fond aléatoire induit par des anomalies causées par ce développement tumoral, est alors une aide indispensable à la compréhension des mécanismes sous-jacents de la tumorogenèse.

Chimiométrie

L'objectif est de trouver des composés permettant de prédire le taux d'infiltration de patients en cellules cancéreuses à partir de données de spectrométrie. Un patient sain a 0% de cellules cancéreuses dans une biopsie tandis qu'un patient malade aura dans une biopsie un pourcentage d'autant plus élevé que l'échantillon contient des cellules cancéreuses. L'intérêt de cette expérience est d'essayer de réduire considérablement le temps d'analyse des biopsies en évitant d'avoir recours à un comptage par un médecin spécialiste en anatomie et cytologie pathologique. Une difficulté statistique supplémentaire apparaît dans l'exemple présenté dans Bertrand *et al.* (2013b) : il y a plus de variables (180) que d'individus (80). Plus de détails sur le protocole expérimental suivi sont disponibles dans l'article dans lequel ce jeu de donnés a déjà été initialement publié Piotto *et al.* (2012).

6.5.5 Bilan

Notre objectif a été de proposer une extension de la régression PLS aux modèles de régression Bêta, puis de la mettre à la disposition des utilisateurs du langage libre `R`. Nous offrons ainsi la possibilité de travailler, pour modéliser des taux ou des proportions, avec des prédicteurs colinéaires, difficulté inévitable dans le cas de la modélisation des mélanges ou lors de l'analyse de spectres, de l'étude de données génétiques, protéomiques ou métabolomiques.

De plus, la régression Bêta PLS peut être aussi appliquée à des jeux de données incomplets. Il est également possible dans ce cas, comme dans celui des données complètes, de sélectionner le nombre de composantes par validation croisée *repeated k-fold cross-validation*. Enfin, nous proposons des techniques *bootstrap* afin

de, par exemple, tester la significativité de chacun des prédicteurs présents dans le jeu de données et ainsi valider les modèles construits. L'étude de deux jeux de données réels a permis aux outils proposés de démontrer leur efficacité.

Plusieurs extensions ont été envisagées : approches robuste, parcimonieuse, inflation de zéro, réponses multivariées (Bry *et al.* (2013)), sélection du nombre de composantes à l'aide du critère d'arrêt introduit dans Magnanensi *et al.* (2017). Il faudrait également réaliser une étude de l'influence de la proportion et du type de valeurs manquantes similaire à celle de Nengsih *et al.* (2019).

6.6 Données de survie

6.6.1 Motivation et premiers résultats

Le point de départ de ces travaux est la volonté d'analyser un jeu de données original sur la prévision de la survie de patients atteints de cancer du côlon à l'aide de données d'allélotypage. L'allélotypage consiste à rechercher le statut normal ou altéré d'un ensemble prédéfini de microsatellites, séquence non-codante de l'ADN, dans une cellule cancéreuse. La survie devait être étudiée en fonction de 33 microsatellites simultanément, voir section 6.2, et du stade de cancer. Un des intérêts spécifique de la régression PLS, souvent apprécié dans l'étude des jeux de données génomiques ou protéomiques, est la détermination des composantes. Elles peuvent représenter des associations, ici entre microsatellites, qui seront interprétables par le biologiste ou le médecin typiquement comme l'altération d'une fonction.

Nous avons commencé par implémenter différents modèles existants à ce jour dont les plus récents venaient d'être présentés dans l'article de Bastien (2008).

- COX-PLS, un modèle de Cox sur des composantes créées à partir de la régression PLS de la durée de survie par rapport aux variables explicatives ;
- LASSO-LARS DR, un modèle de Cox par rapport aux variables explicatives sélectionnées lors de l'application du LASSO aux résidus de la déviance calculés pour un modèle de Cox sans variable explicative.
- PLSDR, un modèle de Cox par rapport aux composantes PLS sélectionnées lors de l'application d'une régression PLS aux résidus de la déviance calculés pour un modèle de Cox sans variable explicative.
- DKPLSDR, un modèle de Cox par rapport aux composantes PLS sélectionnées lors de l'application d'une régression Kernel PLS aux résidus de la

déviance calculés pour un modèle de Cox sans variable explicative.

Nous les avons comparés entre eux et à un modèle introduit entre temps et qui partageait la même finalité à savoir proposer des modèles de Cox en présence d'un très grand nombre de variables explicatives. En effet, l'article Kim *et al.* (2008) a permis à Sohn *et al.* (2009) de développer un modèle qui s'est montré particulièrement efficace : le « L^1 *penalized Cox PH model using the generalized lasso algorithm* ». La qualité prédictive des modèles avait été comparée à l'aide de courbes ROC adaptées aux données censurées introduites par Heagerty *et al.* (2000) et Heagerty et Zheng (2005).

Les deux jeux de données réels, qui ont servi à la comparaison, sont : le jeu de données d'allélotypage mentionné ci-dessus et un jeu de données comportant des observations recueillies à l'aide de puces à ADN. Ce dernier avait été introduit dans Alizadeh *et al.* (2000) et étudié par de nombreux autres auteurs, par exemple dans Rosenwald *et al.* (2002), Bastien (2008) et Sohn *et al.* (2009). Il a été obtenu en utilisant des « *lymphochips* » : des puces conçues pour les pathologies lymphoïdes et qui comprennent environ 18000 clones d'ADN complémentaires. Il sert généralement à comparer les capacités prédictives de modèles de survie dans le cas d'un patient atteint par un lymphome diffus à grandes cellules B. Ce jeu de données comporte un total de 240 patients atteints de LMNH-B dont 138 patients sont décédés pendant le suivi, durée médiane avant décès de $2{,}8$ ans, et 30% de temps de survie censurés à droite. Les « *lymphochips* » contiennent 7399 zones qui représentent 4128 gènes.

6.6.2 Approches parcimonieuses

Le second jeu de données a fait très nettement ressortir l'intérêt d'intégrer une étape de sélection de variable pour faciliter l'interprétation des résultats par la communauté médicale. C'est de ce point de départ qu'est partie l'idée de notre collaboration avec Philippe Bastien pour le développement de méthodes rapides de sélections de variable pour le modèle de Cox. En effet, estimer des modèles de régression PLS ou sPLS, éventuellement à noyaux, sur les résidus d'un modèle de Cox permet de combiner un temps d'exécution rapide, pour pouvoir traiter des jeux de données de grande dimension comportant des centaines ou des milliers de variables, avec toutefois une prise en compte de la nature censurée des données.

L'importance de ce type de jeux de données transparaît au travers d'une vaste littérature depuis les années 2000 qui est consacrée aux relations entre les profils géniques et le temps, pour un sujet, de survie ou de rechute de son cancer. La dé-

couverte de biomarqueurs à partir de données de grande dimension, telles que les profils transcriptomiques ou SNP, constitue un défi majeur dans la recherche de diagnostics plus précis. Le modèle de régression à risque proportionnel suggéré par Cox, 1972, pour étudier la relation entre le temps avant événement et un ensemble de covariables en présence de censure est le modèle le plus couramment utilisé pour l'analyse des données de survie.

Cependant, comme pour la régression multivariée, cela suppose qu'il existe plus d'observations que de variables, des données complètes et des variables non fortement corrélées. En pratique, lorsqu'il s'agit de données de grande dimension, ces contraintes ne sont pas vérifiées. La colinéarité engendre des problèmes de sur-ajustement et de mauvaise identification des modèles. La sélection de variables peut améliorer la précision de l'estimation en identifiant efficacement le sous-ensemble de prédicteurs pertinents et en améliorant l'interprétabilité du modèle avec une représentation parcimonieuse.

Depuis l'article de référence de Tibshirani (1997), de nombreuses méthodes basées sur les modèles à risques proportionnels de Cox pénalisés par *lasso* ont été proposées. La régularisation pourrait également être effectuée à l'aide d'une réduction de dimension, comme c'est le cas pour la régression moindres carrés partiels. Nous avons proposé deux algorithmes originaux nommés sPLSDR et son pendant non linéaire à noyau, DKsPLSDR, voir Rosipal et Trejo (2002), Tenenhaus *et al.* (2007) pour les approches à noyau dans d'autres contextes, en utilisant une régression PLS parcimonieuse (sPLS) basée sur les résidus de déviance. Nous avons comparé leurs performances en termes de prévision avec les meilleurs algorithmes disponibles à l'époque à l'aide d'une vaste campagne de simulations intégrant des jeux de données simulés et des jeux de données réels de référence.

Le résultat de cette campagne de simulation est que les deux nouvelles méthodes proposées sPLSDR et DKsPLSDR ont obtenu des résultats comparables voire meilleurs que les autres méthodes en termes de temps de calcul, de pouvoir prédictif et de sensibilité. De plus, de par leur nature même de régression PLS, elles offrent des possibilités additionnelles intéressantes comme les représentations en *biplot* ou la capacité de s'accommoder naturellement de valeurs manquantes.

Ces nouvelles méthodes, que nous avons considérées comme un ensemble d'outil pertinent pour les utilisateurs du très répandu modèle de Cox, ont été publiées dans Bastien *et al.* (2015).

Outre le fait de rendre accessible non seulement tous ces nouveaux développements méthodologiques aux utilisateurs de `R` mais aussi d'autres approches plus anciennes comme larsDR, coxPLS, PLScox, les points forts de l'implémentation du package `plsRcox`, Bertrand et Maumy-Bertrand (2018c), tiennent au fait que l'utilisateur a la possibilité de faire de la validation croisée, d'utiliser des techniques *bootstrap* et de travailler avec des jeux de données présentant des valeurs manquantes. Ce *package* a été présenté à la conférence internationale des utilisateurs du langage `R`, User ! 2014, Bertrand *et al.* (2014a).

Chapitre 7

Éléments de projet de recherche

7.1 Introduction

Comme je l'ai déjà mentionné dans le chapitre 1, depuis le début de mes travaux de recherche, j'ai abordé plusieurs problématiques statistiques aussi bien d'un point de vue théorique que d'un point de vue applicatif :

1. le processus empirique composé et en particulier la régression non paramétrique,

2. la régression pénalisée,

3. la régression des moindres carrés partiels et certaines de ses extensions.

Depuis quatre ans, j'ai ajouté une thématique d'apprentissage statistique aux trois précédentes. Mes principaux domaines d'application de mes travaux de recherche, comme nous avons pu le constater à travers les chapitres précédents sont essentiellement la biostatistique et la statistique industrielle. Les impulsions à l'origine de mes développements méthodologiques peuvent se classer en deux catégories :

1. soit dans un but d'améliorer ou de compléter une méthodologie existante mais qui a montré ses limites. Je redonne les exemples que j'ai déjà présentés dans le chapitre 1 :

   (a) traitement des valeurs manquantes et des problèmes de colinéarité en régression logistique ou pour les modèles de Cox,

   (b) influence des valeurs manquantes en régression des moindres carrés partiels,

(c) critère de choix de variables ou du nombre de composantes en régression des moindres carrés partiels et ses extensions modèles linéaires généralisés,

(d) cas des modèles parcimonieux.

2. soit même du à la nécessité de concevoir des outils spécifiques pour des expériences innovantes pour lesquelles il faut créer une solution faite sur mesure pour ce nouveau jeu de données. À nouveau, je mentionne les exemples que j'avais cités dans le chapitre 1 :

(a) l'inférence temporelle de réseaux de gènes, inférence conjointe de réseaux aussi bien au niveau population qu'au niveau individuel, détermination de cibles optimales pour prédire une intervention dirigée fiable dans un réseau).

J'essaye, dans la mesure du possible, d'assurer un aller-retour constant entre des développements généraux et des applications de ceux-ci dans l'un des projets interdisciplinaires auquel je participe. Dès le début de mes travaux de recherche, j'ai manifesté un intérêt pour les applications de la statistique et je ne vois pas cette discipline comme étant une discipline « hors-sol ».
Ainsi, très rapidement, je me suis rendue compte que les hypothèses communément faites pour permettre une exploitation scientifiquement rigoureuse des modèles statistiques ne sont qu'exceptionnellement compatibles avec des jeux de données réels, même s'ils sont collectés avec les meilleurs protocoles expérimentaux puisque les problèmes rencontrés tiennent généralement à la nature même des observations et non à la méthodologie de mesure. Or ce sont ces hypothèses qui permettent une gestion rigoureuse des risques d'erreur qui apparaissent dans la théorie des tests de significativité ou des niveaux de confiance présents lors de la construction d'intervalles de confiances. En outre, les séries statistiques ou les échantillons réels présentent souvent des problématiques additionnelles comme l'absence de certaines valeurs, présence de valeurs manquantes avec des mécanismes d'apparition plus ou moins complexes en fonction de l'appareil ou de la méthodologie de collecte des mesures, ou la présence de valeurs atypiques qui peuvent grandement influencer les outils statistiques utilisés.
Ainsi, après une première phase de développement d'un outil dans un cadre classique, typiquement celui de la régression parcimonieuse ou de la régression des moindres carrés partiels, je m'intéresse à étudier ses propriétés en présence de valeurs manquantes ou à en proposer une extension robuste.

7.2 Modélisation des courbes de croissance pour des fœtus à problème

Le Professeur Bernard Foliguet, fœtopathologiste au centre hospitalier régional universitaire de Nancy, a collecté pendant plusieurs années des données biométriques (poids du bébé, longueur vertex coccyx, longueur du pied du bébé, masse de certains organes du bébé) qui proviennent d'environ 600 fœtus humains récupérés au sein du secteur de placentologie et de fœtopathologie à la Maternité régionale de Nancy, secteur unique en France comme le mentionne l'article l'article de l'Est Républicain (Baret-Idatte (2013)).

Le Professeur Bernard Foliguet cherchait à comprendre pourquoi ces 600 fœtus qui ne présentaient aucune pathologie au niveau des échographies de contrôle ne sont pas arrivés à terme. Pour cela, il s'était penché sur les modèles polynomiaux d'ordre deux pour modéliser par exemple le poids en fonction de l'âge du fœtus qu'il peut dater afin de voir si des « ruptures » de pentes s'observaient au sein de la courbe de croissance.

Malheureusement, ces modèles sont trop généralistes et les méthodes statistiques avancées qu'envisageait le Professeur Bernard Foliguet n'étaient pas à sa portée de compréhension (je ne fais que citer ses paroles lors de notre première entrevue). C'est pourquoi il a fait appel d'abord à Sandie Ferrigno puis à moi. Avec Sandie Ferrigno, nous nous sommes intéressées dans un premier temps sur la modélisation de la masse du fœtus en fonction des semaines d'aménorrhée (voir section 3.4). Afin de pouvoir comparer les méthodes que nous voulions développer et pour pouvoir comparer les données obtenues par le Professeur Bernard Foliguet, nous avons eu la chance de rencontrer Barbara Heude (chargée de recherche dans l'équipe de recherche sur les déterminants précoces de la santé de l'Inserm) qui nous a permis l'accès à la base de données EDEN. Nous avons appliqué, au travers d'un projet de troisième année d'école d'ingénieurs et d'un stage de six mois de deuxième année de master de statistique de l'université de Strasbourg financé par un PEPS1 de l'AMIES, des méthodes paramétriques classiques qui nécessitaient certaines hypothèses qui malheureusement n'étaient pas toujours vérifiées. Nous avons donc ensuite envisagé de résoudre le problème à partir de méthodes non paramétriques, en particulier l'estimation polynomiale locale qui s'adapte bien à ce type de données (voir Ferrigno *et al.* (2011), Ferrigno *et al.* (2014) et Ferrigno *et al.* (2015)) et la thèse de Mint El Mouvid (Mint El Mouvid (2000)). Nous pouvons aussi mentionner l'article de Ledolter pour de plus amples informations (Ledolter (2013)).

En effet, elle permet, par le biais de l'utilisation d'un noyau symétrique par exemple et d'une fenêtre d'ajustement bien choisie, de ne considérer qu'une certaine proportion des données autour desquelles nous souhaitons réaliser l'estimation et de donner un poids plus important aux données qui se rapprochent du paramètre ciblé.
L'estimation de la fonction de répartition conditionnelle $F(y|x) = \mathbb{P}(Y \leqslant y|X = x)$ par des méthodes non paramétriques permet de capter toute l'information liant par exemple le poids fœtal à la semaine d'aménorrhée.
Cela permet ensuite d'estimer des courbes de régression, de variance et les quantiles associés à la modélisation recherchée ainsi que des bandes de confiance (voir Ferrigno *et al.* (2014) et Ferrigno *et al.* (2015)). Le travail d'estimation a pu être affiné grâce à l'étude du meilleur choix de noyau et de fenêtre afin d'obtenir des courbes qui reflètent au mieux la réalité.

Les pistes de recherche que Sandie Ferrigno et moi envisageons actuellement sont les suivantes :

- nous souhaiterions pouvoir appliquer les méthodes d'estimation des quantiles non paramétriques que nous avons développées à l'estimation des centiles d'un ensemble de mesures concernant les fœtus et les enfants ayant un poids de naissance dans EDEN (`http://eden.vjf.inserm.fr/index.php/fr/`) afin de comparer les courbes que nous avons obtenues à des courbes de référence déjà existantes obtenues avec des méthodes d'estimation différentes. Le but ultime est de voir si les ruptures de pente pressenties par le Professeur Bernard Foliguet sont bien présentes et si nous pourrions les détecter précocement, ce qui éviterait peut-être la mort du fœtus in utero.

- D'autre part, comme je le disais dans le chapitre3, une élève ingénieure de l'école des Mines de Nancy, Dounia Essaket, réalise son stage d'été de deuxième année sous notre direction afin de regrouper tous les résultats obtenus sur la base EDEN et sur la base du Professeur Bernard Foliguet à la fois à l'aide des méthodes paramétriques et non paramétriques pour les implémenter dans un *package* de `R` (Notre référence est le *package* `refcurv`).

- Enfin, une étude transversale pourrait être envisagée sur des mesures concernant par exemple la masse totale et la masse du placenta sur les données fœtales. Cela permettrait de construire un outil étiologique et d'identifier des fenêtres de variabilité afin de comprendre l'évolution de ces mesures concernant les données de l'étude du Professeur Bernard Foliguet.

7.3 Test A/B

- Nous avons observé avec les expérimentations de *CTREE-UCB* et *DBA-CTREE-UCB* que l'apport sur le gain moyen était remarquable comparé aux approches alternatives, limitant l'utilisation total des données en question (c'est-à-dire des covariables) et/ou imposant un temps de réponse inacceptable pour certaines problématiques industrielles. Si la garantie théorique sur le regret de *CTREE-UCB* a été montrée, celle de *DBA-CTREE-UCB* n'a pu être réalisée pendant la les travaux de thèse d'Emmanuelle Claeys. La transformation par *clustering* d'une covariable évolutive rend la preuve plus complexe. Pourtant, les recherches sur les garanties théoriques des algorithmes de *machine learning* proposent des pistes intéressantes (Bubeck (2010); Gentile *et al.* (2014)), notamment si la fonction décrivant la covariable évolutive est lipschitzienne (Magureanu *et al.* (2014)), ou encore à travers l'étude de *DTW* Cuturi et Blondel (2017). Ainsi, une vision future pourrait être de fournir une borne théorique sur le regret de *DBA-CTREE-UCB*.

- Le deuxième axe de recherche se pose sur la succession de tests A/B pour évaluer la présence de phénomène causaux. Si les approches causales classiques utilisent des réseaux bayésiens, les travaux de Pearl et Robbins (Robins (1987); Pearl (2009)) proposent d'intervenir sur le monde observé à travers une intervention contrefactuelle. Les expérimentations que nous avons menées avec Emmanuelle Claeys ont montré qu'une pré-segmentation avant de démarrer le test identifiait des groupes persistants dans le temps. Ces groupes sont plus ou moins sensibles au test. La mise en évidence de groupes insensibles au tests, mais interprétables par l'utilisateur[1] permet de le guider vers un nouveau test, exclusivement conçu pour ce groupe. Ainsi l'utilisateur continuera à faire des tests successifs jusqu'à l'identification d'une variation déclenchant un effet causal direct sur ce groupe. Cette approche se différencie de celles qui ont été proposées jusqu'à présent par (Bottou *et al.* (2013)).

- Le problème majeur qui freine l'utilisation des méthodes de bandit dans la pratique médicale est la réponse instantanée de la récompense après l'attribution d'une variation. Si cette contrainte ne pose aucun problème dans le *e-commerce*, elle se heurte à l'application médicale (ou l'observation d'une guérison prend plusieurs jours, voir plusieurs mois). Portées par la nécessité de tenir compte de cette contrainte, des méthodes proposent d'intégrer

[1] Nous entendrons par interprétable la description d'un groupe par un ensemble de caractéristiques précis et limité, en opposition aux modèles dits « boîte noire ».

un retard connu ou inconnu dans l'apprentissage du modèle, à travers par exemple, une descente de gradient pour débiaiser les estimateurs (Li *et al.* (2018)). D'autres approches pourraient cependant être utilisées, incluant de l'intelligence artificielle symbolique, comme les méthodes de logique floue. Ce terrain est encore peu exploré et peut être une approche intéressante.

7.4 Régression par les moindres carrés partiels

Au cours des dernières années, j'ai contribué, avec Frédéric Bertrand, de différentes manières à l'enrichissement des modèles de régression par les moindres carrés partiels (PLS). Considérons les variables centrées y, x_1, ..., x_j, ..., x_p. Soit X la matrice des prédicteurs x_1, ..., x_j, ..., x_p. La régression PLS est bien connue et décrite de manière exhaustive notamment par Höskuldsson (Höskuldsson (1988) et Wold *et al.* (Wold *et al.* (2001)). D'une part, la régression PLS est souvent présentée sous forme d'un algorithmique. D'autre part, la régression PLS est un modèle non linéaire qui permet de construire des composantes orthogonales notées souvent t_h obtenues en maximisant les quantités $cov(y,t_h)$. Soit T la matrice formée de ces composantes, nous avons :

$$y = T^t c + \epsilon, \tag{7.1}$$

où ϵ est le vecteur des résidus et ${}^t c$ le vecteur des coefficients des composantes, ^{t} désignant la transposée.
En posant $T = XW^*$, où W^* est la matrice des coefficients des variables x_j dans chaque composante t_h, nous avons l'expression directe de la réponse y à l'aide des prédicteurs x_j :

$$y = XW^{*t} c + \epsilon. \tag{7.2}$$

En développant le membre de droite de l'équation (7.2), nous obtenons pour chaque coordonnée y_i du vecteur y :

$$y_i = \sum_{h=1}^{H} \left(c_h w^*_{1h} x_{i1} + \cdots + c_h w^*_{ph} x_{ip} \right) + \epsilon_i, \tag{7.3}$$

H étant le nombre de composantes retenues dans le modèle final avec $H \leqslant \mathrm{rang}(X)$, H étant en général très inférieur au rang de la matrice X et p étant égal au nombre de prédicteurs contenus dans la matrice X. Les coefficients $c_h w^*_{jh}$, où $1 \leqslant j \leqslant p$, suivant la notation avec $*$ de Wold *et al.* (Wold *et al.* (2001)), traduisent la relation entre le vecteur réponse y et les variables x_j à travers les composantes

t_h.

Frédéric Bertrand et moi avons créé un *package* pour le langage `R`, `plsRglm` (Bertrand et Maumy-Bertrand (2020d); Bertrand *et al.* (2014c)) reprenant les extensions proposées par Bastien *et al.* (Bastien *et al.* (2005)). Nous avons également participé à un premier travail pour étudier ces modèles dans le cas des données d'allélotypages (Meyer *et al.* (2010)). Nous avons alors introduit de nouvelles extensions des modèles PLS, à la régression Bêta dans (Bertrand *et al.* (2013b); Bertrand et Maumy-Bertrand (2020f, 2018b)) et, avec Philippe Bastien, aux modèles de Cox dans (Bastien *et al.* (2015); Bertrand *et al.* (2014a); Bertrand et Maumy-Bertrand (2018c)).

L'objectif suivant a été de chercher des critères de choix du nombre de composantes, étape cruciale, pour ces modèles. Elle a été franchie avec Jérémy Magnanensi dans Magnanensi *et al.* (2016a), Magnanensi *et al.* (2017). Nous sommes actuellement en train de nous intéresser au cas des modèles PLS parcimonieux et je souhaite achever le travail débuté dans Magnanensi *et al.* (2016b).

Une étude plus complète de l'influence des valeurs manquantes en PLS linéaire a été menée avec Nengsih dans la publication Nengsih *et al.* (2019). Il serait pertinent de mener une étude similaire sur l'influence des valeurs manquantes dans le contexte de la régression PLS généralisée, typiquement pour les extensions GLM, Bêta et de Cox de la PLS.

J'aimerais maintenant formaliser des travaux que nous avons entrepris sur les méthodes à noyaux en PLS GLM, Bertrand *et al.* (2012b) et en PLS Bêta Bertrand *et al.* (2012a), puis développer des approches robustes pour les extensions GLM, Bêta et de Cox de la PLS afin d'obtenir des modèles qui limiteront l'influence des données atypiques.

D'un point de vue de l'implémentation, Frédéric Bertrand et moi aimerions ajouter l'utilisation de techniques parallèles et/ou GPU, ainsi qu'explorer la piste du langage Julia (Bezanson *et al.* (2017) pour permettre le traitement de jeux de données de taille plus importante.

7.5 Fouille de processus

Ce projet s'inscrit dans le cadre d'une collaboration de recherche entre la société *Your Data Consulting*, jeune entreprise innovante basée à Paris, Frédéric Bertrand et moi-même, sur la thématique de la science des données et plus particulière-

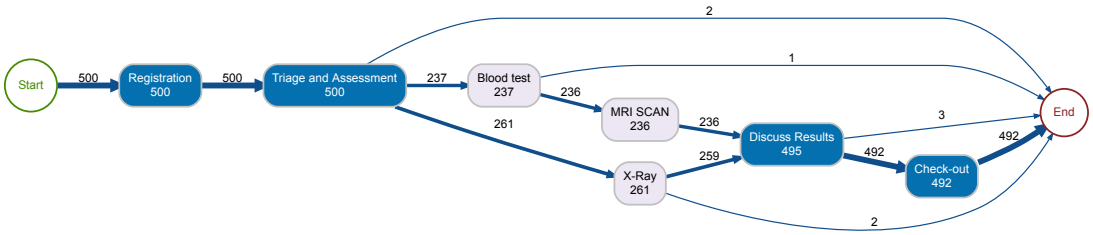


Figure 7.1 : *Visualisation d'un exemple de process.*

ment de l'intelligence artificielle.

Your Data Consulting est propriétaire de la plate-forme *SaaS LiveJourney* qui est présentée à l'adresse suivante : `https://www.livejourney.com`
Cette plate-forme permet aux entreprises de vente par correspondance ou aux entreprises de production ou celles de livraison de visualiser et d'analyser de façon dynamique les parcours de leurs clients, ou de leurs produits ou de leurs colis. De plus, elle permet aux entreprises d'anticiper des événements qui pourraient ralentir le bon déroulement du processus et de détecter les goulots d'étranglement.

Plusieurs thématiques ont été sélectionnées afin d'être particularisées au contexte du *process mining*. Elles ont commencé à faire l'objet d'un travail de recherche post-doctoral (débuté en décembre 2019 par Emmanuelle Claeys qui a été, depuis, recrutée comme maître de conférences à l'Université de Toulouse et qui commencera sa nouvelle fonction le 01/09/2020) et de recherches doctorales (un sujet de thèse pour Yoann Valero a été déposé afin d'obtenir un financement CIFRE par l'ANRT). Les voici :

- *root cause analysis*,
- prédiction des processus,
- *clustering* de séries temporelles,
- causalité et intelligence artificielle.

Un exemple de *process* est représenté à la Figure 7.1.

7.6 Exploitation des données cliniques en radiothérapie. Approches multicentriques et causales

Cet axe de recherches s'inscrit dans le contexte d'un financement d'une thèse de doctorat en mathématiques appliquées soutenue par un co-financement région Grand Est et de l'Université Technologique de Troyes, co-encadrée par Frédéric Bertrand, Professeur à l'Université Technologique de Troyes et moi-même. Un co-encadrement pour les aspects physique médicale et radiothérapie sera assuré par Christophe Mazzara, responsable du service physique médicale et radioprotection, membre du laboratoire des sciences de l'ingénieur, de l'informatique et de l'imagerie (ICube), UMR 7357, de l'Université de Strasbourg. Christophe Mazzara se fera accompagné par un physicien médical (Philippe Meyer) et par deux radiothérapeutes (Jean-Baptiste Clavier et Sébastien Guihard) comme mentionné au chapitre1.

Je présente encore une fois le résumé du sujet de thèse qui occupera Mélanie Piot à compter du mois de septembre 2020 après la fin de son stage de fin d'études d'ingénieur qu'elle réalise déjà dans le contexte de ce projet.
La radiothérapie externe est une des techniques utilisées au Centre Paul Strauss de Strasbourg pour la prise en charge des patients atteints d'un cancer. Elle consiste à irradier la tumeur au moyen d'un faisceau de rayons X de quelques méga volts, ce qui est mis en œuvre à l'aide de données dosimétriques en suivant un *Treatment Planning Systems*. L'évolution de la radiothérapie tend vers une personnalisation accrue des traitements, et donc une prise en compte massive des données disponibles. Il faudra pour cela enrichir les données dosimétriques par des données cliniques capturées lors des consultations médicales durant le traitement et au-delà.

Le département de radiothérapie du Centre Paul Strauss (Institut Régional du Cancer) s'est engagé sur cette voie depuis trois ans, et a mis en place une méthode de saisie des données cliniques au moyen d'un *Oncology Information System*. Environ 10 000 données cliniques structurées sont produites par mois et une base de données de plus de 1 600 000 données cliniques sont disponibles et exploitables à Strasbourg. Un déploiement multicentrique est en cours sur trois centres français. Cette base de données permet actuellement de répondre à des questions médicales précises. Au moyen d'un modèle prédictif simple, elle permet également d'identifier les patients qui présentent une dynamique singulière d'apparition d'une complication lors de son traitement. La prochaine étape que nous envisageons de franchir est la recherche de liens multiples entre les effets secondaires en utilisant et en développant des outils spécifiques d'extraction de

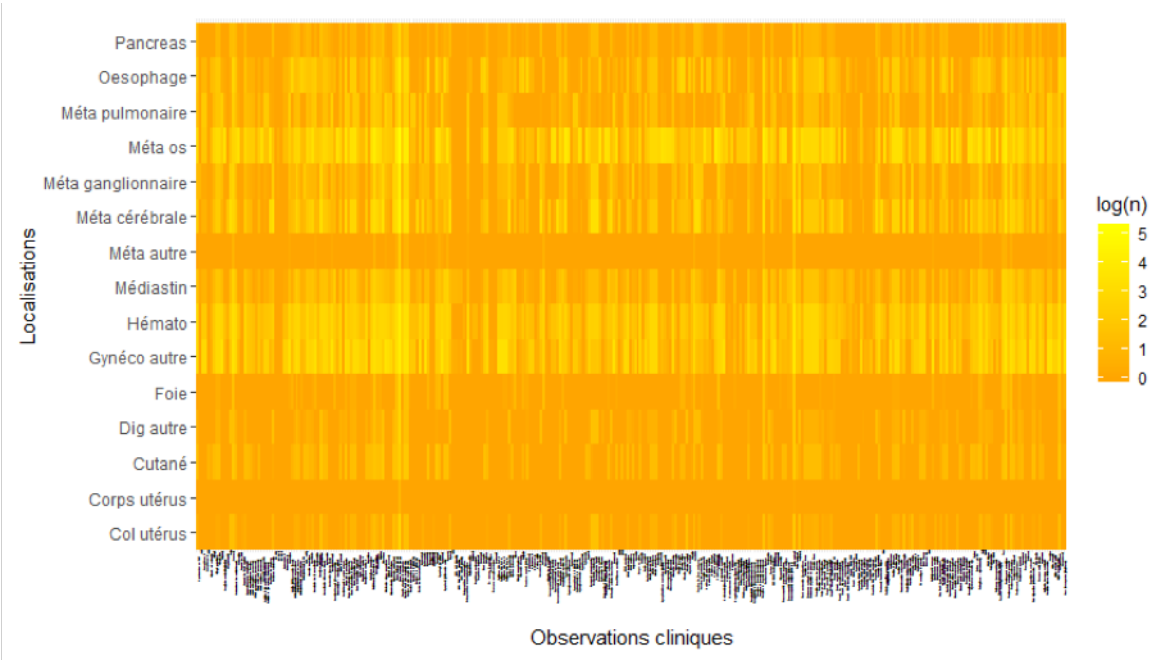


Figure 7.2 : *Cartographie des données disponibles.*

connaissance. C'est ce travail qui sera mené dans un contexte d'interdisciplinarité : deux mathématiciens modélisateurs, deux physiciens médicaux et deux radiothérapeutes. Ces derniers enrichiront l'analyse statistique par les relations entre les paramètres qui sont déjà connues dans la littérature médicale. La cartographie (voir Figure 7.2) donne le nombre de données disponibles par localisation anatomique et par complication clinique. Elle illustre la richesse de la base de données qui a été constituée par l'équipe de Christophe Mazzara et son potentiel prédictif.

Ce travail permettra de construire les premiers modèles causaux qui seront d'un intérêt premier pour la compréhension des phénomènes. Des premiers résultats ont été obtenus à l'aide des réseaux bayésiens lors du travail de stage de fin d'études de Mélanie Piot. Ils ont été accueillis très positivement par les médecins.

Bibliographie

Abidon, C. et Malblanc, S. [2018]. Etude des paramètres culturaux d'un réseau de parcelles de vigne en lien avec l'expression pluriannuelle d'Esca/BDA en Alsace. Dans *Les Maladies du Bois de la Vigne*. Institut Français de la Vigne et du Vin.

Abramowitz, M. et Stegun, I. A. [1972]. *Handbook of Mathematical Functions With Formulas, Graphs and Mathematical Tables - Tenth Printing*. Dover Publications Inc. ISBN 0486612724.

Akaike, H. et Akaïke, H. [1974]. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), 716–723. ISSN 0018-9286. URL `http://ieeexplore.ieee.org/document/1100705/`.

Alawieh, H., Bertrand, F., Maumy-Bertrand, M., Wicker, N. et Al Ayoubi, B. [2018]. A random model for multidimensional fitting method. URL `http://arxiv.org/abs/1810.05042`.

Alawieh, H., Wicker, N., Al Ayoubi, B. et Moulinier, L. [2017]. Penalized multidimensional fitting for protein movement detection. *Journal of Applied Statistics*, **44**(15), 2697–2715. ISSN 0266-4763. URL `https://www.tandfonline.com/doi/full/10.1080/02664763.2016.1261811`.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O. et Staudt, L. M. [2000]. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**(6769), 503–511. ISSN 0028-0836. URL `http://www.nature.com/articles/35000501`.

Allen, D. M. [1971]. *The prediction sum of squares as a criterion for selecting predictor variables*. Technical report 23 - Department of Statistics. University of Kentucky.

Anderson, T. W. [1946]. The Non-Central Wishart Distribution and Certain Problems of Multivariate Statistics. *The Annals of Mathematical Statistics*, **17**(4), 409–431. ISSN 0003-4851. URL http://projecteuclid.org/euclid.aoms/1177730882.

Aouadi, I., Jung, N., Carapito, R., Vallat, L., Bahram, S., Maumy-Bertrand, M. et Bertrand, F. [2018]. selectBoost : a general algorithm to enhance the performance of variable selection methods in correlated datasets. URL http://arxiv.org/abs/1810.01670.

Auer, P., Cesa-Bianchi, N. et Fischer, P. [2002]. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, **47**(2-3), 235–256. ISSN 0885-6125. URL https://doi.org/10.1023/A:1013689704352.

Bach, F. R. [2008]. Bolasso : model consistent lasso estimation through the bootstrap. Dans *Proceedings of the 25th international conference on Machine learning*, 33–40. ACM.

Bair, E., Hastie, T., Paul, D. et Tibshirani, R. [2006]. Prediction by Supervised Principal Components. *Journal of the American Statistical Association*, **101**(473), 119–137. ISSN 0162-1459. URL http://www.tandfonline.com/doi/abs/10.1198/016214505000000628.

Barabási, A.-L. [2003]. Emergence of scaling in complex networks. Dans Bornholdt, S. et Schuster, H. G., éditeurs. *Handbook of graphs and networks : from the genome to the internet*, 69–84. Wiley-VCH, Weinheim.

Baret-Idatte, C. [2013]. Nancy : Fœtus à la loupe. URL https://www.estrepublicain.fr/actualite/2013/03/01/foetus-a-la-loupe. Édition du 1er mars 2013.

Bastien, P. [2008]. Deviance residuals based PLS regression for censored data in high dimensional setting. *Chemometrics and Intelligent Laboratory Systems*, **91**(1), 78–86.

Bastien, P., Bertrand, F., Meyer, N. et Maumy-Bertrand, M. [2015]. Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data. *Bioinformatics*, **31**(3), 397–404. ISSN 14602059.

Bastien, P., Vinzi, V. E. et Tenenhaus, M. [2005]. PLS generalised linear regression. *Computational Statistics & Data Analysis*, **48**(1), 17–46. ISSN 01679473. URL http://www.sciencedirect.com/science/article/pii/S0167947304000271.

Bellusso, P., Maumy-Bertrand, M., Desnos, Y. et Segond, H. [2014]. Intérêts de la psychothérapie à médiation sensorielle dans le cadre de la prise en charge des troubles de la relation et de la communication chez des enfants autistes

sévèrement déficitaires II : Illustration clinique. *Neuropsychiatrie de l'Enfance et de l'Adolescence*, **62**, 95–101. ISSN 02229617.

Berge, C., Froloff, N., Kalathur, R. K. R., Maumy, M., Poch, O., Raffelsberger, W. et Wicker, N. [2010]. Multidimensional fitting for multivariate data analysis. *Journal of computational biology : a journal of computational molecular cell biology*, **17**(5), 723–32. ISSN 1557-8666. URL `http://www.ncbi.nlm.nih.gov/pubmed/20175691`.

Bertrand, F., Bastien, P., Meyer, N. et Maumy-Bertrand, M. [2014a]. plsRcox, Cox-Models in a high dimensional setting in R. Dans *Proceedings of User2014 !, Los Angeles*, 152.

Bertrand, F., Droesbeke, J.-J., Saporta, G. et Thomas-Agnan, C., éditeurs [2017]. *Model Choice and Model Aggregation*. Technip, Paris.

Bertrand, F., Fredon, D. et Maumy-Bertrand, M. [2016]. *Mathématiques Licence 1 - Exercices et méthodes*. Dunod, Paris.

Bertrand, F., Fredon, D., Rabba-Idi, Y. et Maumy-Bertrand, M. [2018]. *Mathématiques Licence 2 - Exercices et méthodes*. Dunod, Paris.

Bertrand, F., Magnanensi, J., Meyer, N. et Maumy-Bertrand, M. [2014b]. *plsRglm : Algorithmic insights and applications*. Vignette of the package.

Bertrand, F., Magnanensi, J., Meyer, N. et Maumy-Bertrand, M. [2014c]. plsRglm, PLS generalized linear models for R. Dans *Proceedings of User2014 !, Los Angeles*, 150.

Bertrand, F., Maumy, M., Fussler, L., Kobes, N., Savary, S. et Grosman, J. [2008]. Étude statistique des données collectées par l'Observatoire National des Maladies du Bois de la Vigne. *Journal de la Société Française de Statistique*, **149**(4), 73–106.

Bertrand, F. et Maumy, M. [2007]. Développements d'Edgeworth de deux estimateurs d'une proportion de mesures. *Comptes Rendus Mathematique*, **345**(7), 399–404. ISSN 1631073X.

Bertrand, F. et Maumy, M. [2010]. Using Partial Triadic Analysis for Depicting the Temporal Evolution of Spatial Structures : Assessing Phytoplankton Structure and Succession in a Water Reservoir. *Case Studies In Business, Industry And Government Statistics*, **4**(1), 23–43. ISSN 2152-372X. URL `http://journal-sfds.fr/index.php/csbigs/article/view/286`.

Bertrand, F., Maumy, M., Fussler, L., Kobes, N., Savary, S. et Grosman, J. [2007]. Using Factor Analyses to explore data generated by the National Grapevine Wood Diseases Survey. *Case Studies in Business, Industry and Government Statistics*, **1**(2). URL `http://hal.archives-ouvertes.fr/hal-00166970`.

Bertrand, F. et Maumy-Bertrand, M. [2011]. *Maxi fiches de Statistique. En 80 fiches*. Dunod, Paris.

Bertrand, F. et Maumy-Bertrand, M. [2012]. *Mathématiques : Concours des catégories A et B*. Dunod, Paris.

Bertrand, F. et Maumy-Bertrand, M. [2018a]. *Initiation à la statistique avec R : Cours, exemples, exercices et problèmes corrigés*. Dunod, Paris, 3e édition.

Bertrand, F. et Maumy-Bertrand, M. [2018b]. *Partial Least Squares Regression for Beta Regression Models*. URL `http://www-irma.u-strasbg.fr/~fbertran/`. R package version 0.2.3.

Bertrand, F. et Maumy-Bertrand, M. [2018c]. *Partial Least Squares Regression for Cox Models and Related Techniques*. URL `http://www-irma.u-strasbg.fr/~fbertran/`. R package version 1.7.3.1.

Bertrand, F. et Maumy-Bertrand, M. [2018d]. plsRglm : Partial least squares linear and generalized linear regression for processing incomplete datasets by cross-validation and bootstrap techniques with R. URL `http://arxiv.org/abs/1810.01005`.

Bertrand, F. et Maumy-Bertrand, M. [2020a]. Deciphering Biological Networks with Patterned Heterogeneous (multiOmics) Measurements. Dans *Proceedings of User2020 !, St Louis, cancelled*.

Bertrand, F. et Maumy-Bertrand, M. [2020b]. *Network Reverse Engineering with Approximate Bayesian Computation*. URL `https://CRAN.R-project.org/package=networkABC`. R package version 0.7-0.

Bertrand, F. et Maumy-Bertrand, M. [2020c]. networkABC : Network Reverse Engineering with Approximate Bayesian Computation. Dans *Proceedings of User2020 !, St Louis, cancelled*.

Bertrand, F. et Maumy-Bertrand, M. [2020d]. *Partial Least Squares Regression for Generalized Linear Models*. URL `https://CRAN.R-project.org/package=plsRglm`. R package version 1.2.5.

Bertrand, F. et Maumy-Bertrand, M. [2020e]. *Patterns : Deciphering Biological Networks with Patterned Heterogeneous Measurements*. URL `https://CRAN.R-project.org/package=Patterns`. R package version 1.2.

Bertrand, F. et Maumy-Bertrand, M. [2020f]. plsRbeta : Partial Least Squares Regression for Beta Regression Models. Dans *Proceedings of User2020 !, St Louis, cancelled*.

Bertrand, F., Maumy-Bertrand, M. et Aouadi, I. [2020]. *selectboost : A General Algorithm to Enhance the Performance of Variable Selection Methods in Correlated*

Datasets. URL `https://CRAN.R-project.org/package=SelectBoost`. R package version 2.0.0.

Bertrand, F., Maumy-Bertrand, M., Ferrigno, S., Muller-Gueudin, A. et Marx, D. [2013a]. *Mathématiques pour les sciences de l'ingénieur - Tout le cours en fiches*. Dunod, Paris. ISBN 9782100570614. URL `http://www.dunod.com/sciences-techniques/sciences-fondamentales/mathematiques/licence/mathematiques-pour-les-sciences-de-lingenieur-tout-le-c`.

Bertrand, F., Maumy-Bertrand, M. et Meyer, N. [2012a]. Kernel pls beta regressions. Dans *Proceedings of CAC 2012, Paris, France*.

Bertrand, F., Maumy-Bertrand, M. et Meyer, N. [2012b]. Kernel pls glm regressions. Dans *Proceedings of ENBIS 2012, Ljublana, Slovenia*.

Bertrand, F., Maumy-Bertrand, M. et Périnel, E. [2011]. *Économétrie, Statistiques et Probabilités : Concours des catégories A et B*. Dunod, Paris.

Bertrand, F., Meyer, N., Beau-Faller, M., Bayed, K. E., Izzie-J., N. et Maumy-Bertrand, M. [2013b]. Régression Bêta PLS [PLS Beta regression]. *Journal de la Société Française de Statistique*, **154**(3), 143–159.

Bertrand, F., Saporta, G. et Thomas-Agnan, C., éditeurs [2019a]. *Statistique et Causalité*. Technip, Paris. à paraître.

Bertrand, F., Claeys, E. et Maumy-Bertrand, M. [2019b]. *Modélisation statistique par la pratique avec R*. Dunod, Paris. ISBN 9782100793525.

Bezanson, J., Edelman, A., Karpinski, S. et Shah, V. B. [2017]. Julia : A fresh approach to numerical computing. *SIAM review*, **59**(1), 65–98. URL `https://doi.org/10.1137/141000671`.

Bottou, L., Peters, J., nonero Candela, J. Q., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P. et Snelson, E. [2013]. Counterfactual reasoning and learning systems : The example of computational advertising. *Journal of Machine Learning Research*, **14**, 3207–3260. URL `http://jmlr.org/papers/v14/bottou13a.html`.

Boulanger, B., Dewé, W., Gilbert, A., Govaerts, B. et Maumy, M. [2007]. Risk management for analytical methods based on the total error concept : Conciliating the objectives of the pre-study and in-study validation phases. *Chemometrics and Intelligent Laboratory Systems*, **86**(2), 198–207.

Boulesteix, A.-L. [2014]. Accuracy estimation for PLS and related methods via resampling-based procedures. Dans *PLS'14 Book of Abstracts*, 13–14.

Bourgon, R., Gentleman, R. et Huber, W. [2010]. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National*

Academy of Sciences of the United States of America, **107**(21), 9546–51. ISSN 1091-6490. URL http://www.pnas.org/content/107/21/9546.long.

Bourjade, M., Call, J., Pelé, M., Maumy, M. et Dufour, V. [2014]. Bonobos and orangutans, but not chimpanzees, flexibly plan for the future in a token-exchange task. *Animal Cognition*, **17**(6), 1329–1340. ISSN 1435-9448. URL http://link.springer.com/10.1007/s10071-014-0768-6.

Bourjade, M., Thierry, B., Maumy, M. et Petit, O. [2009]. Decision-making in przewalski horses (equus ferus przewalskii) is driven by the ecological contexts of collective movements. *Ethology*, **115**(4), 321–330. ISSN 14390310.

Box, G. E. P. et Cox, D. R. [1964]. An Analysis of Transformations. *Journal of the Royal Statistical Society : Series B (Methodological)*, **26**(2), 211–243. ISSN 00359246. URL http://doi.wiley.com/10.1111/j.2517-6161.1964.tb00553.x.

Bry, X., Trottier, C., Verron, T. et Mortier, F. [2013]. Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm. *Journal of Multivariate Analysis*, **119**, 47–60. ISSN 0047259X. URL http://www.sciencedirect.com/science/article/pii/S0047259X13000407.

Bubeck, S. [2010]. *Bandits Games and Clustering Foundations*. Theses, Université des Sciences et Technologie de Lille - Lille I. URL https://tel.archives-ouvertes.fr/tel-00845565.

Carapito, R., Jung, N., Kwemou, M., Untrau, M., Michel, S., Pichot, A., Giacometti, G., Macquin, C., Ilias, W., Morlon, A., Kotova, I., Apostolova, P., Schmitt-Graeff, A., Cesbron, A., Gagne, K., Oudshoorn, M., Van Der Holt, B., Labalette, M., Spierings, E., Picard, C., Loiseau, P., Tamouza, R., Toubert, A., Parissiadis, A., Dubois, V., Lafarge, X., Maumy-Bertrand, M., Bertrand, F., Vago, L., Ciceri, F., Paillard, C., Querol, S., Sierra, J., Fleischhauer, K., Nagler, A., Labopin, M., Inoko, H., Von Dem Borne, P., Kuball, J., Ota, M., Katsuyama, Y., Michallet, M., Lioure, B., De Latour, R., Blaise, D., Cornelissen, J., Yakoub-Agha, I., Claas, F., Moreau, P., Milpied, N., Charron, D., Mohty, M., Zeiser, R., Socié, G. et Bahram, S. [2016]. Matching for the nonconventional MHC-I MICA gene significantly reduces the incidence of acute and chronic GVHD. *Blood*, **128**(15). ISSN 15280020.

Carapito, R., Aouadi, I., Pichot, A., Spinnhirny, P., Morlon, A., Kotova, I., Macquin, C., Rolli, V., Cesbron, A., Gagne, K., Oudshoorn, M., van der Holt, B., Labalette, M., Spierings, E., Picard, C., Loiseau, P., Tamouza, R., Toubert, A., Parissiadis, A., Dubois, V., Paillard, C., Maumy-Bertrand, M., Bertrand, F., von dem Borne, P. A., Kuball, J. H. E., Michallet, M., Lioure, B., Peffault de Latour, R., Blaise, D., Cornelissen, J. J., Yakoub-Agha, I., Claas, F., Moreau, P.,

Charron, D., Mohty, M., Morishima, Y., Socié, G. et Bahram, S. [2020]. Compatibility at amino acid position 98 of MICB reduces the incidence of graft-versus-host disease in conjunction with the CMV status. *Bone Marrow Transplantation*. ISSN 0268-3369. URL http://www.nature.com/articles/s41409-020-0886-5.

Carbonetto, P. et Stephens, M. [2012]. Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies. *Bayesian Analysis*, **7**(1), 73–108. ISSN 1936-0975. URL http://projecteuclid.org/euclid.ba/1339616726.

Chen, L., Goldgof, D. B., Hall, L. O. et Eschrich, S. A. [2007]. Noise-Based Feature Perturbation as a Selection Method for Microarray Data. Dans Măndoiu, I. et Zelikovsky, A., éditeurs. *Bioinformatics Research and Applications. ISBRA 2007*, Lecture Notes in Computer Science, vol 4463, 237–247, Berlin, Heidelberg. Springer Berlin Heidelberg. URL http://link.springer.com/10.1007/978-3-540-72031-7_22.

Chen, S. S., Donoho, D. L. et Saunders, M. A. [2001]. Atomic Decomposition by Basis Pursuit. *SIAM Review*, **43**(1), 129–159. ISSN 0036-1445. URL http://epubs.siam.org/doi/10.1137/S003614450037906X.

Cheung, L. W.-K. [2012]. Classification Approaches for Microarray Gene Expression Data Analysis. 73–85. Humana Press. URL http://link.springer.com/10.1007/978-1-61779-400-1{_}5.

Chu, W., Li, L., Reyzin, L. et Schapire, R. [2011]. Contextual bandits with linear payoff functions. Dans Gordon, G., Dunson, D. et Dudík, M., éditeurs. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 de *Proceedings of Machine Learning Research*, 208–214, Fort Lauderdale, FL, USA. PMLR.

Chun, H. et Keleş, S. [2010]. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, **72**(1), 3–25. ISSN 1369-7412. URL http://www.ncbi.nlm.nih.gov/pubmed/20107611.

Claeys, E., Gançarski, P. et Maumy-Bertrand, M. [2018]. Approche contextuelle par régression pour les tests A/B. Dans *EGC*, Paris, France.

Claeys, E., Gançarski, P. et Maumy-Bertrand, M. [2020a]. Intégration des séries temporelles dans les A/B-Tests. Dans *EGC*, 121—-132, Bruxelles, Belgique. RNTI, Volume E-36.

Claeys, E. [2019a]. *Clusterisation incrémentale, multicritères de données hétérogènes pour la personnalisation d'expérience utilisateur.* (*Incremental, multi-criteria clustering of heterogeneous data for user experience customization*). Thèse de doctorat,

University of Strasbourg, France. URL https://tel.archives-ouvertes.fr/tel-02525206.

Claeys, E. [2019b]. *Incremental, multi-criteria clustering of heterogeneous data for user experience customization*. Theses, Université de strasbourg. URL https://tel.archives-ouvertes.fr/tel-02435605.

Claeys, E., Gançarski, P., Maumy-Bertrand, M. et Wassner, H. [2017]. Regression Tree for Bandits Models in A/B Testing. Dans Adams, N., Tucker, A. et Weston, D., éditeurs. *Advances in Intelligent Data Analysis XVI. IDA 2017*, Lecture Notes in Computer Science, vol 10584, 52–62, Cham. Springer. URL http://link.springer.com/10.1007/978-3-319-68765-0_5.

Claeys, E., Gançarski, P. et Maumy-Bertrand, M. [2019]. Dynamic allocation optimization in A/B tests using classification-based preprocessing. Dans *Use r!*

Claeys, E., Maumy-Bertrand, M. et Bottou, L. [2020b]. Causalité et apprentissage automatique. Dans Bertrand, F., Saporta, G. et Thomas-Agnan, C., éditeurs. *Causalité et statistique*. Technip édition.

Clauset, A., Newman, M. E. J. et Moore, C. [2004]. Finding community structure in very large networks. *Physical Review E*, **70**(6), 066111. ISSN 1539-3755. URL https://link.aps.org/doi/10.1103/PhysRevE.70.066111.

Cole, T. J. [1988]. Fitting Smoothed Centile Curves to Reference Data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **151**(3), 385. ISSN 09641998. URL https://www.jstor.org/stable/10.2307/2982992?origin=crossref.

Cole, T. J. [1990]. The LMS method for constructing normalized growth standards. *European Journal of Clinical Nutr.*, **44**, 45–60.

Cook, J. R. et Stefanski, L. A. [1994]. Simulation-Extrapolation Estimation in Parametric Measurement Error Models. *Journal of the American Statistical Association*, **89**(428), 1314–1328. ISSN 0162-1459. URL https://www.tandfonline.com/doi/full/10.1080/01621459.1994.10476871.

Couégnas, A. [2011]. *Expression de protéine stop et schizophrénie*. Thèse de doctorat, université de Strasbourg.

Cribari-Neto, F. et Zeileis, A. [2010]. Beta Regression in R. *Journal of Statistical Software*, **34**(2), 1–24.

Cuturi, M. et Blondel, M. [2017]. Soft-dtw : a differentiable loss function for time-series.

Davison, A. C. et Hinkley, D. V. [1997]. *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge.

de Onis, M. [2007]. Development of a WHO growth reference for school-aged children and adolescents. *Bulletin of the World Health Organization*, **85**(09), 660–667. ISSN 00429686. URL http://www.who.int/bulletin/volumes/85/9/07-043497.pdf.

Dettling, M. [2004]. BagBoosting for tumor classification with gene expression data. *Bioinformatics*, **20**(18), 3583–3593. ISSN 1367-4803. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bth447.

Deville, J.-C., Lavallee, P. et Maumy, M. [2005]. Composition, factorisation et conditions d'optimalité (faible, forte) dans la méthode de partage des poids. Applications à l'enquête sur le tourisme en Bretagne. Dans *Journées de Méthodologie Statistique de l'INSEE 2005*, Paris, France. INSEE.

Deville, J.-C. et Maumy, M. [2004a]. A new survey methodology for describing tourism activities and expanses. *Estudios Turísticos*. ISSN 0423-5037.

Deville, J.-C. et Maumy, M. [2004b]. A new survey methodology for describing tourism activities and expanses. Dans *7th International Forum on Tourism Statistics*, Stockholm.

Deville, J.-C. et Maumy, M. [2005]. Extensions de la méthode d'échantillonnage indirect et son application à l'enquête dans le tourisme : M.O.R.G.O.A.T. Dans *Journées de Méthodologie Statistique de l'INSEE 2005*, Paris, France.

Deville, J.-C. et Maumy, M. [2006]. La méthodologie de Morgoat : Enquête tourisme en Bretagne. Dans Lavallée, P. et Rivest, L., éditeurs. *Méthodes d'enquêtes et sondages Pratiques européenne et nord-américaine. Actes du Colloque francophone sur les sondages 2005*, 393–398, Paris. Dunod.

Deville, J.-C. et Maumy-Bertrand, M. [2006]. Extensions de la méthode d'échantillonnage indirect et son application aux enquêtes dans le tourisme. *Techniques d'enquête*, **32**(2), 197–206.

Di Camillo, B., Toffolo, G. et Cobelli, C. [2009]. A gene network simulator to assess reverse engineering algorithms. *Annals of the New York Academy of Sciences*, **1158**, 125–42. ISSN 1749-6632. URL http://www.ncbi.nlm.nih.gov/pubmed/19348638.

Donoho, D. L. et Elad, M. [2003]. Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization. *Proceedings of the National Academy of Sciences*, **100**(5), 2197–2202. ISSN 0027-8424. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0437847100.

Douglas Nychka, Reinhard Furrer, John Paige et Stephan Sain [2017]. fields :

Tools for spatial data. URL https://github.com/NCAR/Fields. R package version 10.3.

Doyen, V., Poirot, A., Maumy-Bertrand, M., Domis, N., Jacob, A., Khayath, N., Corraza, F. et De Blay, F. [2020]. Treg response in mite asthma after allergen exposure. Dans *EAACI Digital Congress 2020*.

Efron, B. [1979]. Bootstrap methods : another look at the jackknife. *The Annals of Statistics*, **7**(1), 1–26.

Efron, B., Hastie, T., Johnstone, I. et Tibshirani, R. [2004]. Least angle regression. *The Annals of Statistics*, **32**(2), 407–499. ISSN 0090-5364. URL http://projecteuclid.org/euclid.aos/1083178935.

Efron, B. et Tibshirani, R. J. [1993]. *An introduction to the bootstrap*, volume 57. Chapman & Hall/CRC, 2000 N.W. Corporate Blvd, Boca Raton, Florida 33431, US.

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

Eklund, M. et Zwanzig, S. [2012]. SimSel : a new simulation method for variable selection. *Journal of Statistical Computation and Simulation*, **82**(4), 515–527. ISSN 0094-9655. URL http://www.tandfonline.com/doi/abs/10.1080/00949655.2010.543981.

Ériksson, L., Byrne, T., Johansson, E., Trygg, J. et Wikström, C. [2006]. *Multi- and Megavariate Data Analysis Basic Principles and Applications*. Umetrics, Umeå, troisième édition. ISBN 978-91-973730-5-0. URL https://webshop.umetrics.com/products/multi-and-megavariate-data-analysis-basic-principles-and-applications-third-revised-edition.

Fan, J. [1997]. Comments on «Wavelets in statistics : A review» by A. Antoniadis. *Journal of the Italian Statistical Society*, **6**(2), 131–138. ISSN 1121-9130.

Fan, J. et Li, R. [2006]. Statistical challenges with high dimensionality : feature selection in knowledge discovery. Dans *International Congress of Mathematicians, ICM 2006*, volume 3, 595–622, Zürich. European Math. Soc.

Fan, J. et Lv, J. [2010]. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, **20**(1), 101–148. ISSN 10170405.

Ferrari, S. L. P. et Cribari-Neto, F. [2004]. Beta Regression for Modeling Rates and Proportions. *Journal of Applied Statistics*, **31**(7), 799–815.

Ferrigno, S., Foliguet, B., Maumy-Bertrand, M. et Muller-Gueudin, A. [2015]. Certainty bands for the conditional cumulative distribution function and ap-

plications. CMStatistics 2015. URL `https://hal.inria.fr/hal-01236952`. Poster.

Ferrigno, S., Foliguet, B., Maumy-Bertrand, M. et Muller-Gueudin, A. [2014]. Certainty bands for the conditional cumulative distribution function and applications. URL `https://hal.archives-ouvertes.fr/hal-01025443`. 25 pages.

Ferrigno, S. et Maumy-Bertrand, M. [2019]. Estimation of reference curves for fetal weight. Dans *CFE-CMStatistics 2019 Book of Abstracts*, 170. ECOSTA ECONOMETRICS AND STATISTICS.

Ferrigno, S., Maumy-Bertrand, M. et Muller, A. [2011]. Uniform law of the logarithm for the conditional distribution function and application to certainty bands. URL `https://hal.archives-ouvertes.fr/hal-00641027`. working paper or preprint.

Fisher, R. [1925]. *Statistical methods for research workers*. Edinburgh Oliver & Boyd.

Fisher, R. [1935]. *The design of experiments. 1935*. Oliver and Boyd, Edinburgh.

Forgione, A., Leroy, J., Cahill, R. A., Bailey, C., Simone, M., Mutter, D. et Marescaux, J. [2009]. Prospective evaluation of functional outcome after laparoscopic sigmoid colectomy. *Annals of surgery*, **249**(2), 218–24. ISSN 1528-1140. URL `http://www.ncbi.nlm.nih.gov/pubmed/19212173`.

Fredon, D., Maumy, M. et Bertrand, F. [2009a]. *Mathématiques L1/L2 : Algèbre/-Géométrie en 30 fiches*. Express Sup. Dunod, Paris.

Fredon, D., Maumy, M. et Bertrand, F. [2009b]. *Mathématiques L1/L2 : Analyse en 30 fiches*. Express Sup. Dunod, Paris.

Fredon, D., Maumy, M. et Bertrand, F. [2009c]. *Mathématiques L1/L2 : Statistique et Probabilités en 30 fiches*. Express Sup. Dunod, Paris.

Friedman, J., Hastie, T. et Tibshirani, R. [2010a]. A note on the group lasso and a sparse group lasso. URL `http://arxiv.org/abs/1001.0736`.

Friedman, J., Hastie, T. et Tibshirani, R. [2010b]. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, **33** (1), 1–22. ISSN 1548-7660. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2929880{&}tool=pmcentrez{&}rendertype=abstract`.

Fussler, L., Kobes, N., Bertrand, F., Maumy, M., Grosman, J. et Savary, S. [2008]. A characterization of grapevine trunk diseases in France from data generated by the National Grapevine Wood Diseases Survey. *Phytopathology*, **98**(5). ISSN 0031949X.

Gardosi, J. et Francis, A. [2000]. Early pregnancy predictors of preterm birth : the role of a prolonged menstruation-conception interval. *BJOG : An International Journal of Obstetrics and Gynaecology*, **107**(2), 228–237. ISSN 1470-0328. URL `http://doi.wiley.com/10.1111/j.1471-0528.2000.tb11694.x`.

Gasser, T., Müller, H. G., Köhler, W., Molinari, L. et Prader, A. [1984]. Nonparametric Regression Analysis of Growth Curves. *The Annals of Statistics*, **12**(1), 210–229.

Gentile, C., Li, S. et Zappella, G. [2014]. Online clustering of bandits.

Gittins, J. et Jones, D. [1974]. A dynamic allocation index for the sequential design of experiments. Dans Gani, J., éditeur. *Progress in Statistics*, 241–266. North-Holland, Amsterdam.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. et Lander, E. S. [1999]. Molecular classification of cancer : Class discovery and class prediction by gene expression monitoring. *Science*. ISSN 00368075.

González, I., Lê Cao, K.-A., Davis, M. J. et Déjean, S. [2012]. Visualising Associations Between Paired 'omics' Data Sets. *BioData Mining*, **5**(1), 1–23.

Good, P. [2005]. *Permutation , Parametric and Bootstrap Tests of Hypotheses*. Springer, New York, third édition.

Grandgeorge, M., Wanless, S., Dunn, T. E., Maumy, M., Beaugrand, G. et Grémillet, D. [2008]. Resilience of the British and Irish seabird community in the twentieth century. *Aquatic Biology*, **4**(2), 187–199. ISSN 18647782.

Grenèche, J., Krieger, J., Bertrand, F., Erhardt, C., Maumy, M. et Tassi, P. [2011a]. Short-term memory performances during sustained wakefulness in patients with obstructive sleep apnea-hypopnea syndrome. *Brain and Cognition*, **75**(1). ISSN 02782626 10902147.

Grenèche, J., Krieger, J., Bertrand, F., Erhardt, C., Maumy, M. et Tassi, P. [2013]. Effect of continuous positive airway pressure treatment on short-term memory performance over 24h of sustained wakefulness in patients with obstructive sleep apnea-hypopnea syndrome. *Sleep Medicine*, **14**(10). ISSN 13899457 18785506.

Grenèche, J., Krieger, J., Bertrand, F., Erhardt, C., Muzet, A. et Tassi, P. [2011b]. Effect of continuous positive airway pressure treatment on the subsequent EEG spectral power and sleepiness over sustained wakefulness in patients with obstructive sleep apnea-hypopnea syndrome. *Clinical Neurophysiology*, **122**(5). ISSN 13882457.

Grosman, J. et Doublet, B. [2012]. Maladies du bois de la vigne. Synthèse des dispositifs d'observation au vignoble, de l'observatoire 2003-2008 au réseau d'épidémio-surveillance actuel. *PHYTOMA - La Défense des Végétaux*, **651**, 31–35.

Grün, B., Kosmidis, I. et Zeileis, A. [2012]. Extended Beta Regression in R : Shaken, Stirred, Mixed and Partitioned. *Journal of Statistical Software*, **48**(11), 1–25.

Guan, Y. et Stephens, M. [2011]. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, **5**(3), 1780–1815. ISSN 1932-6157. URL `http://projecteuclid.org/euclid.aoas/1318514285`.

Haaland, D. M. et Howland, J. D. T. [1998]. Weighted partial least squares method to improve calibration precision for spectroscopic noise-limited data. Dans *The eleventh international conference on fourier transform spectroscopy*, volume 430 de *AIP Conference Proceedings*, 253–256. American Institute of Physics, Melville.

Hastie, Tibshirani et Friedman [2008]. The Elements of Statistical Learning : Data Mining, Inference, and Prediction. 1–763. ISSN 09641998.

Hastie, Tibshirani et Friedman [2009]. The Elements of Statistical Learning : Data Mining, Inference, and Prediction. 1–763. ISSN 09641998.

Heagerty, P. J., Lumley, T. et Pepe, M. S. [2000]. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*. ISSN 0006341X.

Heagerty, P. J. et Zheng, Y. [2005]. Survival model predictive accuracy and ROC curves. *Biometrics*. ISSN 0006341X.

Heude, B., Forhan, A., Slama, R., Douhaud, L., Bedel, S., Saurel-Cubizolles, M.-J., Hankard, R., Thiebaugeorges, O., De Agostini, M., Annesi-Maesano, I., Kaminski, M. et Charles, M.-A. [2016]. Cohort Profile : The EDEN mother-child cohort on the prenatal and early postnatal determinants of child health and development. *International Journal of Epidemiology*, **45**(2), 353–363. ISSN 0300-5771. URL `https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/dyv151`.

Hocking, R. R. [1976]. A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression. *Biometrics*, **32**(1), 1–49. ISSN 0006341X. URL `https://www.jstor.org/stable/2529336?origin=crossref`.

Hoerl, A. E. et Kennard, R. W. [1970]. Ridge Regression : Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**(1), 55. ISSN 00401706. URL `https://www.jstor.org/stable/1267351?origin=crossref`.

Hohnadel, M., Maumy, M. et Chollet, R. [2018]. Development of a micromanipulation method for single cell isolation of prokaryotes and its application in food safety. *PLOS ONE*, **13**(5), e0198208. ISSN 1932-6203. URL https://dx.plos.org/10.1371/journal.pone.0198208.

Höskuldsson, A. [1988]. PLS regression methods. *Journal of Chemometrics*, **2**(3), 211–228.

Hugo, W., Zaretsky, J. M., Sun, L., Song, C., Moreno, B. H., Hu-Lieskovan, S., Berent-Maoz, B., Pang, J., Chmielowski, B., Cherry, G., Seja, E., Lomeli, S., Kong, X., Kelley, M. C., Sosman, J. A., Johnson, D. B., Ribas, A. et Lo, R. S. [2016]. Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell*, **165**(1), 35–44. ISSN 00928674. URL https://linkinghub.elsevier.com/retrieve/pii/S009286741630215X.

Jacobs, A., Maumy, M. et Petit, O. [2008]. The influence of social organisation on leadership in brown lemurs (Eulemur fulvus fulvus) in a controlled environment. *Behavioural Processes*, **79**(2), 111–113. ISSN 03766357. URL https://linkinghub.elsevier.com/retrieve/pii/S0376635708001617.

Jagannathan, R. et Ma, T. [2003]. Risk Reduction in Large Portfolios : Why Imposing the Wrong Constraints Helps. *The Journal of Finance*, **58**(4), 1651–1683. ISSN 00221082. URL http://doi.wiley.com/10.1111/1540-6261.00580.

Jenss, R. et Bayley, N. [1937]. A mathematical method for studying the growth of a child. *Human Biology*, **9**, 556–563.

Johnson, N. L., Kotz, S. et Balakrishnan, N. [1995]. *Continuous Univariate Distributions*, volume 2. Wiley, New York, 2nd édition.

Jung, N., Bertrand, F., Bahram, S., Vallat, L. et Maumy-Bertrand, M. [2014]. Cascade : A R package to study, predict and simulate the diffusion of a signal through a temporal gene network. *Bioinformatics*. ISSN 13674803.

Jung, N., Bertrand, F. et Maumy-Bertrand, M. [2015]. AcSel : selecting variables with accuracy in correlated datasets. URL http://arxiv.org/abs/1512.03307.

Kenett, R., Maumy-Bertrand, M. et Bertrand, F. [2020]. Les réseaux bayésiens en pratique. Dans Bertrand, F., Saporta, G. et Thomas-Agnan, C., éditeurs. *Causalité et statistique*. Technip.

Kim, J., Kim, Y. et Kim, Y. [2008]. A Gradient-Based Optimization Algorithm for LASSO. *Journal of Computational and Graphical Statistics*, **17**(4), 994–1009. ISSN 1061-8600. URL http://www.tandfonline.com/doi/abs/10.1198/106186008X386210.

Kobes, N., Fussler, L., Pleynet, M., Savary, S., Bertrand, F. et Maumy, M. [2007]. Vignes, maladies du bois, des facteurs clefs. Premiers résultats de l'analyse statistique des données de l'Observatoire national. *PHYTOMA - La Défense des Végétaux*, **604**, 33–37.

Kosmidis, I. et Firth, D. [2010]. A Generic Algorithm for Reducing Bias in Parametric Estimation. *Journal of Chemometrics*, **4**, 1097–1112.

Kowarik, A. et Templ, M. [2016]. Imputation with the R package VIM. *Journal of Statistical Software*, **74**(7), 1–16.

Koza, J. R., Bennett, F. H. et Stiffelman, O. [1999]. Genetic Programming as a Darwinian Invention Machine. Dans Poli, R., Nordin, P., Langdon, W. et Fogarty, T., éditeurs. *EuroGP 1999 : Genetic Programming*, Lecture Notes in Computer Science, vol 1598, 93–108. Springer, Berlin, Heidelberg. URL `http://link.springer.com/10.1007/3-540-48885-5_8`.

Kraemer, N. et Sugiyama, M. [2011]. The degrees of freedom of partial least squares regression. *Journal of the American Statistical Association*, **106**(494), 697–705.

Kuhn., M. [2018]. *caret : Classification and Regression Training*. URL `https://CRAN.R-project.org/package=caret`. R package version 6.0-80.

Kuntzmann, P., Barbe, J., Maumy-Bertrand, M. et Bertrand, F. [2013]. Late harvest as factor affecting esca and Botryosphaeria dieback prevalence of vineyards in the Alsace region of France. *Vitis - Journal of Grapevine Research*, **52**(4). ISSN 00427500.

Lai, T. et Robbins, H. [1985]. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, **6**(1), 4 – 22. ISSN 0196-8858.

Lazraq, A., Cleroux, R. et Gauchi, J.-P. [2003]. Selecting both latent and explanatory variables in the PLS1 regression model. *Chemometrics and Intelligent Laboratory Systems*, **66**(2), 117–126.

Lê Cao, K., Rossouw, D., Robert-Granié, C. et Besse, P. [2008]. A Sparse PLS for Variable Selection when Integrating Omics data. *Stat Appl Genet Mol Biol*, **7**, Article 35.

Lê Cao, K.-A., Boitard, S. et Besse, P. [2011]. Sparse PLS discriminant analysis : biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*. ISSN 1471-2105.

Lê Cao, K.-A. A., Martin, P. G. G., Robert-Granié, C. et Besse, P. [2009]. Sparse canonical methods for biological data integration : Application to a cross-platform study. *BMC Bioinformatics*, **10**(1), 34. ISSN 14712105. URL `http://www.biomedcentral.com/1471-2105/10/34`.

Ledolter, J. [2013]. Local Polynomial Regression : A Nonparametric Regression Approach. Dans *Data Mining and Business Analytics with R*, 55–66. John Wiley & Sons, Inc., Hoboken, NJ, USA. URL http://doi.wiley.com/10.1002/9781118596289.ch4.

Leroy, B. et Lefort, F. [1971]. À propos du poids et de la taille des nouveau-nés à la naissance. *Rev Fr Gynecol Obstet*, **66**, 391–396. URL https://pubmed.ncbi.nlm.nih.gov/5112836/.

Lesley Stahl [2008]. The Bypass Effect On Diabetes, Cancer. URL https://www.cbsnews.com/news/the-bypass-effect-on-diabetes-cancer/.

Letac, G. et Massam, H. [2004]. A tutorial on the non central Wishart distribution. URL http://www.math.univ-toulouse.fr/~letac/Wishartnoncentrales.pdf.

Li, B., Morris, J. et Martin, E. [2002a]. Model selection for partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, **64**, 79–89.

Li, B., Morris, J. et Martin, E. B. [2002b]. Model selection for partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*. ISSN 01697439.

Li, B., Chen, T. et Giannakis, G. B. [2018]. Bandit online learning with unknown delays.

Lipshutz, R. J., Fodor, S. P., Gingeras, T. R. et Lockhart, D. J. [1999]. High density synthetic oligonucleotide arrays. *Nature Genetics*, **21**(S1), 20–24. ISSN 1061-4036. URL http://www.nature.com/articles/ng0199supp_20.

Little, R. J. A. et Rubin, D. [2002]. *Statistical analysis with missing data*. ISBN 9780471183860. URL https://www.wiley.com/en-us/Statistical+Analysis+with+Missing+Data{%}2C+2nd+Edition-p-9780471183860.

Lubchenco, L. O., Hansman, C., Dressler, M. et Boyd, E. [1963]. INTRAUTERINE GROWTH AS ESTIMATED FROM LIVEBORN BIRTH-WEIGHT DATA AT 24 TO 42 WEEKS OF GESTATION. *Pediatrics*, **32**(5), 793–800. ISSN 0031-4005. URL https://pediatrics.aappublications.org/content/32/5/793.

Luo, X., Stefanski, L. A. et Boos, D. D. [2006]. Tuning Variable Selection Procedures by Adding Noise. *Technometrics*, **48**(2), 165–175. ISSN 0040-1706. URL http://www.tandfonline.com/doi/abs/10.1198/004017005000000319.

Magnanensi, J., Bertrand, F., Maumy-Bertrand, M. et Meyer, N. [2017]. A new universal resample-stable bootstrap-based stopping criterion for PLS component construction. *Statistics and Computing*, **27**(3). ISSN 15731375.

Magnanensi, J., Maumy-Bertrand, M., Meyer, N. et Bertrand, F. [2016a]. A new Bootstrap-Based stopping criterion in PLS components construction. Dans *Springer Proceedings in Mathematics and Statistics*, volume 173.

Magnanensi, J., Maumy-Bertrand, M., Meyer, N. et Bertrand, F. [2016b]. New developments in Sparse PLS regression. URL http://arxiv.org/abs/1601.03281.

Magureanu, S., Combes, R. et Proutière, A. [2014]. Lipschitz bandits : Regret lower bounds and optimal algorithms. *CoRR*, **abs/1405.4758**. URL http://arxiv.org/abs/1405.4758.

Mamelle, N., Munoz, F. et Grandjean, H. [1996]. Mamelle N, Munoz F, Grandjean H. Croissance foetale à partir de l'étude AUDIPOG. I. Établissement de courbes de référence. *J Gynecol Obstet Biol Reprod*, **25**, 61–70. URL https://pubmed.ncbi.nlm.nih.gov/8901304/.

Manne, R. [1987]. Analysis of two partial-least-squares algorithms for multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, **2**(1), 187–197.

Marbac, M. et Sedki, M. [2017]. Variable selection for model-based clustering using the integrated complete-data likelihood. *Statistics and Computing*, **27**(4), 1049–1063. ISSN 0960-3174. URL http://link.springer.com/10.1007/s11222-016-9670-1.

Mardia, K. V., Kent, J. T. et Bibby, J. M. [1979]. *Multivariate Analysis*. Academic Press, London. ISBN 0124712525.

Maumy, M. [2001]. Le comportement des oscillations du processus empirique composé. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, **333**(12), 1101–1104. ISSN 07644442. URL https://linkinghub.elsevier.com/retrieve/pii/S0764444201021851.

Maumy, M. [2002a]. *Etude du processus empirique composé*. Thèse de doctorat, Université Pierre et Marie Curie. URL https://tel.archives-ouvertes.fr/tel-00002724.

Maumy, M. [2002b]. Sur les oscillations du processus de Poisson composé. *Comptes Rendus Mathematique*, **334**(8), 705–708. ISSN 1631073X. URL https://linkinghub.elsevier.com/retrieve/pii/S1631073X02022938.

Maumy, M. [2004]. Strong approximations for the compound empirical process. *Ann. I.S.U.P.*, **1-2**, 69–83.

Maumy-Bertrand, M., Saporta, G. et Thomas-Agnan, C. [2018]. *Apprentissage statistique et données massives*. Éditions Technip, Paris.

Meinshausen, N. et Bühlmann, P. [2010]. Stability selection. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, **72**(4), 417–473. ISSN 13697412. URL http://doi.wiley.com/10.1111/j.1467-9868.2010.00740.x.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. et Teller, E. [1953]. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, **21**(6), 1087–1092. ISSN 0021-9606. URL http://aip.scitation.org/doi/10.1063/1.1699114.

Meyer, N., Fredon, D., Maumy-Bertrand, M. et Bertrand, F. [2018]. *Toute l'UE4 en fiches. Evaluation des méthodes d'analyse appliquées aux sciences de la vie et de la santé*. Dunod, Paris, 3e édition.

Meyer, N., Maumy-Bertrand, M. et Bertrand, F. [2010]. Comparaison de variantes de régressions logistiques PLS et de régression PLS sur variables qualitatives : application aux données d'allélotypage. *Journal de la Société Française de Statistique*, **151**(2), 1–18.

Mint El Mouvid, M. [2000]. *Sur l'estimateur linéaire local de la fonction de répartition conditionnelle*. Thèse de doctorat, Monpellier 2.

Morgan, D., Tjärnberg, A., Nordling, T. E. M. et Sonnhammer, E. L. L. [2019]. A generalized framework for controlling FDR in gene regulatory network inference. *Bioinformatics*, **35**(6), 1026–1032. ISSN 1367-4803. URL https://academic.oup.com/bioinformatics/article/35/6/1026/5086392.

Morgoat [2005]. Une nouvelle méthode d'enquête et de sondage dans le tourisme. *Vent d'ouest*, **90**, 26–27.

Nadaraya, E. [1964]. On estimating regression. *Theory of Probability and its Applications*, **9**(1), 141–142.

Naes, T. et Martens, H. [1985]. Comparison of prediction methods for multicollinear data. *Communications in Statistics – Simulation and Computation*, **14**, 545–576.

Natarajan, B. K. [1995]. Sparse Approximate Solutions to Linear Systems. *SIAM Journal on Computing*, **24**(2), 227–234. ISSN 0097-5397. URL http://epubs.siam.org/doi/10.1137/S0097539792240406.

Nengsih, T. A., Bertrand, F., Maumy-Bertrand, M. et Meyer, N. [2019]. Determining the number of components in PLS regression on incomplete data set. *Statistical Applications in Genetics and Molecular Biology*, **18**(6). ISSN 1544-6115. URL https://www.degruyter.com/view/j/sagmb.ahead-of-print/sagmb-2018-0059/sagmb-2018-0059.xmlhttp://www.degruyter.com/view/j/sagmb.ahead-of-print/sagmb-2018-0059/sagmb-2018-0059.xmlhttp://www.degruyter.com/view/j/sagmb.2019.18.issue-6/sagmb-2018-0059/sagmb-.

Pearl, J. [2009]. *Causality : Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd édition. ISBN 052189560X, 9780521895606.

Peng, C.-Y. J., Lee, K. L. et Ingersoll, G. M. [2002]. An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*, **96** (1), 3–14. ISSN 0022-0671. URL `http://www.tandfonline.com/doi/abs/10.1080/00220670209598786`.

Perrot, A., Pionneau, C., Nadaud, S., Davi, F., Leblond, V., Jacob, F., Merle-Béral, H., Herbrecht, R., Béné, M.-C., Gribben, J. G., Bahram, S. et Vallat, L. [2011]. A unique proteomic profile on surface IgM ligation in unmutated chronic lymphocytic leukemia. *Blood*, **118**(4), e1–15. ISSN 1528-0020. URL `http://www.ncbi.nlm.nih.gov/pubmed/21602524`.

Perry, P. O. [2015]. *bcv : Cross-Validation for the SVD (Bi-Cross-Validation)*. R package version 1.0.1.

Philip, J.-M. et Maumy, M. [2008]. L'apport des outils décisionnels dans la réalisation des enquêtes des pays en développement : le cas de Madagascar et du Mali. Dans Ruiz-Gazen, A., Guilbert, P., Haziza, D. et Tillé, Y., éditeurs. *Méthodes de sondage. Actes du colloque francophone sur les sondage 2007*. Dunod, Paris, France. ISBN 9782100517770.

Piotto, M., Moussallieh, F.-M., Neuville, A., Bellocq, J.-P., Elbayed, K. et Namer, I. J. [2012]. Towards real-time metabolic profiling of a biopsy specimen during a surgical operation by 1H high resolution magic angle spinning nuclear magnetic resonance : a case report. *Journal of Medical Case Reports*, **6**(1).

Pugliese, R., Maggioni, D., Sansonna, F., Scandroglio, I., Forgione, A., Boniardi, M., Costanzi, A., Citterio, D., Ferrari, G. C., Lernia, S. D. et Magistro, C. [2008]. Laparoscopic Distal Pancreatectomy. *Surgical Laparoscopy, Endoscopy & Percutaneous Techniques*, **18**(3), 254–259. ISSN 1530-4515. URL `http://journals.lww.com/00129689-200806000-00005`.

Rau, A., Gallopin, M., Celeux, G. et Jaffrézic, F. [2013]. Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics*, **29**(17), 2146–2152.

Rau, A., Jaffrézic, F., Foulley, J. L. et Doerge, R. W. [2012]. Reverse engineering gene regulatory networks using approximate Bayesian computation. *Statistics and Computing*. ISSN 09603174.

Rendina, F., Armoogum, J. et Maumy-Bertrand, M. [2018]. Mise au point d'un plan de sondage afin de mesurer la mobilité des non-résidents. Dans *Colloque francophone sur les sondages, Lyon*.

Rendina, F., Maumy-Bertrand, M. et Armoogum, J. [2017]. Estimation of tourism mobility by indirect sampling. Dans *Communication In : 11th International Conference on Transport Survey Methods - ISCTSC*.

Rendina, F., Rabaud, M., Hasiak, F., Maumy-Bertrand, M. et Armoogum, J. [2016]. Choix des variables auxiliaires pour le redressement d'une enquête de mobilité. Dans *Colloque francophone sur les sondages, Gatineau*.

Rigby, R. A. et Stasinopoulos, D. M. [2004]. Smooth centile curves for skew and kurtotic data modelled using the Box–Cox power exponential distribution. *Statistics in Medicine*, **23**(19), 3053–3076. ISSN 0277-6715. URL http://doi.wiley.com/10.1002/sim.1861.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. et Smyth, G. K. [2015]. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**(7), e47.

Robins, J. [1987]. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Diseases*, **40**, 139S – 161S. ISSN 0021-9681. URL http://www.sciencedirect.com/science/article/pii/S0021968187800188.

Rolland, A., Bertrand, F., Maumy, M. et Jacquet, S. [2009]. Assessing phytoplankton structure and spatio-temporal dynamics in a freshwater ecosystem using a powerful multiway statistical analysis. *Water Research*, **43**(13). ISSN 00431354.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., Hurt, E. M., Zhao, H., Averett, L., Yang, L., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Powell, J., Duffey, P. L., Longo, D. L., Greiner, T. C., Weisenburger, D. D., Sanger, W. G., Dave, B. J., Lynch, J. C., Vose, J., Armitage, J. O., Montserrat, E., López-Guillermo, A., Grogan, T. M., Miller, T. P., LeBlanc, M., Ott, G., Kvaloy, S., Delabie, J., Holte, H., Krajci, P., Stokke, T. et Staudt, L. M. [2002]. The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma. *New England Journal of Medicine*, **346**(25), 1937–1947. ISSN 0028-4793. URL http://www.nejm.org/doi/abs/10.1056/NEJMoa012914.

Rosipal, R. et Trejo, L. J. [2002]. Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space. *Journal of Machine Learning Research*, **2**, 97–123. ISSN 15324435. URL http://www.crossref.org/jmlr{_}DOI.html.

Rousseau, T., Ferdynus, C., Quantin, C., Gouyon, J.-B., Sagot, P. et CMPRB [2008]. Poids des nouveau-nés issus de grossesses uniques et non compliquées entre 28 et 42 semaines d'aménorrhée à partir des données du réseau périnatal de la région Bourgogne. *Journal de Gynécologie Obstétrique et Biologie de la Reproduction*, **37**(6), 589–596. ISSN 03682315. URL https://linkinghub.elsevier.com/retrieve/pii/S0368231508000781.

Royston, J. P. [1982]. An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. *Applied Statistics*, **31**(2), 115. ISSN 00359254.

Royston, P. et Wright, E. [1998]. How to construct 'normal ranges' for fetal variables. *Ultrasound in Obstetrics and Gynecology*, **11**(1), 30–38. ISSN 09607692. URL http://doi.wiley.com/10.1046/j.1469-0705.1998.11010030.x.

Rubino, F. [2008]. Is Type 2 Diabetes an Operable Intestinal Disease ? : A provocative yet reasonable hypothesis. *Diabetes Care*, **31**(Supplement 2), S290–S296. ISSN 0149-5992. URL http://care.diabetesjournals.org/cgi/doi/10.2337/dc08-s271.

Salomon, L. J., Bernard, J. P., de Stavola, B., Kenward, M. et Ville, Y. [2007]. Poids et taille de naissance : courbes et équations. *Journal de Gynecologie Obstetrique et Biologie de la Reproduction*, **36**(1), 50–56. ISSN 03682315.

Schwarz, G. [1978]. Estimating the Dimension of a Model. *The Annals of Statistics*, **6**(2), 461–464. ISSN 0090-5364. URL http://projecteuclid.org/euclid.aos/1176344136.

Segal, M. R., Dahlquist, K. D. et Conklin, B. R. [2003]. Regression Approaches for Microarray Data Analysis. *Journal of Computational Biology*, **10**(6), 961–980. ISSN 1066-5277. URL http://www.liebertpub.com/doi/10.1089/106652703322756177.

Silverman, B. W. [1985]. Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting. *Journal of the Royal Statistical Society : Series B (Methodological)*, **47**(1), 1–21. ISSN 00359246. URL http://doi.wiley.com/10.1111/j.2517-6161.1985.tb01327.x.

Simas, A. B., Barreto-Souza, W. et Rocha, A. V. [2010]. Improved Estimators for a General Class of Beta Regression Models. *Computational Statistics & Data Analysis*, **54**(2), 348–366.

Sohn, I., Kim, J., Jung, S.-H. et Park, C. [2009]. Gradient lasso for Cox proportional hazards model. *Bioinformatics*, **25**(14), 1775–1781. ISSN 1367-4803. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp322.

Sra, S. [2012]. A short note on parameter approximation for von Mises-Fisher distributions : and a fast implementation of $I_s(x)$. *Computational Statistics*, **27**(1), 177–190. ISSN 0943-4062. URL http://link.springer.com/10.1007/s00180-011-0232-x.

Stein, R. [2008]. Surgery Shows Promise For Treatment of Diabetes. URL https://www.washingtonpost.com/wp-dyn/content/article/2008/05/03/AR2008050301837.html.

Steinley, D. et Brusco, M. J. [2008]. A New Variable Weighting and Selection Procedure for K -means Cluster Analysis. *Multivariate Behavioral Research*, **43** (1), 77–108. ISSN 0027-3171. URL https://www.tandfonline.com/doi/full/10.1080/00273170701836695.

Tenenhaus, A., Giron, A., Viennet, E., Béra, M., Saporta, G. et Fertil, B. [2007]. Kernel logistic PLS : A tool for supervised nonlinear dimensionality reduction and binary classification. *Computational Statistics and Data Analysis*, **51**(9), 4083–4100. ISSN 01679473.

Tenenhaus, M. [1998]. *La régression PLS, Théorie et pratique*. Technip, Paris.

Tenenhaus, M. [1999]. La regression logistique PLS. Dans *Proceedings of the 32èmes journées de Statistique de la Société française de Statistique*, 721–723. FES.

Tenenhaus, M. [2005]. La régression logistique PLS. Dans Droesbeke, J.-J., Lejeune, M. et Saporta, G., éditeurs. *Modèles statistiques pour données qualitatives*, 263–275. Technip, Paris.

Tetsi, L., Charles, A.-L., Georg, I., Goupilleau, F., Lejay, A., Talha, S., Maumy-Bertrand, M., Lugnier, C. et Geny, B. [2019]. Effect of the Phosphodiesterase 5 Inhibitor Sildenafil on Ischemia-Reperfusion-Induced Muscle Mitochondrial Dysfunction and Oxidative Stress. *Antioxidants*, **8**(4), 93. ISSN 2076-3921. URL https://www.mdpi.com/2076-3921/8/4/93.

Thioulouse, J. et Chessel, D. [1987]. Les analyses multitableaux en écologie factorielle. I. De la typologie d'état à la typologie de fonctionnement par l'analyse triadique. *Acta Oecologica, Oecologia Generalis*, **8**(4), 463–480.

Thode, H. C. [2002]. *Testing for normality*. Marcel Dekker. ISBN 9780824796136. URL https://www.crcpress.com/Testing-For-Normality/Thode/p/book/9780824796136.

Thompson, W. R. [1933]. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, **25**(3-4), 285–294.

Tibshirani, R. [1996]. Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**, 267–288. ISSN 00359246.

Tibshirani, R. [1997]. The lasso method for variable selection in the cox model. *Statistics in Medicine*. ISSN 02776715.

Valko, M., Korda, N., Munos, R., Flaounas, I. et Cristianini, N. [2013]. Finite-time analysis of kernelised contextual bandits. Dans *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, 654–663, Arlington, Virginia, United States. AUAI Press.

Vallat, L., Kemper, C. a., Jung, N., Maumy-Bertrand, M., Bertrand, F., Meyer, N., Pocheville, A., Fisher, J. W., Gribben, J. G. et Bahram, S. [2013]. Reverse-engineering the genetic circuitry of a cancer cell with predicted intervention in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(2), 459–64. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3545767{&}tool=pmcentrez{&}rendertype=abstract`.

Vallat, L. D., Park, Y., Li, C. et Gribben, J. G. [2007]. Temporal genetic program following B-cell receptor cross-linking : altered balance between proliferation and death in healthy and malignant B cells. *Blood*, **109**(9), 3989–3997. ISSN 00064971. URL `http://www.ncbi.nlm.nih.gov/pubmed/10666191http://www.ncbi.nlm.nih.gov/pubmed/17234734http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1874586`.

van Buuren, S. et Groothuis-Oudshoorn, K. [2011]. mice : Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, **45**(3), 1–67. ISSN 1548-7660. URL `http://www.jstatsoft.org/v45/i03/`.

van der Hilst, R. D., de Hoop, M. V., Wang, P., Shim, S.-H., Ma, P. et Tenorio, L. [2007]. Seismostratigraphy and thermal structure of Earth's core-mantle boundary region. *Science (New York, N.Y.)*, **315**(5820), 1813–7. ISSN 1095-9203. URL `http://www.ncbi.nlm.nih.gov/pubmed/17395822`.

Wahl, J.-B., Bertrand, F. et Maumy-Bertrand, M. [2019]. PEPS DataFlow : Analyse de données pour des capteurs fluidiques à haute précision. Dans *SMAI 2019*.

Wakeling, I. N. et Morris, J. J. [1993]. A test of significance for partial least squares regression. *Journal of Chemometrics*, **7**(4), 291–304.

Walh, J., Bertrand, F. et Maumy-Bertrand, M. [2019]. PEPS DataFlow : Analyse de données pour des capteurs fluidiques à haute précision. Dans *9ème Biennale Française des Mathématiques Appliquées et Industrielles, SMAI 2019, du 13 au 17 mai 2019.*

Wang, S., Nan, B., Rosset, S. et Zhu, J. [2011]. Random lasso. *The Annals of Applied Statistics*, **5**(1), 468–485. ISSN 1932-6157. URL `http://projecteuclid.org/euclid.aoas/1300715199`.

Watson, G. [1964]. Smooth regression analysis. *Sankhyā : The Indian Journal of Statistics*, 359–372.

Weber, J. C., Meyer, N., Pencreach, E., Schneider, A., Guérin, E., Neuville, A., Stemmer, C., Brigand, C., Bachellier, P., Rohr, S., Kedinger, M., Meyer, C., Guenot, D., Oudet, P., Jaeck, D. et Gaub, M. P. [2007]. Allelotyping analyses of synchronous primary and metastasis CIN colon cancers identified different subtypes. *International Journal of Cancer*. ISSN 00207136.

Weinstein, J. N., Myers, T. G., O'Connor, P. M., Friend, S. H., Fornace, A. J., Kohn, K. W., Fojo, T., Bates, S. E., Rubinstein, L. V., Anderson, N. L., Buolamwini, J. K., van Osdol, W. W., Monks, A. P., Scudiero, D. A., Sausville, E. A., Zaharevitz, D. W., Bunow, B., Viswanadhan, V. N., Johnson, G. S., Wittes, R. E. et Paull, K. D. [1997]. An information-intensive approach to the molecular pharmacology of cancer. *Science (New York, N.Y.)*, **275**(5298), 343–9. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/8994024.

White, I. R., Royston, P. et Wood, A. M. [2011]. Multiple imputation using chained equations : Issues and guidance for practice. *Statistics in Medicine*. ISSN 02776715.

Wiklund, S., Nilsson, D., Eriksson, L., Sjöström, M., Wold, S. et Faber, K. [2007]. A randomization test for PLS component selection. *Journal of Chemometrics*, **21** (10-11), 427–439. ISSN 08869383. URL http://doi.wiley.com/10.1002/cem.1086.

Wold, H. [1966]. Estimation of Principal Components and Related Models by Iterative Least Squares. Dans Krishnaiah, P. R., éditeur. *Multivariate Analysis*, 391–420. Academic Press, New York.

Wold, S., Martens, H. et Wold, H. [1983]. The multivariate calibration problem in chemistry solved by the PLS method. *Proc. Conf. Matrix pencils*, 286–293.

Wold, S., Ruhe, A., Wold, H. et Dunn, III, W. J. [1984]. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, **5**(3), 735–743. ISSN 0196-5204. URL http://epubs.siam.org/doi/abs/10.1137/0905052.

Wold, S., Sjöström, M. et Eriksson, L. [2001]. PLS-regression : a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, **58**(2), 109–130. ISSN 01697439.

Wu, Y., Boos, D. D. et Stefanski, L. A. [2007]. Controlling Variable Selection by the Addition of Pseudovariables. *Journal of the American Statistical Association*, **102**(477), 235–243. ISSN 0162-1459. URL http://www.tandfonline.com/doi/abs/10.1198/016214506000000843.

Yuan, M. et Lin, Y. [2006]. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, **68**(1), 49–67. ISSN 1369-7412. URL http://doi.wiley.com/10.1111/j.1467-9868.2005.00532.x.

Zhang, C.-H. [2010]. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38**(2), 894–942. ISSN 0090-5364. URL http://projecteuclid.org/euclid.aos/1266586618.

Zhou, X., Carbonetto, P. et Stephens, M. [2013]. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLoS Genetics*, **9**(2), e1003264. ISSN 1553-7404. URL https://dx.plos.org/10.1371/journal.pgen.1003264.

Zhu, J. J., Santarius, T., Wu, X. Y., Tsong, J., Guha, A., Wu, J. K., Hudson, T. J. et Black, P. M. [1998]. Screening for loss of herterozygosity and microsatellite instability in oligodendrogliomas. *GENES CHROMOSOMES & CANCER*. ISSN 1045-2257.

Zou, H. [2006]. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**(476), 1418–1429. ISSN 0162-1459. URL http://www.tandfonline.com/doi/abs/10.1198/016214506000000735.

Zou, H. et Hastie, T. [2005]. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, **67**(2), 301–320. ISSN 1369-7412. URL http://doi.wiley.com/10.1111/j.1467-9868.2005.00503.x.

Les activités de recherche présentées dans ce mémoire d'habilitation à diriger les recherches s'articulent autour de plusieurs méthodes statistiques.

La plus importante d'entre elles est la modélisation statistique dans un contexte de petite ou de grande dimension en présence éventuelle de données manquantes et de censures. Cette thématique traite aussi bien des cas de modélisation paramétrique comme la régression linéaire, la régression linéaire généralisée et d'autres types de régression, comme la régression bêta ou le modèle de Cox, pour lesquels ont été produits non seulement de nouveaux critères de choix de modèle mais aussi de sélection de variables mais aussi des cas de modélisation non paramétrique comme l'estimation de la fonction des quantiles. Les contextes d'application m'ont incité à l'utilisation d'approches de régression pénalisée, comme la régression *lasso*, la régression *ridge* ou la régression *elasticnet* ou encore la régression par les moindres carrés partiels parcimonieuse ou non.

Le mémoire décrit plusieurs approches de modélisation de données provenant de la biologie, du domaine médical ou encore de l'industrie. À titre d'exemple, la plus aboutie parmi celles que j'ai développées permettrait la prévention de risques pour le développement fœtal. À ces sujets de recherche, s'ajoute depuis quatre ans une thématique supplémentaire liée à l'apprentissage statistique et plus particulièrement à la théorie mathématique des bandits manchots. Pour chacune de ces thématiques, plusieurs perspectives de recherches sont détaillées à la fin de ce mémoire.

INSTITUT DE RECHERCHE MATHÉMATIQUE AVANCÉE
UMR 7501
Université de Strasbourg et CNRS
7 Rue René Descartes
67 084 STRASBOURG CEDEX

Tél. 03 68 85 01 29
Fax 03 68 85 03 28

https://irma.math.unistra.fr
irma@math.unistra.fr

Institut de Recherche Mathématique Avancée

IRMA 2020/005
ISSN 0755-3390
https ://tel.archives-ouvertes.fr/tel-020