Contributions à une aide à la décision de confiance

Pierre Lemaire

HABILITATION À DIRIGER DES RECHERCHES DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Informatique et Mathématiques appliquées**

présentée par

Pierre Lemaire

Préparée au sein de **Grenoble INP & G-SCOP (UMR 5272)**

École doctorale MSTII : **Mathématiques, Sciences et Technologies de l'Information, Informatique**

Contributions à une aide à la décision de confiance

soutenue publiquement le **10 novembre 2021**
devant le jury composé de :

Prof. Yves CRAMA
Université de Liège, rapporteur

Prof. Stéphane DAUZÈRE-PÉRÈS
Mines de Saint-Étienne, examinateur

Prof. Clarisse DHAENENS
Université de Lille, rapportrice

Prof. Éric GAUSSIER
Université Grenoble Alpes, président

Prof. Adeline LECLERCQ-SAMSON
Université Grenoble Alpes, examinatrice

Prof. Vincent T'KINDT
Université de Tours, rapporteur

Table des matières

Remerciements

Cette habilitation aurait pu n'être qu'un exercice académique ou un passage presque obligé dans une carrière d'enseignant-chercheur... Elle aura été beaucoup plus que cela, et je le dois avant tout aux six personnes qui m'ont fait le plaisir et l'honneur d'accepter d'être membres de mon jury.

Yves, Clarisse, Vincent, Stéphane, nous nous croisons au gré de conférences et j'estime vos travaux exemplaires ; ils sont de plus complémentaires les uns des autres, à même d'éclairer chaque partie de mon travail. Il me semblait évident de vous inviter et je vous remercie d'avoir acceptée [1] sans hésiter cette invitation. Yves, Clarisse et Vincent : vous méritez doublement ce merci pour avoir aussi acceptée la charge d'être les rapporteurs de ce manuscrit et de l'avoir expertisé avec une exigence bienveillante.

Adeline, Éric, vous avez sans doute été davantage surpris de mon invitation : merci de l'avoir acceptée avec enthousiasme et d'avoir ainsi accepté de prendre le risque de venir à une HDR qui ne relevait pas clairement de votre domaine. Pouvoir bénéficier de ces regards croisés et complémentaires venant d'autres communautés était important pour moi, merci donc de m'avoir apportés vos éclairages sur des aspects que j'explore avec encore beaucoup de prudence.

Chacun et chacune de vous six a ainsi grandement contribué à donner du sens à ce travail par sa présence, son regard, et les discussions qui ont suivi. *Merci*.

Ensuite, je veux dire un grand merci à celles et ceux sans qui mes travaux n'existeraient pas, un grand merci à mes collaborateurs et collaboratrices, pour tous les échanges, les discussions sans fin, les réussites et les erreurs, tous ces moments aussi agréables qu'instructifs — avec un merci spécifique à Raja pour une coopération fructueuse de plus de 15 ans, et un merci particulier à mes doctorants : Olivier, Alexandre, Natalia, Kean, Florian, Mehdi, Étienne, Yuzhen, Nikita, et aux personnes avec qui j'ai eu le plaisir de les encadrer : Éric, David, Jean-Philippe, Van-Dat, Iragaël, Marie-Laure, Philippe, Nadia, Bertrand, Bernard, Alain et Cyril.

Enfin, merci aux membres de l'ex-École des Mines de Nantes avec qui j'ai passées trois belles années, et à tous les membres du G-SCOP et de GI qui me supportent maintenant depuis 10 ans... et spécialement aux membres de l'équipe ROSP : c'est un plaisir et un privilège de travailler avec vous chaque jour. Je remercie également tout le personnel administratif et technique de G-SCOP et GI : vous nous facilitez bien le quotidien et j'apprécie vraiment vos solutions souriantes et efficaces à mes problèmes ! Un grand merci également à David Monniaux, de l'École Doctorale MSTII, et Céline Legrenzi, du Collège Doctoral, dont les réponses rapides et efficaces ont bien simplifiés les aspects administratifs de cette habilitation.

Je voudrais ensuite décerner une mention très spéciale à mon père, à Nadia Brauner et à Iragaël Joly : sans vos encouragements répétés et patients et vos nombreux conseils, je serais sans doute encore en train de trouver «autre chose» à faire avant de passer mon HDR. Merci de votre persévérante prévenance.

Une pensée particulière et émue à deux personnages avec qui je n'aurai pas l'occasion de partager ce travail et qui en ont pourtant été les instigateurs par leurs conseils et leur inspiration, et qui restent bien présents dans ma tête et dans mon cœur : Gerd Finke et Peter Hammer.

S'il y a parfois des bonnes choses qui sortent de ma tête, c'est avant tout grâce à toi, ma Julie.
Tu es extraordinaire.

La rédaction de ce manuscrit et la préparation de la soutenance de cette habilitation auraient été beaucoup moins faciles et plaisants sans les merveilleux outils que sont LaTeX en général et TikZ, PGF, beamer en particulier, ou sans le support auditif d'Amorphis, Insomnium, Therion et de l'inoxydable métal d'Accept.

1. Voir note 3, page 7.

Introduction

Ce document propose une synthèse personnelle de mes travaux de recherche. Au fil de ces travaux, je me suis intéressé à l'optimisation de campagnes de recherche d'exo-planètes et au dimensionnement d'un réseau de chaleur, à la prédiction de la taille adulte de jeunes filles et à l'anticipation d'un flux de production, ou encore à la conception de réseaux de télécommunication et à la modélisation de la topographie de puces électroniques. Certaines pages de ce catalogue de problèmes ne seront pas ou peu feuilletées dans la suite, afin de se concentrer sur les illustrations les plus emblématiques et éviter l'ennui de péripéties que j'ai eu plaisir à vivre mais qui n'apporteraient pas grand chose à ce récit — un récit que je n'ai bien sûr pas écrit seul [2].

Présenter mes travaux correspond davantage à une exploration en largeur qu'en profondeur. On peut toutefois en faire une lecture presque chronologique. Mes contributions à l'ordonnancement (chapitre 1) relèvent d'une recherche opérationnelle conventionnelle et prolongent mes premières amours de chercheur alors doctorant. Les travaux sur l'aide au diagnostic médical (chapitre 2) ont commencé par un hasard heureux poursuivi en post-doctorat et mêlent sciences de données et optimisation combinatoire. Ces deux courants, à l'origine indépendants, ont conflué vers l'analyse et l'optimisation de systèmes complexes (chapitre 3), en croisant recherche opérationnelle, sciences de données et génie industriel d'autant plus naturellement que je rejoignais des établissements dont les objets d'études sont les systèmes de production, à savoir d'abord l'IRCCyN/École des Mines de Nantes puis G-SCOP/Grenoble INP - Génie Industriel.

Cette exploration en largeur rend nécessaire la présentation de problèmes variés, afin de rendre compte de mes contributions, d'une part, et surtout de rendre intelligible le projet proposé, d'autre part. Je m'efforcerai d'être aussi synthétique que possible et d'éviter les détails superflus, tout en privilégiant la clarté et la facilité de lecture plutôt qu'une concision excessive. Telle était ma ligne, et j'espère avoir réussi à la suivre et que vous-même la suivrez avec plaisir. La lectrice [3] inquiète qui aimerait savoir où elle met les yeux, ou le lecteur [3] pressé qui préférerait se contenter de peu, ont à leur disposition une description de l'ensemble de mes travaux résumés en 4 pages (annexe A).

Les trois premiers chapitres sont construits selon le même schéma : une contextualisation des études, une présentation des contributions, puis une discussion sur les perspectives scientifiques et les enjeux. Chacun de ces chapitres propose des outils très différents, mais tous s'inscrivent dans le cadre de l'aide à la décision, et plus précisément ce que je considère être une «aide à la décision de confiance». Plutôt que de risquer une définition artificielle et abstraite de cette notion, je préfère en dessiner ainsi les contours à travers mes travaux avant de terminer par un épilogue qui propose un regard sur les perspectives croisées des chapitres précédents et revient de manière approfondie sur cette notion de confiance.

2. «Je» ou «Nous»? Le singulier témoigne d'un avis personnel ou d'une contribution dont j'estime être le moteur essentiel mais dont des discussions et autres échanges variés ont été le carburant tout aussi essentiel. Le pluriel est plus adapté pour des résultats sur lesquels j'ai eu une contribution significative mais davantage partagée (typiquement, dans le cadre des thèses que j'ai pu encadrées).

3. Afin d'éviter certaines lourdeurs syntaxiques, les noms de métiers ou similaires sont écrits au masculin ou au féminin au hasard du contexte et sans plus de raison que l'équiprobabilité des deux cas. De plus, je me permettrai des accords de majorité ou de proximité, aussi naturels que fréquents en Français avant que quelques grammairiens misogynes les jugent inacceptables, eux-mêmes n'acceptant pas les vérités qu'ils permettaient d'énoncer [Vie17] et je me permettrai d'accorder tous les participes passés avec leur COD, où que ce dernier soit placé, n'ayant jamais trouvée de justification censée pour ne pas le faire. Ces facéties devraient, j'espère, ne pas gêner la lecture et passer essentiellement inaperçues dès cette note terminée.

CHAPITRE 1

Ordonnancement

Une bonne partie de mes travaux relèvent de la théorie de l'ordonnancement, selon deux points de vue complémentaires : des analyses de complexité, qui permettent de caractériser au mieux les problèmes, et des méthodes de résolution dédiées.

La théorie de l'ordonnancement comme la théorie de la complexité sont assez techniques et ont chacune développé leur propre jargon de définitions et de notations. Il est tout aussi indispensable de les maîtriser pour comprendre pleinement ce chapitre, qu'utopique de penser pouvoir les faire comprendre en une demi-page. En guise de compromis, je me contenterai de définitions informelles, suffisantes pour que les novices puissent, peut-être, avoir une compréhension superficielle et que les spécialistes sachent de quoi on parle sans fastidieux rappels. Concrètement, si la phrase «$P||C_{\max}$ est NP-complet» n'est pas une évidence pour vous, alors je vous invite à une lecture légère, un simple grignotage de ce chapitre afin d'en tirer quelques saveurs sans que cela vous reste sur l'estomac. Et si cela vous a ouvert l'appétit, vous pourrez vous mettre à jour en dégustant l'excellent Pinedo [Pin16].

Pour un problème d'ordonnancement, on dispose de *tâches* qui nécessitent des *ressources* (en particulier des machines) pour leur exécution ; l'enjeu est de réussir à allouer au mieux, dans le temps, les ressources aux tâches. Il existe une très grande variété de problèmes d'ordonnancement, selon les caractéristiques des tâches, des ressources et des objectifs visés. Rappelons les caractéristiques classiques, ce qui permettra de définir les notations associées :

- j est l'indice associé à une tâche (ou *job*) ;
- p_j (*processing time*) est la durée de la tâche j ;
- r_j (*release date*) est la date de disponibilité de la tâche j ;
- d_j (*deadline*, *due-date*) est l'échéance de la tâche j ;
- w_j (*weight*) est le poids de la tâche j.

Afin de s'y retrouver dans le foisonnement des problèmes d'ordonnancement, il existe une classification, dont je fais un usage intensif, due à Graham et al. [GLLRK79]. Cette classification permet de relier entre eux les problèmes, de manière très pratique et efficace (voir le *scheduling zoo* [BDJ$^+$]) grâce à trois champs $\alpha|\beta|\gamma$. J'en donne ici les quelques aspects qui sont utilisés dans la suite de ce chapitre :

- α caractérise les machines. En particulier 1 stipule une unique machine ; P plusieurs machines identiques en parallèle ; Pm exactement m machines identiques en parallèle.
- β caractérise les restrictions et contraintes. En particulier r_j ou d_j indiquent respectivement la présence de dates de disponibilité et d'échéances ; $p_j = 1$ signifie que toutes les tâches ont une durée unitaire.
- γ caractérise l'objectif. En particulier $C_{\max}$ (*completion time*) correspond au temps pour terminer l'ordonnancement tandis que $\sum w_j U_j$ est traditionnellement utilisé pour le nombre pondéré de tâches en retard.

Ainsi, le problème $P||C_{\max}$ évoqué tantôt correspond à un problème sur machines parallèles où il faut minimiser la date de fin. De même, $1|r_j|\sum U_j$ indique un ordonnancement sur une seule machine, avec des dates de disponibilité et pour lequel il faut minimiser le nombre de tâches en retard.

1.1 Complexité de problèmes d'ordonnancement

Nous ne considérons que des problèmes *calculables*, c'est-à-dire dont la solution peut être trouvée avec un ordinateur. L'enjeu est alors de déterminer si, pour un problème donnée, une solution peut toujours être trouvée rapidement ou pas. Pour cela, la théorie de la complexité permet de caractériser les problèmes en «problèmes faciles» et en «problèmes difficiles». Je n'entrerai pas dans les détails de cette théorie, que l'on pourra découvrir avec plaisir en compagnie de Garey et Johnson [GJ79] ; comme je l'ai annoncé plus haut, je me contenterai de définitions informelles.

Un problème est *facile* (ou *polynomial*) s'il existe un algorithme qui résout chaque instance rapidement. Un problème est *difficile* (ou *NP-complet*) si on ne peut que vérifier une solution rapidement, mais pas la trouver : aucun algorithme ne peut alors garantir une résolution rapide de chaque instance. Dans ce cas, il faut faire un compromis entre le temps de calcul et la qualité de la solution, c'est-à-dire choisir entre la certitude d'une réponse rapide mais seulement «bonne» ou «approchée» ; ou la certitude d'une réponse «optimale» ou «exacte», mais en un temps indéfini. De plus, parmi les problèmes difficiles on distingue les problèmes *faiblement NP-complets* pour lesquels il existe un algorithme pseudo-polynomial, c'est-à-dire rapide tant qu'on n'utilise que des nombres raisonnables, et les problèmes *fortement NP-complets* qui restent difficiles quoi qu'on fasse.

En recherche opérationnelle, la théorie de la complexité est très souvent utilisée pour qualifier un problème précis, typiquement dans le cadre d'une application, afin de justifier le choix des algorithmes déployés. Classiquement, on commence ainsi par montrer que le problème qui nous préoccupe est difficile, puis on propose soit des heuristiques rapides, soit une méthode exacte laborieuse pour le résoudre.

Dans les travaux présentés ci-dessous, je m'attache davantage a dresser un panorama de la complexité d'une famille de problèmes, c'est-à-dire de déterminer, pour un problème général, les sous-cas faciles et les sous-cas difficiles afin de dessiner la frontière entre ce qui est facile et ce qui ne l'est pas. On parle alors de cas *maximalement polynomiaux* (si on ajoute quelque chose, ils deviennent difficiles) ou *minimalement NP-complets* (si on enlève quelque chose, ils deviennent faciles). Ce genre d'étude se marie très bien avec la classification de Graham et al. [GLLRK79].

Dans la suite de cette section, je détaille deux problèmes : les rangements d'objets multiboîtes, et l'ordonnancement de tâches à durée variable.

1.1.1 Rangements d'objets multiboîtes

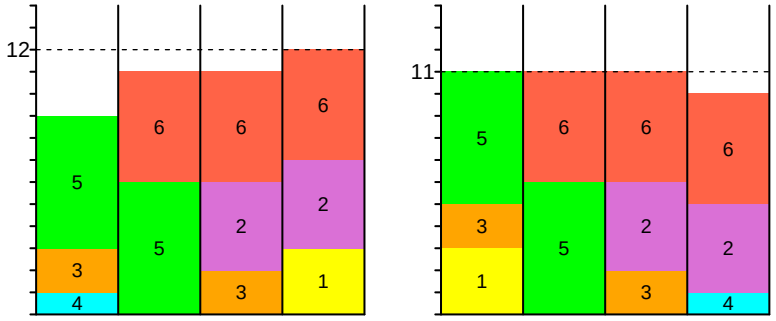
Les «rangements d'objets multiboîtes[1]» ont été le sujet de ma thèse [Lem04], encadrée par Gerd Finke et Nadia Brauner. Il s'agit d'une généralisation des problèmes d'ordonnancement sur machines parallèles (typiquement $P||C_{\max}$), où chaque objet/tâche est disponible en plusieurs exemplaires identiques, chacun devant être exécuté par une machine différente, mais pas nécessairement simultanément. C'est donc, aussi, une relaxation des problèmes d'ordonnancement multi-processeurs [Dro96], où les différents exemplaires d'une tâche doivent être exécutés en parallèle, ou un cas particulier d'ordonnancement avec conflits [BC96, BJW94, EHKR09, KL17].

Afin de décrire facilement les rangements multiboîtes, nous avons étendue la classification de Graham et al. [GLLRK79] :

- (champ α) on utilise la notation B pour indiquer un rangement multiboîte ; les précisions $B5$ et Bm indiquent un nombre fixé de boîtes, respectivement 5 et m ;
- (champ β) on utilise les notations proposées pour l'ordonnancement de tâches multiprocesseurs pour décrire les contraintes de placement, en particulier :
  - $size_j$ indique que chaque objet/tâche j a une largeur (nombre de copies) pré-définies ;
  - set_j indique en plus que seules certaines boîtes sont compatibles avec certains objets ;
  - any_j indique qu'un objet/tâche j peut être rangé dans un nombre quelconque de boîtes, mais que sa hauteur/durée est fonction, à préciser, de ce nombre.
- (champ γ) on utilise les notations suivantes pour décrire l'objectif :
  - $H_{\max}$ pour la minimisation de la hauteur des boîtes (correspond au $C_{\max}$ classique) ;
  - M pour la minimisation du nombre de boîtes (correspond au critère classique pour du *bin-packing*) ;
  - N (resp. W) pour la minimisation du nombre d'objets rangés (resp. nombre pondéré).

À titre d'exemples : le problème $Bm|size_j|H_{\max}$ correspond à une expertise de documents par m spécialistes, où chaque document doit être revu par un nombre défini d'experts selon son importance, et où il faut terminer le plus vite possible ; la Figure 1.1 illustre ce cas. On notera $Bm|set_j|H_{\max}$ si certains documents ne peuvent pas être donnés à certains experts pour des problèmes de compétences ou de conflits d'intérêts. Le problème $B|size_j|M$ correspond, lui, à la sauvegarde de fichiers sur un nombre minimum de disques, avec d'autant plus de copies de chaque fichier que ce dernier est important, tandis que le problème $B|any_j|N$ peut représenter un projet logiciel où la durée de développement de chaque fonctionnalité dépend du nombre de programmeuses qu'on lui alloue, et où il faut terminer le maximum de fonctionnalités dans le temps imparti.

1. Le nom a été proposé dans [FBL01] mais s'est révélé maladroit car l'objectif principalement étudié (minimiser la taille des boîtes) correspond à un problème d'ordonnancement.



objet j	1	2	3	4	5	6
hauteur h_j	3	4	2	1	6	5
largeur $size_j$	1	2	2	1	2	3

FIGURE 1.1 – Une instance de $B|size_j|H_{\max}$ avec une solution réalisable de valeur 12 (gauche) et une solution optimale de valeur 11 (droite).

Les problèmes de rangements multiboîtes généralisent plusieurs problèmes classiques de la recherche opérationnelle comme Partition ($B2|size_j|H_{\max}$), $P||C_{\max}$ ($B|size_j = 1|H_{\max}$), le *bin-packing* ($B|size_j = 1|M$), ou le sac-à-dos ($B1|size_j, h_j = 1|W$). Ces problèmes étant NP-complets, la plupart des cas de rangements multiboîtes le sont eux-aussi. La Figure 1.2 résume les principaux résultats que j'ai obtenus [LFB06].

Les cas polynomiaux correspondent à des algorithmes «à la McNaughton» [McN59] avec éventuellement un tri. Lorsque le nombre de boîtes est fixé (Bm) on peut résoudre les problèmes par programmation dynamique en généralisant l'algorithme classique pour le problème Partition (voir [GJ79] pour cet algorithme). Le cas le plus remarquable est sans doute $B|size_j|\cdot$ dont l'appartenance à NP reste une question ouverte — contrairement aux autres cas, une solution n'est pas un certificat de taille polynomiale (une instance se décrit en $O(n, \log(m))$, une solution en $O(n, m)$), et aucun autre certificat n'a été déterminé.

Chacun de ses problèmes, séparément, n'a qu'un intérêt limité. Pris ensemble, ils dessinent par contre une carte qui fait sens et qui raconte une histoire. Ici, il apparaît clairement qu'avoir des objets de taille unitaire ($h_j = 1$) rend les problèmes faciles tandis qu'avoir des objets de largeur unitaire ($size_j = 1$) ne les simplifie pas. On voit également que, sauf contrainte additionnelle simplificatrice, avoir un nombre fixé de boîtes (Bm) est nécessaire et suffisant pour résoudre en temps pseudo-polynomial. Plus localement, on voit que le critère W est plus difficile que le critère N... mais que cela n'a une incidence que si les objets ne sont pas de hauteur unitaire. Ainsi, cette histoire permet d'aller au delà de la seule difficulté d'un cas et on gagne l'explication de ce qui fait cette difficulté. De plus, cette histoire est beaucoup plus facile à lire et accessible que le récit technique, décousu et parfois confus des démonstrations elles-mêmes [2].

Pour aller plus loin dans l'analyse, je me suis intéressé à des algorithmes à garanties de performances, principalement pour le cas $B|size_j|H_{\max}$. Pour un problème de minimisation, on appelle *α-approximation* un algorithme qui, pour tout instance, renvoie en temps polynomial une solution dont la valeur est au plus α fois la valeur optimale. Par exemple, une 2-approximation garantit de ne jamais faire pire que 2 fois l'optimum.

Le cas $Bm|size_j|H_{\max}$ généralise $Pm||C_{\max}$. Il est donc naturel d'étudier la généralisation des classiques algorithmes de listes de Graham [Gra69]. L'idée de ces algorithmes est de placer chaque objet l'un après l'autre dans les boîtes les moins pleines («meilleur placement»). Dans le cas de $Pm||C_{\max}$ (i.e. $Bm|size_j = 1|H_{\max}$) un tel algorithme est une $(2 - 1/m)$-approximation pour un ordre quelconque des objets/tâches et est une $(4/3 - 1/3m)$-approximation si les objets/tâches sont pris par ordre décroissant des hauteurs/durées [Gra69]. Dans le cas des rangements multiboîtes $Bm|size_j|H_{\max}$, la garantie de $(2 - 1/m)$ pour un ordre quelconque se conserve et les preuves se généralisent facilement. Pour le cas de l'ordre des hauteurs décroissantes des objets, j'ai prouvée une garantie de $4/3$ au prix d'adaptations non triviales des démonstrations [LFB05]. La question reste ouverte de savoir si l'écart de $1/3m$ est inhérent aux problèmes multiboîtes ou est une limitation de la preuve proposée.

Une autre approximation a été proposée, basée sur le principe du «diviser pour régner». L'idée de base de l'algorithme est de fusionner optimalement deux solutions partielles en complétant la boîte la plus remplie de la première solution avec les objets de la boîte la moins remplie de la deuxième solution et ainsi de suite. Si on part de solutions partielles triviales (un objet par solution partielle) et qu'on fusionne les solutions dans un ordre quelconque, on obtient une $(2-1/m)$-approximation [Lem04]. Il est donc tout à fait raisonnable de penser

2. J'aime beaucoup les démonstrations, mais toutes ne sont pas jolies et éclairantes.

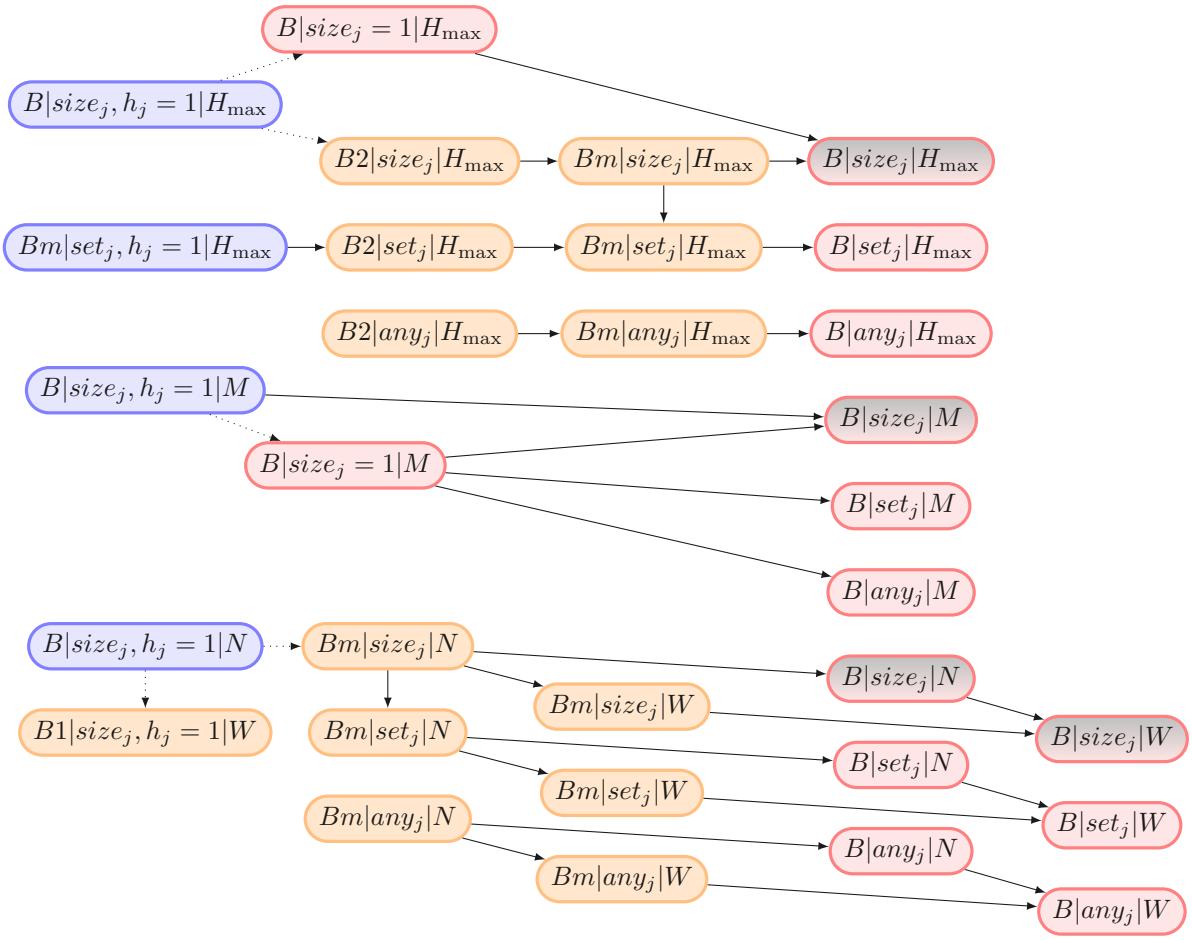


FIGURE 1.2 – Complexité des problèmes de rangements multiboîtes
Problèmes polynomiaux (bleus), NP-complets mais admettant un algorithme pseudo-polynomial (oranges) et NP-complets au sens fort (rouges). L'appartenance à NP des problèmes rouges-gris n'a pas été prouvée. Une flèche $A \to B$ indique que A est un cas particulier de B (trait plein), ou peut se ramener à un cas particulier de B (trait pointillé).

qu'un choix approprié des solutions à fusionner permette d'améliorer la garantie. En effet, cet algorithme est à rapprocher de la ≪méthode par différence≫ de Karp et Karmarkar [KK82], qui a une garantie de $7/6$ pour $m = 2$ [FM87], et une garantie de $(4/3 - 1/3m)$ dans le cas général [MKAL07].

Pour terminer sur les algorithmes d'approximation et sur les rangements d'objets multiboîtes, notons que l'algorithme de ≪meilleur placement≫ permet de construire un schéma d'approximation polynomial (PTAS[3]) tandis que l'algorithme pseudo-polynomial évoqué pour résoudre $Bm|size_j|H_{\max}$ permet de construire un schéma d'approximation fortement polynomial (FPTAS[3]) [LFB05, LFB06].

1.1.2 Ordonnancement de tâches à durée variable

L'ordonnancement de tâches à durée variable a été le sujet de la thèse de Florian Fontan [Fon19], que j'ai encadrée avec Nadia Brauner (G-SCOP). Dans ce problème, chaque tâche j a une échéance d_j et un profit $w_j(p_j)$ qui dépend de la durée p_j de la tâche. L'objectif est alors de maximiser la somme des profits (ce que l'on note $-\sum_j w_j(p_j)$ dans la notation $\alpha|\beta|\gamma$ afin de garder une minimisation de l'objectif). La grande originalité de ce problème est que l'on doit *décider* de la durée des tâches ; il est aussi possible de ne pas ordonnancer certaines (avec la convention qu'une telle tâche a une durée $p_j = 0$).

Bien entendu, la forme des fonctions de profit est essentiel. La Figure 1.3 propose quelques premiers exemples de telles fonctions. Si chaque tâche a un profit de la forme de la Figure 1.3(a) alors le problème

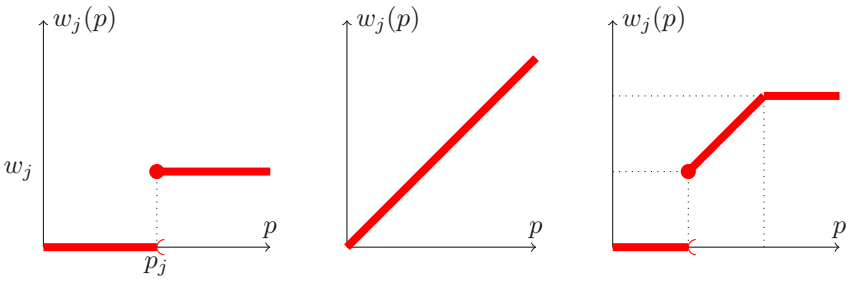


FIGURE 1.3 – Exemples de fonction de profit pour des tâches de durée variable
(a) ordonnancement classique ; (b) profit linéaire ; (c) profit pour l'observation d'objets célestes (section 1.2.1).

$1||-\sum_j w_j(p_j)$ n'est autre que la minimisation du nombre pondéré de tâches en retard, c'est-à-dire $1||\sum w_j U_j$ voire $1||\sum U_j$ si tous les profits sont égaux quand chaque tâche est à sa longueur seuil p_j. Le premier est NP-complet, le second est polynomial. Nous constatons dès à présent que l'on bascule facilement d'un côté ou de l'autre de la frontière des problèmes difficiles ou faciles.

Hormis ce cas particulier de la Figure 1.3(a), on peut avoir intérêt à arrêter une tâche à de multiples instants, selon les autres tâches disponibles. Les solutions ont alors des structures originales et le fait de devoir ainsi décider de la durée des tâches est très inhabituel en ordonnancement. Il y a toutefois des liens avec les problèmes où il s'agit de sélectionner les tâches à effectuer [Slo11, SGK13], ou de trouver un compromis entre un critère régulier et des coûts de rejets [ZLY10, SGY12], ainsi qu'avec les problèmes avec durées contrôlables, pour lesquels la durée d'une tâche peut être diminuée par l'allocation de ressources [SSS18].

Nous nous sommes intéressés à la complexité des problèmes d'ordonnancement avec des durées variables pour différentes formes de fonctions, présentées Figure 1.4 avec les notations associées. Ces formes se généralisent en *Piecewise Linear Profit* (PLP) qui sont des fonctions croissantes, linéaires par morceaux et valant 0 en 0. Une PLP est ainsi définie en K morceaux et le profit sur l'intervalle k, c'est-à-dire pour une durée p telle que $p_j^{k-1} \leq p < p_j^k$, vaut

$$w_j(p) = w_j^k + b_j^k(p - p_j^k).$$

3. (F)PTAS (*(Fully) Polynomial-Time Approximation Scheme*) : $(1+\epsilon)$-approximation pour ϵ fixé ; le temps d'exécution peut croître exponentiellement en $1/\epsilon$ pour un PTAS, mais ne peut croître que polynomialement pour un FPTAS. L'intérêt de ces jolis algorithmes est avant tout théorique car les temps d'exécution réels sont souvent prohibitifs. Les ≪littéraires≫ préféreront peut-être la définition qu'Oscar Wilde a involontairement donnée : ≪If you are not too long, I will wait for you all my life.≫

Toutes les données (seuils de durées p_j^k, pentes b_j^k, seuil de profit w_j^k) sont supposées entières. Les différentes formes proposées sont intimement liées et sont hiérarchisées en termes de complexité, comme le résume la Figure 1.5.

La Figure 1.6 donne un extrait de la carte de complexité qui a été tracée pour les problèmes à durée variable (voir [Fon19] et [FLBa], en préparation). Cette carte, bien que partielle, est déjà beaucoup plus étendue que celle des problèmes multiboîtes. Il est donc plus difficile d'en raconter l'histoire simplement. On peut toutefois en tirer quelques enseignements. Le premier, c'est que la complexité de ces problèmes est très sensible au moindre paramètre. Chaque type de fonction de profit étant définie par plusieurs paramètres, cela engendre un foisonnement de cas entre lesquels serpentent les lignes de démarcation entre problèmes polynomiaux, faiblement NP-complets et fortement NP-complets. Ainsi, chaque cas mérite une attention particulière car il est difficile d'en deviner à l'avance la complexité précise. On peut toutefois remarquer qu'au delà de un ou deux degrés de liberté (*i.e.* paramètres distincts pour les tâches), les problèmes sont en général difficiles, voire fortement difficiles si le nombre de machines n'est pas fixé.

Cette variété se retrouve dans les techniques de preuves de beaucoup de ces résultats, en particulier pour la construction d'algorithmes polynomiaux. Par exemple, on peut parfois s'en sortir avec un simple tri ($P|LP$) ou de la programmation linéaire ($1|LBP$), quand d'autres cas nécessitent de concevoir des algorithmes dédiés ($1|LBPST, p_j^{\min} = p^{\min}, b_j = b$) ou implicites en passant par la résolution d'un nombre polynomial de flots de taille polynomiale ($P|LPSTIP, p_j^{\min} = p^{\min}, d_j = d$ ou $P|LPSTIP, p_j^{\min} = p^{\min}, b_j = b$). Le point commun est que, à chaque fois, il faut exploiter finement des propriétés de dominance spécifiques à chaque cas.

Pour en terminer avec les problèmes d'ordonnancement à durée variable, arrêtons nous sur le coin nord-nord-ouest de la carte, qui mérite une attention particulière avec le problème ouvert $1|LP, r_j| - \sum w_j(p_j)$. Ce problème, le premier à considérer quand on ajoute des dates de disponibilité, semble se situer sur une ligne de crête particulièrement étroite entre problèmes faciles et difficiles. En effet, sans dates de disponibilité le problème est trivialement polynomial, même sur machines parallèles (il suffit de placer chaque tâche, par ordre de pente (b_j) décroissante, dès qu'une machine est disponible et jusqu'à son échéance). De plus le problème $1|LP, r_j| - \sum w_j(p_j)$ reste polynomial si $b_j = b$ ou si la préemption est autorisée. Par contre la généralisation au profit de forme LBP est déjà fortement NP-complète, même lorsqu'un paramètre est fixé ($b_j = b$ ou $w_j^{\max} = w^{\max}$). Ainsi le profit linéaire, qui semble commode, est au contraire la cause de l'ambiguïté du problème car il est compliqué de décider quand arrêter (voire commencer) une tâche, sauf à imposer la moindre restriction, qui rend alors le problème facile. Au final, l'étude de ce problème est aussi intéressante qu'intrigante car, malgré un énoncé très simple du problème, les solutions optimales ne semblent pas avoir de structure particulière et ne ressemblent à rien de classique.

1.2 Résolutions *ad hoc*

Après avoir présenter des études de complexité, qui permettent de caractériser au mieux un problème ou une famille de problèmes, je vais présenter deux cas de résolutions, selon deux approches qui n'ont guère de commun que la volonté de répondre de la manière la plus adaptée possible à une question initiale.

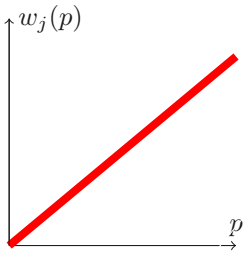
1.2.1 Observation d'objets célestes

Le premier cas est en fait le pendant pratique, et l'origine, de l'étude des ordonnancements de tâches à durée variable (section 1.1.2). Le problème nous a été soumis par des astro-physiciens et apparaît lors de l'observation d'objets célestes, en particulier pour la recherche d'exo-planètes. Dans ce cadre chaque tâche j (une observation possible d'un objet céleste) est caractérisée par un poids w_j (l'importance scientifique de l'objet), une durée p_j (le temps d'observation nécessaire) ainsi que plusieurs fenêtres de disponibilité $[r_j^k, d_j^k]$ (correspondant à la fenêtre de visibilité de l'objet j pendant la nuit k). On dispose d'une unique machine (le *Very Large Telescope* de l'ESO, au Chili) et l'enjeu est de maximiser le poids total des tâches ordonnancées. Une instance typique correspond au placement de quelques centaines d'observations potentielles sur quelques dizaines de nuits, à raison de 5 à 10 observations effectives par nuit.

Pour résoudre ce problème, les astro-physiciens disposent d'un logiciel, SPOT, qui implante une méta-heuristique de type recuit simulé. Ce logiciel trouve des solutions d'assez bonne qualité, mais relativement lentement. De plus, il ne tire pas partie de certaines particularités du problème et il est difficile à adapter à certains aspects qui ne sont pour l'instant pas pris en compte.

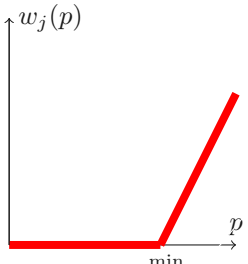
Linear Profit (LP)

$$w_j(p) = b_j p$$



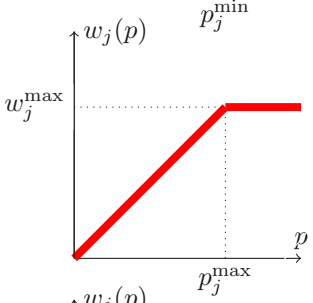
Linear Profit with Setup Time (LPST)

$$w_j(p) = \begin{cases} 0 & \text{for } p < p_j^{\min} \\ b_j(p - p_j^{\min}) & \text{for } p \geq p_j^{\min} \end{cases}$$



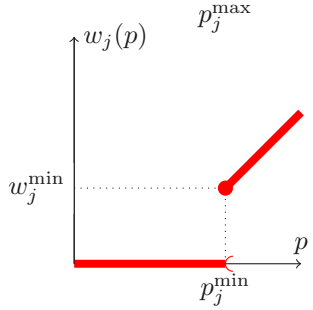
Linear Bounded Profit (LBP)

$$w_j(p) = \begin{cases} b_j p & \text{for } p < p_j^{\max} \\ b_j p_j^{\max} & \text{for } p \geq p_j^{\max} \end{cases}$$



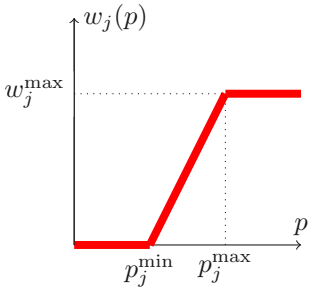
Linear Profit with Setup Time and Initial Profit (LPSTIP)

$$w_j(p) = \begin{cases} 0 & \text{for } p < p_j^{\min} \\ w_j^{\min} + b_j(p - p_j^{\min}) & \text{for } p \geq p_j^{\min} \end{cases}$$



Linear Bounded Profit with Setup Time (LBPST)

$$w_j(p) = \begin{cases} 0 & \text{for } p < p_j^{\min} \\ b_j(p - p_j^{\min}) & \text{for } p_j^{\min} \leq p < p_j^{\max} \\ b_j(p_j^{\max} - p_j^{\min}) & \text{for } p \geq p_j^{\max} \end{cases}$$



Linear Bounded Profit with Setup Time and Initial Profit (LBPSTIP)

$$w_j(p) = \begin{cases} 0 & \text{for } p < p_j^{\min} \\ w_j^{\min} + b_j(p - p_j^{\min}) & \text{for } p_j^{\min} \leq p < p_j^{\max} \\ w_j^{\min} + b_j(p_j^{\max} - p_j^{\min}) & \text{for } p \geq p_j^{\max} \end{cases}$$

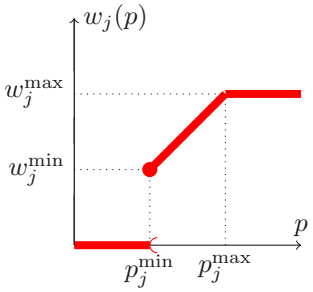


FIGURE 1.4 – Fonctions de profits pour des durées variables.

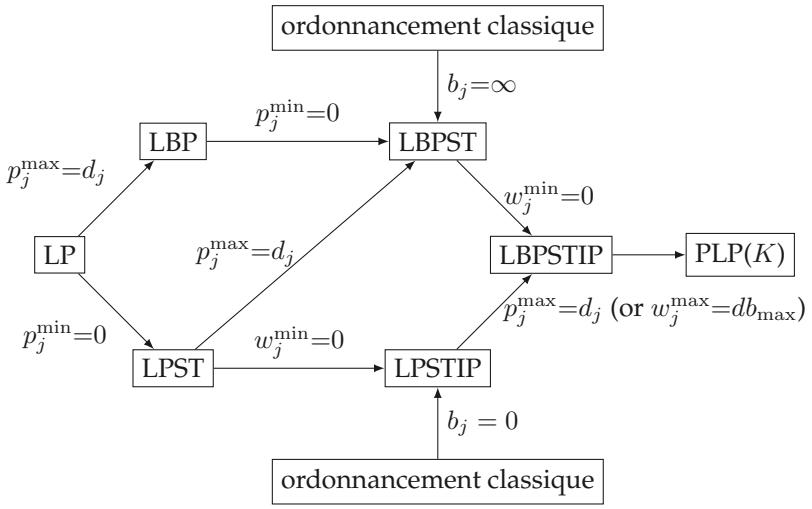


FIGURE 1.5 – Hiérarchie de complexité des profits pour des durées variables.
Une flèche de A à B signifie que A est un cas particulier de B, et le texte indique la restriction correspondante. $PLP(K)$ est le cas particulier de PLP pour lequel le nombre de morceaux est borné par K.

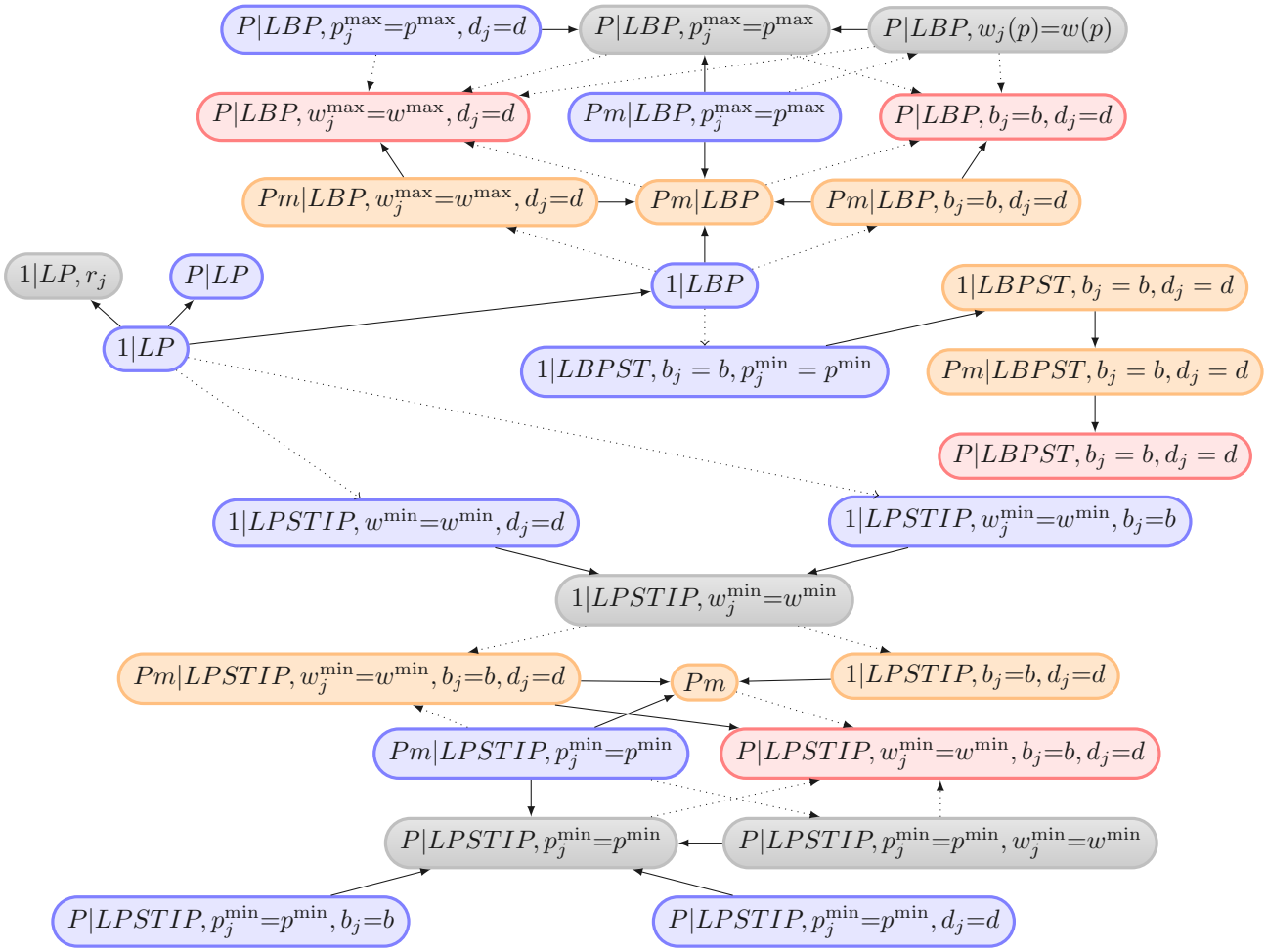


FIGURE 1.6 – Quelques résultats de complexité pour des durées variables.
Problèmes polynomiaux (bleus), NP-complets mais admettant un algorithme pseudo-polynomial (oranges) et NP-complets au sens fort (rouges). La complexité des problèmes gris n'est pas connue. Une flèche $A \to B$ indique que A est un cas particulier de B (trait plein), ou est similaire à un cas particulier de B (trait pointillé).

La première particularité est une relation particulière entre durée d'observation et fenêtre de visibilité : en effet, il faut observer un objet céleste autour d'un instant précis, son méridien, ce qui fait que la durée d'observation est au moins égale à la moitié de la durée de la fenêtre de visibilité. Cette propriété structure très fortement les solutions car l'ordre des tâches pour une nuit donnée est ainsi imposé. Nous avons alors montré que l'optimisation d'une nuit donnée reste un problème NP-complet mais peut se résoudre par une programmation dynamique extrêmement efficace pour les tailles d'instances pratiques, qui peut être utilisée au sein d'une recherche locale d'une part, et d'un algorithme de type *branch-and-price* [CCB+16] très performant mais incapable d'incorporer certaines caractéristiques du problème réel (observations spéciales de calibrage ou observations ne respectant pas la contrainte de méridien), d'autre part. Incorporée à SPOT, l'optimisation d'une nuit a tout de même permis une amélioration pratique des résultats [LRB+16].

Pour aller plus loin et répondre au mieux aux besoins des astro-physiciens, plusieurs points d'amélioration par rapport à SPOT ont été pris en compte dans le cadre de la thèse de Florian Fontan [Fon19] :

— améliorer la qualité des solutions et les temps de calcul ;
— prendre en compte différents usages : du très court terme (ré-optimisation d'une nuit en cas d'intempérie), du moyen terme (planification d'une campagne de quelques jours), et du long terme (anticipation pour la réservation des étoiles à observer) ;
— utiliser la possibilité de réduire la durée d'observation même si cela dégrade un peu la qualité et donc l'intérêt, selon des fonctions similaires à la Figure 1.3(c).

(C'est bien sûr ce dernier aspect qui nous a conduits à mener l'étude de complexité présentée section 1.1.2.)

La méthode proposée est une recherche à voisinage large [PR10] : une solution courante est améliorée itérativement par un procédé de destruction/reconstruction. Dans notre cas, il s'agit de vider des nuits puis de les remplir au mieux avec les observations encore disponibles. La clé de cette procédure est l'algorithme qui remplit *une* nuit et qui va pouvoir être adapté finement aux contraintes considérées.

Ainsi, utiliser la programmation dynamique de [CCB+16] permet de résoudre le problème «pur», ce qui permet d'évaluer les performances de la méthode en comparant aux résultats du *branch-and-price*. De plus, on peut l'adapter aux contraintes du problème telles que prises en compte par SPOT. Pour prendre en compte la réduction possible des durées, nous avons proposés d'une part un autre algorithme de programmation dynamique permettant de décider optimalement entre une durée minimale et une durée maximale, et d'autre part une méthode gloutonne permettant un choix continu des durées.

Ces différentes versions permettent une analyse des résultats selon plusieurs critères. D'une part, en termes de performances, la recherche à voisinage large comble la moitié de l'écart qui restait entre les meilleures solutions proposées par la recherche locale et les bornes calculées par *branch-and-price* de [CCB+16] ; on a donc bien une méthode plus performante que les algorithmes existants. D'autre part, cela permet de mesurer l'intérêt pratiques des durées variables : on améliore systématiquement les solutions obtenues, même si cela reste assez limité avec seulement quelques observations supplémentaires effectuées.

Du point de vue des astro-physiciens, cela permet d'une part d'avoir des solutions en la qualité desquelles on peut avoir confiance ; et d'autre part de décider en connaissance de cause si compliquer la gestion opérationnelle avec des durées variables en vaut la peine.

1.2.2 Production de puces électroniques

Le deuxième cas de résolution *ad hoc* s'inscrit dans le cadre d'une collaboration industrielle avec le CEA-Leti. Ce laboratoire fabrique des puces électroniques en phase de pré-industrialisation : il s'agit donc de s'approcher au mieux d'une efficacité industrielle tout en supportant les nombreuses incertitudes de processus encore peu ou pas matures.

Chaque produit doit passer sur différentes machines selon une gamme de fabrication linéaire et connue et l'enjeu initial était de limiter les retards. Vue sous l'angle de la théorie de l'ordonnancement, il s'agit d'un $J|r_j, chain|\sum w_j T_j$ avec en plus des contraintes sur la disponibilité de ressources additionnelles, les opérateurs. Ce problème est NP-complet au sens fort et a beaucoup de points communs avec le problème couplé d'affectation de personnel et d'ordonnacement que nous avions résolu dans le cadre de la thèse d'Olivier Guyon, co-encadrée par Éric Pinson et David Rivreau (LISA, Angers), et pour lequel nous avions proposées différentes modélisations en programmation linéaire en nombres entiers et une méthode de type *branch-and-cut* [GLPR09, GLPR10, GLPR14]. En fait, bien qu'il existe déjà une abondante littérature sur ce problème et ses voisins, il y aurait de la place pour une n-ième heuristique sophistiquée adaptée à ce cas particulier. Sauf que ce n'est pas ce dont avait besoin les gens du CEA-Leti. En effet, la difficulté réelle n'est pas que le problème soit NP-complet, mais qu'il faille avant tout tenir compte de très nombreuses incertitudes aussi bien sur la

disponibilité des machines et des opératrices que sur les durées de process. Concrètement, il s'agit en fait de planifier une fabrication de produits qui s'étale sur plusieurs mois, où chaque étape peut prendre plusieurs heures, sans pouvoir savoir avec certitude ce qui pourra être fait ne serait-ce que le lendemain.

Face a un tel problème, j'ai rangée ma boîte à outils de recherche opérationnelle, et j'ai adoptée une démarche de génie industriel qui intègre de manière prépondérante les conditions d'usage des solutions proposées. Ce faisant, il s'agissait de tenir compte autant des incertitudes liés à l'ordonnancement lui-même, que de la capacité de faire évoluer les pratiques : une solution incomprise ou que les opérateurs n'arrivent pas à s'approprier serait parfaitement inutile.

Au final, ces travaux n'ont fait l'objet d'aucune publication. Non pas pour des raisons de secret industriel ou de confidentialité, mais simplement parce qu'il n'est ni original ni subtil de gérer des retards en donnant priorité selon les échéances. Les résultats sont ailleurs et, pour moi, plus importants que des indicateurs bibliométriques. Le logiciel que j'ai développé pour l'occasion tourne quotidiennement en production et permet de planifier chaque matin la production de la journée (plusieurs centaines de produits). La solution est intégrée au système d'information ce qui l'a rendue transparente pour les opératrices tout en permettant un meilleur respect des échéances. La planification a été automatisée, tout en laissant la possibilité d'intervenir pour chaque cas jugé particulier, et cela a libéré du temps aux responsables de production qui peuvent le consacrer à leur métier premier — la résolution des problèmes pendant la production —, en disposant d'une meilleure information. Cela signifie un gain de sens et de sérénité pour les personnes impactées. Cela a aussi été l'occasion de rendre les règles de planification plus simples et plus cohérentes et de faire prendre conscience du besoin de compétences spécifiques en gestion de production (besoin qui s'est traduit par une embauche pérenne).

Pour conclure sur une note plus scientifique, notons que le système de production est maintenant beaucoup plus mûr et permet d'envisager des approches plus ambitieuses (ma boîte à outils de recherche opérationnelle n'est jamais loin). Deux problèmes ont été identifiés. L'un, assez classique, vise à améliorer la gestion des pannes et autres absences. L'autre, plus original, vise à améliorer la prédiction et la prise en compte des avances ou des retards, malgré les incertitudes sur les temps, afin d'avoir une exécution plus régulière et fluide.

1.3 Discussion

Ce chapitre m'a permis de décrire plusieurs de mes contributions en ordonnancement. Parmi les perspectives de recherche, j'en retiens deux que j'estime pouvoir être aussi amusantes que formatrices. La première concerne les algorithmes d'approximation pour les rangements multiboîtes, pour lesquelles de meilleures garanties pourraient être prouvées ou des versions améliorées pourraient être proposées, par exemple en s'inspirant de travaux récents tels que [DCST19, DCS20]. La deuxième, et principale, est de poursuivre l'étude des problèmes d'ordonnancement de tâches à durée variable : la complexité d'un certain nombre de cas reste ouverte, et tout est à faire du côté des approximations. Une attention particulière est à porter à $1|r_j, LP| - \sum w_j(p_j)$: ce cas est véritablement très intrigant et particulier, et il faudra sans doute faire preuve d'originalité pour le résoudre.

Au delà de ces perspectives, je voudrais maintenant prendre un peu de recul quant au deux aspects de l'ordonnancement présentés : l'analyse de la complexité des problèmes d'une part et la proposition de méthodes de résolution sur mesure d'autre part.

En premier lieu, je veux revenir sur le lien entre ces deux aspects. Comme je l'ai signalé en introduction de la section 1.1, connaître la complexité d'un problème permet de s'orienter vers des méthodes de résolution adaptées en conséquences avec, en particulier, un compromis à accepter entre qualité de la solution et vitesse de résolution, lorsque le problème est NP-complet. Les études détaillées de complexité, comme celles présentées, permettent d'aller plus loin. Ainsi, dans le cas des rangements multiboîtes, les algorithmes d'approximations contribuent concrètement à une meilleure résolution : ce sont des heuristiques efficaces et même asymptotiquement optimales — il suffit donc d'avoir assez d'objets pour qu'elles soient terriblement efficaces ou, pour le dire autrement, plus le problème est gros, plus il est facile de le résoudre de manière quasi optimale. Ces approximations peuvent aussi s'incorporer facilement comme briques de base pour construire une méta-heuristique efficace, comme cela a été illustré dans ma thèse [Lem04].

D'un autre côté, force est de constater qu'aucun des algorithmes proposés pour l'analyse des tâches à durée variable n'a finalement servi pour la planification des observations d'objets célestes. Une analyse en pure perte ? Pas vraiment. Premièrement, les algorithmes pseudo-polynomiaux développés dans [CCB+16] se sont révélés très efficaces en pratique et on a ainsi eus des éléments concrets de résolution. Deuxièmement,

l'analyse de complexité permet de mettre en avant des cas proches de celui étudié, mais plus faciles. Elle permet alors d'orienter la résolution : on sait ce qu'il serait souhaitable de simplifier. Cela peut prendre deux formes : soit aboutir à un meilleur algorithme d'un problème dégradé (c'est ce qui a été fait en implémentant une version qui n'autorise que 2 durées possibles), ce qui peut aboutir à d'excellent résultats par la réduction de l'espace de recherche ; soit plus prosaïquement discuter avec les usagers du problème traité pour s'assurer qu'on ne s'attaque pas à plus général que nécessaire et pour voir si certaines simplifications, dont on sait qu'elles seraient profitables à l'optimisation, ne seraient pas légitimes.

Cela nous amène au deuxième point de cette discussion : que veut dire ≪résoudre un problème≫ ? En mathématiques, on nomme d'abord conjecture une question bien posée, puis théorème lorsqu'on a trouvée une démonstration adéquate. Problème résolu. Lorsqu'on fait de l'aide à la décision, la tentation est grande, et parfois souhaitable, d'adopter une approche similaire. On ne garde alors qu'une version idéalisée d'un problème, qui bien souvent n'était pas le nôtre, et on y apporte une réponse bien jolie mais peu pratique pour les personnes réellement concernées. La tentation est aussi grande, et souvent aussi souhaitable, d'adopter une approche opposée : renoncer à une solution mathématiquement étayée pour se concentrer sur des heuristiques de bon sens. L'opposition n'est en fait que rhétorique car les deux approches se complètent et s'enrichissent mutuellement et apporter une aide à la décision adéquate revient souvent à trouver le bon équilibre entre apporter des solutions exactes à une réalité approchée, et apporter des solutions approchées à une réalité exacte. Dans tous les cas, il ne faut pas perdre de vue pour qui on résout un problème et tenir compte des conditions d'usages — mais sans s'y restreindre outre mesure, afin de ne pas confondre les habitudes des usagers avec les contraintes du système.

Enfin, je voudrais insister sur le fait que ≪résoudre≫ ne signifie pas nécessairement proposer une méthode de résolution, même si ça en est l'expression la plus courante. Parfois, la méthode pré-existe et ce qui manque, pour une résolution en confiance, c'est la caractérisation de la méthode — en d'autres termes, un théorème. Comme illustration, je prendrais un problème que je n'ai pas détaillé précédemment : l'ordonnancement stochastique avec abandons. Le cas typique est le ≪tri≫ de patients aux urgences. Lorsque des personnes arrivent, un premier diagnostic est fait pour évaluer la gravité immédiate et les risques de complication et, sur cette base, on décide de l'ordre de prise en charge. De telles conditions nécessitent des décisions rapides ; par conséquent des règles de priorités sont très opérationnelles et sont donc souvent utilisées. La question qui se pose est alors de mesurer l'efficacité de cette organisation ou au moins d'éclairer sur les conséquences d'un tel choix. Ainsi, les travaux sur l'ordonnancement stochastique avec impatience ou abandon, menés avec Jean-Philippe Gayon et commencés dans le cadre de la thèse d'Alexandre Salch, montrent que les conditions pour qu'une telle gestion soit optimale sont très fortes [SGL13, CGL][4] et rarement réunies dès qu'il y a une certaine diversité des patients. Ce genre de résultat permet de questionner les solutions mises en œuvre pour les améliorer ou, au moins, les utiliser en connaissance de cause.

Pour terminer cette discussion, j'aimerais revenir sur la notion de complexité. La théorie de la complexité est pleine de jolies surprises mais aussi d'écueils et de bizarreries — la première, et non des moindres, et qu'on ne sait toujours pas si les problèmes difficiles sont réellement différents des problèmes faciles (conjecture P $\neq$ NP). C'est toutefois les limites pratiques de la notion de ≪complexité d'un problème≫ que je veux examiner ici.

Le premier travers vient de la notion de problème : celle-ci est générique et concerne l'ensemble de toutes les instances possibles — un ensemble démesuré mais rendu nécessaire par notre incapacité à mieux qualifier les instances qui nous intéressent, au delà de restrictions relativement triviales. L'analyse effectuée étant au pire cas, on se retrouve avec une vision horriblement pessimiste de la difficulté concrète d'un problème. Un exemple est fourni par le plus simple des problèmes difficiles, Partition ($P2||C_{\max}$). Les petites instances sont résolues extrêmement efficacement par programmation dynamique, et les grandes instances par recherche locale ; ce qui est réellement compliqué, en pratique, est en fait de trouver une instance difficile ! Ainsi, pour une résolution pratique, la complexité de problème ne répond pas tout à fait à la bonne question. Ce que l'on voudrait savoir avant tout c'est si les instances qui nous concernent sont compliquées à résoudre ou pas. Il y a derrière cela deux perspectives : la première, être capable de caractériser voire définir les instances qui nous concernent ; la deuxième, être capable de deviner voire d'apprendre si une instance donnée est compliquée. Je reviendrai là dessus dans la conclusion de ce document.

4. Au passage, signalons que le problème stochastique décrit dans [SGL13] se ramène à un problème déterministe sous des hypothèses classiques de durées exponentielles et d'indépendance. Le problème d'ordonnancement qui en résulte propose un critère d'optimisation régulier mais très inhabituel (où se mélangent sommes et produits), pour lequel la complexité est ouverte.

Le deuxième travers est d'ordre philosophique : la théorie de la complexité est faite du point de vue d'une machine qui calcule une réponse, pas de l'être humain qui va utiliser ou subir cette réponse. Ainsi, la notion de complexité ne prend pas du toute en compte le caractère compréhensible ou explicatif d'un algorithme ni même de la solution proposée. En cela un algorithme est fondamentalement différent d'un expert ; étrangement on se défie plus communément d'un expert que d'un algorithme. Certains algorithmes, et donc les problèmes qu'ils résolvent, sont «faciles» car la plupart des gens les conçoivent naturellement et sont capables de les mettre en œuvre simplement. D'autres algorithmes sont «difficiles», avec toute une gradation de complexité selon ce que l'on conçoit ou comprend ou exécute aisément[5].

Derrière ces propos, il y a deux enjeux. D'une part une confiance étayée, nécessaire au bon usage des algorithmes et de leurs solutions. D'autre part un enjeu pédagogique car avoir des algorithmes capables de calculer devrait aussi rendre les personnes plus intelligentes ou autonomes afin de ne pas se tromper sur qui, de l'humain ou de la machine, est au service de l'autre.

5. Ce terrain a été exploré à travers plusieurs projets d'élèves de L3 ou M1 que j'ai encadrés, portant sur la conception d'instances de difficulté contrôlée pour le jeu Pilzegal (`http://www.kamick.org/lemaire/pilzegal.html`) ou pour le voyageur de commerce. Cela rejoint aussi certains travaux en psychologie [MO96, GJP00].

CHAPITRE 2

Aide au diagnostic médical

Ce chapitre est indépendant du précédent. Si celui-là relevait complètement de la recherche opérationnelle, celui-ci s'inscrit avant tout dans les sciences de données. Dans un premier temps (section 2.1), je présente les principaux résultats que j'ai obtenus dans le cadre de l'aide au diagnostic médical, d'un point de vue applicatif. Je détaille ensuite (section 2.2) les contributions méthodologiques qui ont été développées pour obtenir certains de ces résultats ; ces contributions relevant de l'Analyse Logique de Données, une méthode combinatoire d'apprentissage automatique, nous retrouvons alors des liens avec l'optimisation.

Avant d'entrer dans le vif du sujet, définissons rapidement le vocabulaire et les objets que nous allons utiliser dans ce chapitre, et précisons le contexte des études.

Nous considérons des données médicales, qui peuvent être assimilées à des tableaux de valeurs numériques (y compris booléennes). Un *jeu de données* est ainsi une matrice dont les lignes correspondent aux *observations* (les patients) et les colonnes correspondent aux *variables* ou *attributs* (les caractéristiques de ces patients). Nous nous plaçons dans le cadre de l'apprentissage supervisé : l'une des variables est la *variable à expliquer*, tandis que les autres sont les *variables explicatives* ; il s'agit alors de qualifier au mieux la variable à expliquer grâce aux variables explicatives. La Figure 2.1 donne un exemple d'un tel jeu de données ; sur cet exemple, la variable à expliquer est la malignité (variable Cl) d'un cancer du sein.

La qualification commence par la détection de *marqueurs*, c'est-à-dire de variables qui renseignent sur la variable à expliquer. Sur l'exemple, CT peut être considéré comme un marqueur car $\mathrm{CT} \geq 8$ implique un cancer malin ; au contraire M n'est pas un marqueur car il est non discriminant (constant pour l'ensemble du jeu de données). Les marqueurs, pris séparément, ne suffisent pas à poser un diagnostic mais sont autant de symptômes qui aident le praticien dans cette tâche ; ils sont aussi des effets qui permettent à la chercheuse de mieux comprendre une pathologie. Au delà des marqueurs, on cherche en général un *modèle* de la variable à expliquer en fonction des variables explicatives, c'est-à-dire une formule de celle-là en fonction de celles-ci. Toujours sur l'exemple, la formule « $\mathrm{CT} \geq 8$ ou $\mathrm{MA} \geq 8$ » est un modèle de la malignité — un modèle qui semble excellent puisqu'il ne fait aucune erreur, les observations satisfaisant cette formule étant très exactement celles correspondant à un cancer malin ($\mathrm{CL} = 4$). Lorsque la variable à expliquer est une variable ne comportant qu'un tout petit nombre de valeurs (en général 2), il s'agit d'un problème de *classification* ; lorsque la variable à expliquer prend des valeurs réelles, il s'agit d'un problème de *régression*. La détection de marqueurs met avant tout en jeu des outils statistiques [Sap11]. La construction de modèles relève de l'apprentissage automatique (voir [Tuf17, WHFH16] pour les algorithmes classiques).

Les problèmes sur lesquels j'ai travaillé ont quelques spécificités. Ainsi, les jeux de données sont numériques, donc simples, mais peuvent comporter des données manquantes voire des erreurs. Ensuite, les données

id	CT	UCSi	UCSh	MA	SECS	BN	BC	NN	M	Cl
1000025	5	1	1	1	2	1	3	1	1	2
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
733639	3	1	1	1	2	?	3	1	1	2
563649	8	8	8	1	2	?	6	10	1	4
601265	10	4	4	6	2	10	2	3	1	4
606140	1	1	1	1	2	?	2	1	1	2
606722	5	5	7	8	6	10	7	4	1	4

Signification	
id	*Sample code number (identfiant)*
CT	*Clump Thickness (1–10)*
UCSi	*Uniformity of Cell Size (1–10)*
UCSh	*Uniformity of Cell Shape (1–10)*
MA	*Marginal Adhesion (1–10)*
SECS	*Single Epithelial Cell Size (1–10)*
BN	*Bare Nuclei (1–10)*
BC	*Bland Chromatin (1–10)*
NN	*Normal Nucleoli (1–10)*
M	*Mitoses (1–10)*
Cl	*Class (2 for benign, 4 for malignant)*

FIGURE 2.1 – Exemple de jeu de données.
Onze lignes du jeu de données *Breast Cancer Wisconsin*, disponible sur UCI Machine Learning Repository.

sont de petite taille, en particulier le nombre d'observations qui est de l'ordre de quelques dizaines, ce qui ne va pas sans poser des soucis de validation des résultats. Enfin, les résultats produits doivent être facilement compréhensibles et exploitables par des médecins, ce qui impose des contraintes forte quant à la forme des modèles, parfois au détriment de la performance. Dans tous les cas, on est à l'opposé du *deep-learning* aussi volontiers opaque que gourmand en données.

Dans la suite de ce chapitre, je présente les applications médicales sur lesquelles j'ai travaillé (section 2.1) puis mes contributions méthodologiques à l'apprentissage automatique (section 2.2).

2.1 Études de troubles de la croissance

2.1.1 Contexte et spécificités des études

Toutes les applications sur lesquelles j'ai travaillé ont trait à des problèmes de croissance. Deux questions récurrentes sont de savoir la taille que l'enfant fera une fois adulte, ou plus exactement la perte de taille provoquée par sa condition, et de savoir s'il y a un déficit ou un besoin en hormone de croissance. L'enjeu médical est double. D'une part, éviter à des personnes d'avoir une taille anormalement petite, avec les conséquences psychologiques et sociales que cela implique, sans compter les autres retards de développement qui peuvent être liés. D'autre part, éviter d'avoir recours inutilement à de l'hormone de croissance, le traitement nécessitant un suivi sur le long terme, relativement cher, et pouvant avoir des effets secondaires sérieux pour un gain de l'ordre de 0 à 2 écarts-types (de 0 à 11 cm) selon les pathologies [DC11].

Ces travaux sont le fruit d'une collaboration de plus de dix ans avec Raja Brauner, pédiatre et endocrinologue (Assistance Publique-Hôpitaux, Université Paris-Descartes). Ils s'inscrivent dans le cadre d'une recherche médicale ; du point de vue des mathématiques appliquées et de l'informatique, il s'agit surtout d'ingénierie ou de valorisation et on est avant tout sur des *usages* des sciences de données. De ce fait, la méthodologie CRISP-DM [She00, IBM11] est adaptée pour décrire et discuter de manière synthétique la méthodologie de recherche utilisée dans le cadre des études menées ; c'est aussi l'occasion de mieux préciser les spécificités des problèmes étudiés. On distingue ainsi 6 phases (Figure 2.2) :

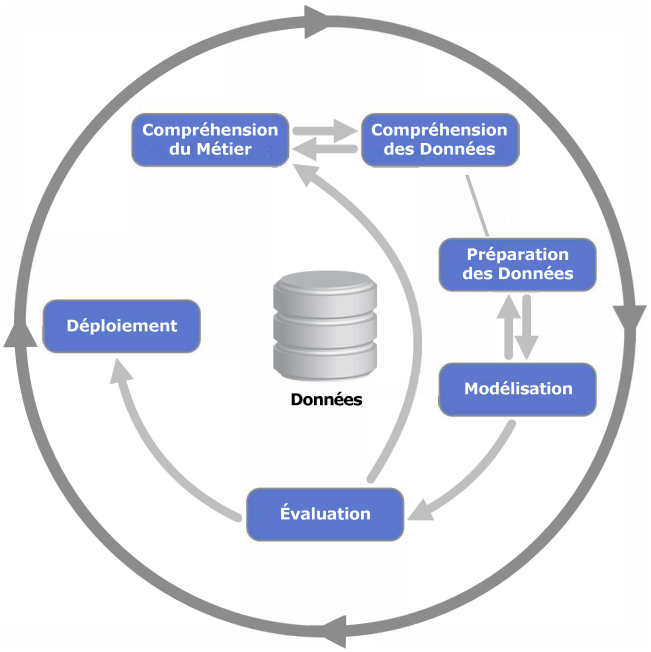


FIGURE 2.2 – Processus CRISP-DM
Source : Abdessamad DERRAZ (2019), CC-BY-SA 4.0

1. **Compréhension du métier**. L'enjeu pour chaque application, a été de répondre à la question posée par les médecins, et non pas de répondre à mes propres questions. Les développements méthodologiques (section 2.2) sont donc, ici, hors sujet. Deux aspects sont à prendre en compte particulièrement. Tout d'abord, la nécessité de construire un vocabulaire partagé voire une culture commune. Toute personne ayant eu des aventures pluridisciplinaires sait que cette phase est tout à la fois indispensable et fastidieuse, longue et délicate, enrichissante et frustrante, source d'erreurs et d'enseignements. Cette compréhension mutuelle est essentielle à une collaboration sereine où chaque partie comprend les responsabilités, les expertises mais aussi les incompétences de chacun [1]. Le deuxième aspect, beaucoup plus spécifique, tient à la forme des résultats proposés : on s'adresse à des praticiens-chercheurs en médecine. Ainsi, marqueurs ou modèles doivent être perçus à la fois comme des hypothèses de recherche claires et comprises, et comme des outils utilisables dans

1. ≪Comme expliquez-vous ce phénomène ?≫ : lorsque cette question me fut posée, il me fallut quelques secondes pour me rendre compte qu'on me demandait un avis médical (!) et que la seule bonne réponse était de dire que je ≪voyais≫ le phénomène dans les données, mais que j'étais bien incapable de le comprendre ou de l'expliquer. Ce genre de mises au point, dans un sens comme dans l'autre, est fréquent.

le cadre d'une consultation pour aider à poser un diagnostic que l'on puisse justifier. Il s'ensuit qu'il ne faut pas trop bousculer les habitudes ou, au moins, être capable de remettre les résultats dans une forme facile à appréhender — en bref, un bon modèle dont on se sert comme il faut est préférable à un excellent modèle dont on ne fait rien ou n'importe quoi.

2. **Compréhension des données**. Les pathologies étudiées ne sont, heureusement, pas trop fréquentes, ce qui fait qu'on ne dispose que de quelques cas chaque année, pour aboutir à des jeux de données relativement petits, quelques dizaines d'observations recueillies sur plusieurs années. Le recueil des données est manuel, long, fastidieux (lettres aux anciennes patientes pour connaître leur taille adulte) et source de nombreuses erreurs (changement des méthodes de dosages, saisies manuelles). Il y a des données manquantes (taille des parents d'un enfant adopté, ou dosage qu'on ne prescrivait pas les premières années [2]). De plus, nous étudions du vivant dans le cas de pathologies encore mal comprises ; cela signifie qu'il y a beaucoup d'inconnues et, en général, une grande variabilité. Concrètement, les données utilisées concernent de cinq à une vingtaine d'attributs, essentiellement des données cliniques (taille, âge, poids...), des données génétiques (taille des parents...) et des données d'analyses (dosages de différentes hormones).
3. **Préparation des données**. Les données sont relativement simples et limitées, mais elles nécessitent tout de même un nettoyage méticuleux, notamment du fait des saisies manuelles, ce qui nécessite de nombreuses discussions pour éviter les erreurs. Quelques attributs calculés sont ajoutés, selon l'expertise des médecins. La plupart des données physiologiques doivent être normalisées en fonction de l'âge, à partir de tables de références.
4. **Modélisation**. Afin de répondre aux enjeux métiers, les méthodes utilisées doivent être des ≪boîtes blanches≫ proposant un modèle explicite ou facilement représentable par un graphique. Cette contrainte peut paraître surprenante mais a été confirmée par l'usage. C'est une contrainte forte qui nous a amené à utiliser essentiellement des régressions linéaires ; la taille des jeux de données nous y incitant également, des méthodes plus élaborées ayant facilement tendance au sur-apprentissage. Il y a deux exceptions notables à cela, avec des contributions méthodologiques à l'analyse logique de données, qui seront expliquées dans la section 2.2, et dont les résultats résumés sur de simples graphiques peuvent d'ores et déjà être regardés sur les figures 2.3 et 2.4.
5. **Évaluation**. Les critères utilisés sont la corrélation linéaire au sens de Pearson et l'erreur absolue (modèles de régression) ou le taux d'erreur (modèles de classification), qui sont adaptés aux problèmes considérés et bien compris par les médecins. L'évaluation des capacités de prédiction des modèles est faite en validation croisée [WHFH16]. Les valeurs obtenues pourront paraître parfois atrocement faibles avec, par exemple, des corrélations inférieures à 0,5 qui seraient incongrus dans d'autres situations. Ici, il s'agissait déjà de valider la pertinence et l'utilité de certains marqueurs des pathologies, ne serait-ce que comme hypothèse de recherche. Il était de ma responsabilité de faire comprendre les usages acceptables d'un modèle ; ainsi un modèle peut être assez bon pour confirmer que telle caractéristique est utile, mais trop mauvais pour que ses prédictions soient utilisées autrement que comme une tendance imprécise.
6. **Déploiement**. Le déploiement et la mise à disposition des modèles faisait partie intégrantes des enjeux métiers. Les modèles produits sont ainsi facilement et directement exploitables, soit sous le forme d'un graphique sur lequel il suffit de placer son patient (de nouveau, voir l'exemple de la figure 2.3) ou sous la forme d'une formule simple, explicite... et mise en ligne pour encore plus de praticité [Lem18] [3]).

2.1.2 Contributions

Venons-en maintenant aux principaux résultats obtenus. La présentation n'est pas exhaustive et vise avant tout à dresser un panorama des questionnements et des réponses apportées. Toutes les études portent, rappelons-le, sur des problèmes de croissance d'enfants. Un court glossaire s'impose pour présenter les principales données utilisées :

— *Âge* est l'âge lors de la première consultation.

2. La méconnaissance des sciences de données, a amené aussi à ce que certaines caractéristiques ≪typiques des malades≫ ne soient collectées que pour ces derniers, rendant toute comparaison impossible. Un peu de pédagogie permet de faire comprendre le problème, mais ne permet pas de récupérer des données exploitables.

3. `http://www.kamick.org/lemaire/med/`. Ces modèles sont consultés quelques dizaines de fois par mois, si j'en crois les statistiques peu détaillées de mon fournisseur d'accès.

— *T (taille)* est la taille de l'enfant lors de la première consultation.
— *TC (taille cible)*, dite aussi taille génétique, est la taille adulte théorique de l'enfant, établie en fonction de la taille de ses parents [4].
— *TF (taille finale)* est la taille finale, une fois l'âge adulte atteint et la croissance terminée.
— *VC (vitesse de croissance)* est la croissance lors de l'année précédant la consultation.

D'autres variables ont été considérées (aussi variées que l'âge osseux, le retard intra-utérin, les taux de différentes hormones, *etc.*) mais ne sont pas nécessaires à l'exposé synthétique des travaux — à quelques exceptions près qui seront exposées au moment opportun.

La plupart des valeurs sont exprimées en *SDS (score en déviations standards)*, c'est-à-dire en nombre d'écarts-types par rapport à l'ensemble de la population, une population qu'il faut comprendre comme l'ensemble des enfants du même âge et sexe. Ainsi une enfant avec une taille de 0 SDS a exactement la taille moyenne des filles de son âge, tandis qu'une enfant de taille -2 SDS a une taille inférieure de 2 écarts-types par rapport aux filles de son âge (elle fait donc partie des 5% des filles les plus petites, avec une hypothèse raisonnable de distribution normale).

La première étude porte sur l'amélioration du diagnostic de déficits idiopathiques [5] en hormone de croissance. Les recommandations, issues d'un consensus international [GH 00], étaient de considérer un cas suspect selon trois critères : (1) taille inférieure à -3 SDS ($T \leq -3$) ; (2) différence entre la taille cible et taille inférieure à -1,5 SDS ($T - TC \leq -1,5$) ; (3) vitesse de croissance inférieure à -2 SDS ($VC \leq -2$). La validation de l'un de ses critères entraîne une suspicion ; l'enfant subit alors un ≪test de sommeil≫ pour confirmer ou infirmer le diagnostic. Les tests de sommeil sont des examens lourds qui nécessitent une hospitalisation. L'enjeu était donc d'améliorer le diagnostic parmi les cas suspects afin de limiter au maximum les tests de sommeil, mais sans rater aucun cas de déficit. Au passage, il s'agissait de confirmer le rôle d'une hormone, IGF-I, dans l'établissement du diagnostic : cette hormone avait été récemment identifiée comme un marqueur a priori fiable et facile à mesurer [FSM+06].

Différents modèles ont été testés, notamment à base de règles ou d'arbre de décision. Malgré leur caractère explicite et de bonnes capacités de diagnostic, ces modèles ont été jugés trop complexes pour être utilisables [6] : le modèle devait pouvoir se ≪résumer en un graphique≫. J'ai donc proposé une extension à l'analyse logique de données pour construire des motifs synthétiques (voir section 2.2 et [Lem11]) et ainsi pouvoir améliorer les performances tout en permettant une représentation graphique aisée pour des modèles à 2 (!) variables. Les résultats médicaux, mis à jour par rapport à la version publiée [LBH+09], sont visibles sur la Figure 2.3. On constate que le modèle est effectivement très facile à utiliser et, malgré le compromis nécessaire pour cette facilité, 55% des cas sont traités (86 sur 155). Seuls les patients de la zone grise ont besoin d'un test de sommeil. Le lecteur attentif aura remarqué deux erreurs (points rouges au dessus de N23) : il s'agit de patients particulièrement atypiques souffrant de multiples pathologies et malformations. À noter que d'autres ≪erreurs≫ faites par les modèles ont permis de questionner les données et de corriger deux diagnostics erronés. Au passage, l'intérêt de IGF-I est confirmé avec une très nette amélioration de son utilisation : la règle proposée par [FSM+06] ne permet de traiter que 3% des cas (5 sur 155).

La deuxième étude porte sur les conséquences d'une irradiation sur la croissance. Des enfants de 2 à 12 ans ont dû subir une irradiation, typiquement lors d'une thérapie contre un cancer, ce qui perturbe grandement leur croissance. Un des effets secondaires est alors une taille adulte anormalement petite, effet qui peut être limité grâce à un traitement par hormone de croissance. Les deux questions qui se posent au médecin, face à un nouveau patient, sont alors : quelle perte de croissante peut-on anticiper ? et est-ce qu'un traitement par hormone de croissance est pertinent ?

Le jeu de données est particulièrement petit (32 patients avec 6 variables explicatives) avec beaucoup de variance, rendant la construction d'un modèle précis et robuste impossible. Nous nous sommes donc concentrés sur la recherche de marqueurs, afin de caractériser au mieux les paramètres les plus influents sur la perte de taille, en croisant les résultats de différents modèles (régressions linéaires, régression LOESS et régression combinatoire développée pour l'occasion, voir section 2.2 et [Lem11]). Ainsi, comme l'illustre la Figure 2.4,

4. Il s'agit de la taille moyenne des deux parents corrigée de +6,5cm pour les garçons, -6,5cm pour les filles. C'est une estimation peu précise mais qui donne une cible pour anticiper un éventuel retard de croissance (voir Taille cible, Encyclopédie Larousse).

5. C'est-à-dire sans cause connue. Il existe d'autres formes de déficit mais dont la cause est connue et facile à déterminer, comme par exemple une malformation de l'hypophyse, visible sur un IRM.

6. J'avoue avoir été perplexe face à cette réaction : il était extrêmement facile d'automatiser le calcul avec un simple tableur ! A posteriori, je pense qu'il s'agissait avant tout d'un problème culturel et de méfiance face à des pratiques peu dans les habitudes. Je pense que la situation a beaucoup évolué.

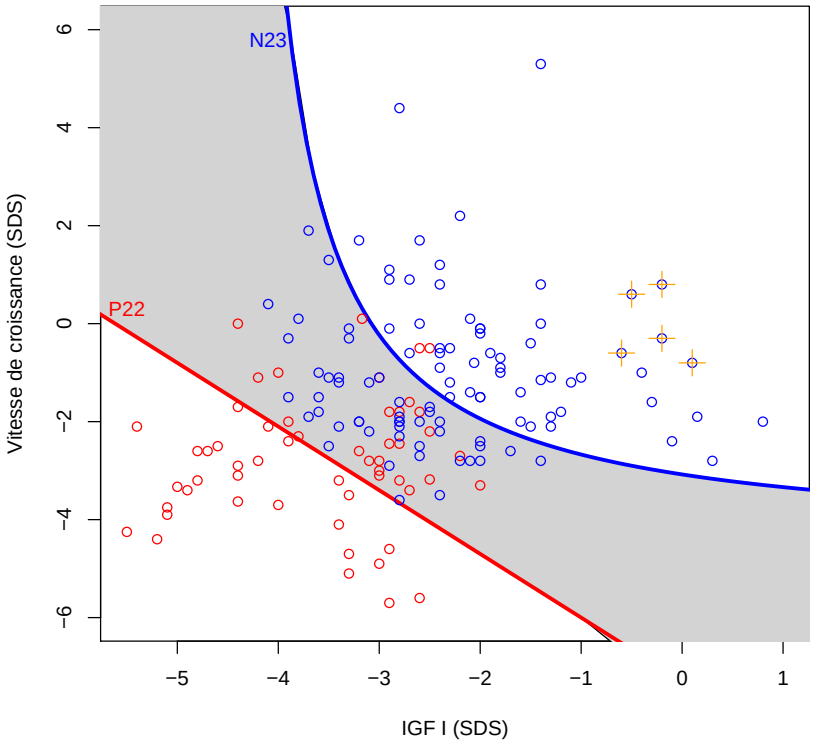


FIGURE 2.3 – Modèle de diagnostic de déficience en hormone de croissance (adaptée de [LBH+09])
Usage : en dessous de la frontière P22, le patient est ≪malade≫ ; au dessus de la frontière N23 il est ≪non malade≫ ; dans la zone grise, des tests complémentaires sont nécessaires. Les points correspondent aux patients connus : malades (rouges) ou non-malades (bleus). Les croix oranges correspondent aux cas détectés ≪non malades≫ par la règle de [FSM+06].

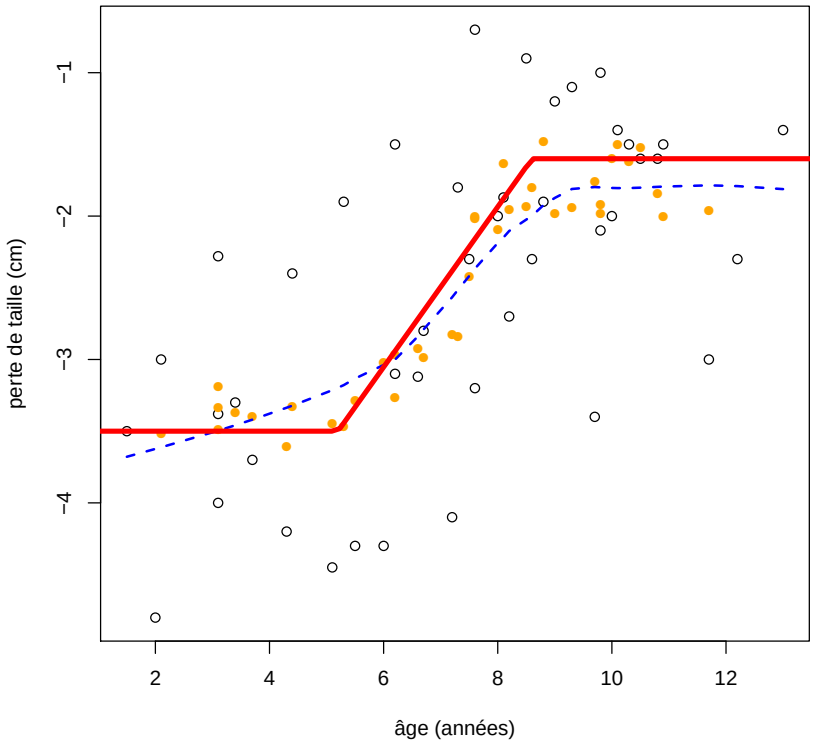


FIGURE 2.4 – Perte de taille en fonction de l'âge à l'irradiation (adaptée de [CSTE+06, Lem11])
Les points noirs correspondent aux données réelles. La courbe rouge est la tendance déduite de nos travaux. Les points oranges (moyennes mobiles sur 5 points) et la courbe bleue (approximation LOESS) viennent confirmer la cohérence de cette tendance.

l'importance de l'âge, déjà connue [FCKK94], a pu être confirmée et mieux caractérisée : nous avons ainsi déterminé que l'âge avait une influence linéaire entre 6 et 8 ans, mais n'avait plus d'influence avant ou après ces âges. Nous avons aussi pu déterminer que le traitement était modérément efficace. À noter que seules ces tendances ont été publiées dans une revue médicale [CSTE+06], afin de ne pas présenter un modèle impossible à valider avec les données disponibles et donc potentiellement trompeur.

La troisième étude se déroule en plusieurs actes et porte sur les conséquences pour des jeunes filles d'une puberté dite précoce, c'est-à-dire commencée avant 8 ans. Une puberté prématurée implique un arrêt rapide de la croissance et une taille finale (adulte) potentiellement très petite. La principale question est alors de prédire la taille finale des jeunes filles. Bayley et Pinneau [BP52] ont proposé une méthode de prédiction de la taille adulte d'un enfant en fonction de la taille de l'enfant et de son âge osseux. Cette méthode, classique et rodée, est réputée la plus fiable pour estimer la taille hors pathologie particulière [Roc93] mais a aussi été utilisée dans le cas de pubertés précoces [BLS+95]. De Ridder et al. [dRSHK07] revendiquent le premier modèle de prédiction de la taille finale pour le cas de pubertés précoces avant un éventuellement traitement ; il s'agit d'une régression linéaire multiple impliquant notamment, après différentes transformations, la taille, la taille cible, le genre, l'âge osseux, la réponse à l'hormone de croissance.

Nous avons établis un modèle de la taille finale [ALCS+11], que nous avons amélioré et affiné dans [GLB15]. Il s'agit d'une régression linéaire multiple relativement simple, puisqu'elle implique, sans transformation, uniquement l'âge, la taille, la taille cible et le dosage de deux hormones. Elle donne toutefois des résultats supérieurs à ceux de [dRSHK07] (corrélation de 0,75 contre 0,53)[7]. Le principal défaut de ce modèle (comme des autres que j'ai pu croiser en endocrinologie pédiatrique) est qu'il n'avait été construit et validé que sur un unique jeu de données : les patientes d'une seule médecin ! Pour tempérer son enthousiasme, j'ai réussi à la convaincre de discuter avec d'autres collègues pour tester notre modèle sur leurs données. Les résultats, publiés dans [LDdBM+18], ont été plus que rassurants, comme le montre la Figure 2.5. En effet, la qualité des

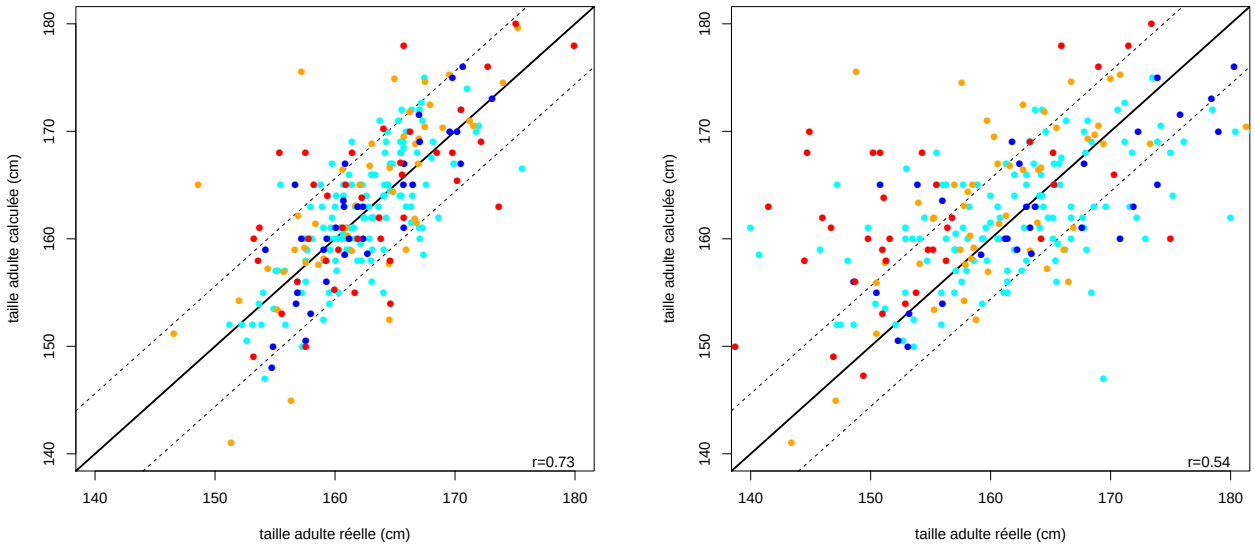


FIGURE 2.5 – Taille finale calculée par la méthode de [GLB15] (gauche) et par la méthode Bayley-Pinneau (droite)

Les couleurs correspondent au jeu de données : français/original (cyan), utilisé pour établir le modèle ; allemand (rouge), néerlandais (orange), français (bleu), utilisés pour valider le modèle. Les lignes en pointillés correspondent à un écart d'un écart-type (5,6cm).

prédictions a été confirmée par l'ensemble des nouvelles données – et est bien supérieure à ce qu'on obtient par la méthode classique de Bayley-Pinneau. Cette validation a permis que le modèle, proposé dans [GLB15] comme hypothèse de recherche, devienne dans [LDdBM+18] un ≪petit morceau de vérité≫.

7. Sur la base des résultats annoncés dans l'article ; il est impossible de faire une vérification plus précise car (1) leurs données sont indisponibles et (2) leurs modèles utilisent des variables que nous n'avons pas. L'inverse est aussi vrai : nos données ne sont pas plus disponibles et nos modèles utilisent des variables qu'ils n'ont pas.

Pour être complet sur le sujet, notons que nous avons aussi proposé un modèle de taille finale dans le cas de pubertés avancées (débutées entre 8 et 10 ans) [LPBB14] ainsi que différents modèles de l'âge des premières règles [ALCS+11, GLB15, TLBTB18], qui sont les premiers et les seuls modèles du genre à ma connaissance. Tous ces modèles sont disponibles en ligne [Lem18] [8], conformément aux besoins de déploiement des modèles.

2.2 Analyse logique de données

En appui des applications médicales décrites ci-dessus, j'ai contribué à la méthode de l'analyse logique de données, en développant un outil logiciel, et deux extensions méthodologiques. Tout d'abord, commençons par expliquer ce qu'est cette méthode.

2.2.1 Principes de l'analyse logique de données

L'Analyse logique de Données (*Logical Analysis of Data*, LAD) est une méthode d'apprentissage automatique supervisée, basée sur l'étude des fonctions pseudo-booléennes et avec des liens et des similitudes avec d'autres méthodes classiques comme les règles d'association ou les arbres de décisions. Les prémices comme les premiers développements méthodologiques sont dus à Peter L. Hammer et ses collaborateurs [CHI88, BHIK97, BHI+00], avec des applications dans le domaine médical [HB06].

Le LAD est conçu, à l'origine, pour les problèmes de classification binaire sur des données binaires — ce dernier aspect ayant été rapidement dépassé pour traiter des problèmes de classification binaire sur des données numériques quelconques [BHIK97]. Dans le cadre du LAD on a l'habitude de parler des observations *positives* et des observations *négatives* pour distinguer les deux classes.

La brique de base du LAD est le *motif* (*pattern*). Un motif est une conjonction de conditions sur variables (par exemple : $P_1 = (x_1 \leq 10) \wedge (x_5 \geq 2)$) suffisamment caractéristique, comme cela va être précisé. Un motif *couvre* une observation si celle-ci satisfait les conditions de celui-là. La *prévalence positive* (resp. négative) d'un motif est la proportion des observations positives (resp. négatives) qu'il couvre. Un motif est *positif* si sa prévalence positive est plus grande que sa prévalence négative ; son *homogénité* est la proportion des observations positives parmi celles couvertes. Des définitions symétriques valent pour les motifs *négatifs* et leur homogénité. Un candidat motif sera un motif si ses prévalences et homogénéités sont assez grandes (au dessus de seuils pouvant dépendre de l'application).

L'apprentissage automatique se déroule en deux phases : (1) générer (et sélectionner) des motifs positifs et négatifs caractéristiques du jeu de données ; puis (2) agréger les motifs en un *discriminant*, c'est-à-dire une somme pondérée des motifs, dont le signe donne la prédiction de classe d'une (nouvelle) observation. Un exemple est donné Figure 2.6.

Il existe différentes méthodes de génération, de sélection et d'agrégation des motifs [AH06b, AH06a, BHK08]. Dans tous les cas, un des gros intérêts du LAD est de proposer des modèles explicites avec des motifs qui sont autant de règles permettant d'expliquer les prédictions. Le LAD reste toutefois très ancré dans le cadre booléen, ce qui le rend peu adapté aux problèmes de régression ; de plus, les motifs sont des «rectangles» qu'il faut parfois multiplier de manière peu naturelle et excessive pour couvrir certains jeux de données.

2.2.2 Contributions

J'ai découvert le LAD lors de mon post-doctorat à Rutcor, avec Peter Hammer. Ma première contribution a alors été le développement de `ladoscope` [Lem05], un logiciel libre proposant plusieurs implémentations pour les différentes étapes du LAD. Ce logiciel a servi pour mes propres besoins, bien entendu, mais a également été utilisé par d'autres chercheurs pour des travaux applicatifs (notamment [RWY+08, BRS+10, BRT+10, SSH+17]) ou méthodologiques (par exemple [SSAH09, AH12]). Mes propres contributions méthodologiques ont été faites à l'occasion d'applications médicales ([LBH+09, CSTE+06], décrites ci-dessus) et ont été publiées dans [Lem11].

La première contribution vise à rendre les motifs plus naturels et moins nombreux en éliminant les «escaliers», comme celui formé par les motifs bleus de la Figure 2.6, pour aboutir à un motif synthétique, comme le motif orange de figure 2.7. La procédure proposée est la suivante :

1. choisir 2 variables x et y et construire les motifs maximum (au sens de l'inclusion) sur x et y ;

8. `http://www.kamick.org/lemaire/med/`

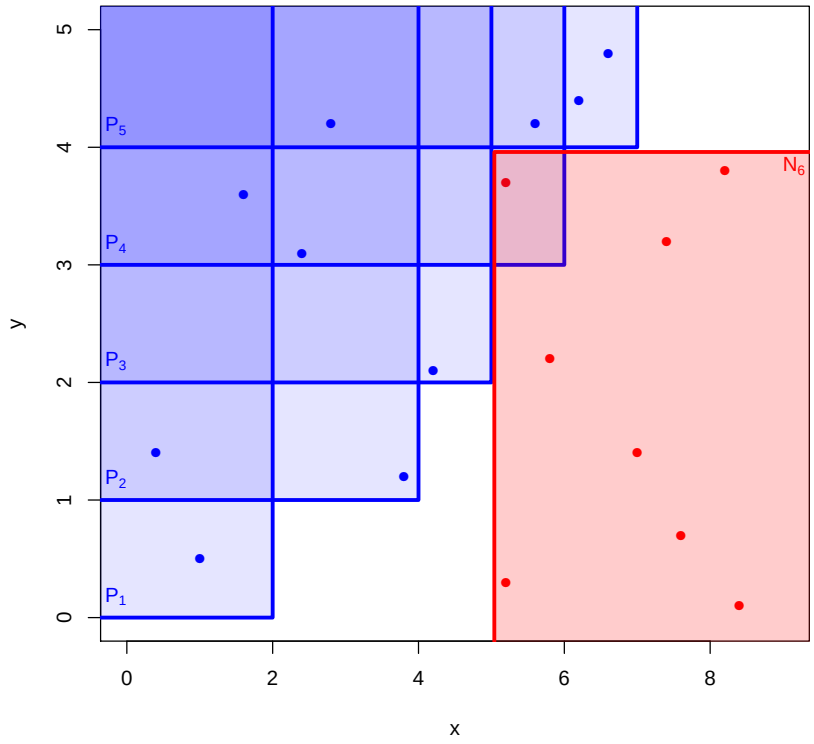


FIGURE 2.6 – Exemple de modèle avec le LAD
On a cinq motifs positifs (bleus) et un motif négatif (rouge) :

$$\begin{aligned} P_1 &\quad x \leq 2 \wedge y \geq 0 \\ P_2 &\quad x \leq 4 \wedge y \geq 1 \\ P_3 &\quad x \leq 5 \wedge y \geq 2 \\ P_4 &\quad x \leq 6 \wedge y \geq 3 \\ P_5 &\quad x \leq 7 \wedge y \geq 4 \\ N_6 &\quad x \geq 5 \wedge y \leq 3 \end{aligned}$$

Le discriminant

$$\Delta(z) = \sum_{i=1}^{5} \frac{1}{5} P_i(z) - N_3(z)$$

(où z est une observation) permet de classer correctement chaque point.

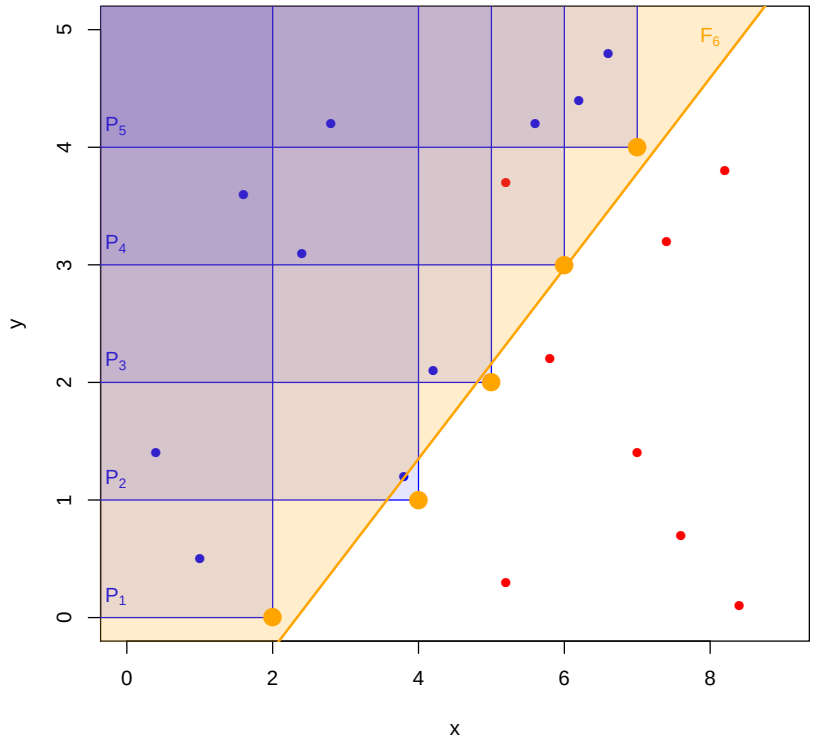


FIGURE 2.7 – Exemple de motif synthétique
Les cinq motifs bleus sont remplacés par un motif synthétique :

$$F_6 \quad y \geq -1,89 + 0,81x$$

La frontière de F_6 est la droite de régression aux moindres carrés sur les points extrêmes (points oranges) des motifs.

2. si tous les motifs ont la même forme, typiquement $x \leq \alpha_k \wedge y \leq \beta_k$, alors construire la régression $y = f(x)$ sur les points $(\alpha_k; \beta_k)$;
3. remplacer les motifs sur x et y par le motif synthétique, typiquement $y \leq f(x)$.

L'étape 2 peut échouer si les motifs ont des formes variées ; en pratique, avoir des motifs de même forme revient à avoir une monotonie par rapport aux variables considérées, ce qui est fréquent. Un autre cas d'échec possible est s'il y a trop peu de motifs sur x et y ; dans ce cas là, construire un motif synthétique est tout simplement inutile. Quant à la fonction f, on peut se permettre n'importe quelle forme paramétrique dont on sait estimer les paramètres ; le meilleur ajustement sera alors conservé.

En pratique, ces motifs synthétiques ont été utilisés pour diagnostiquer les déficits en hormone de croissance [LBH+09], ce qui a permit d'obtenir la Figure 2.3 vue précédemment. Sur cette figure, P22 ($VC \leq -5,2 - IGF$) est un motif synthétique qui remplace 4 motifs et N23 ($VC \geq -4,5 + \frac{6,4}{IGF+4,5}$) est un motif synthétique qui en remplace 5. Le modèle ainsi obtenu n'a pas perdu en qualité, mais est plus facile et naturel à comprendre. Testés sur d'autres jeux de données classiques, les motifs synthétiques ont permis d'obtenir des modèles 4 fois plus petits (en nombre de motifs), sans perte de qualité de prédiction.

La deuxième contribution méthodologique permet une extension du LAD aux problèmes de régression. De tels problèmes ne sont pas naturels pour le LAD, aussi ont-ils été très peu étudiés. On ne trouve en fait que deux occurrences. Dans le cadre d'une application en finance [HKL07], Hammer, Kogan et Lejeune ont construit un modèle binaire puis ont détourné simplement le discriminant en profitant du fait que la réponse à prédire se limite à quelques valeurs entières. D'un point de vue méthodologique, Bonates et Hammer [BH07] ont proposée la «régression pseudo-booléenne» qui consiste à minimiser l'erreur absolue moyenne pour une fonction exprimée dans l'espace des motifs, ces derniers étant générés par une procédure de génération de colonnes ; la technique est assez lourde dans ses calculs et dans la forme de ses résultats (motifs très nombreux) et semble propice au sur-apprentissage.

Comme alternative pragmatique, j'ai proposé la «régression combinatoire» qui revient à

1. choisir k seuils sur la variable à prédire ;
2. construire un modèle binaire pour chaque seuil ;
3. regrouper les motifs ainsi générés en un discriminant avec des poids adaptés.

Chaque étape est résolue par des heuristiques, pour éviter le sur-apprentissage, mais aussi par simplicité — il y a là, clairement, matière à améliorations. Utilisé dans le cadre de [CSTE+06], la procédure a donné des résultats très satisfaisants avec des corrélations et des erreurs comparables à d'autres méthodes (régression linéaire, perceptron, SVM).

Avec une telle procédure, le nombre de motifs reste élevé, ce qui fait que l'interprétabilité du modèle reste limitée. Pour palier cela, j'ai proposé d'utiliser une approximation linéaire : le modèle peut être vu comme une fonction pseudo-booléenne, dont on sait calculer rapidement la meilleure approximation linéaire au sens de la norme L_2 [HH92]. On obtient ainsi une nouvelle fonction qui est une somme pondérée de monomes, c'est-à-dire de conditions ne portant que sur une unique variable. Dans le cadre de l'application [CSTE+06], on obtient ainsi la formule de perte de taille suivante ΔT :

$$\begin{aligned}\Delta T \quad = \quad & -0.39(age < 2.1) - 0.59(age < 5.2) - 0.13(age < 6.4) - 0.16(age < 7.0) \\ & -0.15(age < 7.6) - 0.14(age < 8.6) - 0.11(age < 9.5) + 0.15(age < 12.6) \\ & -0.45(traitement = non) \\ & +0.16(poids < 7.5) + 0.21(poids < 9.8) + 0.15(poids < 10.8) \\ & -0.34(poids < 11.3) - 0.36(poids < 15.3) - 0.15(poids < 19.5)\end{aligned}$$

où on arrive facilement à reconstruire l'influence des différentes variables ; on voit ainsi facilement que le traitement est bénéfique mais de manière modéré, et que la perte de taille est une fonction croissante de l'âge (on peut même retrouver la tendance dessinée sur la Figure 2.4).

2.3 Discussion

Maintenant que j'ai exposé le cadre et mes contributions pour l'aide au diagnostic médical il est nécessaire de prendre un peu de recul et de mettre en questions de tels travaux.

Ces travaux sont-ils éthiques ? Pour l'ensemble des travaux présentés dans ce chapitre, nous sommes loin de la rassurante neutralité d'un théorème. Au contraire, il s'agit de la santé de personnes vulnérables et influençables. Ceci est particulièrement vrai pour le cas des pubertés précoces, comme le discute très bien Hayes [Hay16] : en effet, il s'agit de donner (ou pas) un traitement loin d'être anodin à des jeunes filles en parfaite santé. Une puberté précoce a des conséquences sur la taille adulte et elle peut être mal vécue, mais ce n'est pas une pathologie. Comme [Hay16] le rappelle justement, l'âge de la puberté et la taille adulte devraient être considérées en terme de diversité et non de normalité, et un accompagnement de la personne (et de son entourage) serait sans doute plus bénéfique qu'un traitement. Ce cas des pubertés précoces illustre très bien les enjeux de médication de la normalité, où les déviations à la moyenne sont considérées comme anormales et devant être corrigées. Ainsi un traitement est commencée parce que l'âge de la jeune fille est en dessous d'un seuil «normal», mais il est en général poursuivi au delà de ce seuil et jusqu'à l'âge moyen. On voit là une dérive des usages de modèles ou de statistiques qui, de descriptives, devient prescriptives.

Un exemple typique se trouve dans les conclusions de [LRL+18], qui affirme que le traitement a permis de préserver tout le potentiel génétique de taille, tout en concédant n'avoir aucune fille non traitée pour pouvoir comparer. Cela prend l'allure d'une profession de foi plus que d'un résultat scientifique, puisque l'hypothèse inverse — que le traitement n'a eu aucun effet et que les jeunes filles ont simplement grandi à leur façon, plutôt que selon celle des courbes de carnets de santé établies 20 ou 30 ans auparavant —, est tout aussi tenable avec les données proposées. Ce qui nous amène à la question suivante.

Ces travaux sont-ils pertinents ? La forme des résultats, en particulier, questionne et est à mettre en regard de l'usage qui en est fait. Était-il justifié de se contenter de modèles «bons», alors que nous en avions établis d'autres proposant de meilleures prédictions, juste pour que ces modèles puissent prendre la forme d'un joli dessin ou d'une bête page web ? Ce que j'ai pu voir de l'usage des résultats, notamment à travers les études citant les travaux auxquels j'ai participé, m'a convaincu que oui.

Ainsi, reprenons la surprenante étude [LRL+18] qui s'inscrit explicitement dans la suite de [GLB15] et se propose de confirmer ou infirmer nos modèles sur une cohorte de 48 jeunes filles brésiliennes. Après avoir constaté que le modèle de [GLB15] est effectivement meilleur que la formule de Bayley-Pinneau au moment de la consultation initiale, les auteurs mettent explicitement en doute son utilité car il est moins bon que cette même formule utilisée après un traitement de 3 ans environ. Les auteurs n'envisagent jamais que prédire la taille adulte d'enfants est, par nature, plus facile quand celles-ci ont 11 ans plutôt que 8 ; ils n'envisagent pas, non plus, l'opportunité de tester notre modèle sur ces nouvelles données, comme il le font avec la formule de Bayley-Pinneau. Au passage, comme cela a été mentionné, ils concluent à l'efficacité du traitement, sans argument, et en contradiction avec nos propres conclusions.

Pour moi, cet exemple doit être analysé en termes d'incompréhension, pas d'incompétence. Je mettrai même au crédit de ces auteurs d'avoir osé et essayé de s'approprier le modèle mathématique. Dans les autres études citant nos travaux et que j'ai lues, il semble que le fait que nous proposions une *formule* soit passé inaperçu pour ne retenir que des marqueurs. Il faut dire que, dans cette spécialité (pour ce que j'en ai vu), les modèles sont extrêmement rares, et la plupart des travaux visent à établir des corrélations — ce qui n'est déjà pas simple à faire et explique, sans doute, de ne pas s'embarrasser avec des mises en relation plus riches mais aussi plus complexes comme peut l'être une «simple» régression linéaire.

Il apparaît que la forme des résultats est essentiel. Cependant, contrairement à d'autres domaines, le risque n'est pas que l'outil ou le résultat ne soient pas utilisés mais au contraire qu'ils soient *mal* utilisés. Les erreurs et les mauvais usages sont de la responsabilité des médecins mais encore faut-il leur avoir fourni des outils adaptés avec des modes d'emploi clair. Il serait hypocrite de se cacher derrière quelques équations et sa propre compréhension de celles-ci. L'enjeu n'est donc pas de publier un résultat dont on a exprimé les limites et les hypothèses, c'est de s'assurer que ces limites et hypothèses seront bien comprises — quitte à se censurer. Ainsi, autant je suis content des développements méthodologiques, mêmes incomplets, qui ont permis de faire la figure 2.4, autant je suis rassuré qu'elle n'ait pas été publiée dans une revue médicale. Ce qui nous amène à la question suivante.

Ces travaux sont-ils valables ? Pour ce qui concerne le diagnostic médical, aucun des résultats présentés n'est avéré. Chacun n'est qu'une hypothèse, établie sur la base d'un jeu de données limité, et qui nécessite confirmation. Cette aspect est curieusement mal compris. Plus généralement on peut regretter une certaine inculture en sciences des données, dommageable dans un domaine où les travaux s'appuient avant tout sur des données. Je passerai sur les nombreuses maladresses qui diminuent fortement la qualité et donc le potentiel de ces données pour en venir à la question de l'évaluation.

Certains aspects sont inhérents aux applications traitées. Les jeux de données sont de «mauvaise qualité» : petits, bruités, avec des données manquantes ; il faut faire avec. Un partage des données entre chercheur serait une solution, mais on s'aperçoit qu'on ouvre là un sujet très sensible pour les patients (il s'agit de données personnelles sensibles) mais aussi pour les médecins ; la collecte est longue, fastidieuse et les données sont un trésor dont le partage en diminue la valeur par la perte d'une sorte d'avantage concurrentiel. Au delà de ses aspects légaux, économiques voire personnels, le partage technique n'est lui-même pas évident. Dans de nombreux cas, il est difficile de tester les résultats d'autres travaux parce que les données originales ne sont pas disponibles, donc, mais aussi parce que les données dont nous pourrions disposer ne sont pas compatibles en raison de données non récoltées (pour un marqueur nouvellement identifié) ou de nombreux biais plus ou moins explicites, que ce soit dans les valeurs de référence (les valeurs «normales» ne sont pas les mêmes d'un pays à l'autre) ou dans les mesures (selon l'opérateur ou la technique de dosage utilisée). Curieusement, le besoin de reproductibilité des résultats ne semble pas particulièrement ancré. J'ai ainsi dû discuter longuement — et je suis fier d'avoir réussi à convaincre — de l'utilité de la validation internationale. Ce qui pourrait n'être qu'un problème de personne est plus général, comme en témoigne la réaction de l'éditeur de la première revue à laquelle nous avons proposés ces résultats, qui les a refusés sous le prétexte que, étant une validation, «ils n'apportaient rien de nouveau».

Pour terminer sur ce sujet, il me semble que le manque de maîtrise des fondamentaux des sciences de données [9] est non seulement préjudiciable à la qualité des résultats, mais aussi à l'efficacité de la recherche. Certains travaux, que je n'ai même pas pris la peine de détailler ci-dessus [LCST+11, PLH+13, CLBTB20], auraient été plus vite et mieux menés si, au lieu travailler avec moi, les médecins avaient eues quelques compétences de base et le support léger d'un ingénieur d'étude, comme ils ont déjà le support de laborantins. Ce qui nous amène à la question suivante.

Quel bilan et quelles perspectives ? Ces travaux sur le diagnostic médical s'inscrivent dans une collaboration de 15 ans. Les débuts ont été laborieux car il n'est pas évident de construire un vocabulaire, une compréhension et des problématiques communes lorsqu'on vient, pour moi, d'une culture des mathématiques appliquées et, pour les autres, de la médecine clinique. Comme toute personne ayant travaillé dans un tel contexte, je peux témoigner que l'effort conséquent se retrouve largement dans la curiosité, l'ouverture d'esprit, et les questionnements qu'il provoque. Cette discussion en dessine un dénouement — il reste toutefois une ou deux applications à traiter, qu'il serait sans doute dommage de laisser dans un tiroir maintenant que nous nous comprenons plutôt bien. L'expérience aura été enrichissante et formatrice — pour cela, travailler sur des données «sales» empêche beaucoup de facilités et invite à une prudence salutaire — et m'aura permis de développer des compétences en sciences des données que j'ai pu transposer à l'analyse et au suivi des systèmes de productions, que nous aborderons au chapitre suivant. D'un point de vue méthodologique, je serais plus proactif quant aux perspectives.

L'étude des petits jeux de données est un problème en soi qui, bien qu'identifié depuis longtemps, est encore relativement peu étudié. Les méthodes classiques de *bootstrap* ont un effet limité. Une piste serait de réussir à générer des données cohérentes pour enrichir le jeu de données, comme proposé récemment par [LLC+18].

Plus proche de mes préoccupations, de nombreux développements seraient à faire sur le LAD. En premier lieu, mon bon vieux `ladoscope` ne tient pas compte des développements les plus récents, et une réimplémentation et une intégration à des bibliothèques bien connues comme `weka` [FHW16] ou `scikit-learn` [PVG+11], seraient bienvenues. Ce serait l'occasion de divers développements, aussi bien sur la méthodologie de calcul de modèles que sur ses usages.

Tout d'abord, si le LAD a eu de beaux succès sur des jeux de données de taille raisonnable, il est encore peu adapté au *big data* ; la génération des motifs, en particulier, est très gourmande en temps de calcul. Dans cette perspective, il y a clairement des idées à adapter d'algorithmes classiques comme ceux dédiés aux arbres de décision, ou d'algorithmes comme MOCA-I [JTD+15], une méta-heuristique de génération de règles utilisée avec succès pour du diagnostic médical [JMHLF+20] pour des jeux de données très déséquilibrés et de grande taille.

Ensuite, les contributions que j'ai proposées dans [Lem11] sont loin d'avoir été pleinement exploitées. Premièrement, afin d'améliorer l'expressivité des modèles tout en gardant leur compréhensibilité, les motifs synthétiques sont une piste à creuser. Un catalogue de fonctions plus étendu et des approches moins heuristiques amélioreraient les résultats, et il y aussi besoin de valider la robustesse de l'approche, en particulier

9. Que j'ai pu observer dans le domaine très spécialisé dans lequel s'inscrivent mes travaux ; dans d'autres spécialités médicales on trouve, au contraire, d'excellents statisticiens.

lorsque les motifs similaires sont peu nombreux. En second lieu, la régression combinatoire reste trop heuristique et gagnerait à intégrer davantage les idées de la régression pseudo-booléenne [BH07] ou de la très récente Régression-LAD [KYB21], en prenant garde d'arriver à équilibrer la qualité de prédiction avec la complexité du modèle, afin de ne pas perdre l'un des gros intérêts du LAD.

Beaucoup de propriétés des fonctions pseudo-booléennes, à l'origine du LAD, n'ont pas été exploitées, en particulier les reformulations. Ainsi, l'approximation linéaire utilisée dans le cadre de la régression combinatoire peut également être utilisée pour approcher le discriminant d'une classification et fournir, ainsi, une expression simplifiée du modèle. De manière plus spéculative, les techniques de quadratisation des fonctions pseudo-booléennes [BCRH20] permettraient de limiter les modèles à des règles simples de degré 2 (*i.e.* 2 variables), au prix de variables additionnelles qu'il faudrait interpréter.

Enfin, le caractère explicatif et explicite du LAD est loin d'être complètement exploité, notamment pour apporter des garanties ou des mesures de confiance sur les prédictions.

Premièrement, le LAD n'a pas besoin d'avoir une observation complète pour faire une prédiction : on peut donc chercher l'≪explication≫ minimale d'une prédiction en calculant le sous-ensemble minimal (en termes de variables utilisées) tel que la prédiction d'une observation ne change pas [10]. On identifierait ainsi les variables nécessaires et suffisantes pour caractériser cette observation. Deuxième, on peut chercher la ≪contradiction≫ la plus proche d'une observation, c'est-à-dire en construisant l'observation classée différemment la plus proche possible. Ces idées sont à rapprocher des exemples et autres explications contradictoires (*counterfactual explanations*, *adversarial examples*, voir par exemple [Mol21]), avec des procédures dédiées au LAD qui pourraient s'inspirer de ce qui existe pour les arbres et forêts [PV21].

Une autre exploitation des bonnes propriétés d'un modèle LAD serait de chercher l'ordre optimal dans lequel considérer les variables pour aboutir à un diagnostic de moindre coût. Par exemple, sur la Figure 2.6, si je sais déjà que $x = 2$, il est inutile de mesurer y pour savoir que l'observation n'est pas rouge (dans ce cas, il subsiste tout de même un doute qu'elle soit bleue). Ce problème peut être vu comme un problème d'ordonnancement.

Ces derniers problèmes, très combinatoires, n'ont pas été étudiés à ma connaissance mais ont un intérêt aussi bien théorique que pratique.

10. Ainsi défini, le problème est ambigu quant aux variables exclues : la prédiction doit-elle être valable si ces variables n'ont pas de valeur, ou bien quelles que soient ces valeurs ?

CHAPITRE 3

Analyse et optimisation de systèmes complexes

Les travaux présentés dans les deux chapitres précédents concernaient des systèmes «simples» dans le sens où le cadre était relativement bien défini, les enjeux univoques et les inconnues limitées. Dans le cadre de l'ordonnancement (chapitre 1), les travaux s'inscrivent complètement dans une recherche opérationnelle classique, où les problèmes sont formellement et complètement définis et les résultats se mesurent à l'aune de critères mathématiques bien identifiés. La seule exception serait la production de puces électroniques (section 1.2.2), qui est très proche de ce qui va être discuté dans ce chapitre (en particulier section 3.1), mais que les contraintes de mise en œuvre ont, au final, beaucoup épuré. Dans le cadre de l'aide au diagnostic (chapitre 2), les aspects méthodologiques concernent eux aussi des problèmes complètement définis et des critères bien identifiés, tandis que les applications se focalisent sur un aspect précis et restreint, essentiellement la construction d'un modèle prédictif, qui fait que le problème se résume au jeu de données dont on dispose.

Les travaux présentés dans le présent chapitre concernent, au contraire, des systèmes «complexes». Wikipedia propose la définition appropriée suivante : «un *système complexe* est un ensemble constitué d'un grand nombre d'entités en interaction dont l'intégration permet d'achever une mission commune. Les systèmes complexes sont caractérisés par des propriétés émergentes qui n'existent qu'au niveau du système et ne peuvent pas être observées au niveau de ces constituants.» [1] Pour préciser et souligner les différences par rapport aux systèmes simples précédents, il s'agit de systèmes aux contours imprécis, avec des incertitudes mais aussi des inconnues et l'influence d'éléments extérieurs, et pour lesquels il faut prendre de multiples décisions évaluées selon des modalités multi-critères encore à construire et dont les conséquences ne sont ni finement ni explicitement connues.

En terme de systèmes complexes, je me préoccupe de systèmes de production avec des enjeux d'organisation, de planification ou de pilotage, et des objectifs d'optimisation. D'une certaine manière, il s'agit de problèmes que l'on aimerait bien capturer dans un joli modèle de programmation mathématique, mais la méconnaissance qu'on en a l'en empêche. L'apport de sciences de données permet alors de compenser et d'enrichir l'analyse pour des décisions mieux caractérisées.

Dans la suite de ce chapitre, je présente deux cas d'étude : le premier présente des travaux autour de la variabilité des flux de production en micro-électronique, le second présente des travaux autour de l'optimisation des réseaux énergétiques.

3.1 Variabilité des flux de production en micro-électronique

La modélisation de la variabilité des flux de production en fabrication micro-électronique a été le sujet de la thèse CIFRE de Kean Dequéant [Deq17], que j'ai co-encadrée avec Marie-Laure Espinouse (G-SCOP) ainsi que Philippe Vialletelle (STMicroelectronics).

Mis à part quelques géants de la micro-électronique qui produisent des volumes tels qu'ils peuvent avoir des usines dédiées à quelques produits comme des mémoires ou des processeurs, les fabricants comme STMicroelectronics doivent proposer un catalogue conséquent de produits, ce qui rend particulièrement complexe le système de production. En particulier :

— il y a une grande variété de produits en petits volumes (*High Mix, Low Volume*) et la fabrication d'un produit nécessite plusieurs centaines d'étapes et prend plusieurs semaines ;
— les machines, extrêmement coûteuses, ne peuvent être dédiées ni à certains produits ni à certaines étapes et il faut donc gérer des qualifications (certaines machines sont incompatibles avec certains produits) et de la ré-entrance (un même produit peut passer plusieurs fois sur la même machine).

1. Wikipedia, *Système complexe* (2021-07-15).

De nombreuses étapes de fabrications sont assimilables à des problèmes d'ordonnancement et sont traitées comme telles avec des outils de programmation linéaire ou programmation par contraintes, par exemple. De fait, la plupart des étapes sont bien maîtrisées et l'on peut dire, à titre d'illustration, qu'à chaque étape, tout ce passe comme prévu dans 99% des cas. Cependant, comme chaque lot subit plusieurs centaines d'étapes, cela veut dire que 99% des lots ne sont pas produits comme prévu, ce qui implique des retards, des accumulations, des goulots, *etc.*, autant d'effets qui viennent dégrader l'efficacité globale du système. Dans le cadre de nos travaux, nous nous sommes justement intéressés à cet aspect systémique en s'efforçant de qualifier d'abord puis quantifier la variabilité des flux.

La qualification de la variabilité a été faite par un état de l'art [DLEV16b] qui a permis de mettre en avant les causes principales et secondaires de la variabilité, ainsi que ses conséquences sur la production en général et sur le temps de cycle[2] en particulier, ce dernier étant un indicateur particulièrement sensible et regardé dans l'industrie micro-électronique. Deux aspects, que je développe par la suite, ont été mis en avant : l'importance du mix de produits pour une utilisation efficace des équipements, et l'incidence de la dépendance entre événements.

Une source importante d'écart entre les temps de cycle réels et les temps planifiée et la production effectuée est la difficulté de connaître la capacité réelle de production d'un équipement ou d'un groupe d'équipements. Le problème est en général abordé comme une file d'attente de type «G/G/m» (lois générales pour les arrivées et les durées d'exécution, plusieurs machines) et dans le cadre de laquelle Hopp et Spearman [HS08] ont proposée une approximation du temps de cycle très utilisée dans l'industrie :

$$CT = \left(\frac{C_a^2 + C_e^2}{2}\right)\left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)}\right)t_e + t_e$$

où $t_e = PT/A$ avec PT le temps de process moyen, A la disponibilité moyenne des équipements, C_a le coefficient de variation des temps inter-arrivées, C_e le coefficient de variation des temps de process effectifs, m le nombre d'équipements en parallèle et u le taux d'utilisation de ces équipements. Cette formule a été critiquée pour la difficulté d'évaluer ses paramètres [EVL+11] et pour ses hypothèses simplificatrices [SDZ07, ATDS09] ; des corrections ont été proposées mais celles-ci tendent à ne corriger qu'un problème à la fois et sont spécifiques [Kin09].

Dans notre cas, un biais supplémentaire et essentiel provient de la diversité des produits et des équipements, ce qu'une telle formule ignore complètement. La capacité du système est alors fortement sur-estimée car des équipements peuvent se retrouver en attente ou chargés seulement partiellement faute de produits compatibles. Ce genre de soucis se répercutent et s'amplifient sur les étapes suivantes, ce qui rend rapidement impossible d'anticiper avec assez de finesse l'arrivée de produits ; on doit faire en «temps réel» avec les produits disponibles. À défaut de pouvoir mettre en œuvre des approches de recherche opérationnelle qui permettraient de planifier des chargements optimisés, nous nous sommes tournés vers des sciences de données pour déduire le comportement futur du système de son comportement passé.

Pour cela nous avons défini le *WIP*[3] *concurrent* d'un lot i sur un groupe d'équipements G comme étant la somme des temps de process des lots traités sur l'un des équipements de G, et dont la prise en charge par un équipement s'est effectuée entre l'arrivée du lot i sur le groupe G et la prise en charge du lot i sur l'un des équipements de G [DLEV16a]. Dans cette définition, utiliser les «temps de process» plutôt que le nombre de lots permet d'homogénéiser la diversité des produits. Le WIP concurrent est une mesure qui rend bien compte du temps de cycle (Figure 3.1). C'est une mesure a posteriori, mais elle permet de déduire des informations essentielles quant au comportement du système. En particulier il permet de déterminer la capacité «réelle» d'un groupe d'équipements selon le niveau de WIP en attente.

Ainsi, la Figure 3.2(a) représente la capacité d'un système simulé, telle que vue par chaque produit lors de son attente, ce qui permet par des régressions de type LOESS, de déduire une capacité moyenne (ou limite) selon le niveau de WIP en attente. La capacité ainsi calculée a été utilisée pour déduire le temps d'absorption d'un WIP pour le même système simulé. La Figure 3.2(b) représente les trajectoires d'absorption moyenne (bleue) et limites (rouge, verte) prévues, tandis que les multiples trajectoires grises correspondent à autant de simulations du système ; on constate la qualité des prédictions, les trajectoires couvrant l'intervalle prévu ([DLEVa], en préparation).

2. On appelle *temps de cycle* d'un produit le temps entre son arrivée à un premier jalon et son départ d'un deuxième jalon ; ce temps inclut donc les temps de process, mais aussi d'attentes voire de transport si on regarde le temps de cycle sur plusieurs étapes.
3. WIP (*Work In Progress*) : nom habituellement donné à l'encours, c'est-à-dire aux produits en cours de fabrication.

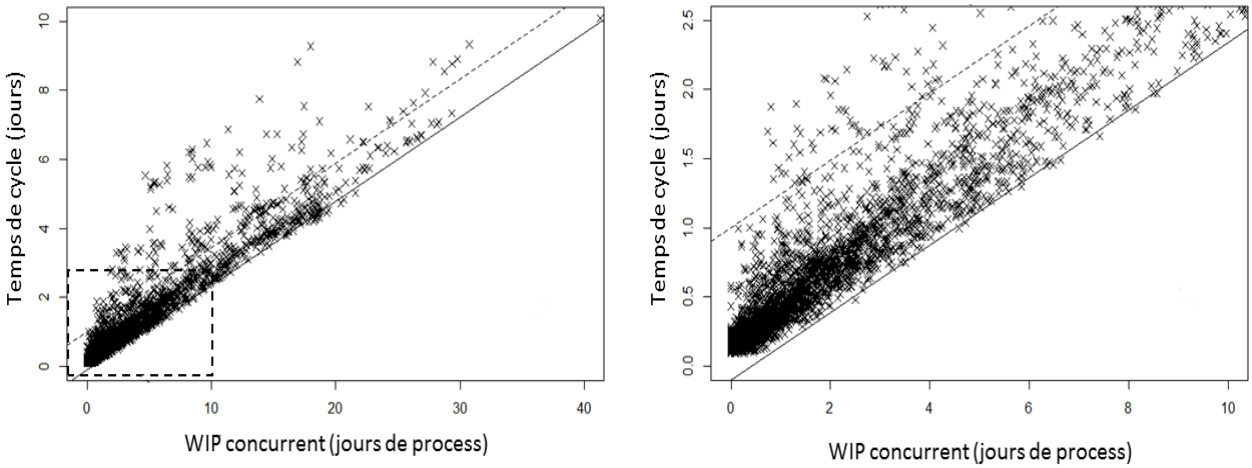


FIGURE 3.1 – Relation entre WIP concurrent et temps de cycle [DLEV16a].
Mesures sur des données réelles de STMicroelectronics (gauche) et zoom sur les temps courts (droite).

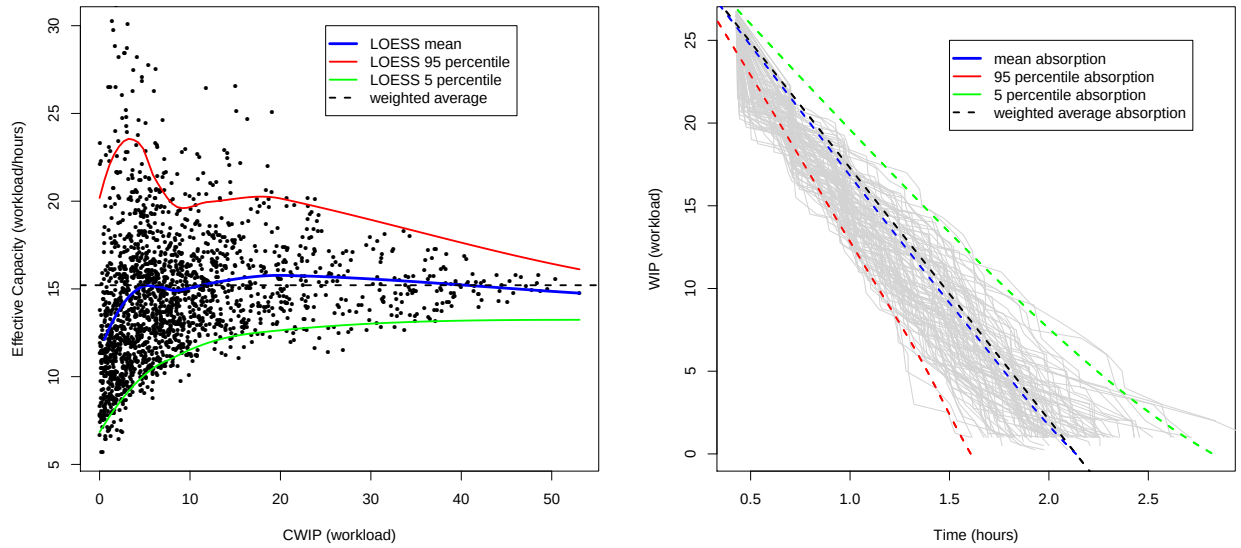
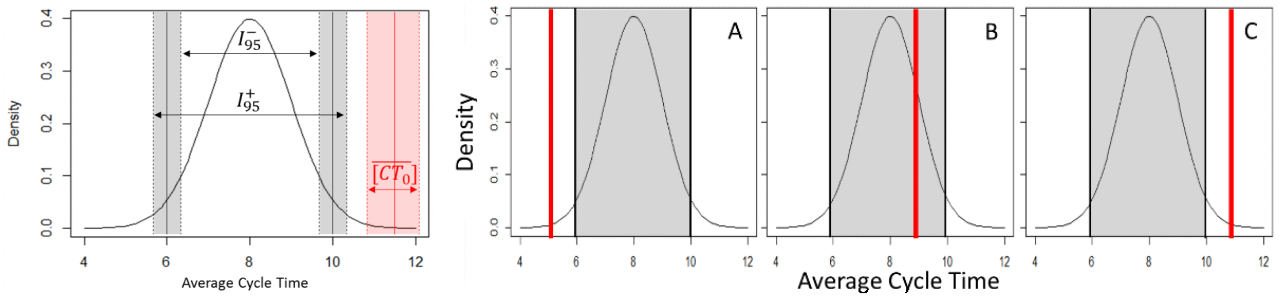


FIGURE 3.2 – Estimation de la capacité réelle d'un système.
Capacité réelle d'un système simulé (gauche) et absorption de WIP de ce même système (droite).

En plus de l'effet de la diversité de produits sur la variabilité, nous nous sommes aussi intéressés à celui de la dépendance entre événements. L'hypothèse d'indépendance est très souvent faite en planification même si beaucoup d'éléments du terrain viennent la contredire. Il y a pour cela deux raisons principales : la première c'est que l'indépendance des événements simplifie grandement formules et calculs ; la deuxième est que la dépendance est difficile à qualifier, en particulier pour des flux de production aussi complexes où les dépendances sont multiples et à peu près impossible à modéliser. L'enjeu de la dépendance des événements a déjà été identifié et des méthodes de corrections proposées [ATDS09] pour des systèmes relativement simples. Dans notre cas, nous nous sommes posés la question de l'incidence de ces dépendances. Pour le dire naïvement : il est inutile de s'en faire quant aux dépendances si celles-ci n'ont pas d'effet sur le temps de cycle.

Nous avons donc proposée une méthodologie qui mesure l'effet de la dépendance [DLEV17]. Prenons l'exemple des arrivées de produits. Pour une étape donnée, nous disposons de la séquence réelle (historique) des arrivées, S_0. Cette séquence est utilisée comme entrée d'un système simulé qui nous donne un temps de cycle moyen de référence, CT_0. Nous «cassons» alors la dépendance éventuelle en générant de multiples séquences S_i, permutations uniformes de S_0, ce qui permet d'obtenir le temps de cycle d'une séquence quelconque, que l'on caractérise par un intervalle de confiance empirique à 95%, I_{95}. Notez que ces transformations de S_0 ne changent pas les caractéristiques synthétiques de la séquence, ce qui signifie qu'une formule comme celle de Hopp et Spearman est complètement aveugle à de telles modifications. Pour tenir compte de la variabilité due aux autres paramètres (par exemple les temps de process), les simulations sont répétées plusieurs fois afin de déterminer les intervalles d'incertitude empiriques autour de CT_0 et I_{95} (Figure 3.3(a)). Il suffit alors de comparer CT_0 et I_{95} pour connaître l'effet de la dépendance, comme l'illustre la Figure 3.3(b). Testée sur 19 équipements de STMicroelectronics, cette méthode a permis de mettre en évidence des effets très changeants selon les équipements : pour un tiers d'entre eux, l'effet est limité (de 5% à 20–30% de différence de temps de cycle) et même parfois bénéfique dans 2 cas ; pour un tiers, l'effet est assez conséquent (30% à 100%) ; et pour un petit tiers il est majeur, avec des temps de cycles sous-estimés par des facteurs 4, 5 et jusqu'à plus de 10 !



(a) (b)

FIGURE 3.3 – Incidence de l'indépendance sur le temps de cycle
(a) Évalutalion du temps de cycle de référence CT_0, de l'intervalle de confiance sur le temps de cycle sans dépendance I_{95} et des incertitudes sur ces valeurs. (b) Comparaisons des valeurs : la dépendance diminue le temps de cycle (A), l'augmente (C) ou n'a pas d'effet significatif (B).

Certes, une telle analyse ne donne aucune solution. Elle ne permet que de cibler les équipements pour lesquels on va avoir des surprises, ce qui serait déjà beaucoup dans une usine avec des centaines d'équipements si ces équipements n'étaient pas déjà identifiés, comme c'est en général le cas en pratique. Pour les équipements dont on sait déjà que le temps de cycle est mal estimé, cette approche reste néanmoins intéressante car elle permet de quantifier la contribution de la dépendance des arrivées (ou des pannes ou de tout autre événement) dans le biais constaté, et ainsi de pouvoir agir sur les bons leviers afin de mieux maîtriser le temps de cycle.

3.2 Optimisation de réseaux énergétiques

Le second cas d'étude porte sur l'optimisation des réseaux énergétiques et est le sujet de la thèse d'Étienne Cuisinier (en cours), que je co-encadre avec Bernard Penz (G-SCOP), Alain Ruby et Cyril Bourrasseau (CEA-Liten). Plus précisément, le sujet initial est le dimensionnement des réseaux énergétiques. Le cas typique est l'approvisionnement en chaleur et électricité d'un quartier, pour lequel il faut décider de l'implantation ou pas

de différents moyens de production (chaudières au bois, chaudière au gaz, panneaux solaires, éoliennes...) et de stockage. Il s'agit alors de décider du choix des technologies employées et de leur dimensionnement afin d'avoir un réseau énergétique qui satisfasse la demande à moindres coûts (euros et carbone) d'investissement et d'opération.

Une particularité essentielle des systèmes envisagés et que l'on permet des stockages de long terme qui permettent d'accumuler de la chaleur en été pour la restituer en hiver. Par un état de l'art conséquent [CBR+21], nous avons établi que de tels stockages n'étaient que peu ou pas pris en compte dans la littérature, même en restant à un niveau de pilotage du système. Or, sans un bon modèle de pilotage, il est impossible d'évaluer correctement les coûts opérationnels, qui dépendent fortement du bon usage du réseau. Il nous a donc fallu construire un modèle d'optimisation correspondant au pilotage sur une année d'un réseau. Concrètement, nous avons proposé un programme linéaire en nombres entiers que nous avons résolu par des techniques d'horizon glissant avec des pas de temps variables et différentes modalités d'agrégation des données. Cela permet en fait de simuler un pilotage «optimal» du système, comme l'illustre la Figure 3.4. C'est de la jolie recherche opérationnelle ([CLP+], soumis), mais ce n'est «que» de la recherche opérationnelle, aussi je ne m'y attarde pas davantage dans ce chapitre.

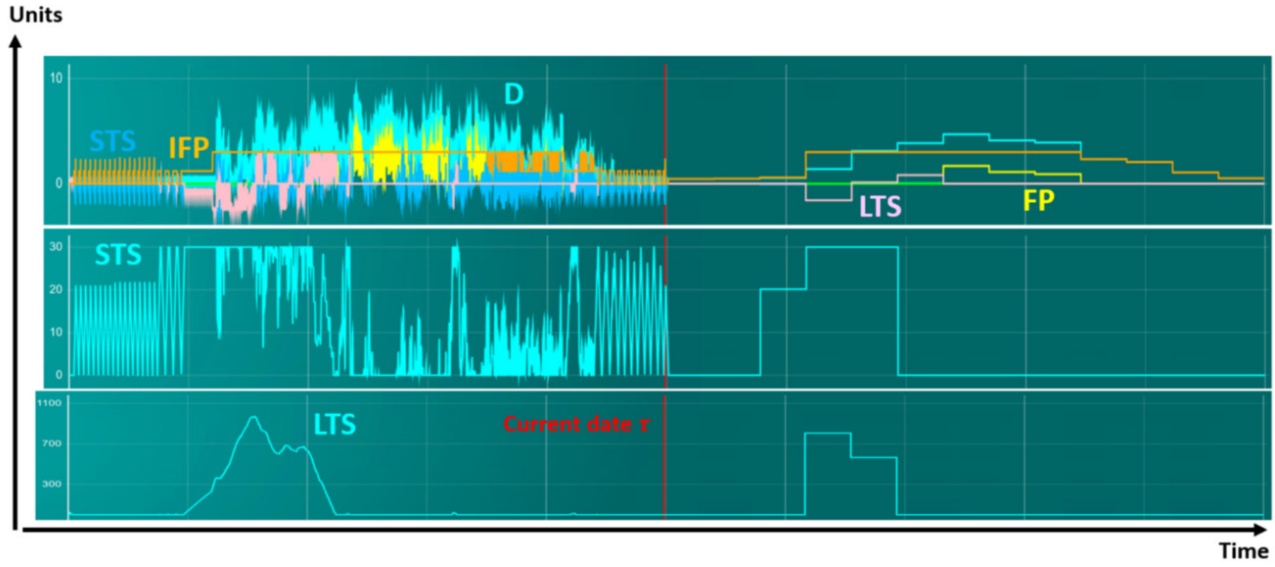


FIGURE 3.4 – Simulation d'un pilotage optimal d'un réseau de chaleur.
La demande (D) est satisfaite grâce à une chaudière responsable mais contraignante (IFP, *InFlexible Production unit*), une chaudière pratique mais polluante (FP, *Flexible Production unit*), un stockage court terme (STS, *Short Term Storage*) et un stockage long terme (LTS, *Long Term Storage*). Les deux graphiques du bas montrent les niveaux des stocks. La ligne rouge (*current date*) indique le présent de la simulation ; à gauche de la ligne, on voit le réalisé, à droite le planifié. Le graphique complet correspond à 2 années.
(Note : ce graphique est dû à Étienne Cuisiniser)

La question qui m'intéresse ici est plutôt de se demander ce que vaut un tel modèle. Pour être plus clair : peut-on lui faire confiance pour calculer un bon dimensionnement ? Cette question couvre plusieurs aspects et autant de sources de remise en cause :

— Le premier doute, sans doute le plus évident, vient des données. Ainsi la demande ne se reproduit pas exactement d'une année sur l'autre, au gré des hivers rudes et des mois de juillet pluvieux ; les prix de l'énergie évoluent et on peut espérer une électricité moins carbonée dans 10 ans qu'aujourd'hui.
— Le deuxième doute vient de la modélisation du système physique. Le réseau implique avant tout des productions, des transferts et des pertes de chaleur ou d'électricité. À chacun de ces phénomènes sont associés différents paramètres qui sont difficiles à évalués tant que le système n'existe pas — ce qui est bien sûr le cas lorsqu'on veut dimensionner le réseau ; cela rejoint le premier doute sur les données. Surtout, tous ces phénomènes sont non linéaires si on les modélise finement. Ils sont donc approchés par des fonctions linéaires par morceaux avec un compromis entre précision et complexité, c'est-à-dire entre réalisme et temps de calcul, qu'il est difficile d'estimer.
— Le troisième doute vient de la modélisation opérationnelle : les décisions de pilotage proposées se veulent optimisées, mais sont-elles adéquates ? Par exemple, la Figure 3.4 montre un cyclage de l'IFP

sur les premières semaines, ce qui implique des arrêts et des démarrages pouvant entraîner une usure non prise en compte dans le modèle. Certaines décisions, trop finement optimisées, pourraient aussi être incomprises voire contre-intuitives, ou simplement très difficiles à mettre en œuvre effectivement créant un décalage préjudiciable entre le pilotage optimal et le pilotage réel.

Ces trois doutes sont trois aspects d'un même enjeu : le réalisme des solutions produites. Ce réalisme s'immisce dans chaque élément qui conduit à la solution. Ainsi, utiliser un horizon glissant n'est pas qu'un enjeu de temps de calcul : résoudre le programme linéaire directement sur un an se fait plutôt bien, ça ne coûte que quelques jours de calcul, c'est-à-dire une broutille pour un investissement sur 20 ou 40 ans. Par contre, procéder ainsi reviendrait à croire qu'on connaît précisément la météo, heure par heure, un an à l'avance — une hypothèse difficile à qualifier de «réaliste» et qui mènerait à sous-estimer fortement les coûts opérationnels.

Les doutes sur les données peuvent être traités grâce à une analyse par scénarios sur la demande, les prix, *etc*. Pour la modélisation physique du système, on peut mesurer les écarts entre les modèles fins et les approximations utilisées, même s'il est difficile d'en tirer des conclusions quant aux conséquences en termes de décisions opérationnelles ; on peut surtout tester différentes approximations, plus ou moins fines, afin de déterminer un niveau de finesse suffisant. Tester différents scénarios et différentes modélisations est typiquement ce que nous sommes en train de faire dans le cadre de la thèse d´Étienne Cuisinier, afin de déterminer les éléments prépondérants, les plus influants. Ce faisant, on se contente en fait de faire une analyse de sensibilité des solutions aux choix d'optimisation du décideur. La mesure du réalisme se limite à vérifier la cohérence des variations — on utilise moins d'électricité quand celle-ci est plus chère, ou davantage de stockage quand les pertes de celui-ci sont minimes.

Pour conclure remarquons que ce réalisme, qui semble difficile à cerner, est aussi nécessaire que superflu ! Il est indispensable pour avoir une estimation correcte des coûts avant de décider de la construction du réseau. Il est indispensable si on veut mettre en œuvre les décisions préconisées, que ce soit automatiquement (elles doivent être réellement réalisables) ou intelligemment (elles doivent faire sens et donner confiance). Par contre, si on revient à la question initiale d'un dimensionnement optimal, le réalisme n'est pas utile. En effet, reprenons notre casquette brodée RO : étant donné la taille et la complexité du système, on ne peut guère imaginer comme méthode d'optimisation qu'une décomposition en un problème maître (le choix du dimensionnement) et un problème esclave (l'évaluation du pilotage de ce dimensionnement), le tout étant traité au sein d'une méta-heuristique. Dans un tel cadre, il suffit que l'évaluation du pilotage classe dans le bon ordre les choix de dimensionnement — le plus vite étant le mieux, afin d'évaluer un plus grand nombre de solutions. On a donc en fait tout intérêt à avoir deux modèles : le premier, utile pour l'optimisation, sera aussi rapide que possible et simplement capable de montrer la bonne direction ; le second, utile pour qualifier la solution finale, pourra se permettre de prendre son temps, mais devra être réaliste.

3.3 Discussion

Vous l'aurez sans doute senti à la lecture, les travaux présentés dans ce chapitre sont encore «en cours». C'est évident pour les travaux sur l'optimisation des réseaux énergétique et les éléments proposés en conclusion de la section 3.2 ne sont ni plus ni moins que les éléments de réflexion issus de nos dernières réunions de travail, et dessinent le plan de travail des derniers mois de la thèse d'Étienne Cuisinier — un calendrier bien court pour une telle feuille de route, mais qui se prolonge aussi bien chronologiquement que scientifiquement grâce à une nouvelle thèse, elle aussi financée par le CEA, et que je dirigerai, en collaboration avec Mathieu Vallée et Alain Ruby (CEA-Liten), sur la «contribution des méthodes d'apprentissage et de modélisation probabilistes des incertitudes à l'optimisation du dimensionnement et du pilotage des réseaux d'énergie multi-vecteur».

Les travaux sur la variabilité des flux de production ne sont, eux non plus, pas terminés. Le premier enjeu est de dépasser le cadre d'étude de l'entreprise d'accueil afin de mieux qualifier les méthodologies et leurs usages (articles [DLEVb] et [DLEVa], en préparation). À plus long terme, la question des usages est particulièrement judicieuse pour le WIP concurrent qui définit un nouveau point de vue, la focale n'étant plus centrée sur le serveur, dont on est loin d'avoir tirées toutes les conséquences pratiques et théoriques. Quant à l'étude de la dépendance, elle trouverait des prolongements naturels par une mise en saine concurrence voire collaboration avec des méthodes qui essaient de modéliser une partie de la dépendance pour des systèmes «pas si complexes» (comme par exemple [SNWL13]), ou qui s'efforcent d'apprendre le comportement du système pour en déterminer les paramètres adéquats [TK21].

Au delà des deux cas d'études présentés, ce chapitre a été l'occasion d'un positionnement scientifique différent des deux précédents.

L'optimisation et les sciences de données étaient déjà présentes dans les deux premiers chapitres de ce manuscrit : l'optimisation lorsqu'il s'agit de calculer un ordonnancement ou un modèle d'analyse logique de données, les sciences de données lorsqu'il s'agit d'évaluer les performances empiriques d'une heuristique ou d'un modèle de diagnostic. Les deux disciplines étaient mises à contribution, mais en un dialogue timide, chacune restant essentiellement cantonnée à des étapes spécifiques, clairement identifiées et séquentielles. Il s'agissait plus de prises de paroles successives que de véritables discussions. Les cadres et les enjeux des études justifiaient cela et il eût été maladroit de faire autrement.

Au contraire, les travaux présentés dans ce chapitre prônent une intrication de la recherche opérationnelle, du génie industriel et des sciences de données, aussi bien pour définir les enjeux que pour les outils à mobiliser. Une dichotomie entre calcul et évaluation serait inappropriée, au mieux stérile, au pire trompeuse. Ainsi, les jolies formules de calcul de temps de cycle sont inopérantes si elles ignorent des phénomènes majeurs (section 3.1), et les préconisations de pilotage sont inutilisables si les modèles n'intègrent pas aussi bien la physique du réseau que le comportement de ses opérateurs (section 3.2).

Les disciplines sont joyeusement contraintes à se comprendre au delà d'une collaboration pluridisciplinaires entre chercheurs. On ne peut plus se contenter, pour les uns de poser des questions, pour les autres de proposer des réponses. Chaque élément de réponse devient une question en retour, une remise en cause au sein d'une conversation permanente, une querelle constructive afin de réussir à faire émerger, voire réussir à converger vers, un compromis entre usages, performances, compréhension, temps de calcul, réalisme...

La liste des critères pourraient être très longue, mais arrêtons nous sur ce dernier item, sans doute le plus important. Dans le cadre des réseaux énergétiques, j'ai mis en évidence la nécessité du réalisme aussi bien que sa futilité, selon les usages. Le phénomène n'est, bien entendu, aucunement lié à une application particulière. Certes, il est admis depuis longtemps que tous les modèles sont faux[4], mais pour que certains soient utiles encore faut-il être capables de les qualifier.

Le monde est euclidien lorsque je le parcours à pied. De même, des systèmes de production simples seront tout à fait correctement représentés par des problèmes d'ordonnancement, par exemple si le système est suffisamment isolé de l'extérieur (observations d'objets célestes, section 1.2.1) ou si les décisions sont limitées dans les choix et dans leurs effets (production de puces électroniques, section 1.2.2). Au delà, un ≪réalisme suffisant≫ ne peut pas être une hypothèse sans être consciencieusement étayée et mise à l'épreuve. L'enjeu est simple : savoir ce qui est validé et ce qui ne l'est pas ; savoir ce en quoi on peut avoir confiance, et ce dont il faut se méfier.

Contrairement à beaucoup de modèles physiques, la confrontation à la réalité n'est pas toujours possible. Même en supposant l'hypothèse que l'on maîtrise suffisamment le système pour pouvoir mener une expérimentation — hypothèse irréaliste pour des systèmes complexes —, STMicroelectronics ne pourrait pas se permettre de sacrifier quelques jours de production pour valider la modélisation d'un flux qui aura changé le mois suivant, et l'on ne construira un système énergétique qu'après l'avoir dimensionné. Quant au diagnostic médical, on imagine facilement des expériences aussi informatives qu'inacceptables sur un plan éthique.

À défaut de pouvoir expérimenter, un premier niveau de qualification du réalisme est donné par une analyse de sensibilité aux données ou aux hypothèses. Le premier enjeu est de reconnaître un manque de réalisme et d'en mesurer les effets potentiels, afin de ≪ne pas se préoccuper des souris quand on est à côté de tigres≫ [Box76] ; c'est, par exemple, ce que nous avons proposé dans le cas de la mesure de l'effet de l'indépendance (section 3.1) — un cas facile car on ne se préoccupe de l'effet que sur une grandeur, le temps de cycle. Dans des cas plus compliqués, comme le dimensionnement des réseaux énergétiques, la difficulté est de ne pas se contenter de la sensibilité de la valeur optimale, comme le fait une majeure partie de la littérature. Malencontreusement, aller au delà est ardu et souvent spécifique. Pour les réseaux énergétiques on peut facilement analyser la sensibilité des solutions de dimensionnement, par exemple vérifier qu'on utilise davantage le stock si son rendement augmente, mais il est déjà extrêmement malaisé de comparer les solutions de pilotage : est-il plus différent d'avoir allumé la chaudière un peu plus fort ou un peu plus tôt ? La littérature sur la robustesse pourrait apporter des éléments de réponse. Quoi qu'il en soit, il y a derrière cette qualification du réalisme des questions avec des enjeux importants pour toute recherche appliquée.

Pour conclure, rappelons que, dans cette quête de réalisme, il faut se méfier de modèles de plus en plus compliqués qui sont non seulement en contradiction avec le principe de parcimonie d'Ockham, mais dont les solutions deviennent incompréhensibles. Ceci est particulièrement sensible pour des modèles d'aide à la

4. ≪*All models are wrong, but some are useful*≫, aphorisme attribué à George Box.

décision de systèmes complexes : ils intègrent aussi, pour partie, des comportements sociaux ou individuels avec le danger que la réalité s'adapte au modèle [5] par le pouvoir prescriptif des solutions et décisions proposées.

5. ≪In cases of major discrepancy it's always reality that's got it wrong.≫ affirme humouristiquement Douglas Adams dans son *Hitchhicker's guide to the galaxy*.

Chapitre 4

Épilogue

À travers ce manuscrit, je me suis efforcé de transcrire un portrait de mes travaux représentatif à défaut d'être — et heureusement — exhaustif. J'ai préféré donner les idées plutôt que les résultats techniques ; tant pis pour les quelques élégants théorèmes et autres jolis algorithmes qu'on aurait pu croiser, les publications leur siéent davantage. Je ne reviendrai pas sur les conclusions et les perspectives déjà proposées à la fin de chaque chapitre, afin de me concentrer sur la dynamique d'ensemble qui m'a permis de mener ces travaux et surtout préside à leurs suites. Aussi, avec ce chapitre conclusif, je veux présenter les perspectives qui ne s'inscrivent pas le cadre limité d'un des chapitres précédents (section 4.1) avant de conclure sur ce qu'est, selon moi, une «aide à la décision de confiance» (section 4.2).

4.1 Recherche opérationnelle || Sciences des données

Recherche opérationnelle et sciences de données sont deux piliers complémentaires de l'aide à la décision. La première a une vocation essentiellement prescriptive tandis que la seconde est plutôt descriptive ou prédictive. Les liens sont forts et anciens, que ce soit dans l'utilisation de statistiques pour analyser les performances d'algorithmes d'optimisation, ou dans l'utilisation d'algorithmes d'optimisation (souvent continue) pour calibrer des modèles d'apprentissage automatique. Jusqu'à récemment, toutefois, les deux disciplines avaient plus tendance à se côtoyer qu'à s'entre-mêler.

Mes travaux illustrent cela : j'ai moi-même développées des compétences propres dans ces domaines, deux «jambes». La première est profondément drapée dans la recherche opérationnelle (chapitre 1) ; la deuxième est habillée de sciences de données (chapitre 2). Cependant, après avoir appris à marcher, c'est-à-dire à faire avancer ces deux jambes l'une après l'autre en cohérence (chapitre 3), il s'agit maintenant de courir et de sauter, c'est-à-dire de renforcer la convergence et l'intrication entre recherche opérationnelle et sciences de données. Plusieurs voies sont ouvertes pour explorer la croisée de ces deux mondes.

La première direction est l'utilisation des sciences de données pour la recherche opérationnelle. L'objectif reste alors l'optimisation d'un problème «simple» (par opposition à complexe, pas à difficile), et les sciences de données sont un des moyens mis à contribution pour y parvenir. Cette optimisation guidée par les données (et en particulier l'apprentissage automatique) revêt plusieurs aspects.

Tout d'abord, les sciences de données permettent de mieux caractériser les solutions, les instances et les algorithmes. Ainsi Arnold et Sörensen [AS19] décrivent et reconnaissent les bonnes solutions pour des tournées de véhicules, tandis que Bostel, Castagliola, Dejax et Langevin [BCDL14] approchent la valeur optimale d'une instance pour le problème du postier chinois. De tels résultats peuvent alors servir à guider des heuristiques ou des méthodes exactes. Ils permettent aussi d'établir la sensibilité aux données, et donc de gagner en robustesse, par exemple pour anticiper le coût d'une tournée dont on ne connaît pas précisément à l'avance les points de passage, comme cela avait été étudié dans la thèse de Natalia Duarte-Ferrin (co-encadrée avec Van-Dat Cung (G-SCOP) et Iragaël Joly (GAEL) ; thèse abandonnée pour raisons personnelles) [DFLCJ14, DFLCJ15]. Enfin, l'apprentissage automatique permet d'anticiper les performances d'algorithmes et donc d'utiliser l'heuristique la plus appropriée ou avec le bon paramétrage [BLP20] ; on pourrait ainsi arrêter une méta-heuristique quand le gain marginal d'une minute de calcul supplémentaire devient trop faible, ou ne lancer qu'une heuristique : celle qui nous donnera presque sûrement la meilleure solution. Enfin, l'apprentissage automatique ouvre de nouveaux paradigmes de résolution, comme celui de transformer une instance d'un problème difficile en une instance d'un problème facile «de même solution», comme cela a été fait très récemment pour des problèmes d'ordonnancement par Parmentier et T'Kindt [PT21].

Deux remarques s'imposent. La première est qu'être capable d'anticiper la performance d'un algorithme pour une instance donnée répond en grande partie aux limites pratiques de la théorie de la complexité, discutées à la fin du chapitre 1. Rappelons que cette théorie s'attache aux pires cas des problèmes, ce qui est

particulièrement pessimiste dans la réalité. En anticipant les performances des algorithmes sur une instance donnée, on arrive à anticiper la difficulté à résoudre l'instance qui nous intéresse, ce qui permet une résolution beaucoup plus à-propos.

La deuxième remarque est que de tels travaux dépendent, par définition, des données utilisées, un aspect en général éludé bien qu'il soit primordial. De un cadre d'optimisation, les données et autres instances sont en général générées ou extraites de dépôts reconnus (comme la VRP lib pour les problèmes de tournées de véhicule). Dans tous les cas, elles sont choisies, bien qu'il y ait rarement discussion quant à ce choix. On peut pourtant imaginer au moins deux objectifs antagonistes : d'une part avoir des instances les plus typiques possibles et donc relativement semblables, au risque d'obtenir des résultats peu généralisables et enfermés dans une «bulle de filtre» ; d'autre part avoir des instances les plus informatives possibles et donc très diversifiées, au risque d'une généricité peu performante. Ces deux objectifs se traduisent, chacun, en une question de recherche fondamentale : comment décider des meilleures données pour un bon apprentissage ? et comment vérifier qu'une instance peut être utilisée sereinement avec un modèle d'apprentissage ?

Ces questions peuvent surprendre, surtout la première, quand on a plutôt tendance à utiliser toutes les données disponibles et à compter sur les capacités de généralisation de la méthode d'apprentissage pour faire le tri. Cependant une telle approche, aussi frontale, implique des modèles avec un nombre gigantesque de paramètres et donc par nature peu compréhensibles. L'alternative est d'avoir des modèles spécialisés au champ d'application restreint.

Un tel cas s'est présenté dans les perspectives de la thèse CIFRE de Mehdi Kessar [Kes21] (co-encadrée par Bertrand Le Gratiet, STMicroelectronics) et portant sur la modélisation de la topographie de puces électroniques à partir des données de design (*i.e* de conception). En effet, d'une part nous avons établi que des modèles simples (linéaires) étaient suffisants pour modéliser la topographie, mais se généralisaient mal si les données d'apprentissage étaient trop diversifiées ; d'autre part nous avons proposée une mesure de distances entre puces qui se corrèle bien à la qualité des prédictions. De là, l'idée prospective d'incorporer une étape de sélection des données dans la construction d'un modèle de topographie pour une nouvelle puce, c'est-à-dire de sélectionner les données de manière à fabriquer une puce fictive proche de la nouvelle puce et de calculer le modèle sur cette puce fictive. Cela n'est pas sans poser des questions quant à la validation d'une telle procédure et sur ce qu'on modélise réellement ainsi.

Une deuxième direction de recherche liant recherche opérationnelle et sciences de données est la symétrique de celle qui vient d'être discutée. L'objectif est alors l'apprentissage automatique, et la recherche opérationnelle est un moyen d'améliorer celui-ci.

Bien entendu, les premières perspectives portent sur développements de l'analyse logique de données, aussi bien sur des aspects méthodologiques que pratiques, afin de se rapprocher des performances des méthodes «boîtes noires» tout en conservant le caractère compréhensif et explicatif. Cela a déjà été discutée au chapitre 2.

Dans le cadre de la thèse de Yuzhen Wang (thèse en cours, co-dirigée par Iragaël Joly (GAEL) et Nadia Brauner (G-SCOP)), nous nous intéressons à la méthode des plus proches voisins (NN, *Nearest Neighbors*) et en particulier au calcul d'une distance optimale pour cet algorithme [WLJB21] — un sujet qui n'a été, curieusement, que très partiellement traité (voir [HM15]). Comme pour l'analyse logique de données, l'objectif est double avec une amélioration des performances prédictives comme explicatives : les coefficients de la «bonne» distance pouvant être interprétés en terme d'importance des différentes variables, comme cela est communément fait pour une régression linéaire.

La troisième direction de recherche ambitionne une intrication plus serrée entre recherche opérationnelle et sciences de données, sans que l'une soit au service de l'autre, mais avec pour enjeu de questionner les usages de l'aide à la décision, en particulier pour les systèmes complexes. Comme nous avons pu le voir dans le chapitre 3, pour de tels environnements, l'information est nécessairement partielle et changeante, mais il faut néanmoins être capable de prendre des décisions rapides et justes pour anticiper les problèmes sans en créer de nouveaux.

Il y a tout d'abord une opportunité : comme le remarquent Khakifirooz, Fathi et Wu [KFW19] pour le cas de la micro-électronique, recherche opérationnelle et sciences de données sont bien connues dans l'industrie mais encore utilisées de manière relativement séparées et seulement pour une petite partie de leurs applications potentielles, en se concentrant sur la gestion de production et l'ordonnancement. Cela signifie qu'il y a un environnement mature et propice aux innovations. Leurs conclusions rejoignent tout à fait ce que j'ai pu observer chez STMicroelectronics dans le cadre des thèses de Kean Dequeant (section 3.1) et de Mehdi Kessar.

Au delà d'un certain opportunisme, la pénétration de plus en plus profonde des sciences de la décision au sein de ces systèmes complexes invite à questionner ce que valent modèles et décisions. Comme je l'ai discuté à la fin du chapitre 3, la confrontation au réel pour de tels systèmes n'est pas évidente. Il est donc fondamental d'apprendre à distinguer ce qu'on peut effectivement modéliser ou résoudre avec telle ou telle procédure. Pour illustrer cela, je m'appuie sur la thèse de Nikita Gusarov (thèse en éconmie, en cours, dirigée par Iragaël Joly (GAEL) et que je co-encadre) : pour modéliser des choix de transport, la littérature propose plusieurs modèles correspondant à des hypothèses de comportements différentes ; testés sur des jeux de données contrôlés, il apparaît que tous les modèles s'ajustent aux données, c'est-à-dire que les indicateurs statistiques sont bons même quand le comportement décrit par le modèle est faux ! La procédure de calcul du modèle n'est ainsi pas capable de se rendre compte de sa propre erreur et il donc nécessaire d'avoir d'autres garde-fous — cela est d'autant plus nécessaire quand ce genre de modèles est utilisé pour des décisions lourdes, comme la construction d'infrastructures telles une nouvelle autoroute...

4.2 Aide à la décision de confiance

La conclusion du paragraphe précédent pourrait être prise comme exemple typique de mésusage des sciences de la décision lors d'une prise de décision : mal comprises, les «aides» à la décision se révèlent trompeuses.

Les sciences de la décision sont conçues pour répondre aux problèmes d'*autres* personnes. Je n'observe pas d'exo-planètes, je ne fabrique pas de puces électroniques, je ne soigne pas des malades. Le scientifique de la décision évolue dans un environnement fondamentalement pluridisciplinaire ; c'est à lui d'établir les théorèmes et de concevoir les algorithmes indispensables mais que n'auront à comprendre ni les astrophysiciennes, ni les médecins, ni les micro-électroniciennes, ni les énergéticiens, *etc.* La scientifique de la décision n'est pas en première ligne ; elle fournit les outils mais doit aussi fournir les modes d'emploi et assumer sa part de responsabilité si les outils sont malencontreusement mal employés.

À travers mes travaux, des pratiques ont évolué à la satisfaction des praticiens. J'ai à peu près réussi à faire comprendre l'intérêt d'une validation internationale à quelques médecins qui étaient loin de telles préoccupations (section 2.1) ; j'ai participé au changement des habitudes et des approches de gestion de production au sein du Leti (section 1.2.2). Il a fallu pour cela se comprendre et apporter du sens en plus des outils. Sens et outils : les deux composantes, pratiques et théoriques, d'une «aide à la décision de confiance», dont j'ai essayé de dessiner un panorama à travers ce qui précède, et qui sert de boussole à mes recherches.

Les deux derniers paragraphes suffisent pour entendre ma définition de l'«aide à la décision» ; c'est, de plus, une notion suffisamment partagée pour ne pas s'y attarder. Le qualificatif «de confiance» se décline, lui, sous de multiples aspects et mérite d'être développé.

La confiance commence par se construire avec une bonne compréhension mutuelle avec les personnes décisionnaires. Il s'agit avant tout de faire sien leur problème, de se l'approprier sans le corrompre, de le questionner, afin qu'à leur tour elles puissent faire siennes nos réponses et s'approprier nos solutions sans se tromper. Il faut donc éviter les idées ou les approches pré-conçues, se permettre de faire un simple tri ou au contraire se l'interdire et bousculer les habitudes, selon ce qui est le plus adéquat. La mesure de cette adéquation se construit aussi bien par le dialogue que par une compréhension des problèmes et de leur structure, une compréhension des données et de leurs propriétés, une compréhension des méthodes, de leurs limites et hypothèses. Les théorèmes établissant la complexité d'un problème, la garantie d'une approximation, ou les conditions d'optimalité d'un algorithme peuvent être aussi nécessaires qu'insuffisants — tout comme les performances expérimentales, pour des raisons symétriques — et il faut s'adapter aux besoins.

Après avoir établi la confiance entre les personnes, il faut établir la confiance avec l'outil. Pour cela, les interfaces et l'intégration sont essentiels. Comme le recommande [KFW19] il faut «intégrer le comportement des décideurs humains dans les solutions». Ces humains doivent être à l'aise, les aides à la décision doivent être pratiques. Cette praticité est aussi indispensable que dangereuse car elle permet aussi de se servir d'un outil n'importe comment. Des bibliothèques comme `scikit-learn` [PVG+11] illustrent très bien cela : elles permettent à n'importe quelle personne disposant de données de s'en servir sans avoir la moindre idée de ce qui est réellement fait ou de ce que peuvent valoir les modèles ainsi créés. Pour un modèle comme la Figure 2.3, l'extrême simplicité d'utilisation ne doit pas faire oublier les hypothèses sous-jacentes et que ce modèle n'est *pas* adapté pour des enfants sans troubles de la croissance. Il y a donc une formation nécessaire aux outils, afin que non seulement les décideurs puissent utiliser l'outil en confiance, mais pour qu'on ait aussi confiance

en l'utilisation de l'outil par les décideurs. Comme le notent Benhamou et Janin [BJ18] : «une grande partie des travailleurs vont mobiliser des dispositifs à base d'IA [1] sans forcément savoir qu'il s'agit d'IA : l'enjeu est ici celui de la formation à la bonne utilisation des outils. Il concerne l'essentiel des conseillers bancaires, des personnels médicaux mais aussi les chauffeurs ou les réparateurs.» Établir la confiance avec l'outil relève donc pour une grande part de la formation de tout un chacun, mais ne dispense pas les scientifiques de la décision de vigilance et de pédagogie.

Le troisième niveau de confiance, sans doute le plus important et sujet de beaucoup de recherches actuelles, est la confiance en les décisions. La personne décisionnaire doit être en mesure de comprendre la décision voire de l'expliquer — même si ces décisions n'impactent que de vulgaires produits, c'est une condition *sine qua none* pour garder le contrôle et pouvoir agir de manière autonome sur un système. Dans d'autres contextes, comme le rappelle le rapport Villani [Vil18] : «à long terme, l'explicabilité de ces technologies est l'une des conditions de leur acceptabilité sociale. S'agissant de certains sujets, c'est même une question de principe : on ne peut admettre, en tant que société, que certaines décisions importantes puissent être prises sans explication. En effet, sans possibilité d'expliquer les décisions prises par des systèmes autonomes, il apparaît difficile de les justifier. Or, comment accepter l'injustifiable dans des domaines aussi décisifs pour la vie d'un individu que l'accès au crédit, à l'emploi, au logement, à la justice ou à la santé ? Cela paraît inconcevable.»

Cette confiance dans les décisions passe notamment par des méthodes d'apprentissage «boîtes blanches», comme l'analyse logique de données, qui justifient leurs prédictions par construction. De côté de l'optimisation, elle passe par des méthodes dont les solutions sont compréhensibles et donc étayées par tous les indicateurs nécessaires à cela. La confiance passe aussi par une évaluation claire des algorithmes et des modèles qui mettent en avant leurs biais et leurs limites : le danger n'est en effet pas d'être faux ou mauvais, c'est de l'être sans le savoir et d'être utilisé en dehors de ce qui légitime ou raisonnable. À noter que si les «boîtes blanches» proposent des justifications de décisions individuelles qu'il est facile de s'approprier, l'évaluation des algorithmes et modèles permet une qualification de leur comportement générique plus difficile à appréhender et plus facile à ignorer ou oublier. La mise en œuvre doit tenir compte de cela — on rejoint la confiance avec l'outil —, ne pas perdre de vue les usages et faire en sorte que les usagers soient éclairés. Quitte à «dégrader» la décision, en ne prenant pas de décisions incertaines (zone grise de la Figure 2.3) ou en bridant l'usage (la production du Leti n'est réellement automatisée qu'à 90% ; le diagnostic médical pas du tout).

Ces trois niveaux de confiance — avec les personnes, les outils, les décisions — sont indispensables pour une aide à la décision de confiance. Les deux premiers niveaux sont avant tout applicatifs mais le troisième doit être bien présent à l'esprit des concepteurs d'algorithmes, de modèles et autres méthodologies. «En premier lieu, il faut accroître la transparence et l'auditabilité des systèmes autonomes d'une part, en développant les capacités nécessaires pour observer, comprendre et auditer leur fonctionnement et, d'autre part, en investissant massivement dans la recherche sur l'explicabilité» [Vil18].

Dans un processus intégrant de l'aide à la décision, il est facile de diluer les responsabilités. La personne qui a rassemblées données et instances et celle qui a implémentés les algorithmes ; celle qui a calibrés les algorithmes sur les données ; celle qui a obtenues des solutions et celle qui s'en est servi. Pourtant une décision, même automatisée ou partagée, reste une décision humaine. Pour paraphraser Sartre : décider ne pas choisir, c'est déjà une décision [2]. Pour cela, les sciences de la décision ont besoin d'une réflexion éthique propre, encore largement à construire, qui aille au delà des codes de déontologie actuels comme, par exemple, celui pourtant très bon de l'ACM (*Association for Computing Machinery*) [ACM].

Comme l'a très bien documenté O'Neil [O'N16], les outils d'aides à la décision se transforment facilement en armes de destructions matheuses. Certes, le fabricant d'une kalachnikov peut moins prétendre à l'innocence que celui d'un couteau de cuisine, même si les deux outils permettent de tuer, mais la distinction n'est pas si simple [3] et cette comparaison triviale ne doit pas masquer de nombreuses tensions éthiques dont voici quelques exemples.

Quelles décisions aider ? Dans *Hooked* [Eya14], Eyal explique comment créer des applications [4] addictives

1. Par «IA», comprendre tout système automatisé d'aide à la décision ; l'apprentissage automatique en fait partie, la recherche opérationnelle aussi.
2. «Ne pas choisir, c'est encore choisir.»
3. Pour illustrer l'ambiguïté de bien des travaux, je renvoie à la biographie de Fritz Haber. Ce chimiste allemand a permis la synthèse de l'ammoniac, pour laquelle il a reçu le prix Nobel et qui a permis le développement des engrais essentiels pour éviter de terribles familles au XXè siècle. Il a poursuivi ses travaux avec enthousiasme sur la conception et l'usage de gaz de combat, abondamment déversés pendant la première guerre mondiale... D'un point de vue purement comptable, le macabre bilan de Fritz Haber serait pourtant certainement positif si les engrais ammoniaqués n'étaient pas, aujourd'hui, identifiés comme hautement toxiques pour l'environnement.
4. Pensez à n'importe quel réseau social grand public...

afin de conditionner ses utilisateurs à prendre les ≪bonnes≫ décisions. Il ne s'agit ni plus ni moins que de développer des systèmes artificiels qui entraînent des intelligences humaines à ≪prendre≫ sans s'en rendre compte des décisions prescrites, ce qu'Eyal justifie, sans la moindre ironie, par l'assurance que les décisions sont ≪bonnes≫ pour l'utilisateur pour la seule raison que, sinon, ça serait mal et donc on ne le ferait pas, et par l'affirmation confortable que l'utilisateur, même dupé, reste entièrement responsable [5].

À quel prix aider une décision ? Certaines applications sont tout à fait acceptables, mais le coût de l'aide à la décision peut être jugé prohibitif. Par exemple, est-il équilibré de proposer une aide au covoiturage (avec de jolis problèmes de calculs d'itinéraires mêlant optimisation et apprentissage...) en contre-partie d'un traçage systématique de tous les déplacements et toutes les localisations quotidiennes ? La question est d'autant plus importante pour des données personnelles sensibles comme les données de santé où la vie privée de chaque individu est mise en balance avec l'intérêt général. Comme l'indique le Comité consultatif national d'éthique : ≪une voie de passage éthique doit être trouvée entre l'impératif de protection des données de santé et la nécessité de leur partage pour renforcer la qualité et l'efficience de notre système de santé. [...] La diffusion du numérique en santé peut induire des effets potentiellement importants au regard des inégalités de santé, dans le sens de leur réduction ou, dans certains cas, de leur élargissement≫ [Com18b]. Même si le *Health Data Hub* français existe depuis un an et demi, le débat est loin d'être clos.

Quand utiliser un système d'aide à la décision ? L'aide à la décision est utilisée même pour des décisions critiques, qui peuvent décider sur de mauvaises bases de la vie ou de la trajectoire de vie de personnes. Ainsi, le journal Propublica avait montré que le logiciel d'estimation des risques de récidives utilisé par la justice américaine était raciste et inefficace, conduisant à la détention abusives de prisonniers noirs et la libération de blancs dangereux [O'N16, Vil18]. Cet exemple avait fait dire à O'Neil que de tels algorithmes ne devraient pas être utilisés pour faire des prédictions, mais pour chercher les biais dans les données utilisées.

Même sans biais, la question de la légitimité de l'utilisation d'un système d'aide à la décision est difficile à affirmer en particulier quand elle touche la personne humaine, à moins que les décisions soient compréhensibles. Ainsi, le Comité consultatif national d'éthique ≪propose que soit inscrit au niveau législatif le principe fondamental d'une garantie humaine du numérique en santé, c'est-à-dire la garantie d'une supervision humaine de toute utilisation du numérique en santé, et l'obligation d'instaurer pour toute personne le souhaitant et à tout moment, la possibilité d'un contact humain en mesure de lui transmettre l'ensemble des informations la concernant dans le cadre de son parcours de soins≫ [Com18a].

Enfin, il convient de ne pas oublier que les systèmes d'aide à la décision ont aussi des conséquences pour les personnes qui doivent mettre en œuvre les décisions, avec là encore des risques : ≪susceptible de favoriser une meilleure coordination des organisations, l'IA peut aussi conduire à un plus grand isolement des travailleurs. [...] Il conviendra donc de ne pas sous-estimer les risques liés au déploiement des outils IA en matière de conditions de travail (perte d'autonomie, intensification du travail, *etc.*)≫ [BJ18].

Les sciences de la décision, en particulier en associant recherche opérationnelle et sciences de données, sont extrêmement riches de verrous scientifiques et de perspectives théoriques et applicatives, comme j'ai pu l'illustrer au travers de ce document. Les territoires à explorer sont immenses, parmi lesquels je privilégie les approches explicatives, dussent-elles être moins ≪efficaces≫.

Le développement de l'intelligence artificielle — sciences de données comme recherche opérationnelle — ne doit pas être aussi celui de l'abêtissement naturel. La réflexion et l'engagement doivent donc aussi porter sur l'éthique et ne pas se limiter aux seuls aspects mathématiques et algorithmiques, afin de développer une science de la décision au service de l'humain, de tous les êtres humains.

≪*Models, despite their reputation for impartiality, reflect goals and ideology. [...] Models are opinions embedded in mathematics.*≫ — Cathy O'Neil.

5. Les 15 pages (sur plus de 200) tardivement consacrées à cette question révèlent une pensée un peu plus nuancée, mais dont les méandres permettent avant tout de mieux cerner le marché de son produit.

Bibliographie

[ACM] ACM code of ethics and professional conduct. https://www.acm.org/code-of-ethics.

[AH06a] Gabriele Alexe and Peter L. Hammer. Spanned Patterns in the Logical Analysis of Data. *Discrete Applied Mathematics*, 154(7) :1039–1049, 2006.

[AH06b] Sorin Alexe and Peter L. Hammer. Accelerated Algorithm for Pattern Detection in Logical Analysis of Data. *Discrete Applied Mathematics*, 154(7) :1050–1063, 2006.

[AH12] Juan Felix Avila Herrera. Integer programming applied to rule based systems. *Procedia Computer Science*, 9 :1553–1562, 2012. Proceedings of the International Conference on Computational Science, ICCS 2012.

[ALCS+11] Slimane Allali, Pierre Lemaire, Ana-Claudia Couto-Silva, Géraldine Prété, Christine Trivin, and Raja Brauner. Prediction of differences between adult and target heights and time between puberty onset and first menstruation in 122 girls with precocious puberty after a positive review. *Medical Science Monitor*, 17(4) :PH41–48, 2011.

[AS19] Florian Arnold and Kenneth Sörensen. What makes a vrp solution good? the generation of problem-specific knowledge for heuristics. *Computers & Operations Research*, 106 :280–288, 2019.

[ATDS09] R. Akhavan-Tabatabaei, S. Ding, and J. G. Shanthikumar. A method for cycle time estimation of semiconductor manufacturing toolsets with correlations. In *Proceedings of the 2009 Winter Simulation Conference*, MASM'09, pages 1719–1729, 2009.

[BC96] Brenda S. Baker and Edward G. Coffman, Jr. Mutual exclusion scheduling. *Theoretical Computer Science*, 162(2) :225–243, 1996.

[BCDL14] Nathalie Bostel, Philippe Castagliola, Pierre Dejax, and André Langevin. Approximating the length of chinese postman tours. *4OR*, 12(4) :359–372, 2014.

[BCF+03] Nadia Brauner, Yves Crama, Gerd Finke, Pierre Lemaire, and Christelle Wynants. Approximation Algorithms for the Design of SONET Networks. *RAIRO Operations Research*, 37(4) :235–247, 2003.

[BCRH20] Endre Boros, Yves Crama, and Elisabeth Rodríguez-Heck. Compact quadratizations for pseudo-boolean functions. *Journal of Combinatorial Optimization*, 2020.

[BDJ+] Peter Brucker, Christoph Dürr, Sven Jäger, Sigrid Knust, Damien Prot, Rob van Stee, and Óscar C. Vásquez. The scheduling zoo. http://www-desir.lip6.fr/~durrc/query/.

[BH07] Tibérius Bonates and Peter Hammer. Pseudo-Boolean Regression. Technical Report RRR 3-2007, Rutcor, Rutgers University, New Jersey, January 2007.

[BHI+00] Endre Boros, Peter L. Hammer, Toshihide Ibaraki, Alexander Kogan, Eddy Mayoraz, and Ilya Muchnik. An Implementation of Logical Analysis of Data. *IEEE Transactions on Knowledge and Data Engineering*, 12(2) :292–306, 2000.

[BHIK97] Endre Boros, Peter L. Hammer, Toshihide Ibaraki, and Alexander Kogan. Logical Analysis of Numerical Data. *Mathematical Programming*, 79 :163–190, 1997.

[BHK08] Tibérius Bonates, Peter Hammer, and Alexander Kogan. Maximum Patterns in Datasets. *Discrete Applied Mathematics*, 156(6) :846–861, 2008.

[BJ18] Salima Benhamou and Lionel Janin. *Intelligence artificielle et travail*. France Stratégie, 2018.

[BJW94] Hans L. Bodlaender, Klaus Jansen, and Gerhard Woeginger. Scheduling with incompatible jobs. *Discrete Applied Mathematics*, 55(3) :219–232, December 1994.

[BLP20] Yoshua Bengio, Andrea Lodi, and Antoine Prouvost. Machine learning for combinatorial optimization : a methodological tour d'horizon, 2020.

[BLS+95] A. Bar, B. Linder, E. H. Sobel, P. Saenger, and J. DiMartino-Nardi. Bayley-pinneau method of height prediction in girls with central precocious puberty : Correlation with adult height. *Journal of Pediatrics*, 126(6) :955–958, 1995.

[Box76] George E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356) :791–799, 1976.

[BP52] N. Bayley and S.R. Pinneau. Tables for predicting adult height from skeletal age : revised for use with the greulich-pyle hand standards. *Journal of Pediatrics*, 40(423–441), 1952.

[BRS+10] A. Rose Brannon, Anupama Reddy, Michael Seiler, Alexandra Arreola, Dominic T. Moore, Raj S. Pruthi, Eric M. Wallen, Matthew E. Nielsen, Huiqing Liu, Katherine L. Nathanson, Börje Ljungberg, Hongjuan Zhao, James D. Brooks, Shridar Ganesan, Gyan Bhanot, and W. Kimryn Rathmell. Molecular stratification of clear cell renal cell carcinoma by consensus clustering reveals distinct subtypes and survival patterns. *Genes & Cancer*, 1(2) :152–163, 2010.

[BRT+10] Marie-Luise Brennan, Anupama Reddy, W. H. Wilson Tang, Yuping Wu, Danielle M. Brennan, Amy Hsu, Shirley A. Mann, Peter L. Hammer, and Stanley L. Hazen. Comprehensive peroxidase-based hematologic profiling for the prediction of 1-year myocardial infarction and death. *Circulation*, 122 :70–79, 2010.

[CBR+21] Etienne Cuisinier, Cyril Bourasseau, Alain Ruby, Pierre Lemaire, and Bernard Penz. Techno-economic planning of local energy systems through optimization models : a survey of current methods. *International Journal of Energy Research*, 45(4) :4888–4931, 2021.

[CCB+16] Nicolas Catusse, Hadrien Cambazard, Nadia Brauner, Pierre Lemaire, Bernard Penz, Anne-Marie Lagrange, and Pascal Rubini. A branch-and-price algorithm for scheduling observations on a telescope. In *IJCAI'16, Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 3060–3066, New-York, USA, 9-15 July 2016.

[CCBL21] Hadrien Cambazard, Nicolas Catusse, Nadia Brauner, and Pierre Lemaire. Teaching OR : automatic evaluation for linear programming modelling. *4OR : A Quarterly Journal of Operations Research*, July 2021.

[CGL] Gang Chen, Jean-Philippe Gayon, and Pierre Lemaire. Stochastic scheduling with abandonment : Necessaryand sufficient conditions for the optimality of a strictpriority policy. *Operations Research*. (seconde révision).

[CHI88] Yves Crama, Peter L. Hammer, and Toshihide Ibaraki. Cause-effect relationships and partially defined boolean functions. *Annals of Operations Research*, 16(1-4) :299–325, 1988.

[CLBTB20] Victoria Corvest, Pierre Lemaire, Sylvie Brailly-Tabard, and Raja Brauner. Puberty and inhibin b in 35 adolescents with pituitary stalk interruption syndrome. *Frontiers in Pediatrics*, 8, 2020.

[CLP+] Etienne Cuisinier, Pierre Lemaire, Bernard Penz, Cyril Bourasseau, and Alain Ruby. New rolling horizon optimization approaches to balance short-term and long-term decisions : an application to energy planning. *Energy*. (accepté).

[Com18a] Comité Consultatif National d'Éthique. *AVIS 129 - Contribution du Comité Consultatif National d'Éthique à la révision de la Loi de bioéthique*, 2018. https://www.ccne-ethique.fr/sites/default/files/avis_129_vf.pdf.

[Com18b] Comité Consultatif National d'Éthique. *Numérique & santé : quels enjeux éthiques pour quelles régulations ?*, 2018. https://www.ccne-ethique.fr/fr/actualites/numerique-sante-quels-enjeux-ethiques-pour-quelles-regulations-0.

[CSTE+06] Ana-Claudia Couto-Silva, Christine Trivin, Hélène Esperou, Jean Michon, André Baruchel, Pierre Lemaire, and Raja Brauner. Final height and gonad function after total body irradiation during childhood. *Bone Marrow Transplant*, 38(6) :427–432, 2006.

[DC11] Annalisa Deodati and Stefano Cianfarani. Impact of growth hormone therapy on adult height of children with idiopathic short stature : systematic review. *British Medical Journal*, 342 :c7157, 2011.

[DCS20] Federico Della Croce and Rosario Scatamacchia. The longest processing time rule for identical parallel machines revisited. *Journal of Scheduling*, 2020.

[DCST19] Federico Della Croce, Rosario Scatamacchia, and Vincent T'kindt. A tight linear time $\frac{13}{12}$-approximation algorithm for the $p2||c_{\max}$ problem. *Journal of Combinatorial Optimization*, 2019.

[Deq17] Kean Dequeant. *Modélisation de la variabilité des flux de production en fabrication microélectronique*. PhD thesis, Univ. Grenoble Alpes, 2017.

[DFLCJ14] Natalia Duarte-Ferrin, Pierre Lemaire, Van-Dat Cung, and Iragaël Joly. An economic efficiency analysis of the capacitated vehicule routing problem. In *27th Conference of the European Chapter on Combinatorial Optimization*, ECCO, München, Germany, 1-3 May 2014.

[DFLCJ15] Natalia Duarte-Ferrin, Pierre Lemaire, Van-Dat Cung, and Iragaël Joly. Analyse économétrique des solutions d'un cvrp. In *16ème conférence de la Société Française de Recherche Opérationnelle et Aide à la Décision*, ROADeF, Marseille, France, 25-27 February 2015.

[DLEVa] Kean Dequeant, Pierre Lemaire, Marie-Laure Espinouse, and Philippe Vialletelle. Concurrent WIP : a tool to analyze manufacturing systems subject to complex variability. (en préparation).

[DLEVb] Kean Dequeant, Pierre Lemaire, Marie-Laure Espinouse, and Philippe Vialletelle. A study of variability induced by events dependency in microelectronic production. (en préparation).

[DLEV16a] Kean Dequeant, Pierre Lemaire, Marie-Laure Espinouse, and Philippe Vialletelle. Le WIP concurrent : une proposition de file d'attente du point de vue du produit pour caracteriser le temps de cycle. In *MOSIM'16, 11th International Conference on Modeling, Optimization & SIMulation*, Montreal, Canada, 22-24 August 2016.

[DLEV16b] Kean Dequeant, Pierre Lemaire, Marie-Laure Espinouse, and Philippe Vialletelle. A literature review on variability in semiconductor manufacturing : The next forward leap to industry 4.0. In T.M.K Roeder, P.I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, editors, *Proceedings of the 2016 Winter Simulation Conference*, MASM'16, pages 2598–2609, Washington D.C, U.S.A, 11-14 December 2016.

[DLEV17] Kean Dequeant, Pierre Lemaire, Marie-Laure Espinouse, and Philippe Vialletelle. A study of variability induced by events dependency in microelectronic production. In *IESM 2017 (7th IESM Conference)*, Saarbrücken, Germany, 11–13 October 2017.

[Dro96] Maciej Drozdowski. Scheduling multiprocessor tasks - an overview. *European Journal of Operational Research*, 94 :215–230, 1996.

[dRSHK07] Maria A. J. de Ridder, Theo Stijnen, and Anita C. S. Hokken-Koelega. Prediction of Adult Height in Growth-Hormone-Treated Children with Growth Hormone Deficiency. *The Journal of Clinical Endocrinology & Metabolism*, 92(3) :925–931, 03 2007.

[EHKR09] Guy Even, Magnús M. Halldórsson, Lotem Kaplan, and Dana Ron. Scheduling with conflicts : online and offline algorithms. *Journal of Scheduling*, 12(2) :199–224, 2009.

[EVL+11] L. F. P. Etman, C. P. L. Veeger, E. Lefebre, I. J. Adan, and J. E. Rooda. A method for cycle time estimation of semiconductor manufacturing toolsets with correlations. In *Proceedings of the 2011 Winter Simulation Conference*, 2011.

[Eya14] Nir Eyal. *Hooked : How to Build Habit-Forming Products*. Portfolio Penguin, 2014.

[FBL01] Gerd Finke, Nadia Brauner, and Pierre Lemaire. Packing of multi-bin objects. In *New trends in scheduling for parallel and distributed systems*, Marseilles, 1-5 October 2001.

[FCKK94] B.D. Fletcher, D.B. Crom, R.A. Krance, and R.E. Kun. Radiation-induced bone abnormalities after bone marrow transplantation for childhood leukemia. *Radiology*, 191 :231–235, 1994.

[FHW16] Eibe Frank, Mark A. Hall, and Ian H. Witten. *Data Mining : Practical Machine Learning Tools and Techniques*, chapter The WEKA Workbench. Online Appendix. Morgan Kaufmann ; 4e édition, 2016. `https://www.cs.waikato.ac.nz/ml/weka/`.

[FLBa] Fontan Florian, Pierre Lemaire, and Nadia Brauner. Complexity of scheduling tasks with processing time dependent profit. (en préparation).

[FLBb] Fontan Florian, Pierre Lemaire, and Nadia Brauner. A fast, efficient, simple and versatile large neighborhood search algorithm to schedule star observations on the very large telescope. (en préparation).

[FLPQ09] Gerd Finke, Pierre Lemaire, Jean-Marie Proth, and Maurice Queyranne. Minimizing the number of machines for minimum length schedules. *European Journal of Operational Research*, 199(3) :702–705, 2009.

[FM87] Matteo Fischetti and Silvano Martello. Worst-case analysis of the differencing method for the partition problem. *Mathematical Programming*, 37 :117–120, 1987.

[Fon19] Florian Fontan. *Theoretical and practical contributions to star observation scheduling problems*. PhD thesis, Univ. Grenoble Alpes, 2019.

[FSM+06] G. Federico, M.E. Street, M. Maghnie, M. Caruso-Nicoletti, S. Loche, S. Bertelloni, and S. Cianfarani. Assessment of serum igf-i concentrations in the diagnosis of isolated childhood-onset gh deficiency : A proposal of the italian society for pediatric endocrinology and diabetes (siedp/isped). *Journal of Endocrinological Investigation*, 29 :732–737, 2006.

[GH 00] GH Research Society. Consensus Guidelines for the Diagnosis and Treatment of Growth Hormone (GH) Deficiency in Childhood and Adolescence : Summary Statement of the GH Research Society. *The Journal of Clinical Endocrinology & Metabolism*, 85(11) :3990–3993, 11 2000.

[GJ79] Michael R. Garey and David S. Johnson. *Computers and Intractability (A Guide to the Theory of NP-Completeness)*. W.H. Freeman And Company, 1979.

[GJP00] Scott M. Graham, Anupam Joshi, and Zygmunt Pizlo. The traveling salesman problem : A hierarchical model. *Memory & Cognition*, 28(7) :1191–1204, 2000.

[GLB15] Éloïse Giabicani, Pierre Lemaire, and Raja Brauner. Models for predicting the adult height and age at first menstruation of girls with idiopathic central precocious puberty. *PLOS ONE*, 10(3) :e0120588, April 2015.

[GLLRK79] Ronald R. Graham, Eugene L. Lawler, Jan Karel Lenstra, and Alexander H. G. Rinnooy Kan. Optimization and approximation in deterministic sequencing and scheduling theory : A survey. *Annals of Discrete Mathematics*, 5 :287–326, 1979.

[GLPR09] Olivier Guyon, Pierre Lemaire, Éric Pinson, and David Rivreau. Near optimal and optimal solutions for an integrated employee timetabling and production scheduling problem. In *INCOM'09, 13th IFAC Symposium on Information Control Problems in Manufacturing*, Moscow, Russia, 3-5 June 2009.

[GLPR10] Olivier Guyon, Pierre Lemaire, Éric Pinson, and David Rivreau. Cut generation for an integrated employee timetabling and production scheduling problem. *European Journal of Operational Research*, 201(2) :557–567, 2010.

[GLPR14] Olivier Guyon, Pierre Lemaire, Éric Pinson, and David Rivreau. Solving an integrated job-shop problem with human resource constraints. *Annals of Operations Research*, 213(1) :147–171, February 2014.

[Gra69] Ronald L. Graham. Bounds On Multiprocessor Timing Anomalies. *SIAM Journal of Applied Mathematics*, 17(2) :416–429, March 1969.

[Hay16] Peter Hayes. Early puberty, medicalisation and the ideology of normality. *Women's Studies International Forum*, 56 :9–18, 2016.

[HB06] Peter L. Hammer and Tibérius Bonates. Logical Analysis of Data : From Combinatorial Optimization to Medical Applications. *Annals of Operations Research*, 148 :203–225, 2006.

[HH92] Peter L. Hammer and Ron Holzman. Approximations of pseudo-boolean functions; applications to game theory. *ZOR-Methods and Models of Operations Research*, 36(1) :3–21, 1992.

[HKL07] Peter Hammer, Alexander Kogan, and M.A. Lejeune. Reverse-engineering bank's financial strength ratings using logical analysis of data. Technical Report RRR 10-2007, Rutcor, Rutgers University, New Jersey, January 2007.

[HM15] Jens Hocke and Thomas Martinetz. Maximum distance minimization for feature weighting. *Pattern Recognition Letters*, 52 :48–52, 2015.

[HS08] Wallace J. Hopp and Mark L. Spearman. *Factory Physics : Foundations of Manufacturing Management*. New-York : Irwin/McGraw-Hill (third edition), 2008.

[IBM11] IBM Corporation. *IBM SPSS Modeler CRISP-DM Guide*, 2011. `ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf`.

[JMHLF+20] Julie Jacques, Hélène Martin-Huyghe, Justine Lemtiri-Florek, Julien Taillard, Laetitia Jourdan, Clarisse Dhaenens, David Delerue, Arnaud Hansske, and Valérie Leclercq. The detection of hospitalized patients at risk of testing positive to multi-drug resistant bacteria using MOCA-I, a rule-based "white-box" classification algorithm for medical data. *International Journal of Medical Informatics*, 142 :104242, 2020.

[JTD+15] Julie Jacques, Julien Taillard, David Delerue, Clarisse Dhaenens, and Laetitia Jourdan. Conception of a dominance-based multi-objective local search in the context of classification rule mining in large and imbalanced data sets. *Applied Soft Computing*, 34 :705–720, 2015.

[Kes21] Mehdi Kessar. *Sciences de données pour la microélectronique : analyse de topographie*. PhD thesis, Univ. Grenoble Alpes, 2021.

[KFW19] Marzieh Khakifirooz, Mahdi Fathi, and Kan Wu. Development of smart semiconductor manufacturing : Operations research and data science perspectives. *IEEE Access*, 7 :108419–108430, 2019.

[Kin09] J. F. C. Kingman. The first erlang century—and the next. *Queueing Systems*, 63(1) :3–12, 2009.

[KK82] Narendra Karmarkar and Richard M. Karp. The differencing method of the set partitioning. Technical Report UCB/CSD 82/113 Computer SCi. Division (EECS), University of California, Berkeley, California, 1982.

[KL17] Daniel Kowalczyk and Roel Leus. An exact algorithm for parallel machine scheduling with conflicts. *Journal of Scheduling*, 20(4) :355–372, 2017.

[KYB21] Ramy M. Khalifa, Soumaya Yacout, and Samuel Bassetto. Developing machine-learning regression model with logical analysis of data (lad). *Computers & Industrial Engineering*, 151 :106947, 2021.

[LBH+09] Pierre Lemaire, Nadia Brauner, Peter Hammer, Christine Trivin, Jean-Claude Souberbielle, and Raja Brauner. Improved screening for growth hormone deficiency using logical analysis of data. *Medical Science Monitor*, 15(1) :MT5–10, 2009.

[LCST+11] Sylvie Laporte, Ana-Claudia Couto-Silva, Séverine Trabado, Pierre Lemaire, Sylvie Brailly-Tabard, Hélène Espérou, Jean Michon, André Baruchel, Alain Fischer, Christine Trivin, and Raja Brauner. Inhibin b and anti-müllerian hormone as markers of gonadal function after hematopoietic cell transplantation during childhood. *BMC Pediatrics*, 11(20), 2011.

[LDdBM+18] Pierre Lemaire, Gwénaëlle Duhil de Bénazé, Dick Mul, Sabine Heger, Wilma Oostdijk, and Raja Brauner. A mathematical model for predicting the adult height of girls with idiopathic central precocious puberty : A european validation. *PLOS ONE*, 13(10) :1–10, 10 2018.

[Lem04] Pierre Lemaire. *Rangements d'objets multiboîtes : modèles et algorithmes*. PhD thesis, Université Joseph Fourier (Grenoble 1), Grenoble, France, September 2004.

[Lem05] Pierre Lemaire. The `ladoscope` gang : tools for the logical analysis of data. `http://www.kamick.org/lemaire/software-lad.html`, 2005. OCaml, GNU General Public License.

[Lem11] Pierre Lemaire. Extensions of logical analysis of data for growth hormone deficiency diagnoses. *Annals of Operations Research*, 186(1) :199–211, 2011.

[Lem18] Pierre Lemaire. Mathematical models for growth issues. `http://www.kamick.org/lemaire/med/`, 2010–2018. html/javascript, CC-BY-SA.

[Lem20] Pierre Lemaire. Ordonnancement - l'art de faire chaque chose en son temps. *Tangente*, HS75, 2020.

[Lem21] Pierre Lemaire. Pilzegal - la recherche opérationnelle par le jeu. *Bibliothèque Tangente*, à paraitre, 2021.

[LFB05] Pierre Lemaire, Gerd Finke, and Nadia Brauner. The Best-Fit Rule for Multibin Packing : An extension of Graham's list algorithms. In Graham Kendall, Edmund Burke, Sanja Petrovic, and Michel Gendreau, editors, *Multidisciplinary Scheduling ; Theory and Applications*, pages 269–286. Springer, 2005.

[LFB06] Pierre Lemaire, Gerd Finke, and Nadia Brauner. Models and Complexity of Multibin Packing Problems. *Journal of Mathematical Modelling and Algorithms*, 5(3) :353–370, 2006.

[LLC+18] Der-Chiang Li, Wu-Kuo Lin, Chien-Chih Chen, Hung-Yu Chen, and Liang-Sian Lin. Rebuilding sample distributions for small dataset learning. *Decision Support Systems*, 105 :66–76, 2018.

[LPBB14] Pierre Lemaire, Delphine Pierre, Jean-Baptiste Bertrand, and Raja Brauner. A mathematical model for predicting the adult height of girls with advanced puberty after spontaneous growth. *BMC Pediatrics*, 14(1) :172–178, 2014.

[LRB+16] Anne-Marie Lagrange, Pascal Rubini, Nadia Brauner, Hadrien Cambazard, Nicolas Catusse, Pierre Lemaire, and Laurence Baude. Spot : an optimization software for dynamic observation programming. In *SPIE Astronomical Telescopes + Instrumentation*, Edinburgh, United Kingdom, 18 July 2016.

[LRL+18] Mateus Cavarzan Lopes, Carolina Oliveira Ramos, Ana Claudia Latronico, Berenice B. Mendonça, and Vinicius N. and Brito. Applicability of a novel mathematical model for the prediction of adult height and age at menarche in girls with idiopathic central precocious puberty. *Clinics*, 73 :e480, 2018.

[McN59] R. McNaughton. Scheduling with deadlines and loss functions. *Management Sciences*, 6 :1–12, 1959.

[MKAL07] Wil Michiels, Jan Korst, Emile Aarts, and Jan van Leeuwen. Performance ratios of the karmarkar-karp differencing method. *Journal of Combinatorial Optimization*, 13(1) :19–32, 2007.

[MO96] J.N. MacGregor and T. Ormerod. Human performance on the traveling salesman problem,. *Perception & Psychophysics*, 58(4) :527–539, 1996.

[Mol21] Christoph Molnar. *Interpretable Machine Learning*. leanpub.com, 2021. https://christophm.github.io/interpretable-ml-book/.

[O'N16] Cathy O'Neil. *Weapons of Math Destruction*. Crown Books, 2016.

[Pin16] Michael L. Pinedo. *Scheduling : Theory, Algorithms, and Systems*. Springer ; 5ème édition, 2016.

[PLH+13] Luu-Ly Pham, Pierre Lemaire, Annie Harroche, Jean-Claude Souberbielle, and Raja Brauner. Pituitary stalk interruption syndrome in 53 postpubertal patients : Factors influencing the heterogeneity of its presentation. *PLOS ONE*, 8(1) :e53189, January 2013.

[PR10] David Pisinger and Stefan Ropke. Large neighborhood search. In *Handbook of metaheuristics*, pages 399–419. Springer, 2010.

[PT21] Axel Parmentier and Vincent T'Kindt. Learning to solve the single machine scheduling problem with release times and sum of completion times, 2021.

[PV21] Axel Parmentier and Thibaut Vidal. Optimal counterfactual explanations in tree ensembles. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8422–8431. PMLR, 18–24 Jul 2021.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011. https://scikit-learn.org/stable/.

[Roc93] P. Rochiccioli. Intérêt des méthodes de prédiction de la taille adulte. *Journal de Pédiatrie et de Puériculture*, 6(7) :401–405, 1993.

[RWY+08] Anupama Reddy, Honghui Wang, Hua Yu, Tiberius O. Bonates, Vimla Gulabani, Joseph Azok, Gerard Hoehn, Peter L. Hammer, Alison E. Baird, and King C. Li. Logical analysis of data (LAD) model for the early diagnosis of acute ischemic stroke. *BMC Medical Informatics and Decision Making*, 8(1), 2008.

[Sap11] Gilbert Saporta. *Probabilités, analyse des données et Statistique*. Technip, 3e édition révisée, 2011.

[SDZ07] J. George Shanthikumar, Shengwei Ding, and Mike Tao Zhang. Queueing theory for semiconductor manufacturing systems : A survey and open problems. *IEEE Transactions on Automation Science and Engineering*, 4(4) :513–522, 2007.

[SGK13] Dvir Shabtay, Nufar Gaspar, and Moshe Kaspi. A survey on offline scheduling with rejection. *Journal of Scheduling*, 16(1) :3–28, 2013.

[SGL13] Alexandre Salch, Jean-Philippe Gayon, and Pierre Lemaire. Optimal static priority rules for stochastic scheduling with impatience. *Operations Research Letters*, 41(1) :81–85, 2013.

[SGY12] Dvir Shabtay, Nufar Gaspar, and Liron Yedidsion. A bicriteria approach to scheduling a single machine with job rejection and positional penalties. *Journal of Combinatorial Optimization*, 23(4) :395–424, 2012.

[She00] Colin Shearer. The CRISP-DM model : The new blueprint for data mining. *Journal of Data Warehousing*, 5(4) :13–22, 2000.

[Slo11] Susan A. Slotnick. Order acceptance and scheduling : A taxonomy and review. *European Journal of Operational Research*, 212(1) :1–11, 2011.

[SNWL13] Songsak Sriboonchitta, Hung T. Nguyen, Aree Wiboonpongse, and Jianxu Liu. Modeling volatility and dependency of agricultural price and production indices of thailand : Static versus time-varying copulas. *International Journal of Approximate Reasoning*, 54(6) :793–808, 2013.

[SSAH09] M. Subasi, E. Subasi, M. Anthony, and P.L. Hammer. Using a similarity measure for credible classification. *Discrete Applied Mathematics*, 157(5) :1104–1112, 2009.

[SSH+17] Ersoy Subasi, Munevver Mine Subasi, Peter L. Hammer, John Roboz, Victor Anbalagan, and Michael S. Lipkowitz. A classification model to predict the rate of decline of kidney function. *Frontiers in Medicine*, 4 :97, 2017.

[SSS18] Akiyoshi Shioura, Natalia V. Shakhlevich, and Vitaly A. Strusevich. Preemptive models of scheduling with controllable processing times and of scheduling with imprecise computation : A review of solution approaches. *European Journal of Operational Research*, 266(3) :795–818, 2018.

[TK21] Barış Tan and Siamak Khayyati. Supervised learning-based approximation method for single-server open queueing networks with correlated interarrival and service times. *International Journal of Production Research*, 0(0) :1–26, 2021.

[TLBTB18] Jérémie Tencer, Pierre Lemaire, Sylvie Brailly-Tabard, and Raja Brauner. Serum inhibin b concentration as a predictor of age at first menstruation in girls with idiopathic central precocious puberty. *PLOS ONE*, 13(12) :1–14, 12 2018.

[Tuf17] Stéphane Tufféry. *Data Mining et Statistique décisionnelle : La science des données*. Technip, 5e édition revue et augmentée, 2017.

[Vie17] Eliane Viennot. *Non, le masculin ne l'emporte pas sur le féminin !* Editions IXe ; 2ème édition, 2017.

[Vil18] Cédric Villani. *Donner un sens à l'intelligence artificielle*. Mission parlementaire, 2018.

[WHFH16] I Witten, H., Elbe Frank, and Mark A. Hall. *Data Mining : Practical Machine Learning Tools and Techniques*. Morgan Kaufmann ; 4e édition, 2016.

[WLJB21] Yuzhen Wang, Pierre Lemaire, Iragaël Joly, and Nadia Brauner. Programmation linéaire pour améliorer la performance de k-nn avec la distance euclidienne pondérée. In *22ème conférence de la Société Française de Recherche Opérationnelle et Aide à la Décision*, ROADeF, Mulhouse, France, 26-30 April 2021.

[ZLY10] Liqi Zhang, Lingfa Lu, and Jinjiang Yuan. Single-machine scheduling under the job rejection constraint. *Theoretical Computer Science*, 411(16) :1877–1882, 2010.

Annexe A

Synthèse des activités de recherche

A.1 Activités de recherche présentes et passées

Mes activités de recherche relèvent un premier lieu de la recherche opérationnelle (en particulier l'étude de la complexité de problèmes et la conception de méthodes de résolutions), mais aussi des sciences de données (en particulier l'apprentissage automatique), avec une conjonction des deux domaines autour de la modélisation et l'analyse des systèmes complexes de production. Mes travaux mélangent aussi bien des aspects théoriques que des applications, avec une constante : proposer des solutions adaptées, que ce soit un théorème technique ou une heuristique simple mais efficace, selon le contexte (souvent pluridisciplinaire). La description de ces travaux est organisée autour de 3 axes principaux : l'ordonnancement, l'apprentissage automatique pour le diagnostic médical, les systèmes complexes de production.

A.1.1 Ordonnancement

L'axe principal de mes recherches s'inscrit dans la théorie de l'Ordonnancement. Parmi eux, une grande partie de mes travaux académiques portent sur des analyses de complexité et la conception d'algorithmes de résolution efficaces.

Ces travaux ont commencé lors de ma thèse [Lem04], encadrée par G. Finke et N. Brauner, sur des problèmes de ≪rangements multiboîtes≫, qui correspondent en fait à des problèmes d'ordonnancement sur machines parallèles où chaque tâche existe en plusieurs copies qu'il faut placer sur des machines différentes, avec diverses contraintes possibles (machines dédiées, incompatibilités...). Nous avons proposés différents modèles et analysé leur complexité et approximabilité [LFB06], et proposé des algorithmes d'approximation généralisant les classiques ≪algorithmes de liste≫ de Graham [LFB05]. À la fin de ma thèse, j'ai profité d'une visite de M. Queyranne pour étudier un problème original où il s'agit de minimiser les ressources (les machines) sous contraintes de ne pas rallonger la durée de l'ordonnancement. Nous avons fait l'analyse de complexité des principaux cas [FLPQ09].

Plus récemment, j'ai participé avec d'autres membres de mon équipe (N. Brauner, H. Cambazard, N. Catusse, B. Penz), à une collaboration scientifique avec AM. Lagrange et P. Rubini de l'Institut de Planétologie et d'Astrophysique de Grenoble (IPAG). L'objet était l'ordonnancement d'observations sur le Very-Large-Telescope, dans le cadre de la recherche d'exo-planètes. Nous avons formalisé le problème, déterminé sa complexité et proposé de premières approches exactes et heuristiques [CCB+16]. Ces travaux ont été poursuivis dans le cadre de la thèse de F. Fontan (co-encadrée avec N. Brauner) avec une étude théorique détaillée de la complexité avec différents algorithmes polynomiaux reposant sur des techniques variées (règles de priorités, flots...). Nous avons également proposé une méthode de résolution pour le problème réel, intégrant toutes les spécifications données par les astrophysiciens, et qui améliore les résultats existants et permet plus de flexibilité quant à la définition du problème. Ces travaux sont en train d'être finalisés pour publication [FLBa, FLBb].

Auparavant, après avoir rejoint l'IRCCyN et l'École des Mines de Nantes, j'avais co-encadré la thèse d'O. Guyon (avec É. Pinson et D. Rivreau), sur des méthodes exactes pour la résolution de problèmes couplés de planification et ordonnancement. Dans ce cadre, nous avons proposé différentes modélisations en programmation linéaire en nombres entiers, ainsi que différentes relaxations lagrangiennes et générations de coupes pour la résolution [GLPR10, GLPR14]. Une version ≪heuristique≫ des méthodes, permettant une résolution efficace et rapide, a été récompensée à INCOM09 [GLPR09].

Le dernier aspect de mes travaux académiques en ordonnancement relève de l'optimisation stochastique. Plus précisément, j'étudie des ordonnancements pour lesquels les tâches ont des durées aléatoires et peuvent abandonner (avec un coût) selon leur impatience, elle aussi aléatoire. Ces travaux ont été initiés dans le cadre de la thèse de A. Salch (co-encadrée avec JP. Gayon), pendant laquelle nous avons déterminées des conditions nécessaires et suffisantes pour que des politiques de priorité stricte soient optimales dans le cas dit ≪sta-

tique» [SGL13]. Ces travaux se poursuivent sur le cas dit «dynamique», pour lequel nous avons déterminé de nouvelles conditions nécessaires et suffisantes [CGL] (en cours de révisions).

Mes travaux en ordonnancement ne se sont pas limités à des aspects académiques. En particulier, j'ai travaillé avec le CEA-Leti pour les aider à organiser la production de leurs puces électroniques. Cette production est compliquée car elle relève de la «pré-industrialisation», c'est-à-dire qu'il y a déjà des flux conséquents (un en-cours de plusieurs centaines de produits), mais avec de petites séries de produits très variés, qui sont autant de prototypes pour lesquels tout ne se passe pas comme prévu... Avec les ingénieurs du Leti, nous avons re-défini les priorités et les conditions de l'ordonnancement, et intégré tout cela dans un logiciel que j'ai entièrement spécifié et implémenté. Ce logiciel, qui tourne quotidiennement, permet de définir en quelques secondes les tâches prioritaires à avancer et anticiper les retards à venir, et permet de soulager les équipes qui peuvent se consacrer à leur métier : la résolution des problèmes liés à la fabrication, et pas à l'organisation.

A.1.2 Aide au diagnostic médical et apprentissage automatique

Le deuxième grand axe des travaux relève des sciences de données, plus particulièrement l'apprentissage automatique, avec des applications au diagnostic médical, en particulier pour des troubles de la croissance. Ces travaux ont commencé lorsque je suis parti en post-doc à Rutcor (Rutgers University, NJ), et se poursuivent depuis plus de 15 ans avec des chercheurs-practiciens de l'Assistance-Publique - Hôpitaux de Paris (AP-HP).

Pour partie, ces travaux n'ont pas de rapport direct avec ma discipline de base (la recherche opérationnelle). Ce sont des «applications» pour lesquelles il faut mobiliser des compétences ingénieuses en mathématiques appliquées et en informatique, sans pouvoir se prévaloir de contributions pour ces domaines. Toutefois, ces travaux nécessitent une compréhension fine des outils mathématiques, mais aussi des enjeux et des usages médicaux : il ne s'agit pas de développer des boîtes noires, mais de mieux comprendre des pathologies, et la collaboration ne se résume pas à une simple mise en œuvre. Les résultats obtenus [CLBTB20, TLBTB18, GLB15, LPBB14, PLH+13, LCST+11, ALCS+11] n'en ont pas moins une grande valeur pour la recherche médicale, comme en témoigne la qualité des revues dans lesquelles ils ont été publiés, et une contribution personnelle significative, comme en témoigne le fait d'être «premier auteur» de plusieurs de ces articles (ce qui est révélateur dans cette communauté).

Ces travaux ne sont pas restés au stade d'«applications» et m'ont permis de développer une compétence et une réflexion propres en sciences de données, en particulier sur ce qu'est un «bon» modèle, en tenant compte des usages des modèles produits. Cela a amené des contributions méthodologiques à différents niveaux, le dernier chronologiquement étant celui d'une validation médicale cohérente des résultats [LDdBM+18]. Avant cela, mes travaux en sciences des données ont débuté sur des aspects méthodologique lors de mon post-doc à Rutcor, pendant lequel j'ai travaillé sur l'Analyse Logique de Données (*Logical Analysis of Data*, LAD), une méthode combinatoire d'apprentissage automatique. C'est à cette occasion que j'ai développé `ladoscope` [Lem05], un logiciel utilisé dans différentes publications (pas seulement les miennes). Sur cette base, j'ai ensuite proposé deux extensions de la méthodologie pour répondre à deux problèmes médicaux différents [LBH+09, CSTE+06]. Les contributions originales, relevant bien de la recherche opérationnelle cette fois, ont été détaillées dans un article dédié [Lem11].

A.1.3 Analyse et optimisation de systèmes complexes

Le troisième grand axe de mes recherche permet une synthèse des deux premiers : il s'agit d'analyser et d'optimiser des systèmes de production (dans un sens très large) en mêlant d'une part les approches algorithmiques et d'optimisation que l'on retrouve dans l'ordonnancement, et d'autre part les sciences de données que l'on retrouve dans l'aide au diagnostic médical. Il s'agit en particulier d'étudier des systèmes réels pour lesquels des multiples décisions doivent être prises, selon des points de vue divers, et pour lesquels on ne dispose pas de spécifications complètes, que l'on remplace par des données historiques. Ces travaux se placent naturellement dans le cadre de collaborations pluridisciplinaires.

Les premiers travaux conséquents sur le sujet ont été faits dans le cadre de la thèse de K. Dequeant (CIFRE avec STMicroelectronics, co-encadrée avec ML. Espinouse et P. Vialletelle). Il s'agissait d'étudier la variabilité des flux de productions à travers un état de l'art conséquent quant aux sources de variabilités dans la fabrication microélectronique [DLEV16b], complété par un nouvel outil méthodologique d'analyse des flux basé sur l'historique de production [DLEV16a] ainsi que par une étude originale sur l'importance de l'indépendance

des événements pour la qualité des prédictions de planification [DLEV17]. Ces deux derniers aspects sont en train d'être renforcés pour faire l'objet de publications dans des revues [DLEVa, DLEVb].

Toujours dans le cadre de la fabrication microélectronique, mais sur un autre sujet : la thèse M. Kessar (CIFRE avec STMicroelectronics, co-encadrée avec B. Le Gratiet). Il s'agit cette fois de caractériser la topographie réelle d'une puce afin d'anticiper des difficultés de fabrication ou de détecter des défauts. Pour le premier cas, il s'agit de prédire à partir des données de conception d'une puce quelle sera sa topographie réelle ; dans le deuxième cas, il s'agit de transformer des mesures pour faire ressortir les défauts (thèse en cours).

Enfin, sur un tout autre domaine, la thèse d'É. Cuisinier (thèse CEA, co-encadrée avec B. Penz, A. Ruby, C. Bourasseau) étudie l'optimisation de la planification énergétique du point de vu opérationnel et des investissements. Un panorama des méthodes existantes a été publié [CBR+21], et une nouvelle approche basée sur des horizons glissants originaux a été proposée [CLP+] (travaux soumis).

A.1.4 Autres activités de recherche

Pour que cette présentation de mes travaux soit complète, je signale un premier article sur l'optimisation de réseaux de télécommunication [BCF+03], sujet sur lequel je ne travaille plus.

Par contre, j'ai commencé récemment à concrétiser une activité de recherche qui rejoint mes préoccupations d'enseignant : d'une part en publiant plusieurs articles de vulgarisation de la recherche opérationnelle pour le magasine *Tangente* [Lem21, Lem20], et d'autre part avec une contribution sur l'évaluation de programmes linéaires dans le cadre de l'enseignement, récemment publiée dans une revue reconnue de recherche opérationnelle [CCBL21].

A.2 Projet de recherche

Mon projet de recherche s'inscrit dans la continuité de la dynamique créée ces dernières années, en poursuivant les travaux en cours et en renforçant la convergence entre recherche opérationnelle et sciences de données.

Il s'agit tout d'abord de poursuivre les travaux académiques en ordonnancement. En particulier, les thèses de A. Salch et F. Fontan ont laissé plusieurs questions ouvertes quant à la complexité et l'approximabilité de certains cas. Dans les deux cas, ces modèles épurés ont le double intérêt d'être à la fois très proches de problèmes bien connus, et pourtant de présenter des structures très particulières.

Il s'agit ensuite de poursuivre les collaborations industrielles autour de l'analyse et de l'optimisation des systèmes complexes. Ainsi, d'une part, les travaux en cours avec le CEA se poursuivent avec une nouvelle thèse pour l'étude d'approches couplées pour le dimensionnement et le pilotage des réseaux d'énergie. D'autre part, les résultats obtenus lors des deux thèses CIFRE avec STMicroelectronics (qui doivent bien sûr être finalisés pour des contributions académiques) ouvrent plusieurs pistes quant à l'analyse de systèmes de production, au delà du secteur de la micro-électronique. Dans ces environnements extrêmement complexes, l'information est nécessairement partielle et changeante, mais il faut néanmoins être capable de prendre des décisions rapides et justes pour anticiper les problèmes sans en créer de nouveaux. Ces travaux ont en effet mis en avant la nécessité de développements pour faire co-exister et collaborer les méthodes d'optimisation et les méthodes d'apprentissage pour que chacune vienne compléter l'autre dans l'analyse des systèmes comme dans la prise de décision.

Enfin, il s'agit de contribuer à une ≪intelligence artificielle de confiance≫ où, pour ce qui me concerne, la recherche opérationnelle et l'apprentissage automatique sont intimement liés (contrairement au paragraphe précédent où ils sont essentiellement juxtaposés). D'une part, il s'agit d'améliorer la confiance en l'optimisation en comprenant mieux ce qui fonctionne et les limites des solutions proposées. Cela passe par une intégration d'apprentissage dans les méthodes d'optimisation afin d'anticiper les performances et de gagner à la fois en efficacité, en choisissant les paramétrages adaptés, et en robustesse en connaissant les conséquences de petites variations. D'autre part, il s'agit de développer la confiance en la fouille de données, avec des modèles plus compréhensibles et plus explicatifs. L'expérience acquise, notamment dans le domaine médical et l'analyse logique de données, permet de spécifier les besoins, et la recherche opérationnelle de trouver des solutions. Ces travaux s'inscrivent dans une collaboration initiée avec I. Joly, économiste au laboratoire GAEL (deux thèses en cours) ainsi que dans le cadre de la chaire *AI for data-driven and self-configurable supply chains* de l'institut MIAI.

A.3 Synthèse des activités de recherche

Cette section propose un synthèse en une page de mes activités de recherche. Sur tous les aspects, des détails et compléments sont disponibles dans mon CV détaillé.

Publications J'ai publié 19 articles dans des revues scientifiques, 11 articles dans des conférences internationales avec actes. La table A.1 donne un aperçu des plus «significatives», selon le facteur d'impact (IF) rapporté par WebOfScience et les classements de journaux et conférence établis par CORE et SCImago Journal Rank (quartiles et «SJR score»). Toutes les données ont été récupérées en avril 2021.

Publications	**Journal (ou conference)**	**IF**	**CORE**	**SJR**
[GLPR10, FLPQ09]	European Journal of Operational Research	4.2	A	Q1 (2.36)
[CCB+16]	IJCAI conference	—	A*	— (1.21)
[GLPR14, Lem11]	Annals of Operations Research	2.5	B	Q1 (1.12)
[SGL13]	Operations Research Letters	0.8	—	Q2/Q1 (0.70)
[CBR+21]	International Journal of Energy Research	3.7	—	Q2/Q1 (0.79)
[TLBTB18, LDdBM+18, GLB15, PLH+13]	PLoS One	2.7	—	Q1 (1.02)
[CLBTB20]	Frontiers in Pediatrics	2.6	—	Q1 (0.87)
[LPBB14, LCST+11]	BMC Pediatrics	1.9	—	Q1 (0.85)

TABLE A.1 – Sélection de publications significatives

Encadrements J'ai co-encadré ou dirigé 5 thèses déjà soutenues et je co-dirige 3 thèses en cours. J'ai été l'encadrant principal de plusieurs de ces thèses. J'ai également encadré une dizaine d'étudiants de Master, dont plusieurs en co-encadrement avec des universités étrangères (Eindhoven Univ. of Technology, Virginia Tech University) et une petite quinzaine d'élèves sur des projets d'initiations à le recherche (niveau M1) dont plusieurs ont permis de poser les bases de collaborations ou de travaux en cours.

Collaborations et rayonnement En plus des collaborations au sein des laboratoires dont j'ai fait partie (G-SCOP, Rutcor/Rutgers Univ., IRCCyN), mes collaborations les plus significatives sont avec des médecins de Assistance Publique–Hôpitaux de Paris et Université Paris-Descartes (10 articles publiés) ainsi qu'avec STMicroelectronics (2 thèses CIFRE, 2 articles en préparation). Je travaille également avec un économiste du laboratoire GAEL (2 thèses en cours). Plus récemment, j'ai entamé une collaboration avec le CEA-Liten (1 thèse en cours, 1 article publié, 1 article soumis). J'ai aussi été impliqué dans différents travaux relevant davantage de la valorisation : en particulier, j'ai participé à plusieurs études logistiques avec le Conseil Général et la Chambre d'Agriculture de l'Isère, et j'ai mené un projet d'ordonnancement avec le CEA-Leti.

Je suis, classiquement, arbitre pour différents journaux et conférences. J'ai été examinateur pour 4 jurys de thèse, membre du jury du prix Gaspard Monge pour l'optimisation, et expert scientifique pour des dossiers CIFRE.

Enfin, je fais de la vulgarisation scientifique en participant au dispositif *Partenaire scientifique pour la classe* et en publiant des articles dans le magasine *Tangente*.