Facultés Universitaires Catholiques de Mons
Département de Sciences de Gestion – Louvain School of Management

INTÉGRATION DU FACTEUR HUMAIN DANS LA GESTION DES RESSOURCES PARTAGÉES APPLIQUÉE À LA PROGRAMMATION OPÉRATOIRE

Thèse présentée en vue de l'obtention du grade de
Docteur en Sciences Economiques et de Gestion par

BENOÎT ROLAND

Membres du jury:

Pr. Patrick Scarmure	Facultés Universitaires Catholiques de Mons, Président
Pr. Fouad Riane	Facultés Universitaires Catholiques de Mons, Superviseur
Pr. Michel Gourgand	Université Blaise Pascal (France), Comité
Pr. Laurence Wolsey	Université catholique de Louvain, Comité
Pr. Marc Lambrecht	Katholieke Universiteit Leuven
Pr. Marc Pirlot	Université de Mons
Pr. Philippe Wieser	Ecole Polytechnique Fédérale de Lausanne (Suisse)

Mons, Juin 2010

Remerciements

Le temps est venu pour moi de remercier ceux qui, d'une façon ou d'une autre, ont contribué à l'accomplissement de cette thèse.

Tout d'abord, je tiens à remercier très sincèrement la personne sans qui ceci n'aurait jamais été écrit ! Il s'agit bien sûr de mon promoteur, le Professeur Fouad Riane. Malgré son exil au Maroc, il m'a toujours aidé, soutenu et guidé dans mes recherches, et m'a sans cesse témoigné une grande confiance (avait-il seulement le choix ?) pour me permettre de réaliser cette thèse. Cependant, outre ses qualités scientifiques indéniables, c'est surtout la personne plus que le professeur que je souhaite saluer. Ce sont très certainement ses qualités humaines qui ont permis l'aboutissement de ce travail et qui ont fait naître de notre collaboration une amitié sincère qui, paradoxalement, me ferait presque regretter d'être arrivé au terme de ma recherche ! Mais qui sait ce que nous réserve le futur... J'en profite également pour remercier son épouse, Meriem, qui, contrainte et forcée, m'a accueilli durant trois semaines alors que je terminais la rédaction de ce document (merci Eyjafjöll).

Je profite également de ces quelques lignes pour remercier les autres membres du jury. En premier lieu, les Professeurs Michel Gourgand et Laurence Wolsey pour avoir accepté de faire partie de mon comité d'accompagnement. Ils m'ont épaulé durant ces années de thèse, montrant intérêt et enthousiasme pour mes travaux. Leur expertise m'a été d'une aide incommensurable et a indéniablement contribué à améliorer la qualité de mes travaux. En particulier, je voulais souligner la patience et la disponibilité dont a fait preuve le Professeur Wolsey, m'offrant un soutien providentiel lorsque Fouad était à l'étranger.

Graag wil ik Professor Marc Lambrecht persoonlijk bedanken om, ondanks de communautaire spanningen, te hebben aanvaard om jurylid van dit Frans thesis te zijn. Ik dank hem voor de interessante vragen en discussies die mee hebben bijgedragen tot de verbetering van dit werk.

Les Professeurs Marc Pirlot et Philippe Wieser doivent également être remerciés pour avoir consenti à être membres du jury. Qui plus est, leurs remarques éclairées ont sans nul doute contribué à la qualité de ce travail. Enfin, je remercie le Professeur Patrick Scarmure pour avoir assuré la présidence de ce jury avec tout l'enthousiasme qu'on lui connaît.

A titre purement formel, je tiens à exprimer ma reconnaissance à Messieurs Jean-Sébastien Tancrez et Jean-Philippe Cordier, deux collègues et amis avec qui j'ai eu la chance de collaborer sur les aspects stochastiques entrepris dans nos travaux et qui ont chacun contribué à faire du septième chapitre de ce document ce qu'il est. Quand je repense au chemin parcouru depuis la Chine...

Durant mes années de thèse, j'ai eu l'occasion de côtoyer bon nombre de personnes avec lesquelles j'ai vécu des moments inoubliables. L'ambiance qui régnait au sein des FUCaM restera

très certainement à tout jamais inégalée. Je pense évidemment à Jean-Philippe II (mon binôme) et Jean-Sébastien (mon autre binôme... oui je sais, ça fait beaucoup), mais également à Jean-Philippe premier du nom (on peut le dire ? Oui allez, maintenant on peut le dire : beloteur hors pair !), Mathieu (l'indécis asymétrique), Tonton (beloteur débutant... depuis douze ans), Fabian (l'avorton chimiste), François (le double père), Olivier (la culture incarnée), David (la rigueur incarnée), Guillaume (tel...), Jean-Christophe (the OR-reader). Avant qu'on ne me traite de phallocrate, et simplement parce que je garde le meilleur pour la fin, voici les demoiselles qui ont revêtu une importance particulière durant mon séjour dans le monde académique : droit d'aînesse oblige, commençons par les ancêtres Karine (Danish bar), Sabine (sexy-taxi) et Valérie (pause-métronome), Emerence, Marie et Caroline (Marketing... forever !), Caroline (Barbie Science-Po), les jeunes pousses prometteuses Daisy et Orane, Géraldine (t'as-des-nouvelles-de-Fouad ?), mais également les organisatrices en chef Vanessa et Diana, Sonia (reproductrice en chef), Binsa (biblio-bus),... J'espère que les éventuel(le)s oublié(e)s ne m'en tiendront pas rigueur !

Enfin, je remercie profondément mes parents, mes deux frères adorés ainsi que mes amis les plus proches qui n'ont pas forcément toujours compris ce que je faisais (win-for-four ?) mais qui ont toujours été là pour me soutenir et pour me détourner du droit chemin quand j'en avais grandement besoin (merci les gars !).

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Durant les dernières décennies, l'accès aux soins médicaux pour l'ensemble des citoyens a guidé les politiques de santé de la plupart des pays européens. Les établissements hospitaliers, subventionnés par les états, ont joui durant ces années d'une grande liberté d'actions, notamment grâce à une situation budgétaire peu contraignante. Cette volonté d'améliorer la couverture des soins de santé a propulsé nos systèmes hospitaliers européens parmi les plus performants du monde. Malheureusement le temps de l'insouciance est révolu, en particulier en Belgique où le trou de la sécurité sociale nous rappelle chaque jour l'ampleur des dépenses du secteur médical. Qui plus est, ces dépenses augmentent avec le temps, notamment en raison de techniques médicales plus élaborées, plus efficaces et plus coûteuses, ou d'un recours abusif aux médications de la part d'une patientèle peu soucieuse de sa consommation. Du point de vue des centres hospitaliers, c'est principalement l'inefficacité avec laquelle les ressources médicales sont gérées qui alourdit la note. Pour résorber la dette sanitaire, les états ont pris des mesures drastiques afin de réduire les subsides octroyés. En Belgique, le système est supporté par la sécurité sociale, elle-même alimentée par les cotisations sociales ou autres impôts ; la principale source de financement est donc constituée des organismes assureurs. La rationalisation des dépenses fédérales de santé a poussé le gouvernement belge à mettre en place, depuis 2002, un système de financement des hôpitaux basé sur leur activité. Pour chaque type de pathologie, en fonction de la sévérité du cas, l'hôpital se voit subsidier un certain nombre de journées dites justifiées, calculées sur base de la moyenne nationale nécessaire au traitement de cette pathologie. Si le séjour du patient devait dépasser cette durée justifiée, l'excédent serait à la charge de l'hôpital. La bonne gestion de l'activité médicale conditionne dès lors la santé financière de l'établissement, rendant la recherche de performance indispensable.

L'hôpital est un système à haute valeur ajoutée humaine : ses principales ressources, essentielles à son activité, sont le personnel. Les ressources médicales, qu'elles soient humaines ou matérielles sont mutualisées au sein de ce système, rendant leur organisation complexe. En effet, s'il n'est pas efficient, le partage des ressources provoquera une dégradation des performances générant surcoût, insatisfaction du personnel et diminution de la qualité des soins. Notamment, en tant qu'unité principale de la création de soins, la planification de l'activité chirurgicale au sein du quartier opératoire détermine l'utilisation de ces ressources, conditionnant les performances de celui-ci, et donc celles de l'hôpital. Le quartier opératoire est donc un point stratégique du système hospitalier qui, étant donné le système de financement en vigueur, dicte autant les dépenses que les revenus de l'établissement. Notamment, les gestionnaires des unités chirurgicales affrontent une grande complexité de management, des ressources humaines en particulier, notamment en raison d'un nombre considérable de contraintes légales, économiques et humaines à prendre en compte, tout en assurant de la bonne qualité des soins prodigués. Ces difficultés d'organisation, couplées

à la nécessité de réduire les dépenses, incitent les hôpitaux à adopter des techniques de gestion autrefois réservées au secteur des entreprises, en les adaptant aux spécificités de la production de soins.

En réponse aux besoins de rationalisation managériale exprimés par les centres hospitaliers, et principalement des gestionnaires de quartier opératoire, nous ambitionnons, dans cette recherche, de développer des outils d'aide à la décision permettant d'optimiser le processus de programmation opératoire, et donc la gestion des ressources impliquées dans ce processus. Sur le plan tactique et opérationnel, les procédures développées devront prendre en compte un maximum de contraintes liées à l'activité chirurgicale, en particulier concernant les ressources opératoires, afin d'automatiser la construction des plannings opératoires. L'importance des ressources humaines étant une caractéristique propre au processus étudié, nous envisageons également d'inclure une dimension humaine trop peu présente dans les approches de planification et d'ordonnancement que l'on rencontre classiquement dans la littérature qui s'y rapporte. Au niveau stratégique et tactique, notre intention est de mettre à la disposition des décideurs un outil leur permettant de mesurer l'impact de la variabilité inhérente à l'environnement médical sur les choix relatifs au dimensionnement et à l'organisation du quartier opératoire.

Ce document est organisé comme suit : le chapitre 1 décrit succinctement le contexte hospitalier actuel, sa complexité et la place du quartier opératoire dans le processus de soins.

Le chapitre 2 dresse l'état de l'art des travaux portant sur la planification et l'ordonnancement des opérations chirurgicales au sein des blocs opératoires. Les approches constituant la littérature du domaine y sont classées selon différents points de vue : en fonction de la typologie mise en place, des contraintes prises en compte, des objectifs poursuivis et des outils appliqués pour y parvenir. Cette étude de la littérature met en évidence certaines questions de recherche que nous abordons par la suite.

Le chapitre 3 expose la problématique telle que nous l'avons traitée, définit le cadre pratique de la recherche et résume l'ensemble des notations utilisées dans la suite du document. Il détaille les étapes constituant le déroulement d'une opération chirurgicale ainsi que les contraintes à considérer lors du processus de programmation opératoire. La problématique du quartier opératoire y est également illustrée par la description et l'analyse de données réelles provenant d'un hôpital belge de la région bruxelloise, mettant en exergue les spécificités pratiques de cette unité.

La littérature recense essentiellement deux types de blocs opératoires : ceux qui sont dédiés au traitement exclusif de certaines pathologies (pour raisons structurelles, organisationnelles ou logistiques) et ceux qui sont polyvalents. Nous distinguons également nos approches de résolution selon les caractéristiques des salles composant le quartier opératoire. Le chapitre 4 détaille deux approches de construction automatisée de plannings opératoires, tenant compte notamment des disponibilités des ressources humaines. La première est basée sur un programme linéaire en nombres entiers, permettant d'obtenir des solutions optimales. Confrontée à des données réelles, cette approche exacte montre ses limites d'application pour certains ensembles de test. Afin de contrer la complexité du problème étudié, ce chapitre développe ensuite une approche métaheuristique que nous appliquons pour établir des programmes opératoires de qualité, mais non nécessairement optimums, en des temps de calcul acceptables.

Le chapitre 5 adapte les techniques de résolution du chapitre 4 au contexte de blocs opératoires polyvalents. Cette spécificité engendre un élément nouveau à considérer : la symétrie due à la

nature multifonctionnelle des salles. Nous proposons donc deux approches de programmation en nombres entiers permettant de prendre en compte cette propriété afin de produire des plannings optimums de blocs opératoires polyvalents.

Le chapitre 6 concerne l'intégration du facteur humain au processus de programmation opératoire, et détaille deux approches de modélisation coopératives de l'attribution des plages horaires aux chirurgiens. Toutes deux dérivent de la théorie des jeux : la première s'inspire des concepts d'utilité rencontrés notamment dans les jeux de *Nash bargaining*, la seconde découle de la théorie des enchères.

Le chapitre 7 détermine comment les approches déterministes de programmation opératoires développées dans ce document peuvent être ajustées pour les rendre robustes aux aléas. Nous y détaillons un outil reposant sur la théorie de Markov et permettant, entre autres, d'évaluer quel pourcentage de capacité doit être préservé lors de la construction des plannings opératoires afin d'absorber tout ou partie des événements imprévisibles jonchant le parcours du patient dans le quartier opératoire.

Enfin, ce document s'achève par une conclusion recensant les résultats de notre recherche et par les perspectives qu'il nous paraît intéressant d'inclure à nos travaux.

Chapitre 1

Le contexte hospitalier

A l'instar du monde manufacturier qui, il y a une trentaine d'années, s'est vu contraint de modifier son organisation, ses modes de fonctionnement et ses techniques de gestion pour mieux répondre aux exigences de productivité, de normalisation des dépenses et de satisfaction des clients, le monde hospitalier occidental est en proie, aujourd'hui, à de profonds remaniements. De nombreuses réformes voulues par les pouvoirs publics imposent plus de rigueur dans la gestion des établissements hospitaliers, en vue, d'une part, d'améliorer la qualité de vie des concitoyens en structurant les pratiques médicales, et d'autre part, de rationaliser les dépenses liées aux soins de santé. Cette nécessité d'une gestion plus stricte est renforcée par divers autres facteurs : l'insuffisance de personnel médical, le vieillissement de la population, l'évolution des pathologies ou encore l'exigence accrue d'une patientèle plus avertie des progrès médicaux en sont des exemples. La maîtrise des coûts dans un hôpital n'épargne pas le plateau médico-technique, particulièrement le quartier opératoire. Ce dernier impacte indéniablement la performance globale de l'hôpital. C'est une activité primordiale de la chaîne de création de soins, aux revenus et dépenses importants.

Ce premier chapitre constitue une introduction à la problématique de la gestion hospitalière et a pour but d'éclairer le lecteur sur le fonctionnement de l'hôpital, sur la place du quartier opératoire au sein de celui-ci et sur les ressources partagées dans le domaine des soins de santé. Nous décrivons pourquoi les systèmes hospitaliers sont complexes, quelles sont les étapes d'un processus opératoire chirurgical et pourquoi il y a lieu de s'intéresser à la planification des blocs opératoires.

1.1 L'hôpital

L'hôpital est un maillon central de la production de soins, qui est de plus en plus assurée au moyen de réseaux hospitaliers. C'est un lieu où les relations professionnelles et interpersonnelles entre les gestionnaires, les administratifs, les médecins, les chirurgiens, les infirmiers et les patients sont prépondérantes. Au sein de l'hôpital se juxtaposent diverses compétences et contingences rendant souvent les processus de prise en charge des patients, de même que la gestion du flux, particulièrement complexes à organiser.

Les difficultés d'organisation au sein d'un hôpital sont nombreuses. Elles se manifestent au travers de différents problèmes : délais d'attente, qualité d'accueil, confort insatisfaisant, risques d'erreurs ou d'infections nosocomiales, sous-productivité des ressources médicales, surcoûts de fonctionnement, conditions de travail difficiles et stress pour le corps soignant et le personnel hospitalier. Ceci explique notamment le manque d'engouement pour la carrière médicale, et dès lors la peine que certains établissements éprouvent à engager du personnel de qualité en suffisance.

Les organisations hospitalières ont connu durant les dernières années de profondes mutations à la suite de multiples réformes. Ces réformes traduisent la volonté des pouvoirs publics, d'une part, de garantir l'accès aux soins pour tous, et d'autre part, d'impliquer largement le secteur de la santé dans une dynamique visant l'amélioration de la qualité de vie de l'ensemble des citoyens.

Les hôpitaux sont donc le théâtre de restructurations, de recompositions, de réorganisation en pôles et de mutations des instances. Cette refonte induit, par sa redistribution des responsabilités, des rôles et des fonctions, de nouvelles règles relationnelles entre les acteurs, les services et les unités de soins au sein même d'un établissement hospitalier. Face à une réglementation sans cesse changeante et des restrictions budgétaires sévères, les hôpitaux sont contraints de repenser leurs modes de gouvernance pour améliorer leurs performances.

Ces constantes transformations obligent l'hôpital à s'adapter à un environnement complexe et mouvant, particulièrement dans les pays comme le nôtre où les budgets de fonctionnement octroyés par les Etats sont de plus en plus serrés. Cette adaptation amène nécessairement une réflexion sur l'utilisation rationnelle des ressources de l'hôpital. Les responsables hospitaliers cherchent à optimiser leur gestion pour réduire leurs dépenses tout en garantissant une qualité et une sécurité des soins pour le patient : tout un programme !

1.2 La complexité des milieux hospitaliers

S'il est vrai que l'acte de soin demeure au centre de l'activité hospitalière, l'organisation de l'hôpital, l'organisation logistique, la circulation des patients, des médicaments et des informations diffèrent selon les architectures, la nature des hôpitaux et selon les pays.

Structurellement, l'activité de l'hôpital est multiple et complexe. Elle peut cependant se scinder en deux catégories (Kharraja, 2003), reprenant chacune différents secteurs de l'organisation (Moisdon & Tonneau, 1999) :

1. L'activité orientée patient comprend :
   - Les services cliniques (également appelés unités de soins), qui assurent l'hébergement des patients. Ce secteur nécessite peu de ressources technologiques, le principal des équipements étant constitué des lits d'hospitalisation.
   - Les services de consultation, comme l'accueil aux urgences, qui requièrent également peu de technologie et peuvent être compris dans les unités de soins.
   - Le plateau technique, qui nécessite une technologie importante afin de répondre aux demandes des services cliniques et de consultation. Il comporte trois activités majeures dans la chaîne de création de soins : l'imagerie médicale, la biologie médicale et le plateau médico-technique. S'y déroulent les actes thérapeutiques et les diagnostics médicaux.

2. L'activité d'appoint et de support comprend :
   – Le secteur logistique, qui gère essentiellement les fonctions de pharmacie, de transport et brancardage des patients et des consommables, de blanchisserie et de restauration. Certaines de ces activités peuvent être sous-traitées.
   – Le secteur administratif, qui regroupe la direction générale, le service financier, la gestion des ressources humaines, etc. Il comprend peu de membres comparativement aux entreprises manufacturières de même taille.

La complexité des organisations hospitalières se justifie, selon Rakotondranaivo (2006), par quatre facteurs :

– *Les missions* : la prise en charge des patients fait appel à différents services comme les services administratif et logistique, la pharmacie, le bloc opératoire, les unités de soins, etc. La multiplicité des acteurs et la mise en œuvre de leurs activités impliquent plusieurs processus qui interagissent entre eux. Ces acteurs ont des objectifs propres, souvent variés, et potentiellement contradictoires : enjeux économiques versus enjeux sanitaires, par exemple. La finalité demeure toutefois une offre de soins de qualité mettant le "bien-être" du patient au centre des préoccupations.
– *Les métiers* : L'hôpital est une organisation de professionnels qui couvre une multitude de métiers aussi variés que la prestation de soins, le laboratoire pharmaceutique, la restauration, la maintenance, l'informatique, etc. De plus, bon nombre de ces métiers sont soumis à une constante évolution (due aux avancées technologiques ou aux réformes des Etats), obligeant les professionnels à régulièrement se mettre à jour. Il s'agit donc d'organiser ce milieu de professionnels, hétérogène de par ses intervenants, pour le bon déroulement des processus de soins.
– *Le système de décision* : Le corps administratif ne gère pas l'ensemble de l'organisation de l'hôpital. Entre autres, il incombe aux chirurgiens de gérer leurs propres services, pratiquement en toute indépendance. Ils prennent généralement des décisions locales, sans vision stratégique globale de l'organisation, et peuvent privilégier les objectifs éthiques (bien-être du patient) aux objectifs économiques (réduction des coûts hospitaliers) généralement poursuivis par l'administration.
– *L'environnement* : ce dernier est particulièrement évolutif dans le contexte des soins de santé. Il est continuellement en changement. Que ce soient les contraintes économiques, les réglementations, le manque croissant de personnel médical, l'évolution des pathologies, la multiplicité et l'hétérogénéité du matériel ou encore les impératifs de performance, tout l'environnement hospitalier contribue à rendre son organisation complexe.

A la complexité structurelle, organisationnelle et environnementale des hôpitaux, s'ajoute celle due à la combinaison d'objectifs stratégiques qui peuvent paraître antagonistes. En effet, deux objectifs majeurs poursuivis par les organisations de santé sont, d'une part, de satisfaire aux exigences de bien-être et de qualité des prestations tant pour les patients que pour le personnel médical, et d'autre part de répondre au besoin impératif de rationaliser les dépenses de soins. Or, un soin de qualité se définit par une utilisation appropriée des services, en corrigeant l'éventuelle sous- ou surconsommation des ressources chirurgicales, et en minimisant les risques d'erreurs médicales (Lee, 2006). Les notions d'efficacité opérationnelle et d'efficience économique, qui constituent le moteur de la performance des entreprises manufacturières, peuvent paraître ainsi difficilement défendables dans une optique de soins.

1.3 Enjeux et défis

Malgré la pression financière grandissante, l'hôpital essaie de ne pas perdre de vue son rôle premier : assurer l'excellence des soins dispensés aux patients. Toutefois, la réalité économique est fort marquante. L'académie américaine des sciences souligne que sur 1$ dépensé dans la santé, 0.3$ à 0.4$ représentent des sur- et sous-utilisations des ressources, des utilisations inadaptées des ressources, des duplications d'activités, des défaillances du système de production, d'un manque d'information, d'inefficacité dans l'organisation. Une étude menée par Landry & Beaulieu (2002) pour des hôpitaux de divers pays (France, Pays-Bas, Québec et USA) estime que les dépenses notamment de la logistique hospitalière représentent de 30 à 40% des coûts annuels hospitaliers. Parmi ces coûts logistiques, la pharmacie représente à elle seule plus de la moitié des dépenses et présente donc un potentiel intéressant de gains. Le bloc opératoire consomme quant à lui environ 10% du budget de fonctionnement d'un hôpital.

Dans ce contexte, il apparaît indispensable de porter un regard critique sur la gestion des systèmes hospitaliers et d'envisager une réflexion globale sur leur évolution. Les centres hospitaliers doivent, en effet, faire face à l'accroissement du niveau d'exigence des patients, à l'augmentation des demandes de soins et à l'évolution des pathologies. Il est donc crucial pour ces établissements d'identifier leurs postes de coûts et les sources de gaspillage qui cachent souvent des potentialités de gains substantiels.

Ce sujet a fait l'objet de plusieurs écrits. Le thème de la gestion hospitalière intéresse, en effet, de plus en plus la communauté scientifique en management et en recherche opérationnelle. Le potentiel de développement est conséquent, d'une part parce que la priorité des hôpitaux a été, pendant longtemps, largement consacrée aux questions médicales plutôt qu'aux questions de gestion ; et d'autre part parce que l'hôpital fait émerger des problématiques intéressantes d'optimisation, où l'homme joue un rôle central en tant qu'acteur du processus de soins, ressource critique de la prestation médicale, décideur conciliant coût et qualité des soins, ou en tant que patient client du processus.

1.4 Le quartier opératoire

L'hôpital est une structure de production multi-services qui fait intervenir des éléments technologiques de pointe, le savoir-faire de plusieurs catégories professionnelles, ainsi qu'une forte implication humaine dans la prise en charge du patient et dans le processus de création de soins. Au cœur de ce processus se trouve le quartier opératoire. Son organisation et son fonctionnement sont importants à étudier. Il s'agit effectivement d'une unité majeure du dispositif médical qui conditionne de façon prépondérante la performance globale des établissements hospitaliers. C'est le maillon capital de la chaîne de production de soins, qui génère les principaux revenus et de considérables dépenses ; notamment, il consomme près de dix pourcents du budget global d'un centre hospitalier (Gordon et al., 1988; Macario et al., 1995).

Au même titre que l'obstétrique, la pédiatrie, l'anesthésie-réanimation, l'imagerie et la biologie, la chirurgie fait partie du plateau technique. Le plateau technique de chirurgie, également appelé quartier opératoire, comprend l'ensemble des moyens humains et des équipements (salles, matériel, etc.) mobilisés autour de l'activité chirurgicale ou anesthésique (Baubeau & Thomson,

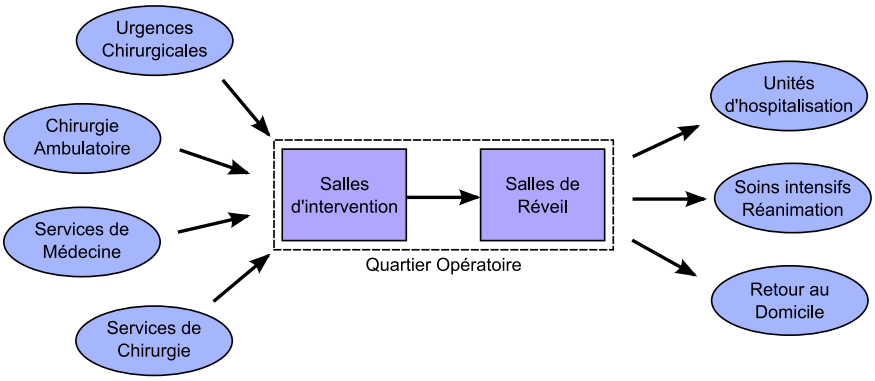

FIG. 1.1 – *Flux entrant et sortant du quartier opératoire (Marcon, 2004)*

2002). A lui seul, il constitue la ressource critique du processus opératoire, notamment en raison de la multiplicité des intervenants (chirurgiens, anesthésistes, infirmières, etc.) aux compétences et disponibilités variées.

Comme l'illustre la figure 1.1, le quartier opératoire se compose de deux structures bien distinctes : les salles d'intervention (également appelées salles d'opération ou blocs opératoires) et les salles de surveillance post-interventionnelle (SSPI), ou salles de réveil. Dans les premières se déroulent les activités chirurgicales et anesthésiques alors que le réveil des patients opérés se fait dans les secondes. Ces dernières se composent essentiellement des lits permettant aux patients de récupérer lors de la phase de réveil postopératoire. Les salles d'opération sont équipées en permanence des outils de contrôle de l'état des patients. Le matériel chirurgical supplémentaire est à préciser par le chirurgien et l'anesthésiste en fonction de la nature de l'intervention chirurgicale à réaliser. Les blocs opératoires peuvent être de deux types : dédiés ou polyvalents. Des salles d'intervention dédiées peuvent l'être pour des raisons logistiques (salles dédiées à certains types de chirurgie en fonction de l'équipement) ou pour des raisons organisationnelles (salles dédiées aux urgences ou à la chirurgie réglée par exemple). Les blocs opératoires polyvalents sont, eux, supposés pouvoir accueillir tout type d'interventions.

La prise en charge du patient lors d'une opération chirurgicale se décompose en trois phases :

1. *La phase préopératoire* débute par les consultations chirurgicale et anesthésique. Suite aux consultations, l'intervention est incorporée au programme opératoire et se voit attribuer une date provisoire de réalisation. Cette phase se termine par la préparation du patient (prémédication, etc.) et par son acheminement au bloc opératoire.
2. *La phase per-opératoire* s'étale entre l'arrivée du patient au bloc opératoire et sa sortie. Le patient y subit d'abord une étape d'induction avant de subir l'opération proprement dite. L'intervention terminée, le patient quitte le bloc opératoire pour immédiatement rejoindre la salle de réveil dans laquelle il demeurera tant que l'anesthésiste en charge ne l'autorise pas à rejoindre son service d'hospitalisation.
3. *La phase postopératoire* dépend de l'état du patient après intervention. Si son état est jugé satisfaisant, le patient est alors transféré vers son lit d'hospitalisation, où lui seront administrés les soins post-interventionnels. Par contre, si le patient présente un état critique suite à l'opération, il est directement admis en soins intensifs ou en réanimation.

Les patients transitant par le quartier opératoire sont issus de différents flux au sein de l'hôpital (voir figure 1.1). A ces flux correspondent trois modes de prise en charge des patients : la chirurgie ambulatoire, la chirurgie réglée et la chirurgie urgente. Ces trois modes de prise en charge nécessitent des ressources différentes et conditionnent le processus de programmation du quartier opératoire. Ces trois modes sont (Trilling, 2006) :

- *La chirurgie ambulatoire*, qui concerne les cas pour lesquels le patient arrive et repart de l'hôpital le jour même. Le nombre de patients traités par jour est généralement plus élevé que pour les chirurgies avec hospitalisation, avec comme inconvénient l'impossibilité de reporter l'intervention au jour suivant.
- *La chirurgie réglée en hospitalisation complète*, qui impose l'hospitalisation du patient, et monopolise donc un lit en unité de soins. Ce type d'intervention est inclus dans le programme opératoire et considéré lors de la planification des ressources. Le chirurgien est assuré ainsi de disposer des ressources humaines et matérielles nécessaires à son intervention .
- *La chirurgie urgente*, qui traite les patients requérant des soins immédiats et non prévus dans le programme opératoire prévisionnel. Les urgences émanent aussi bien de l'extérieur que de l'intérieur du centre hospitalier. Les patients internes peuvent provenir de réanimation mais également des unités de soins si l'état du patient se dégrade de façon inattendue. Les patients se présentant spontanément ou acheminés par ambulances représentent, eux, les urgences externes. Les urgences sont réparties en classes, suivant la pathologie et la criticité de l'état du patient. Ces quatre classes sont (Dexter & O'Neill, 2001) :

  1. Les urgences vitales qui doivent être opérées sur le champ et nécessitent la révision du programme opératoire.
  2. Les cas urgents qui sont à traiter le jour même de leur venue. Ils peuvent être intégrés au programme opératoire établi ou placés en fin de journée afin de ne pas le perturber ; ils génèrent le plus souvent des heures supplémentaires.
  3. Les cas pouvant supporter une journée d'attente avant de se voir opérer. Ils sont incorporés soit au programme opératoire du jour, soit au programme du lendemain, et peuvent causer le report d'interventions déjà programmées.
  4. Les cas pouvant attendre plusieurs jours avant leur traitement, ce qui permet leur insertion dans le programme opératoire sans le perturber. Ces cas ne sont plus considérés comme urgents mais comme programmés.

Selon la configuration du quartier opératoire, les chirurgies ambulatoire, réglée et urgente se réalisent dans des salles d'opération qui leur sont spécifiquement dédiées, auquel cas chacune de ces chirurgies peut se traiter dans des sites différents, de façon totalement autonome. Alternativement, elles s'effectuent dans des salles d'opérations pluridisciplinaires, ou polyvalentes, regroupées sur un unique site et partageant un ensemble de ressources (matérielles et/ou humaines). Ces différents modes de prise en charge induisent des problématiques de planification et d'ordonnancement des blocs opératoires avec un besoin de prise en considération des inévitables perturbations, émanant de la chirurgie urgente notamment. Les gestionnaires des blocs sont donc à la recherche de programmes opératoires robustes, c'est-à-dire capables d'absorber, sans trop de remise en cause, les différents aléas.

La programmation opératoire consiste à agencer, sur un horizon de temps bien défini, les interventions de chirurgie réglée qui font l'objet d'une demande d'inscription au programme opératoire. La décision d'inscrire un patient au programme opératoire émane des consultations chirurgicales et anesthésiques. Le programme opératoire veille à ce que les ressources (humaines et matérielles) adéquates soient disponibles au moment et à l'endroit voulus. Le programme prévisionnel établi est ensuite validé chaque (fin de) semaine pour la semaine à venir. Le patient est alors averti d'une date provisoire d'intervention, en accord conjointement avec le chirurgien et le quartier opératoire, qui peut ou non être modifiée par après. L'inscription au programme opératoire peut se faire selon différentes stratégies qui seront développées par la suite (Dexter & O'Neill, 2001) : une programmation sans dépassement horaire (*fixed hours*), une programmation permettant le choix de la date d'intervention par le chirurgien et le patient (*any workday*), ou une programmation offrant un compromis entre les deux (*reasonable time*). La programmation opératoire offre également plusieurs modes d'affectation (Kharraja, 2003) selon qu'elle s'effectue par salle (*open scheduling*), par allocation de plages horaires (*block scheduling*) ou par allocation de plages horaires avec processus d'ajustement (*modified block scheduling*).

La programmation opératoire dresse donc la liste des patients et leur ordre de passage dans chaque salle d'intervention, et éventuellement dans chaque salle de réveil. Cette planification des blocs opératoires est généralement scindée en deux sous-problèmes séquentiels : l'affectation hebdomadaire des actes chirurgicaux aux différents blocs et l'ordonnancement journalier des interventions au sein d'un bloc. La programmation opératoire doit satisfaire un nombre important de contraintes (Hammami, 2006). La construction du programme opératoire doit, entre autres, tenir compte du nombre de salles d'opération disponibles, des équipements chirurgicaux spécifiques, de la disponibilité des ressources humaines, des horaires d'ouverture des salles d'opération (heures normales ou heures supplémentaires), de l'éventuelle accessibilité limitée de ces dernières due à des niveaux d'équipement différents. Les ressources indispensables aux diverses interventions peuvent être génériques, c'est-à-dire utilisées indépendamment de la pathologie du patient, ou spécifiques, c'est-à-dire affectées en fonction du nom du patient ou du type d'opération requis. Si le médecin anesthésiste et le lit de réveil sont des ressources flexibles qui peuvent intervenir dans tout type de chirurgie, d'autres ressources sont, elles, spécifiques. Un acte chirurgical implique, en effet, un ou plusieurs chirurgiens particuliers et peut exiger du matériel spécial telle une salle bien précise, une prothèse ou autres, sous peine de report de l'intervention. Les ressources critiques dans le processus opératoire semblent toutefois être les ressources humaines, en raison de la multiplicité des compétences nécessaires à l'achèvement des interventions chirurgicales.

Enfin, le programme opératoire établi sera sujet à perturbations, quels que soient le mode et la stratégie de programmation adoptés. En effet, le quartier opératoire évolue dans un environnement incertain, stochastique. Cette incertitude a diverses sources. Notamment, certaines opérations inopinées devront être insérées au programme opératoire, perturbant de la sorte le planning prévisionnel. L'incertitude provient également de la durée des différentes activités qui jalonnent le parcours du patient, cette durée dépendant de l'état de santé du patient, de son âge, du type d'intervention, etc. L'ensemble des perturbations auxquelles le programme opératoire sera soumis imposera un réajustement de ce dernier. Le planning opératoire se doit donc d'être aussi robuste que possible ; en d'autres mots il doit être capable d'absorber des changements de l'environnement sans en être trop grandement affecté.

1.5 Les ressources opératoires

La gestion des ressources opératoires est pointue et dépend du niveau de polyvalence et de la mutualisation de celles-ci. La polyvalence peut se définir comme l'aptitude à effectuer diverses tâches, la capacité à s'adapter. Par exemple, des salles polyvalentes sont équipées de sorte à pouvoir accueillir différents types de chirurgie. Quant à la mutualisation, elle s'explique comme la mise en commun, le partage de moyens afin d'effectuer des activités communes ou indépendantes. Par exemple, les blocs opératoires ou certains types d'infirmières sont des ressources mutualisées au sein du quartier opératoire. La polyvalence et la mutualisation des ressources matérielles sont indispensables dans les quartiers opératoires mono-site pluridisciplinaires, et l'organisation de telles ressources l'est tout autant. Quant aux ressources humaines, elles peuvent être mutualisées, et donc polyvalentes, même pour des infrastructures multi-site aux ressources matérielles dédiées. C'est ce contexte polyvalent et mutualisé qui rend la gestion des ressources chirurgicales si importante et complexe.

La planification des interventions, qui s'opère classiquement en deux phases successives, est un élément déterminant du bon fonctionnement du quartier opératoire. Ce processus génère un planning opératoire qui affecte les différentes opérations aux différentes salles, et détermine donc l'utilisation de ces dernières. Dans la mesure du possible, cette procédure doit aboutir à un programme opératoire robuste permettant de minimiser l'impact des aléas sur le bon déroulement des interventions programmées.

La gestion efficiente du personnel médical attitré au quartier opératoire est un défi majeur pour les hôpitaux, les coûts impliquant l'entièreté des ressources humaines du centre hospitalier, tous secteurs confondus, pouvant représenter 70% de son budget global (Trilling, 2006). Et contrairement à la planification des interventions, la planification du personnel dans les blocs opératoires est peu étudiée par la communauté scientifique. Techniquement, le planning opératoire doit satisfaire à un certain nombre de contraintes législatives ainsi qu'aux diverses préférences du personnel. De la qualité des plannings établis dépendra la satisfaction du personnel, elle-même répercutée sur la qualité des soins prodigués. On distingue deux types de planification d'horaires de travail en fonction du mode de programmation opératoire : les plannings cycliques ou non cycliques. Les plannings cycliques sont d'application pour une période de temps bien déterminée. Il s'agit de reproduire les affectations établies pour chaque horizon de temps (c.-à-d. l'horizon de planification, généralement une semaine) composant la période d'application du planning. Ces plannings se caractérisent par une simplicité de conception mais présente le désavantage de difficilement tenir compte des préférences des personnes impliquées. Les plannings non cycliques sont uniques et refaçonnés à chaque horizon de temps. Ils ne proposent pas de vision à long terme des horaires, mais ils sont plus flexibles et intègrent plus facilement les préférences personnelles que leurs pendants cycliques. Le temps nécessaire à la construction de tels programmes non-cycliques est important, mais leur réalisation s'en verra moins souvent perturbée car ils tiennent compte du facteur humain sous-jacent.

Les ressources du quartier opératoire considérées comme les plus critiques restent les chirurgiens et les salles d'opération. La plupart des approches planifiant les activités du quartier opératoire sont focalisées sur leurs disponibilités. La programmation opératoire fixe non seulement l'utilisation des blocs mais conditionne aussi la programmation des autres ressources chirurgicales. Sur base du planning opératoire sont ensuite construits les emplois du temps des autres

ressources, humaines ou matérielles : anesthésistes, infirmiers, matériels spécifiques, etc. Les chirurgiens sont les principaux acteurs du processus opératoire, après le patient lui-même. C'est sur base de leur emploi du temps que l'activité du quartier opératoire est régulée, Trilling (2006) les comparant même à des "clients". Les ressources du quartier opératoire sont mises à leur disposition, salles comme personnel. La planification des chirurgiens est donc un problème directement lié à la programmation des interventions chirurgicales, ce processus étant contraint par leurs disponibilités.

La description du contexte hospitalier entreprise dans ce chapitre nous confirme la difficulté de la planification des blocs opératoires due, notamment, à l'hétérogénéité et la pluridisciplinarité des ressources impliquées ainsi qu'aux pressions économiques quelque peu contradictoires avec la vocation première de l'hôpital. Il en découle une nécessité, pour les centres hospitaliers généralement, de gérer différemment leurs unités de soins. Rendre plus efficace la gestion d'un quartier opératoire passe par une utilisation plus rationnelle de ses ressources tant humaines que matérielles. Comme nous l'avons souligné, la perception de la qualité des soins par le patient, le taux d'utilisation des ressources chirurgicales et le degré de satisfaction du personnel sont autant d'indicateurs de performance directement conditionnés par le processus de programmation opératoire. Par conséquent, il semble pertinent de s'atteler à améliorer la planification des salles d'opération, afin de construire des plannings plus efficients. Cependant, les seules préoccupations financières ne peuvent suffire à l'établissement de ces programmes opératoires. Nous ne pouvons ignorer l'importance de la place des femmes et des hommes tant dans la réalisation des processus médicaux que dans la planification de leur exécution. Ceci est d'autant plus vrai que l'organisation hospitalière est complexe : organisation complexe de professionnels partageant des ressources critiques, intégrant le patient au processus opératoire et poursuivant des objectifs non nécessairement lucratifs. La prise en compte de la dimension humaine dans les objectifs d'optimisation du fonctionnement du quartier opératoire et de sa programmation est la clef de voûte de cette recherche. Celle-ci propose des outils d'aide à la décision pour une meilleure utilisation des ressources chirurgicales en tenant compte de l'opinion des acteurs de ce processus médical. Ces outils visent à améliorer la satisfaction du staff médical et par conséquent l'efficacité du quartier opératoire. En particulier, l'approche prônée ici consiste en l'intégration des préférences des chirurgiens lors de l'élaboration des programmes opératoires hebdomadaires. Cette approche se base sur des modèles mathématiques pouvant être résolus de façon exacte ou heuristique et permettent la construction de programmes opératoires optimisés. Ceux-ci tiennent compte, entre autres, des disponibilités du personnel et des desiderata des acteurs fondamentaux du processus : les chirurgiens.

Après avoir exposé le contexte dans lequel s'inscrit notre recherche, nous allons à présent passer en revue les travaux de la communauté scientifique ayant proposé des solutions pour améliorer le processus de programmation opératoire. Le chapitre suivant dresse un rapide état de l'art du domaine et recense les principales approches traitant de la programmation opératoire.

Chapitre 2

Etat de l'art

L'hôpital évolue dans un environnement complexe et continuellement en changement. Le chapitre précédent décrit les différents défis auxquels sont confrontés les gestionnaires d'hôpitaux d'aujourd'hui. Nous y avons souligné l'importance du processus de programmation opératoire sur les performances du quartier opératoire, en termes d'utilisation des ressources ou de travail en heures supplémentaires. Nous avons également insisté sur l'importance des ressources humaines, dont la bonne gestion se répercute à la fois économiquement et sur la satisfaction des acteurs concernés, ce qui rejaillit immanquablement sur la qualité des soins prodigués.

La communauté de recherche en gestion hospitalière présente un intérêt sans cesse croissant pour le domaine de la programmation des salles d'opération. Cet état de l'art s'intéresse principalement aux avancées récentes en la matière et se limite aux aspects de planification des ressources et d'ordonnancement des activités chirurgicales. Il s'inspire en partie de l'état de l'art détaillé en la matière proposé par Cardoen et al. (2010). Ne s'y retrouvent pas les problématiques de dimensionnement de ressources ou d'estimation de durées opératoires, par exemple. Cet état de l'art est organisé comme suit. Nous commençons par définir la typologie associée au processus étudié et reprenons en revue les différents modes de programmation opératoire ainsi que les travaux qui s'y rapportent. Ensuite, nous exposons les différentes contraintes les plus rencontrées dans la littérature sur la planification des blocs opératoires. Nous terminons cet état de l'art en évoquant les différents objectifs poursuivis lors de la construction de plannings opératoires, et les différents outils de résolution les plus fréquemment proposés par la littérature pour obtenir de tels plannings.

2.1 Typologie du processus de programmation opératoire

La programmation opératoire, c'est-à-dire la planification des interventions dans le quartier opératoire, est un élément déterminant du bon fonctionnement de celui-ci. Elle comporte différents modes et stratégies qui régissent l'organisation du quartier opératoire. Le choix du mode de programmation et de la stratégie associée engendre des décisions à prendre à différents niveaux, et influence le coût et la qualité de service porté aux patients. Nous dénombrons trois modes de programmation des interventions (Kharraja, 2003).

Open Scheduling. Egalement appelé programmation par salle, ce mode propose un planning vierge, sans contraintes de pré-allocation des salles ou des plages horaires, où les praticiens de toute spécialité chirurgicale peuvent déclarer leurs opérations et les placer de différentes manières (Trilling, 2006) :

- chronologiquement, sur base de la règle premier arrivé, premier servi (FCFS). Cette manière de procéder est simple à mettre en place mais génère une sous-utilisation des ressources et d'importants dépassements horaires.
- de façon concertée, suivant un processus de négociation entre les différents acteurs (chirurgiens, anesthésistes et infirmiers). Cette organisation est souple et adaptable, mais nécessite un temps conséquent à l'élaboration d'un planning acceptable par tous, surtout si le quartier opératoire comporte de nombreuses salles.

Ce mode de programmation peut poser des problèmes de répartition de la charge de travail qui peut s'avérer déséquilibrée sur la semaine, et dès lors causer un manque de lits (de réveil ou d'hospitalisation) pour certaines périodes. La tâche d'ordonnancement journalier des interventions au sein des blocs opératoires se trouve également plus complexe pour le gestionnaire de bloc.

Block Scheduling. Il s'agit d'une programmation par allocation de plages horaires. Ce mode consiste à imposer un squelette (appelé Plan Directeur d'Allocation) du programme opératoire pour la semaine, en pré-allouant les plages horaires (généralement des demi-journées) aux chirurgiens ou aux spécialités. Ensuite, sur base du plan directeur d'allocation, les chirurgiens remplissent à leur guise les plages qui leur sont destinées. Le plan de répartition des plages opératoires peut varier d'une semaine à l'autre mais être cyclique sur une plus longue période comme un trimestre. Au bout de la période d'application du plan, ce dernier peut être amené à subir des modifications, pour tenir compte d'éventuels changements dans les disponibilités du personnel ou pour des raisons de performance. La difficulté majeure de ce mode de programmation réside dans la conception d'un bon plan directeur d'allocation en fonction de l'activité des chirurgiens. Il est également plus délicat de faire respecter les plages allouées aux chirurgiens puisque c'est à eux qu'incombe la tâche de remplir (ni trop, ni trop peu) les heures qui leur sont octroyées. Le gestionnaire de bloc, quant à lui, voit son travail simplifié puisque seule lui revient la vérification de la faisabilité du programme ainsi établi. Pour ce type d'organisation, les performances du quartier opératoire sont donc directement liées à la capacité des chirurgiens à respecter les plages horaires attribuées.

Modified Block Scheduling. Ce dernier mode de programmation est en quelque sorte un compromis entre les deux premiers. Il applique les bases du Block Scheduling en allouant des plages horaires aux chirurgiens, mais offre également la possibilité de réattribuer le temps opératoire non utilisé par d'autres chirurgiens. Il offre plus de flexibilité en permettant de réaffecter ou de libérer des heures allouées au préalable. Ce type de programmation nécessite que le remplissage des plages horaires soit connu suffisamment tôt pour pouvoir redistribuer les heures encore disponibles.

Outre le mode, la programmation opératoire comprend différentes stratégies proposées par Dexter & O'Neill (2001) :

- *Fixed Hours* : cette stratégie n'autorise pas le travail en heures supplémentaires. Les horaires d'ouverture du quartier opératoire sont strictement respectés et aucun dépassement n'est

permis. Elle se focalise sur l'utilisation des ressources, avec l'inconvénient, pour le patient, de reporter des opérations en cas de surcharge du planning.
– *Any Workday* : cette stratégie permet au chirurgien et à son patient de choisir le jour d'intervention selon leurs desiderata. La satisfaction du chirurgien et du patient en est l'objectif premier. Elle nécessite une certaine flexibilité des acteurs pour pallier les différences de charge de travail sur la semaine.
– *Reasonable Time* : compromis entre les deux précédentes, cette stratégie propose de fixer la date d'intervention en tenant compte de la proposition du chirurgien mais sans provoquer de dépassement horaire. Dans la mesure du possible, la date d'opération est arrêtée afin d'être la plus proche de celle suggérée par le chirurgien et son patient.

La conception d'un programme opératoire dépendra donc du mode de programmation appliqué. Les travaux traitant la planification des interventions au sein du quartier opératoire abordent ce problème différemment selon le mode de programmation pratiqué. Les décisions à prendre de même que leur impact sur les performances du quartier opératoire varient selon les approches.

2.1.1 Programmation par Block Scheduling et Modified Block Scheduling

Pour les modes Block Scheduling et Modified Block Scheduling, la littérature décompose généralement le processus de programmation opératoire en trois étapes (Beliën & Demeulemeester, 2007; Testi et al., 2007). La première étape se situe au niveau stratégique et consiste à déterminer combien de plages horaires sont attribuées à chaque chirurgien ou à chacune des spécialités chirurgicales qui se partagent le quartier opératoire (on parle généralement d'un problème dit de *case mix planning*, Blake & Carter, 2002). La deuxième phase est, elle, plus située sur un plan tactique. Elle comprend la construction d'un plan directeur d'allocation, c.-à-d. un calendrier cyclique qui définit quelle salle d'opération est assignée à chacune des plages horaires, et donc à chaque chirurgien ou type de chirurgie (Kharraja et al., 2006; van Oostrum et al., 2008, par exemple). La dernière étape se situe au niveau opérationnel et concerne l'ordonnancement journalier des interventions chirurgicales au sein de chaque salle.

La principale difficulté de ce type de programmation opératoire réside dans la construction d'un bon plan directeur d'allocation (PDA, appelé *master surgical schedule* en anglais). De l'attribution et du dimensionnement des plages horaires dépendront les performances du quartier opératoire en termes d'utilisation des salles et de recourt aux heures supplémentaires. Le PDA est initialement défini sur base de l'activité réelle des chirurgiens et nécessite de bonnes estimations des durées d'opération. La plupart du temps, il concerne la programmation d'une semaine d'ouverture du quartier opératoire, mais peut très bien englober quinze jours. La durée d'application du PDA varie selon les établissements mais est généralement de plusieurs mois (sur base semestrielle ou annuelle le plus souvent). A la fin de cette période, une révision permet d'ajuster le PDA aux évolutions de l'environnement ou sur base des performances atteintes.

L'élaboration du PDA représente la phase de planification des interventions. La dernière étape du processus concerne l'ordonnancement des cas journaliers. Elle consiste à remplir les plages horaires avec les différentes interventions de la semaine, en coordination avec la planification des lits. Cette dernière étape n'est pas cruciale car les opérations sont placées dans des plages horaires auxquelles correspondent déjà un chirurgien et un bloc opératoire. Outre le chirurgien, les autres ressources humaines nécessaires sont, elles aussi, affectées en amont de la phase d'ordon-

nancement. Seul l'ordre dans lequel les interventions sont agencées au sein d'une plage horaire peut avoir un impact sur l'utilisation des ressources mutualisées, les lits de réveil notamment. En raison du caractère secondaire de l'étape d'ordonnancement, la plupart des travaux traitant la programmation par allocation de plages horaires se concentrent sur l'élaboration du plan directeur d'allocation.

Dans leurs travaux respectifs, Blake et al. (2002) et Blake & Donald (2002) proposent une formulation mathématique en nombres entiers et une heuristique pour distribuer le temps total de disponibilité des salles aux différentes spécialités chirurgicales. Ils cherchent à répartir équitablement les plages horaires entre celles-ci, en considérant différents types de blocs opératoires. Ils construisent un plan directeur d'allocation qui minimise la différence entre le temps effectivement octroyé à chaque spécialité et le temps auquel ces dernières peuvent prétendre. Dans la continuité, Kharraja et al. (2006) présentent un programme linéaire en nombres entiers pour construire un PDA minimisant la différence entre la capacité globale du quartier opératoire et la demande émanant des différentes spécialités, tout en tenant compte du type d'intervention et des disponibilités des chirurgiens. Ils considèrent deux approches : une approche centrée sur les chirurgiens individuels et l'autre sur les groupes chirurgicaux. Ils développent ensuite une heuristique afin de répartir les interventions non programmées sur les plages horaires non utilisées.

Hammami (2006) s'intéresse plus particulièrement au mode de programmation Modified Block Scheduling. Elle propose, par l'intermédiaire de modèles linéaires en nombres entiers, une approche en trois phases : la construction du PDA, l'affectation des interventions et l'ajustement du planning en fonction des aléas. La première phase consiste à répartir les chirurgiens en groupes équilibrés et à construire le PDA en minimisant le nombre de plages horaires nécessaires, celles-ci pouvant être communes ou nominatives. La dernière étape est de type réactif : une urgence est introduite dans le programme opératoire, si possible dans les plages horaires du groupe de chirurgiens correspondant sinon dans celles d'un groupe différent ; si besoin est, des plages horaires sont redéfinies.

Beliën et al. (2006) ont élaboré un outil d'aide à la décision permettant de visualiser l'impact d'un PDA sur l'utilisation des ressources. Leur outil permet au gestionnaire de blocs de visualiser directement l'évolution de l'utilisation des ressources au fur et à mesure des allocations de plages horaires. Les décisions se prennent ici au niveau des chirurgiens et non des spécialités chirurgicales.

Testi et al. (2007) font état d'une approche hiérarchique en trois phases pour construire des programmes opératoires. Elle regroupe les différents niveaux de décision. La première phase consiste à répartir le temps total opératoire disponible parmi les différentes disciplines, sur base d'un niveau de priorité qui tient compte de la liste d'attente de chaque discipline et du nombre de ses demandes insatisfaites. La deuxième phase concerne le développement d'un PDA qui détermine quelle salle d'opération est attribuée à quelle plage horaire, en fonction des préférences des chirurgiens quant à la durée de séjour des patients. Ces deux étapes sont résolues chacune par programmation en nombres entiers. La troisième phase consiste à déterminer la plage horaire durant laquelle chaque cas électif sera traité ainsi que la séquence des cas au sein d'une même plage horaire. Cette dernière étape est évaluée à l'aide d'un modèle de simulation.

D'autres travaux traitent de la construction du plan directeur d'allocation pour des stratégies de *block scheduling* (Chaabane et al., 2006; Rohleder et al., 2005; Santibanez et al., 2007) et déterminent un calendrier cyclique définissant le nombre et le type de salles d'opérations disponibles,

les heures durant lesquelles celles-ci sont accessibles et les chirurgiens ou spécialités assignés à chaque plage horaire. Ces approches proposent des décisions essentiellement au niveau du chirurgien ou d'une spécialité.

2.1.2 Programmation par Open Scheduling

C'est le mode de programmation qui nous intéresse. Classiquement, les travaux traitant ce mode de programmation opératoire approchent ce problème en deux temps (Magerlein & Martin, 1978). Une première étape de planification permet d'assigner la date et la salle de chaque intervention. Pour éviter toute confusion, nous nommerons celle-ci *étape d'affectation*. Vient alors l'étape d'ordonnancement à proprement dite qui définit l'ordre de réalisation de chaque intervention au sein d'une même salle. Cette étape sera dénommée *étape d'agencement*. Après avoir défini les plages d'ouverture des salles d'opération, les interventions y sont affectées de manière à maximiser l'utilisation des ressources tout en évitant les dépassements horaires et les reports d'intervention. Ensuite, ces mêmes interventions sont agencées au sein de chaque bloc en tenant compte de la disponibilité de la totalité des ressources (salles, lits de réveil, brancardiers) et en considérant un temps de processus composé d'une durée opératoire et d'une durée de réveil.

Guinet & Chaabane (2003) développent une procédure de résolution basée sur cette décomposition (bien qu'ils se concentrent essentiellement sur la phase d'affectation des opérations). Au moyen d'une heuristique primale-duale, ils minimisent des coûts d'hospitalisation reprenant la durée de séjour des patients et les heures supplémentaires en développant une extension de la méthode hongroise (*Hungarian method*). Dans Jebali et al. (2006), la phase d'affectation s'effectue en minimisant des coûts liés au temps d'attente des patients, aux heures supplémentaires et à la sous-utilisation des salles, tandis que la phase d'agencement minimise uniquement les heures supplémentaires. Ces deux étapes sont formulées à l'aide de programmations en nombres entiers mixtes. Ces travaux incluent tous deux les lits de réveil aux ressources du quartier opératoire.

Dans leur thèse de doctorat respective, Kharraja (2003) aborde le quartier opératoire et les salles de réveil comme k problèmes *flowshop* (chacune des salles de réveil contenant exactement k lits), alors que Chaabane (2004) traite l'ordonnancement des salles d'opération comme un problème *flowshop* hybride à trois étages avec contraintes de précédence et sans temps d'attente. Pham & Klinkert (2008), quant à eux, proposent une extension du problème *job shop*, appelée *multi-mode blocking job shop*, où un mode correspond à un ensemble de ressources. Les ressources utilisées par une activité chirurgicale restent indisponibles tant que celle-ci ne peut accéder à l'étape suivante de son traitement.

Outre les diverses formulations mathématiques, d'autres types d'approches sont également utilisées pour résoudre les deux phases d'affectation et d'agencement des opérations. Fei (2006), par exemple, établit dans ses travaux une procédure de résolution par génération de colonnes, de même qu'une approche méta-heuristique du problème. Elle y couple un algorithme génétique à de la recherche Tabou. Krempels & Panchenko (2006) font également usage d'heuristiques pour résoudre les quatre sous-problèmes identifiés : affectation du personnel aux plages horaires, création des équipes de travail, affectation des opérations aux différentes équipes et affectation des salles aux diverses opérations. Finalement, Sciomachen et al. (2005) développent des modèles de simulation à événements discrets pour l'évaluation de performances tels le taux d'utilisation des salles, le nombre moyen de chirurgies effectuées ou le nombre moyen de cas en dépassement horaire.

Si les approches citées précédemment offrent deux niveaux de décision, permettant de déterminer date, heure et salle d'opération, elles représentent une minorité dans la littérature. En effet, la majorité des travaux scientifiques constituant le domaine se concentrent uniquement sur l'une des deux phases décrites ci-dessus, ne proposant alors qu'un seul niveau de décision.

Etape d'affectation des interventions

La plupart des travaux abordant la phase d'affectation des opérations se placent au niveau du patient (Bowers & Mould, 2005; Guinet & Chaabane, 2003; Hans et al., 2008; Ogulata & Erol, 2003; van Oostrum et al., 2008).

Ce problème d'affectation des patients aux salles d'opérations est modélisé par Marcon et al. (2003) comme un problème de sac à dos multiple. Dans leur étude, ils comparent deux fonctions de coûts qui minimisent, pour la première, la différence de charge horaire entre les salles et, pour la seconde, le risque de non réalisation des opérations chirurgicales établies. Ils supposent la date d'opération connue. Comme le déroulement du programme opératoire sera soumis à des événements aléatoires, ils introduisent une phase dynamique durant laquelle le programme pourra, si nécessaire, être ré-agencé via simulation.

Dexter & Traub (2002) considèrent un service de chirurgie où ce sont les chirurgiens et leurs patients qui conviennent d'un jour d'opération. Ils proposent deux heuristiques afin d'insérer les patients nouvellement arrivés dans un planning pré-existant. L'une place le nouveau cas électif dans la salle d'opération permettant de débuter l'opération au plus tôt. Dans l'autre, le cas électif est ajouté dans la salle proposant le temps de début d'opération le plus tard, mais suffisamment tôt pour que l'opération se déroule probablement durant les heures normales d'ouverture. Ils comparent alors ces deux heuristiques en fonction de l'efficacité en termes de sous- et sur-utilisation des salles d'opérations.

Dans Fei et al. (2008), un ensemble de cas électifs est assigné aux différentes salles d'opérations afin de minimiser les coûts opératoires (également exprimés en termes de sous- et sur-utilisation des salles) en tenant compte de contraintes de capacité des salles et de date limite d'intervention. Dans leur papier, ils développent un algorithme arborescent de type *branch-and-price*, combinant génération de colonnes et *branch-and-bound*.

Etape d'agencement des blocs opératoires

D'autres recherches se concentrent essentiellement sur la phase d'agencement des interventions au sein des blocs opératoires pour une unique journée d'ouverture. C'est le cas notamment de Denton et al. (2007); Dexter (2000); Dexter et al. (2000); Harper (2002); Lebowitz (2003); Marcon & Dexter (2007). Les règles ordonnant les opérations dans les différents blocs opératoires ont un impact important non seulement sur les performances de ceux-ci, mais également sur le fonctionnement des salles de réveil, ou salles de surveillance post-interventionnelle (SSPI).

Ce point est mis en avant par Marcon & Dexter (2006), entre autres. Leur étude vise à évaluer comment les règles d'ordonnancement affectent le nombre de patients présents dans les salles de réveil et, par conséquent, le personnel à y pourvoir. A l'aide d'un modèle de simulation à événements discrets, ils analysent sept règles d'ordonnancement de cas électifs dans les salles

d'opérations. Ces différentes règles sont classées en fonction des performances qu'elles offrent tant au sein des blocs opératoires que des salles de réveil. Les performances y sont mesurées en termes d'heures supplémentaires dans les blocs opératoires, d'heures de personnel nécessaires en salles de réveil et de retard d'admission en salles de réveil.

Dans le même registre, Hsu et al. (2003) décrivent une approche déterministe pour ordonnancer les patients d'un centre chirurgical ambulatoire. Ils déterminent la séquence des patients qui minimise le nombre d'infirmières nécessaires en salles de réveil. Le problème est formulé comme un problème d'ordonnancement de type *process shop* à deux étages. Un algorithme de type recherche tabou est également développé pour une résolution heuristique du problème.

Avec une formulation mathématique en nombres entiers mixtes (*Mixed Integer Programming*, MIP) de même qu'une procédure arborescente de type *branch-and-bound*, Cardoen et al. (2006a,b) proposent une autre approche déterministe pour déterminer la séquence des opérations d'une journée de planning. Leur modèle multi-objectif tente notamment de planifier les actes chirurgicaux impliquant des enfants ou des patients prioritaires aussi tôt que possible le jour de l'opération. Les patients se trouvant à une distance conséquente de l'établissement hospitalier sont, eux, programmés après une certaine heure.

2.2 Contraintes fonctionnelles

Dans cette section, nous recensons les travaux de la littérature selon les types de contraintes dont ils tiennent compte. Nous distinguons des contraintes liées à la disponibilité des ressources (matérielles et humaines), liées à l'incertitude inhérente au problème et liées aux interventions.

2.2.1 Contraintes de ressources

Les ressources nécessaires au bon fonctionnement du processus opératoire sont de deux types : les ressources matérielles et les ressources humaines. L'utilisation de telles ressources constitue une contrainte très restrictive et difficile à satisfaire dans beaucoup d'approches. En effet, ces ressources chirurgicales sont coûteuses et existent en quantité limitée, et par conséquent influencent grandement l'ensemble de solutions réalisables.

Ressources matérielles

Les contraintes de ressources matérielles les plus répandues sont celles liées à la disponibilité des blocs opératoires. Une grande partie de la littérature considère au moins les heures normales d'ouverture du quartier opératoire. Assez étonnamment, nous recensons peu de travaux qui restreignent le nombre d'heures supplémentaires disponibles dans les salles d'opération (Guinet & Chaabane, 2003; Fei et al., 2006; Pham & Klinkert, 2008), d'autres approches se contentant simplement de les minimiser.

Les autres ressources matérielles régulièrement citées sont les services d'hospitalisation, les soins intensifs et les lits de réveil. Concernant les services d'hospitalisation et de soins intensifs, van Oostrum et al. (2008) proposent une approche par génération de colonnes afin de construire

un plan directeur d'allocation qui maximise l'utilisation des salles tout en nivelant les besoins en lits d'hospitalisation et de soins intensifs. Les plages horaires sont dédiées aux différentes spécialités chirurgicales en fonction de la fréquence de la demande émanant des patients. Des plages vides sont réservées afin de garder la probabilité d'avoir recours aux heures supplémentaires sous un seuil acceptable par le gestionnaire. Remarquant que tous les chirurgiens n'ont pas la même productivité, Dexter & Ledolter (2003) présentent un outil d'aide à la décision afin d'améliorer les retours financiers en déterminant les chirurgiens auxquels il est préférable d'allouer plus de temps d'opération. Ils appliquent la programmation linéaire pour estimer la contribution financière marginale de chaque chirurgien lorsqu'on lui attribue une unité de temps opératoire supplémentaire. La programmation linéaire est ici soumise à des contraintes de disponibilité des lits d'hospitalisation et de soins intensifs. Ces estimations étant sujettes à incertitude, ils utilisent un programme quadratique pour déterminer, sur base du portefeuille de chirurgiens, quelles sont les contributions marginales sensibles aux aléas. Santibanez et al. (2007) étudient, à l'aide d'un modèle mathématique en nombres entiers, les conséquences d'un changement simultané des plans directeurs d'allocation de plusieurs hôpitaux sur le nombre de patients traités ou sur l'utilisation des ressources postopératoires (lits d'hospitalisation ou de soins intensifs). Plus spécifiquement, Calichman (2005) part du constat que les lits en services de soins sont, la plupart du temps, sous-utilisés. Afin d'y remédier, il propose une procédure d'optimisation de la construction du PDA. Il relie la programmation opératoire à l'occupation des lits d'hospitalisation afin d'utiliser au mieux, et sur la totalité des jours de la semaine, ces lits. Bien qu'ils fassent partie intégrante du quartier opératoire, très peu de papiers prennent en compte les lits de réveil (Cardoen et al., 2006a; Jebali et al., 2006; Augusto et al., 2008).

Enfin, rares sont les travaux qui intègrent les disponibilités d'autres équipements chirurgicaux tels certains matériels stériles ou autres médicaments. Hormis les unités opératoires et les lits post-interventionnels, tout autre outil d'intervention requis est généralement supposé disponible et présent dans les blocs opératoires. La diversité des équipements nécessaires peut notamment se modéliser en supposant des salles d'opérations dédiées à certains types de chirurgie (Jebali et al., 2006). Nous n'avons pas connaissance de travaux pouvant considérer individuellement des ressources non-renouvelables comme les prothèses ou des équipements chirurgicaux spécifiques.

Ressources humaines

Si tout le monde s'accorde à dire que les ressources humaines sont les plus importantes du quartier opératoire, cela ne transparaît pas à l'analyse des diverses approches proposées dans la littérature. En effet, la majorité des travaux ayant pour objet la programmation opératoire suppose les ressources humaines présentes et disponibles en quantité suffisante. Les chirurgiens sont les ressources humaines les plus fréquemment considérées, au moyen de leurs disponibilités (Jebali et al., 2006; Tan et al., 2007), de leurs préférences (Ozkarahan, 2000; Testi et al., 2007) ou de leurs nombres (Hans et al., 2008). Il est encore plus inhabituel de considérer les disponibilités des anesthésistes (Chaabane et al., 2007) et des infirmières (Beliën & Demeulemeester, 2008). Sur un plan plus stratégique, Trilling (2006) propose, dans ses travaux de thèse, une démarche d'aide à la décision pour la conception de plateaux médico-techniques et le pilotage des ressources humaines mutualisées. Dans un premier temps, elle évalue les courbes de charge exprimant les besoins en personnel à l'aide d'un outil de simulation. Elle aborde ensuite, via un programme linéaire en nombres entiers, la question du dimensionnement du staff médical du plateau par la construc-

tion de vacations couvrant la charge prévisionnelle. Enfin, elle étudie la construction d'horaires de travail pour les infirmiers anesthésistes et les médecins anesthésistes en appliquant des techniques de programmation linéaire mixte et de programmation par contrainte.

2.2.2 Incertitudes

En pratique, le quartier opératoire évolue dans un environnement hautement incertain, ce qui rend sa programmation d'autant plus délicate. Les deux principales sources d'incertitude considérées dans la littérature du domaine sont les arrivées d'urgences et les temps d'opération. En effet, la programmation opératoire sera perturbée par l'arrivé d'événements imprévus tels que des urgences, mais également par des durées opératoires qui s'avèrent stochastiques. Un troisième type d'incertitude envisageable mais très peu étudié est celui lié à la disponibilité des ressources chirurgicales. Grâce à sa flexibilité de modélisation, l'outil privilégié pour étudier de tels systèmes stochastiques est la simulation.

Gerchak et al. (1996) proposent un modèle stochastique de programmation dynamique pour traiter la phase d'affectation de la programmation opératoire. Ils considèrent le temps total journalier utilisé par les urgences et par les opérations électives comme des variables aléatoires. Considérant de nouvelles demandes de programmation de cas électifs survenant le jour-même, leur but est de déterminer le nombre de ces interventions additionnelles à insérer dans le planning journalier. Pour ce faire, ils maximisent une fonction profit tout en pénalisant le recours aux heures supplémentaires et le report d'interventions. Ils évaluent alors la politique d'affectation optimale sur base de cette fonction profit.

Everett (2002) développe un modèle de simulation pour l'affectation des patients sur liste d'attente aux différents hôpitaux d'un système de santé publique en Australie. Les patients à traiter sont catégorisés par niveau d'urgence et par type d'opération. Il suppose que les temps d'opération de même que les durées d'hospitalisation suivent une distribution normale, alors que les urgences arrivent selon un processus de Poisson.

Bowers & Mould (2004) s'attaquent à la gestion d'un bloc opératoire dédié aux urgences orthopédiques. La variabilité de la demande et le contexte incertain engendrent généralement la sous-utilisation des ressources. Considérant des arrivées d'urgences suivant un processus de Poisson et des durées opératoires également stochastiques, ils appliquent un modèle de simulation de type Monte Carlo pour déterminer la durée appropriée de la plage horaire à allouer aux urgences orthopédiques. Cette durée est supposée faire un compromis entre la maximisation de l'utilisation de la salle, la minimisation des heures supplémentaires et la qualité de soins aux patients. Leur simulation examine également l'impact de l'insertion de cas électifs parmi les urgences. Il ressort de leur étude que le débit d'opérations peut être substantiellement amélioré, pour autant que les patients électifs soient enclins à accepter une possible annulation de leur intervention.

Wullink et al. (2007) étudient la meilleure façon de réserver des plages horaires du programme opératoire afin de traiter les opérations urgentes. Ils considèrent deux approches pour réserver cette capacité : soit dédier une ou plusieurs salles aux urgences, soit réserver la capacité requise en la répartissant sur la totalité des salles d'opération. Leur étude se base sur un modèle de simulation à événements discrets, et préconise cette seconde option pour offrir une meilleure réactivité aux urgences. Leur constat s'appuie sur des mesures de performance comme le temps d'attente des urgences, les heures supplémentaires à prester et l'utilisation des salles.

Nous pouvons encore citer les travaux de Lamiri et al. (2008) qui proposent une méthode de planification des interventions (uniquement la phase d'affectation des interventions aux différents jours du planning) tenant compte des urgences. Ils considèrent les urgences comme une durée stochastique à inclure au planning tandis que les temps opératoires des cas électifs sont, eux, déterministes. Ils combinent simulation de Monte Carlo et programmation linéaire en nombres entiers pour minimiser les coûts liés aux heures supplémentaires et des coûts liés à la date d'intervention.

D'autres références traitant de l'influence de l'incertitude de l'environnement hospitalier sur la programmation opératoire sont Dexter et al. (2000); Kim & Horowitz (2002); Lovejoy & Li (2002); Dexter et al. (2003); Marcon & Dexter (2006); Beliën & Demeulemeester (2007) ou encore Denton et al. (2007).

2.2.3 Contraintes sur les interventions

De manière sporadique, nous retrouvons, dans la littérature, des contraintes liées aux opérations chirurgicales à planifier. Celles-ci sont essentiellement de deux types : des contraintes de précédence entre les opérations et des contraintes temporelles sur les échéances à respecter. Certains actes chirurgicaux sont, en effet, soumis à un certain ordre de réalisation (Cardoen et al., 2006a; Ogulata & Erol, 2003). En pratique, il n'est pas rare de programmer les interventions les plus infectieuses en fin de journée. D'autre part, pour une question de sécurité et de confort pour le patient, certains travaux tiennent compte de dates d'échéance à honorer pour chacune des interventions (Guinet & Chaabane, 2003; Krempels & Panchenko, 2006; Fei et al., 2008; Lamiri et al., 2008).

2.3 Objectifs de programmation

Dans cette section, les approches de résolution sont classées en fonction des objectifs poursuivis : maximisation du profit, maximisation de l'utilisation des salles d'opération, équilibrage de la charge de travail, minimisation du temps d'attente ou maximisation de la satisfaction des préférences des ressources humaines.

Objectif financier

Les critères financiers qui sont principalement appliqués dans l'industrie ne constituent pas l'un des objectifs primordiaux de la programmation opératoire. Dexter et al. se sont penchés sur cette alternative aux mesures de performance plus classiques du monde hospitalier (Dexter et al., 2001, 2002; Dexter & Ledolter, 2003; Dexter et al., 2005). Ils étudient comment la programmation opératoire contribue à améliorer une fonction profit définie comme les revenus moins les coûts variables. Notamment, Dexter et al. (2000) développent une stratégie d'ordonnancement des cas excédant les plages horaires allouées aux chirurgiens. Cette stratégie tente d'équilibrer, d'une part, la nécessité pour le gestionnaire de réduire les coûts de personnel et, d'autre part, le besoin de flexibilité pour les patients et les chirurgiens dans le choix de la date d'intervention. Leur stratégie est ensuite évaluée par simulation.

D'autres auteurs se sont également intéressés à ce type de critère. Par exemple, en raison de l'augmentation des cas chirurgicaux, Lovejoy & Li (2002) envisagent un accroissement de capacité en comparant l'opportunité de construire de nouveaux blocs opératoires ou d'étendre les heures d'ouverture des blocs déjà existants. Ils analysent un compromis entre trois critères de performance : le temps d'attente avant la programmation, la fiabilité des temps de début des interventions programmées et le profit de l'hôpital. L'objectif est de déterminer quelle est la meilleure option d'expansion pour l'établissement.

Objectif opérationnel

Le taux d'utilisation des salles d'opération est une mesure de performance régulièrement rencontrée dans la littérature sur la planification des interventions. Entre autres, citons Dexter (2000); Dexter & Traub (2002); Guinet & Chaabane (2003); Marcon et al. (2003), mais également Sciomachen et al. (2005); Marcon & Dexter (2006); Denton et al. (2007); Fei et al. (2008) ou encore Hans et al. (2008); Pham & Klinkert (2008).

Par exemple, Ogulata & Erol (2003) développent trois modèles mathématiques multi-critères pour effectuer les tâches suivantes : la sélection, parmi une liste d'attente, des patients à opérer la semaine à venir, l'attribution des patients aux différents groupes de chirurgiens, et la détermination de la date et de la salle d'opération de chaque intervention. Cette démarche s'effectue en maximisant l'utilisation globale des salles, en équilibrant la charge de travail des différents groupes de chirurgiens et en minimisant le temps d'attente des patients.

Dexter et al. (2003) appliquent un modèle de simulation à événements discrets pour évaluer l'utilisation moyenne des salles d'opération pour des chirurgiens réalisant peu d'interventions. Ils démontrent que pour de tels chirurgiens, les méthodes statistiques classiques ne suffisent pas à estimer précisément le temps nécessaire à allouer à ces chirurgiens, et préconisent l'allocation des plages horaires sur base de l'efficience des salles et non sur base de leur utilisation.

Concernant la phase d'agencement de la programmation opératoire, Jebali et al. (2006) proposent deux programmes en nombres entiers mixtes : le premier affecte les interventions du jour aux différents blocs opératoires en minimisant la sous- et sur-utilisation des salles ; le second agence les interventions dans chacune des salles en minimisant le recours aux heures supplémentaires.

Au même titre que le taux d'utilisation des salles, l'équilibrage de la charge de travail fait partie des objectifs associés au processus opérationnel qui constitue l'activité opératoire. Certains travaux se focalisent sur la bonne répartition de la charge de travail. Principalement, celle-ci peut s'équilibrer sur les blocs opératoires (Beliën et al., 2006; Marcon et al., 2003; Ogulata & Erol, 2003), sur les services de soins (Beliën & Demeulemeester, 2007; Santibanez et al., 2007; Tan et al., 2007; van Oostrum et al., 2008) ou sur les salles de réveil (Beliën et al., 2006; Cardoen et al., 2006a; Marcon & Dexter, 2006).

Temps d'attente

Le temps d'attente des patients ou des chirurgiens est également une mesure commune de performance du programme opératoire (Arenas et al., 2002; Chaabane et al., 2006; Everett, 2002;

Guinet & Chaabane, 2003; Krempels & Panchenko, 2006; Ogulata & Erol, 2003; Testi et al., 2007; Wullink et al., 2007).

Considérant des durées opératoires stochastiques, Denton et al. (2007) mesurent l'impact de l'étape d'agencement sur le temps d'attente des chirurgiens et du staff médical, sur la durée d'inoccupation de la salle et sur l'utilisation d'heures supplémentaires. Ils développent un programme stochastique en nombres entiers mixtes ainsi que des heuristiques pour ordonnancer quotidiennement une unique salle d'opération.

Au niveau du patient, VanBerkel & Blake (2007) utilisent un modèle de simulation à événements discrets pour faciliter les décisions de planification des capacités de lits pour un quartier opératoire multi-site. Ils mettent en rapport la capacité des salles SSPI avec l'évolution du temps moyen d'attente (en jours) avant intervention pour les patients électifs.

Préférences des ressources humaines

Un dernier objectif encore peu abordé par la littérature mais qui nous intéresse au plus haut point concerne la satisfaction des préférences des acteurs du processus de programmation opératoire. Un des outils les plus prisés pour inclure cet objectif semble être la programmation par buts (*goal programming*).

Ozkarahan (2000) applique un modèle de *goal programming* afin de minimiser les temps de non production (temps d'arrêt) et les heures supplémentaires, tout en augmentant la satisfaction du personnel. Dans le cadre d'une stratégie type *block scheduling*, son approche consiste à ordonner les demandes pour un jour particulier en fonction des plages horaires allouées, de l'utilisation des salles, des préférences des chirurgiens et de la disponibilité des lits en soins intensifs. Les buts, ou critères poursuivis, sont : respecter les durées des plages horaires allouées aux chirurgiens, minimiser les sous- et sur-utilisations des salles d'opération, satisfaire les préférences des praticiens quant aux salles d'intervention, respecter les dates d'intervention pour les chirurgies les plus importantes et enfin, veiller à ne pas dépasser la capacité d'accueil de l'unité de soins intensifs. Hormis les blocs opératoires, toute autre ressource est supposée disponible en quantité suffisante et ne constitue pas une contrainte.

Blake & Carter (2002) appliquent cette même approche de *Goal Programming* pour l'allocation de ressources hospitalières. La méthodologie développée utilise deux modèles linéaires, l'un traitant du *case mix* et l'autre des coûts. Le *case mix* est un système de financement et de mesure de performances des hôpitaux, regroupant les patients en catégories distinctes sur base d'un diagnostic requérant des ressources similaires. L'objectif de ce modèle est double. Premièrement, il permet d'identifier la catégorie de patients et le volume traités par chaque spécialiste qui satisfassent aux impératifs économiques. Deuxièmement, le modèle détermine, parmi tous les cas possibles, quelle catégorie de patients (ou *case mix*) se rapproche le plus des souhaits des praticiens. Le modèle de coûts traduit, quant à lui, les décisions du premier modèle en un ensemble commensurable de changements de pratiques que chaque chirurgien doit appliquer afin de coller au mieux à ses desiderata (en termes de volume patients, de temps ou de revenu).

Tan et al. (2007) incluent l'occupation des lits dans la construction de plannings hebdomadaires. Ils développent un modèle mathématique en deux étapes (bien que seule la phase d'affectation des interventions soit explicitée), tenant compte des besoins en lits et des disponibilités du

quartier opératoire, des chirurgiens et des patients. Pour leur modèle multi-objectif, ils s'appuient sur la version lexicographique de la programmation type *Goal Programming*. Leurs trois objectifs principaux sont, par ordre de priorité, la réduction de l'occupation excédentaire des lits, la réduction du temps d'inoccupation et des heures supplémentaires, et le nivellement des volumes de patients sur les différents jours de planification. L'intégration de la dimension humaine passe ici par la minimisation du nombre de jours de présence de chaque chirurgien.

Ce type d'objectif social, prenant en compte les préférences de certains acteurs du processus opératoire, se retrouve également dans les travaux de Cardoen et al. (2006a); Velasquez & Melo (2006); Krempels & Panchenko (2006); Testi et al. (2007).

2.4 Outils de résolution

Les travaux présentés dans cette section sont répertoriés selon les outils de résolution appliqués pour la construction de programmes opératoires. Les principaux sont la programmation mathématique, les approches heuristiques et méta-heuristiques et la simulation.

La programmation mathématique prend une place importante dans la littérature sur le traitement de la programmation opératoire. Si la programmation en nombres entiers mixtes y est la mieux représentée (Jebali et al., 2006; Pham & Klinkert, 2008, entre autres), nous retrouvons également de la programmation linéaire (Guinet, 2001; Dexter et al., 2002), quadratique, qui traite des fonctions objectifs non linéaires (Beliën & Demeulemeester, 2007), ou encore dynamique (Augusto et al., 2008). Pour les problèmes aux objectifs multiples, la littérature s'oriente vers des approches de *goal programming* (Ozkarahan, 2000; Blake & Carter, 2002; Ogulata & Erol, 2003), qui associent à chaque objectif une valeur cible à atteindre, appelée but (*goal*). L'unique fonction objectif consiste alors à minimiser l'écart pondéré entre les valeurs atteintes et leurs cibles.
La résolution d'instances de taille réelle par ces méthodes exactes nécessite bien souvent un nombre important de variables de décision. Certains auteurs s'orientent alors vers des techniques de génération de colonnes qui, au lieu de considérer l'ensemble des variables (ou colonnes, puisqu'à chaque variable correspond une colonne de la matrice de contraintes concaténée à la fonction objectif), génèrent celles-ci en cours de résolution pour optimiser le problème sur une partie seulement des variables (Fei et al., 2006; van Oostrum et al., 2008). La génération de colonnes suppose un problème linéaire non entier. Pour des problèmes d'optimisation en nombres entiers, la génération de colonnes peut également être couplée avec un algorithme *branch-and-bound* pour assurer l'intégrité du vecteur de solution. Cette technique de résolution est reprise, dans la littérature, sous le nom de *branch-and-price* (voir Beliën & Demeulemeester, 2008; Fei et al., 2008, pour des exemples d'application en gestion hospitalière). Enfin, nous distinguons encore les approches du type relaxation lagrangienne qui permettent l'obtention de bornes de qualité sur l'objectif en relaxant les contraintes les plus restrictives du problème, et en les dualisant dans la fonction objectif (Augusto et al., 2008).

Une alternative aux méthodes d'optimisation exactes sont les approches heuristiques. Les méthodes exactes s'attellent à prouver l'optimalité de la solution qu'elles produisent. Or, pour des problèmes complexes d'optimisation, comme peut l'être la programmation opératoire, l'obtention d'une solution optimale peut nécessiter de larges temps de calcul. L'on privilégie alors des techniques heuristiques qui ont la faculté de produire des solutions de qualité, non nécessairement

optimales, en des temps mesurés (dépendant généralement des paramètres associés à ces techniques). Outre les heuristiques classiques (Dexter et al., 2000; Denton et al., 2007; Hans et al., 2008), nous trouvons des approches dédiées aux problèmes difficiles d'optimisation combinatoire pour lesquels aucune autre technique (méthodes exactes ou heuristiques traditionnelles) ne s'est révélée efficace : les méta-heuristiques. Les principales méta-heuristiques dédiées à la construction de plannings opératoires sont : le recuit simulé (voir Kirkpatrick et al., 1983, pour une bonne introduction), méthode appliquée par Beliën & Demeulemeester (2007); Hans et al. (2008), notamment ; la recherche Tabou (voir Glover & Laguna, 1997, pour une bonne introduction), sur laquelle portent les travaux de Fei et al. (2006) qui combinent recherche Tabou et algorithmes génétiques, qui forment la troisième grande classe de méta-heuristiques (voir Goldberg, 1989, pour une bonne introduction aux algorithmes génétiques).

La seconde technique la plus adoptée par la littérature, après la programmation mathématique, est la simulation. Le type de simulation le plus appliqué est la simulation à événements discrets, qui reproduit l'évolution d'un système selon des pas de temps discrétisés (Dexter et al., 2001; Dexter & Traub, 2002; Everett, 2002; Harper, 2002; Bowers & Mould, 2005; Sciomachen et al., 2005; Marcon & Dexter, 2006; Testi et al., 2007; VanBerkel & Blake, 2007). L'on rencontre également des approches de simulation dites de Monte-Carlo qui représentent, elles, l'état du système en un point précis du temps (Bowers & Mould, 2004; Ogulata & Erol, 2003; Hans et al., 2008; Lamiri et al., 2008).

D'autres techniques de résolution, peu représentées dans la littérature, sont par exemple les approches analytiques (Lovejoy & Li, 2002) ou l'intelligence artificielle distribuée comme les systèmes multi-agents (voir Wooldridge, 2002, pour une bonne introduction) utilisés notamment par (Czap et al., 2005).

2.5 Grille synthétique

Le tableau 2.1 résume les différents travaux recensés dans cet état de l'art, et met en exergue les lacunes de la littérature actuelle sur le sujet, en fonction de deux critères : le niveau des décisions prises et l'intégration du facteur humain. Le niveau tactique reprend essentiellement l'étape d'affectation des interventions, pour des décisions prises sur la date et/ou la salle d'opération, mais s'y trouvent également des travaux traitant de la capacité des ressources à un niveau plus stratégique ; le niveau opérationnel concerne l'étape d'agencement des blocs opératoires et génère des décisions essentiellement sur l'heure d'opération. Les approches citées dans la colonne *Tactique & Opérationnel* prennent des décisions sur les date, heure et salle d'opération. Les références qui y sont soulignées font état de travaux intégrant les phases d'affectation et d'ordonnancement en une unique procédure. Du point de vue de l'intégration des ressources humaines, les approches faisant état de contraintes spécifiques à celles-ci ou tenant compte des préférences de certains membres du personnel sont classées dans la partie *Avec* intégration. Les autres, ne faisant pas explicitement état de mesures orientées ressources humaines sont reprises dans la partie *Sans* intégration.

La lecture de cet état de l'art dédié à la programmation opératoire nous révèle essentiellement trois choses. Tout d'abord, les travaux issus de la littérature sur le sujet semblent s'accorder à scinder la construction de plannings d'interventions en plusieurs phases. En effet, quel que soit le mode de programmation opératoire, ce processus comporte au moins deux étapes, l'une veillant

		Niveau décisionnel		
		Tactique	**Opérationnel**	**Tactique & Opérationnel**
Intégration des ressources humaines	**Sans**	Arenas et al. (2002); Blake et al. (2002) Blake & Donald (2002); Bowers & Mould (2004) Bowers & Mould (2005); Dexter et al. (2002) Dexter & Ledolter (2003); Dexter et al. (2005, 2001) Dexter et al. (2003); Dexter & Traub (2002) Everett (2002); Fei et al. (2008) Hammami (2006); Kim & Horowitz (2002) Lamiri et al. (2008); Lovejoy & Li (2002) Marcon et al. (2003); Rohleder et al. (2005) van Oostrum et al. (2008); VanBerkel & Blake (2007) Wullink et al. (2007)	Augusto et al. (2008); Denton et al. (2007) Dexter (2000); Dexter et al. (2000) Harper (2002); Lebowitz (2003) Marcon & Dexter (2006, 2007) Beliën & Demeulemeester (2007)	Beliën et al. (2006); Chaabane et al. (2007) Fei et al. (2006); Sciomachen et al. (2005)
	Avec	Blake & Carter (2002); Guinet (2001) Guinet & Chaabane (2003); Hans et al. (2008) Kharraja et al. (2006); Ogulata & Erol (2003) Santibanez et al. (2007); Ozkarahan (2000) Tan et al. (2007); Trilling (2006)	Beliën & Demeulemeester (2008) Cardoen et al. (2006a,b) Hsu et al. (2003); Velasquez & Melo (2006)	Jebali et al. (2006) Krempels & Panchenko (2006) Pham & Klinkert (2008); Testi et al. (2007)

TAB. 2.1 – *Grille synthétique classant les travaux issus de la littérature sur la programmation opératoire selon deux critères : le niveau décisionnel et la prise en compte des ressources humaines.*

à l'affectation des opérations aux différents jours de planification constituant l'horizon de temps considéré, l'autre déterminant la séquence selon laquelle ces mêmes opérations seront exécutées jour par jour au sein d'un même bloc opératoire. A partir de ce constat, nous pouvons nous poser la question de savoir si cette décomposition est toujours d'actualité, et dans quelle mesure celle-ci n'engendre pas des solutions sous-optimales. Actuellement, les approches proposées par la littérature permettent d'optimiser essentiellement localement les étapes composant le processus de programmation des interventions, alors même que ces étapes interagissent. Les décisions prises à un niveau supérieur conditionnent immanquablement celles des niveaux inférieurs. Notamment, modifier le jour ou la salle d'une intervention a un impact sur la disponibilité des ressources, et donc sur la phase d'ordonnancement des opérations au sein des blocs opératoires. Dès lors, pourquoi ne pas tenter d'intégrer ces étapes en une seule procédure répondant de manière conjuguée aux niveaux de décision tactique et opérationnel ? Dans quelle mesure des modèles intégrés proposant des plannings opératoires complets, mentionnant quelles seront les date, salle et heure de chaque intervention à programmer, peuvent-ils se montrer efficients sur des données de taille réelle ?

La seconde constatation émanant de cette revue de la littérature concerne les ressources opératoires, et plus particulièrement les ressources humaines. Si la problématique hospitalière souligne l'importance des intervenants dans les processus de soins, ceux-ci paraissent quelque peu oubliés dans la majorité des travaux relatés ici. Nous n'avons pu recenser qu'une quinzaine de travaux intégrant réellement les disponibilités du personnel, dont sept seulement considèrent les préférences (sous différentes formes) émises par ces personnes. Dans la littérature classique de la planification et de l'ordonnancement (voir Demeulemeester & Herroelen, 2002; Klein, 2000, par exemple), de même que dans celle dédiée à la programmation opératoire, les ressources humaines sont principalement perçues comme passives : si elles ne sont pas considérées comme éternellement disponibles, seules leurs disponibilités et leur habilité à exécuter certaines tâches sont parfois prises en

compte. Mais ces personnes, qui interviennent directement dans la mise en œuvre de l'objet de la planification, ne participent pas à l'élaboration du planning en question. En particulier, dans le contexte hospitalier qui nous préoccupe, les principales ressources sont des professionnels de la santé ayant des avis critiques vis-à-vis de leurs conditions de travail. Il semble donc opportun de les associer à la prise de décision en leur permettant d'établir leur planning d'activités de manière conjuguée. Ceci pose les questions suivantes : dans quelle mesure peut-on incorporer une dimension coopérative au processus de programmation opératoire, permettant aux acteurs de s'accorder sur l'utilisation des plages horaires ? Quels outils utiliser pour y parvenir ? Est-il possible de définir un protocole de négociation qui permettrait aux chirurgiens de s'échanger des plages horaires ? Pour autant qu'elles répondent favorablement à la première constatation, les méthodes intégrées peuvent-elles également s'appliquer dans ce cadre coopératif ?

Enfin, le dernier enseignement que nous tirons de ce survol de la littérature remet en cause la gestion de l'aléatoire. Bien que le caractère stochastique du quartier opératoire soit reconnu de tous, la majorité des travaux recensés se positionne dans un contexte déterministe. Ces derniers ne sont toutefois pas dénués de valeur puisqu'ils traduisent les pratiques des gestionnaires hospitaliers. En effet, ceux-ci appliquent traditionnellement des approches de programmation opératoire déterministes, en ayant au préalable fixé un taux d'occupation maximum des salles d'opération. Le temps opératoire non alloué sert alors de tampon supposé absorber les aléas inhérents au quartier opératoire. Par contre, dans la littérature, nous n'avons pu trouver aucun travail permettant explicitement de déterminer comment fixer le taux d'occupation à appliquer lors de la planification des blocs opératoires. De plus, les approches stochastiques relatives à la programmation opératoire considèrent essentiellement deux sources d'incertitude : les arrivées aléatoires d'urgences et les temps d'opération variables. Or, le blocage des blocs opératoires lié à la saturation de la salle de réveil est un phénomène qui s'observe en pratique, qui provient directement de la nature stochastique des durées de réveil, mais dont l'étude reste marginale. Dès lors, pourquoi cette source d'aléas est-elle absente de la littérature ? Son impact sur les performances du quartier opératoire est-il minime ? Est-il possible de combiner ces trois grands types d'événements imprévisibles et de mesurer comment ils altèrent le fonctionnement du quartier opératoire ? Compte tenu de diverses sources aléatoires, est-il possible de rationaliser le choix du taux d'occupation des salles en le reliant au nombre d'heures supplémentaires générées par exemple ?

Chapitre 3

Problématique et notations

Le quartier opératoire est le théâtre où se déroulent les actes thérapeutiques et les diagnostics médicaux. Il se trouve au cœur des processus métiers de l'hôpital et constitue, à ce titre, un maillon essentiel du processus opératoire. Son organisation et son fonctionnement sont importants à étudier. Le chapitre 1 a souligné son importance au sein des établissements hospitaliers dont il conditionne de façon prépondérante la performance globale. Le quartier opératoire génère, en effet, les principaux revenus de l'hôpital et constitue, par la même occasion, une des plus importantes sources de dépenses (consommant près de dix pourcents du budget global de tout centre hospitalier).

L'optimisation du bloc opératoire a fait l'objet de plusieurs recherches de la part de la communauté scientifique. Le potentiel de développement reste toutefois conséquent, tant les approches soulignées dans l'état de l'art ne prennent pas en considération le rôle central que joue l'être humain dans le processus de programmation opératoire. L'étude de la littérature portant sur la programmation opératoire a mis en exergue certaines questions qui demeurent ouvertes et sur lesquelles nous basons notre recherche doctorale.

3.1 Cadre pratique de la recherche

Il ressort des chapitres précédents que l'hôpital, par la nature de sa mission, par son organisation complexe, par ses métiers pluridisciplinaires, par sa haute valeur ajoutée humaine, est un système difficile à gérer et dont l'enjeu est primordial : prodiguer des soins de qualité afin de maintenir autant que possible les patients en bonne santé. Dans cette organisation complexe, le quartier opératoire tient un rôle central. Son importance comme unité fondamentale de la production de soins mais également son poids dans la structure économique hospitalière en ont fait un sujet d'étude très prisé.

La planification des blocs et des ressources opératoires est délicate, notamment à cause de l'hétérogénéité du système et des ressources impliquées dans la réalisation des interventions chirurgicales. A cela s'ajoutent des pressions économiques auxquelles l'hôpital doit faire face. Il en découle un besoin évident de développer des outils d'aide à la décision permettant aux centres hospitaliers de gérer plus efficacement leurs unités.

Rendre plus efficace la gestion du quartier opératoire passe par une utilisation plus rationnelle des ressources opératoires. L'utilisation des ressources chirurgicales, de même que d'autres mesures de performance comme la satisfaction du personnel par exemple, sont directement conditionnées par le processus de programmation opératoire. Par conséquent, il semble pertinent de s'atteler à améliorer la planification des salles d'opération, afin de construire des plannings opératoires plus efficients. Les critères économiques ne peuvent cependant constituer le seul objectif poursuivi lors de la constitution de ces programmes opératoires. Ce processus ne peut éclipser l'importance du personnel médical tant dans la réalisation des opérations chirurgicales que dans la planification de leur exécution. Ceci est d'autant plus vrai que l'organisation hospitalière est complexe : mutualisation de ressources critiques partagées par un conglomérat de professionnels privilégiant généralement l'éthique au financier. C'est pourquoi, outre les objectifs économiques, la prise en compte de la dimension humaine demeure une piste intéressante à approfondir, d'autant plus que peu d'études abordent ces questions dans les travaux scientifiques.

Une grande partie de ces travaux semble effectivement avoir occulté certains aspects de la gestion du quartier opératoire. Tout d'abord, dans une optique d'optimalisation du processus de programmation opératoire, la littérature s'obstine à scinder ce dernier en des étapes successives de planification et d'ordonnancement. Or, la tendance aujourd'hui est à l'intégration des modèles afin de prendre des décisions de plus en plus globales sur les systèmes étudiés, notamment dans le domaine de la chaîne logistique. Si les ressources informatiques ne permettaient pas encore cette intégration il y a peu, nous allons, dans un premier temps, vérifier si cette constatation est toujours d'actualité. Nous allons donc, dans ce travail, évaluer l'opportunité d'intégrer les étapes de planification et d'ordonnancement du quartier opératoire, investiguer différentes méthodes de résolution et analyser leur capacité à répondre à la question suivante : est-il possible de développer une approche intégrée efficiente sur des jeux de données de taille réelle ? Par ailleurs, les travaux de référence considèrent usuellement les salles d'opération comme étant identiques et donc polyvalentes. Mais cette réalité n'est pas celle de tous les hôpitaux, notamment en Belgique. Nous allons dès lors également mesurer l'impact de la nature des salles sur les performances des approches de résolution envisagées. Des procédures spécifiques sont-elles à prévoir selon que les blocs opératoires soient dédiés ou polyvalents ?

Le monde hospitalier en général, et particulièrement le quartier opératoire, sont des structures ou l'humain tient une place essentielle. C'est bien l'expertise, pour ne pas dire l'art, des chirurgiens, anesthésistes ou autres infirmières qui permet au centre hospitalier son activité opératoire. Or, la littérature sur la programmation opératoire, processus qui alloue notamment les ressources chirurgicales, porte peu d'attention aux disponibilités ou autres desiderata du personnel médical. Le quartier opératoire est pourtant un environnement nécessitant une grande coopération, entre autres pour l'attribution des plages horaires allouées aux différentes spécialités ou chirurgiens. Outre l'intégration des étapes séquentielles du processus de programmation opératoire, nous envisageons d'incorporer à celui-ci une dimension coopérative. Dans un deuxième temps, nous allons donc apprécier s'il est envisageable de tenir compte des préférences du personnel, ou du moins des chirurgiens, lors de la constitution des plannings opératoires. Et si, à partir des approches déjà développées, il est possible de définir un protocole de négociation permettant à d'éventuels acteurs, mécontents du calendrier opératoire, d'échanger de plage horaire.

Le dernier point qu'il nous semble pertinent d'aborder concerne la gestion des événements survenant inopinément. A nouveau, la littérature s'étend somme toute peu sur le sujet. Pourtant, la

nature stochastique du quartier opératoire n'est un secret pour personne. Certains travaux considèrent des arrivées aléatoires d'urgences ou des durées opératoires variables mais peu tiennent compte simultanément de ces deux sources d'aléas. Qui plus est, d'autres phénomènes de sources imprévisibles perturbent le bon déroulement du programme opératoire. Nous pensons notamment aux patients obligés d'entamer leur phase de réveil postopératoire dans une salle d'intervention, par manque de lits de réveil disponibles en SSPI. A notre connaissance, cette singularité n'est que très épisodiquement abordée dans la littérature, alors qu'elle s'observe régulièrement en pratique. Nous envisageons donc, en troisième lieu, d'estimer l'effet de ces trois sources d'aléas sur les performances du quartier opératoire, en termes de temps d'attente des urgences ou d'heures supplémentaires générées par exemple. Nous souhaitons également évaluer la possibilité de définir un taux d'occupation des salles, c'est-à-dire une marge de sécurité à appliquer lors de la construction des plannings chirurgicaux afin d'augmenter leur robustesse à l'aléatoire, en fonction de ces trois types d'incertitudes.

Les recherches que nous envisageons ont donc comme principal objectif de proposer des outils pour une meilleure utilisation des ressources opératoires. Notamment, dans un contexte économique délicat, il est plus que nécessaire aux hôpitaux en général, et aux quartiers opératoires en particulier, de réduire leurs coûts. A cette logique financière, nous souhaitons ajouter une dimension humaine qui caractérise le milieu médical. En effet, prendre en compte l'opinion des acteurs du processus opératoire devrait améliorer la satisfaction des équipes y intervenant et, par conséquent, l'efficacité du quartier opératoire. Enfin, cette efficacité est également fortement dépendante du caractère stochastique de l'environnement dans lequel se déroule le processus étudié. C'est pourquoi il nous semble nécessaire d'évaluer l'impact qu'a la stochasticité inhérente au quartier opératoire sur ses performances. La méthodologie que nous allons appliquer dans cette recherche pour répondre à toutes ces interrogations comporte exclusivement des approches quantitatives. Elles sont au nombre de trois :

- **Modélisation mathématique.** Tout d'abord, comme nous sommes confrontés à une logique d'optimisation, des coûts notamment, le premier réflexe est de développer une approche exacte de résolution. Nous allons donc, dans un premier temps, modéliser le processus de programmation opératoire à l'aide d'un programme mathématique. L'objectif poursuivi est de déterminer des plannings opératoires minimisant les coûts (en termes d'utilisation des salles d'opération et d'heures supplémentaires générées) tout en satisfaisant l'ensemble des contraintes liées au fonctionnement du quartier opératoire (y compris les disponibilités du personnel généralement supposé présent en quantité suffisante dans la plupart des travaux). Comme il semble plus contraignant de considérer des salles dédiées à certaines chirurgies seulement, nous commencerons par envisager ce cas. Ensuite, nous appliquerons le modèle mathématique au cas de blocs polyvalents et verrons dans quelle mesure il est nécessaire de l'ajuster. Si les approches exactes s'avèrent peu concluantes sur le plan calculatoire, nous examinerons la possibilité d'une approche heuristique, voire méta-heuristique, pour obtenir des plannings optimisés (mais non nécessairement optimums) en des temps plus raisonnables.
- **Théorie des jeux.** Les aspects coopératifs sont présents dans deux domaines bien distincts : la théorie des jeux et l'intelligence artificielle distribuée. La première étant plus en lien avec nos volontés d'optimisation, c'est pour la théorie des jeux que nous allons opter en premier lieu. Nous tenterons d'appliquer des théories comme le *Nash Bargaining* ou les mécanismes d'enchères afin d'inclure un minimum de coopération dans le processus de programmation

opératoire. Si d'aventure ces concepts ne satisfaisaient pas suffisamment à nos attentes, nous nous orienterions alors davantage vers des procédures informatiques comme les systèmes multi-agents.

– **Théorie de Markov.** L'utilisation de la théorie Markovienne se fait dans le cadre de l'étude de la stochasticité du quartier opératoire. En effet, la modélisation mathématique considère des activités déterministes, or la réalité n'est pas prévisible avec exactitude. Les trois techniques les plus utilisées pour traiter ces considérations aléatoires sont la théorie des files d'attente, la théorie de Markov et la simulation. Toujours dans la même logique de résolution, nous privilégions une approche analytique qui permet une certaine diversification des applications, ce que permet difficilement la simulation qui nécessite une connaissance plus approfondie du système étudié et s'applique dès lors plus aisément sur des cas particuliers. Nous entamerons ce pan de nos recherches par le développement d'un modèle de file d'attente ou de Markov, selon les spécificités considérées et l'applicabilité de ces théories. Si ces dernières ne se révélaient pas satisfaisantes, nous nous tournerions alors vers des modèles de simulation qui permettent de détailler plus spécifiquement le système étudié, cependant au prix d'une certaine généralisation de ceux-ci.

Notre objectif est donc d'évaluer l'opportunité d'intégrer les étapes ordinairement séquentielles du processus de programmation opératoire tout en y incorporant une dimension humaine trop rare ; d'analyser l'effet de la variabilité sur les performances du quartier opératoire ; de développer des approches de résolution efficientes permettant de rationaliser la prise de décision, de les appliquer sur des jeux de données réelles et d'analyser et critiquer les résultats obtenus.

3.2 Le cas d'un hôpital belge

L'étude sur la programmation opératoire entreprise dans cette thèse émane d'une collaboration que nous avons développée avec un hôpital belge de la région bruxelloise. Dans l'hôpital investigué, le quartier opératoire est ouvert cinq jours par semaine (excepté pour les urgences) durant huit heures, auxquelles doivent être ajoutées les heures supplémentaires. Le temps opératoire total disponible est partagé par 77 chirurgiens. Le département de chirurgie compte également 29 anesthésistes et 31 infirmières ou instrumentistes. Le jeu de données collecté comporte toute l'activité opératoire réellement observée dans l'hôpital entre le 2 janvier et le 28 février 2006, ce qui représente un total de 757 interventions réalisées sur 42 jours ouvrables, dont 632 sont des opérations électives et 125 sont des urgences. Le nombre d'opérations programmées par semaine (week-end exclu) varie de 60 à 100, alors que 16 opérations sont planifiées par jour, en moyenne. Les durées opératoires des interventions électives varient de 20 minutes à 720 minutes, avec un temps opératoire électif moyen de 153 minutes et un écart-type de 88 minutes, sur base de la fonction de densité de probabilité empirique décrite dans la figure 3.1. Le nombre d'urgences survenant par semaine varie, quant à lui, de 11 à 18, tandis que le nombre d'urgences par jour est de 2.97 en moyenne. Leurs durées opératoires varient de 15 minutes à 390 minutes, avec une moyenne de 87 minutes par urgence et un écart-type de 63 minutes, sur base de la fonction de densité de probabilité empirique également reprise en figure 3.1. Le quartier opératoire est programmé hebdomadairement et comprend sept salles d'opération polyvalentes et une SSPI composée de onze lits de réveil. Les hôpitaux belges sont, en effet, légalement contraints de munir la salle de réveil d'un nombre de lits égal à une fois et demie le nombre de blocs opératoires.

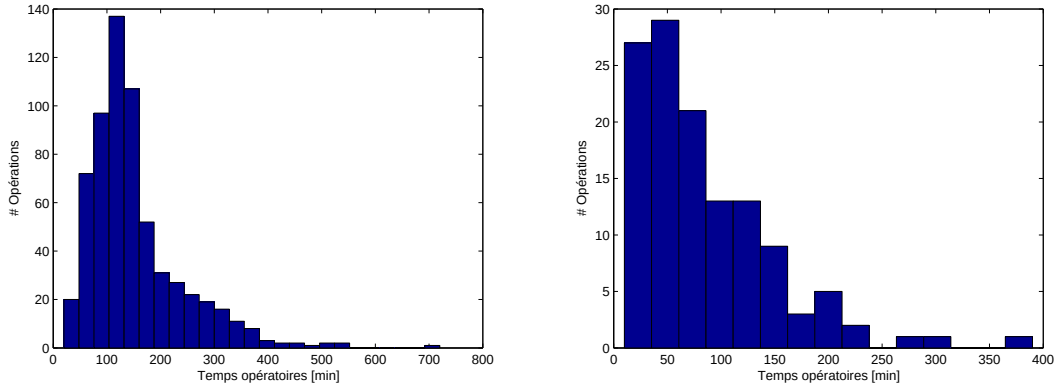

FIG. 3.1 – *Fonction de densité de probabilité empirique des durées opératoires observées, pour les opérations électives (gauche) et pour les urgences (droite).*

L'analyse de ce jeu de données nous en apprend un peu plus sur le fonctionnement, en pratique, du quartier opératoire. Tout d'abord, nous pouvons y observer que l'activité opératoire de l'hôpital se compose de quatre types d'intervention : les opérations électives, les opérations ajoutées, les urgences et les opérations du week-end. Les opérations électives reprennent tous les patients dont les traitements chirurgicaux sont programmés depuis une semaine au moins. Les interventions ajoutées sont des urgences non critiques, c'est-à-dire des opérations non programmées à insérer dans le programme opératoire hebdomadaire et devant être réalisées sous trois jours (elles seront dans un premier temps considérées comme électives, bien que modifiant le planning hebdomadaire). Les urgences sont des événements journaliers survenant inopinément et devant être traités le jour-même. Par conséquent, les urgences sont prioritaires et généralement insérées entre deux opérations électives, perturbant de la sorte le planning initial. La dernière catégorie contient toute l'activité de fin de semaine, période de temps durant laquelle le quartier opératoire est théoriquement fermé. Ce type d'activité ne sera dès lors pas considéré dans la suite de ce travail.

Le quartier opératoire manque cruellement de personnel, en particulier d'anesthésistes et d'infirmières, à tel point que seules six salles peuvent être utilisées simultanément lors de la construction du calendrier opératoire. La gestion de ce personnel s'en trouve d'autant plus aiguë, nous confortant dans notre démarche d'intégration des ressources humaines dans le processus concerné. Nous constatons également que, dû à ce manque de personnel médical, l'hôpital ne peut assurer la présence de trois infirmières par opération, le minimum légal. Par contre, les gestionnaires tentent de respecter le minimum communément admis de deux infirmières par bloc opératoire. Les données à disposition nous montrent que, en pratique, certaines infirmières ne prennent pas part à l'entièreté d'une intervention en cours mais la quittent pour assister un autre patient.

Les données reflètent également la nature stochastique du quartier opératoire. Les temps opératoires d'un même type de traitement diffèrent d'une occurrence à l'autre. Les temps mentionnés dans ce travail représentent le temps global d'occupation de la salle d'intervention. Ceci comprend le temps de préparation nécessaire à l'achèvement des actes pré-chirurgicaux, le temps d'induction nécessaire à l'anesthésie du patient et le temps d'opération à proprement parler. Ces durées opératoires fluctuent également en fonction du type d'intervention. En effet, nous avons observé que la durée d'une urgence est, en moyenne, inférieure à celle d'une opération programmée (87 minutes

contre 153 minutes). Ce caractère aléatoire du quartier opératoire, encore renforcé par l'arrivée d'urgences à traiter prioritairement, joue donc un rôle non négligeable dans la génération d'heures supplémentaires (celles-ci s'élevant à sept heures en moyenne, par jour), situation à éviter autant que possible pour d'indéniables raisons économiques.

Théoriquement, un patient opéré est supposé rejoindre la SSPI avant d'être dirigé, sur base de son état de santé, soit en chambre d'hospitalisation soit aux soins intensifs. Cependant, sur base de nos données, en pratique cela se passe quelque peu différemment. Après opération, le patient est admis dans l'une des quatre destinations suivantes.

- SSPI. Si le nombre de lits de réveil le permet, le patient joint directement la salle de réveil. Le taux de patients électifs admis en salle de surveillance post-interventionnelle est de 87%, alors que ce taux est de 52% pour les urgences.
- Soins intensifs. Les cas les plus critiques sont directement transférés dans l'unité de soins intensifs sans passer par la salle de réveil, supposant que l'état de santé du patient le justifie. Les soins intensifs sont un département indépendant du quartier opératoire et de son organisation. Le taux de patients électifs y étant admis est de 10% alors qu'il s'élève à 12% pour les urgences.
- Bloc opératoire. Le nombre de lits de réveil composant la SSPI étant limité par des contraintes physiques ou financières, il se peut qu'à la fin d'une opération aucun d'eux ne soit disponible. Comme un patient ne peut attendre avant de se voir prodiguer les soins post-opératoires, ce dernier reste dans la salle d'opération qu'il occupe jusqu'à ce qu'il se réveille ou qu'un lit se libère en SSPI. Dans ce cas, le bloc opératoire est immobilisé et les opérations électives suivantes sont retardées, ce qui perturbe le planning opératoire. Le taux de patients électifs immobilisant une salle est de 2.8% tandis que le taux d'urgences bloquant une salle est de 35%.
- Morgue. Malheureusement, certains patients ne survivent pas à leur intervention, ne nécessitant plus de passer par la salle de réveil, et sont transférés directement à la morgue. Le taux de patients électifs décédant est très faible à 0.2% alors que celui des urgences est légèrement plus élevé à 1%.

3.3 Description du problème envisagé et notations

Sur base du cas d'étude illustré ci-dessus, nous décrivons en détail le problème tel que nous l'envisageons. Nous nous intéressons à la gestion d'un quartier opératoire composé d'un nombre fixé de salles nécessaires à la réalisation des opérations chirurgicales. Plus précisément, nous nous focalisons sur le processus de programmation opératoire, c'est-à-dire la planification en horizon de temps restreint de ces tâches chirurgicales (nous considérons typiquement des programmes opératoires d'une semaine). Cet horizon se définit par un nombre D de jours d'ouverture du quartier opératoire (notés $d = 1, \ldots, D$), chaque journée étant subdivisée en T périodes de temps. Ces divisions temporelles, représentées par $t = 1, \ldots, T$ seront par exemple des quarts d'heure. Quelle que soit l'approche de résolution adoptée, il va de soi que plus la discrétisation temporelle est fine, plus les temps de calcul risquent d'être importants (par exemple, la discrétisation temporelle conditionnerait en partie le nombre de variables et de contraintes d'un éventuel modèle mathématique).

Les interventions façonnant une journée en blocs opératoires sont de deux types : il y a d'une part les actions programmées prévues par le planning et d'autre part les urgences qui sont, elles, le fruit du hasard. Ces dernières étant des événements aléatoires, nous n'en tiendrons pas compte, dans un premier temps, lors de la conception du calendrier opératoire. Dans ce contexte déterministe, l'objectif est donc de planifier les différentes tâches programmées en leur assignant les ressources disponibles (humaines et matérielles) nécessaires, tout en satisfaisant des contraintes économiques et temporelles. Les approches de programmation opératoire développées étant déterministes, il nous faudra, par après, évaluer l'impact de la variabilité sur ce processus.

L'ensemble des tâches médicales (ou opérations) à ordonnancer se compose de J traitements chirurgicaux (indicés $j = 1, \ldots, J$) qualifiés de "non-préemptifs", ce qui signifie qu'une fois débutés, ils ne peuvent être interrompus. Chaque opération j est définie par un temps d'exécution déterministe p_j et par certains besoins en ressources indispensables à son accomplissement, notés r^{ρ}_{jk} ou r^{ν}_{jk} selon que la ressource k soit renouvelable ($k \in \mathcal{K}^{\rho}$) ou non ($k \in \mathcal{K}^{\nu}$). Ces consommations r^{ρ}_{jk} et r^{ν}_{jk} sont d'application durant la totalité de l'opération j. Les ressources renouvelables (comme le personnel par exemple) se caractérisent par des disponibilités ou capacités R^{ρ}_{kd} périodiques constantes par journée d, tandis que les ressources non renouvelables (comme les médicaments ou le matériel stérile) se distinguent par des disponibilités R^{ν}_{kd} globales pour toute une journée. Par souci de sécurité et de confort pour le patient, chaque intervention j doit impérativement être traitée dans un intervalle temporel défini par un jour de début au plus tôt et un jour de début au plus tard, désignés par ES_j et LS_j respectivement.

Notre recherche s'axant sur l'intégration du facteur humain dans la planification, nous mettons ici l'accent sur les disponibilités et les préférences des ressources humaines. Nous distinguons deux catégories d'individus parmi les membres des équipes médicales : nous avons d'une part les personnes tels les anesthésistes ou les infirmières que nous qualifierons d'"anonymes" dans le sens où elles peuvent intervenir indépendamment du nom du patient. D'autre part, nous avons les chirurgiens qui ont leurs patients propres et réalisent donc des traitements qui leur sont spécifiquement réservés.

En Belgique, une intervention chirurgicale nécessite la présence d'au moins un chirurgien, un anesthésiste et de trois infirmières. Comme nous l'avons vu dans la description des données liées au cas d'étude bruxellois, l'hôpital ne peut assurer la présence de trois infirmières par opération. Par contre, les gestionnaires essaient de respecter le minimum communément admis de deux infirmières par salle d'intervention. Les données à disposition nous montrent également que, en pratique, certaines infirmières n'assistent pas à la totalité d'une intervention mais se retirent pour prendre soin d'un autre patient. Afin de coller à cette réalité, due à un manque persistant de personnel, nous supposons qu'une ressource humaine dite anonyme est susceptible de passer d'un bloc opératoire à un autre avant la fin de l'intervention. Pour modéliser cet état de fait, outre la consommation en ressources r^{ρ}_{jk} qui concerne l'entièreté de l'intervention, nous définissons pour chaque opération j une extra-consommation Γ^{ρ}_{jk} d'application pour les $\lambda_{jk} < p_j$ premières périodes de temps que durera cette opération (avec $\lambda_{jk} \geq 0$). Cette mesure s'applique également aux anesthésistes dont le nombre est généralement insuffisant pour couvrir l'ensemble des salles d'opération, et peut s'étendre à un modèle plus général où les variations du niveau d'utilisation des ressources ne se produit pas qu'en consommant depuis le début jusqu'à un moment donné.

Chacun de ces traitements chirurgicaux est assigné à un chirurgien $c = 1, \ldots, C$ particulier et l'ensemble des tâches dévolues au chirurgien c est défini par $O(c)$. Les chirurgiens, quant à

eux, spécifient les demi-journées durant lesquelles ils sont disponibles au moyen d'une matrice binaire $R^C \in \{0,1\}^{C\times 2D}$. Cette matrice est construite de telle sorte que l'entrée $R^C_{c,2d-1} = 1$ si le chirurgien c est disponible le matin du jour d, et $R^C_{c,2d} = 1$ s'il l'est l'après-midi du même jour.

Le quartier opératoire est équipé de S salles d'opération notées $s = 1, \ldots, S$. Pour chacune de ces salles s, nous définissons une disponibilité régulière D^N_{sd}, pour le jour d, représentant le nombre de périodes de temps en heures normales d'ouverture. Et puisque nous autorisons les dépassements horaires, chaque salle s se caractérise également par une disponibilité maximale $D^N_{sd} \leq D^M_{sd} \leq T$ exprimant le nombre total de périodes de temps disponibles le jour d (reprenant donc heures normales et heures supplémentaires). Comme précisé dans la description du contexte hospitalier, la littérature recense deux types de salles d'intervention (Trilling, 2006) :

- **Les blocs opératoires dédiés** sont généralement pourvus différemment et ne peuvent supporter tout type de chirurgie. Ces salles d'opérations sont réservées uniquement à certaines spécialités chirurgicales, essentiellement pour des raisons logistiques (le matériel requis n'étant pas toujours transportable d'une salle à l'autre) ou organisationnelles (certaines salles pouvant être réservées au traitement exclusif des urgences, par exemple). Dans de telles conditions, une intervention ne peut dès lors s'accomplir que dans un ensemble fixé de salles dédiées, représenté par la matrice $Q \in \{0,1\}^{J\times S}$ définie de sorte que l'entrée q_{js} de cette matrice vaille 1 lorsque l'activité chirurgicale j peut avoir lieu dans le bloc s.
- **Les blocs opératoires polyvalents**, ou identiques, permettent, eux, de traiter toutes les pathologies et sont donc équipés en conséquence. Pour ce type de salles, il n'est plus nécessaire de définir une telle matrice Q vu l'absence de restriction sur leur accessibilité.

Le quartier opératoire se compose également d'une ou plusieurs SSPI (Salles de Surveillance Post-Interventionnelle) contenant les lits de réveil. Son opération terminée, le patient quitte le bloc opératoire pour joindre immédiatement la salle de réveil, où celui-ci va se rétablir de son intervention avant de rejoindre, suivant son état de santé, son lit d'hospitalisation ou le service des soins intensifs. Par manque d'autonomie respiratoire, un patient ne peut attendre de recevoir les soins postopératoires. Par conséquent, la phase de réveil débute dès la première période de temps suivant la fin de l'opération. Le temps nécessaire au patient dans la SSPI (c'est-à-dire le temps d'occupation du lit de réveil) est estimé à b_j périodes. Le nombre total R^B de lits de réveil disponibles est généralement fixé par rapport au nombre de blocs opératoires. La règle en vigueur préconise un nombre de lits équivalent à une fois et demi le nombre de blocs (voir notamment Dexter & Tinker, 1995, qui suggèrent une proportion allant de 1.5 à 2).

Finalement, un bon planning opératoire se doit d'utiliser au mieux les ressources à disposition. En particulier, sur le plan économique, il est préférable de réduire le nombre de salles d'opération utilisées, et ceci pour deux raisons. D'une part, minimiser le nombre de blocs opératoires en activité permet de préserver des salles libres afin d'absorber certains aléas comme l'arrivée d'urgences ou de gros retards engendrés par des complications. Ceci devrait éviter une trop grande perturbation du programme opératoire et par conséquent préserver les patients, comme le personnel, d'un certain mécontentement. D'autre part, nécessiter moins de salles évite d'inévitables coûts fixes liés à leur utilisation (en termes de personnel infirmier, de consommation énergétique, etc.). Soit C^{open}, le coût correspondant à l'ouverture d'une salle d'opération. Pour d'évidentes raisons financières, un bon planning opératoire doit également restreindre autant que possible le travail en heures supplémentaires. Soit C^{over}, le coût d'une période de temps d'utilisation des salles en dehors des heures normales d'ouverture. D'un point de vue gestionnaire, l'objectif poursuivi est simple : minimiser les frais de fonctionnement du quartier opératoire.

TEMPS	
D	Nombre de jours d'ouverture d sur l'horizon considéré
T	Nombre de périodes de temps t composant un jour d'ouverture
OPÉRATIONS	
J	Nombre total d'opérations à programmer
p_j	Durée opératoire de l'opération j
b_j	Durée d'occupation d'un lit de réveil par l'opération j
r^{ρ}_{jk}	Besoin en ressource renouvelable k pour l'opération j
r^{ν}_{jk}	Besoin en ressource non-renouvelable k pour l'opération j
Γ^{ρ}_{jk}	Extra-consommation en ressource renouvelable k pour l'opération j
ES_j	Jour de début au plus tôt de l'opération j
LS_j	Jour de début au plus tard de l'opération j
CHIRURGIENS	
C	Nombre total de chirurgiens
$O(c)$	Ensemble des opérations du chirurgien c pour l'horizon considéré
$R^C \in \{0,1\}^{C \times 2D}$	Disponibilités des chirurgiens exprimées par demi-journée
BLOCS OPÉRATOIRES	
S	Nombre total de salles s composant le quartier opératoire
D^N_{sd}	Disponibilité normale du bloc s le jour d
D^M_{sd}	Disponibilité maximale du bloc s le jour d
$Q \in \{0,1\}^{J \times S}$	Accessibilité des blocs dédiés pour chaque opération
R^B	Nombre total de lits de réveil du quartier opératoire
COÛTS	
C^{open}	Coût fixe associé à l'ouverture d'un bloc opératoire
C^{over}	Coût associé à une période de temps en heures supplémentaires

TAB. 3.1 – *Tableau récapitulatif de l'ensemble des notations utilisées dans la description du problème*

Chapitre 4

Méthodes de résolution pour des blocs opératoires dédiés

Ce chapitre développe deux approches destinées à optimiser la programmation d'un quartier opératoire composé de salles dédiées, c'est-à-dire n'accueillant que certaines chirurgies. L'objectif est double. Tout d'abord, il s'agit d'aider le gestionnaire à mieux prendre en compte les contraintes de fonctionnement des blocs opératoires et à en améliorer, par conséquent, le processus de programmation. Ensuite, nous avons pour but d'inclure le facteur humain dans la construction des programmes opératoires. A cette fin, nous avons apporté un soin particulier à la modélisation des disponibilités des ressources humaines (mais également matérielles) indispensables au bon déroulement des interventions. Nous débutons ce chapitre par la description d'un programme linéaire en nombres entiers mixtes, approche exacte permettant la construction de plannings optimums. Celle-ci s'avérant relativement gourmande en temps de calcul, nous décrivons ensuite une approche méta-heuristique, et plus précisément un algorithme génétique, ayant la caractéristique d'offrir des solutions de qualité en des temps plus raisonnables. Enfin, nous évaluons l'applicabilité de ces deux techniques de résolution sur un jeu de données réelles provenant d'un hôpital belge de la région bruxelloise. Ces expériences numériques nous apprennent que la programmation opératoire hebdomadaire peut être résolue à l'optimum pour autant que le temps de calcul ne soit pas une ressource critique. L'approche méta-heuristique, quant à elle, offre nettement plus de réactivité que son pendant exact et produit néanmoins une programmation hebdomadaire des salles d'opération de qualité, malgré des temps de calcul beaucoup moins importants.

4.1 Approche exacte : formulation mathématique en nombres entiers

Nos travaux de recherche ont pour principal objectif d'aider les gestionnaires hospitaliers à programmer les interventions chirurgicales au sein du quartier opératoire, sur un court horizon de temps de typiquement une semaine. Comme mentionné précédemment, la littérature dévouée à ce domaine tend à scinder cette tâche en deux sous-problèmes distincts. Une première étape de planification (ou étape d'affectation des interventions) fixe le jour et la salle d'opération pour chaque intervention. Vient ensuite une étape d'ordonnancement (ou étape d'agencement des blocs opératoires) qui détermine les temps de début de chaque opération programmée une même jour-

née, en tenant compte des contraintes de mise en œuvre comme veiller à la disponibilité des ressources par exemple. Cependant, ces deux étapes interagissent et les solutions proposées par la littérature peuvent s'avérer sous-optimales. En effet, modifier le jour d'intervention a un impact notamment sur la disponibilité des ressources et, par conséquent, peut également influencer la valeur objectif de l'ordonnancement. Dans cette section, nous décrivons une modélisation mathématique intégrant ces deux aspects en une procédure unique reflétant à la fois la problématique de la planification et celle de l'ordonnancement. Le modèle proposé se présente sous la forme d'un programme d'optimisation en nombres entiers mixtes (*Mixed-Integer Programming*, ou MIP, voir Wolsey, 1998; Nemhauser & Wolsey, 1999, pour de bonnes introductions à la programmation en nombres entiers) qui permet au gestionnaire de considérer tout un ensemble de contraintes liées au fonctionnement de ses blocs opératoires.

Notre approche de modélisation repose sur un modèle très répandu de gestion de projet, le Resource-Constrained Project Scheduling Problem (Brucker et al., 1999; Kolisch & Hartmann, 2005; Demeulemeester & Herroelen, 2002), ou RCPSP. Un projet se définit dans ce contexte comme un ensemble de tâches, souvent reliées par des contraintes de précédence, à réaliser à l'aide d'un nombre limité de ressources tout en poursuivant un ou plusieurs objectifs donnés. L'ordonnancement de projet revient donc à déterminer les dates d'exécution de chaque activité, en tenant compte des disponibilités des ressources, afin de satisfaire au mieux l'objectif fixé (généralement la minimisation de la durée du projet, ou "makespan"). Dans un tel problème, une tâche se définit par un temps d'accomplissement et par des consommations en ressources. Ces ressources sont dites "renouvelables", ce qui signifie que leurs quantités sont constantes et renouvelées à chaque période de temps. Pour des raisons techniques ou temporelles par exemple, certaines activités sont également soumises à des contraintes de précédence par rapport à d'autres. Notons cependant que, dans le cadre de l'optimisation du quartier opératoire qui nous préoccupe ici, nous ne ferons pas usage de relation de précédence entre les opérations chirurgicales ; nous introduisons en revanche un second type de ressources dites "non-renouvelables", comme peuvent l'être les médicaments par exemple, ayant une capacité globale pour l'entièreté d'un projet (ici pour un jour entier), c'est-à-dire que toute unité de ressource consommée est perdue et n'est plus jamais accessible.

Le problème de planification de blocs opératoires dédiés, tel que décrit dans la section 3, peut se formuler mathématiquement au moyen du programme linéaire en nombres entiers suivant. Tout d'abord, rappelons que les indices j, s, d et t représentent respectivement les opérations chirurgicales à traiter, les blocs opératoires, les jours de planification et les périodes de temps opératoire. La formulation comporte trois types de variables. Premièrement, les variables binaires x_{jsdt} sont dédiées aux activités chirurgicales et prennent la valeur 1 uniquement lorsque l'opération j débute dans le bloc s, le jour d au moment t. De la même manière, les variables binaires z_{sd} sont vouées aux blocs opératoires, dont l'accessibilité y est ici supposée restreinte, et valent 1 si l'on procède à l'ouverture du bloc s le jour d. Finalement, notre modèle intègre un troisième type de variables entières positives, les variables l_{sd}, qui comptabilisent le nombre de périodes en heures supplémentaires dans la salle s le jour d.

L'objectif premier vise la minimisation des frais de fonctionnement du quartier opératoire. Le modèle se décrit dès lors par la fonction objectif suivante :

$$\min_{z,l} \sum_{s} \sum_{d} [\, C^{open}\, z_{sd} + C^{over}\, l_{sd}\,] \tag{4.1}$$

sujette, pour des blocs opératoires dédiés, aux contraintes (4.2)–(4.16) reprises dans le tableau 4.1.

$$\sum_{s}\sum_{d}\sum_{t} x_{jsdt} \;=\; 1, \qquad \forall j, \tag{4.2}$$

$$\sum_{j}\sum_{s}\left[r^{\rho}_{jk} \sum_{\substack{\tau=t-p_j+1\\ \tau>0}}^{t} x_{jsd\tau} + \Gamma^{\rho}_{jk} \sum_{\substack{\tau=t-\lambda_{jk}+1\\ \tau>0}}^{t} x_{jsd\tau} \right] \leq\; R^{\rho}_{kd}, \qquad \forall d, \forall t, \forall k \in \mathcal{K}^{\rho}, \tag{4.3}$$

$$\sum_{j}\sum_{s} \sum_{\substack{\tau=t-(p_j+b_j)+1\\ \tau>0}}^{t-p_j} x_{jsd\tau} \leq R^{B}, \qquad \forall d, \forall t, \tag{4.4}$$

$$\sum_{j}\sum_{s}\sum_{t} r^{\nu}_{jk}\, x_{jsdt} \;\leq\; R^{\nu}_{kd}, \qquad \forall k \in \mathcal{K}^{\nu}, \forall d, \tag{4.5}$$

$$x_{jsdt} = 0, \forall c, \forall j \in O(c), \forall s, \forall d | R^{C}_{c,2d-1} = 0, \forall t \leq \frac{T}{2}, \tag{4.6}$$

$$x_{jsdt} = 0,\; \forall c, \forall j \in O(c), \forall s, \forall d | R^{C}_{c,2d} = 0, \forall t > \frac{T}{2}, \tag{4.7}$$

$$\sum_{s}\sum_{j\in O(c)} \sum_{\substack{\tau=t-p_j+1\\ \tau>0}}^{t} x_{jsd\tau} \;\leq\; 1, \qquad \forall t, \forall d, \forall c, \tag{4.8}$$

$$\sum_{j} \sum_{\substack{\tau=t-p_j+1\\ \tau>0}}^{t} x_{jsd\tau} \;\leq\; z_{sd}, \qquad \forall s, \forall d, \forall t, \tag{4.9}$$

$$\sum_{d}\sum_{t} x_{jsdt} \;\leq\; Q_{js}, \qquad \forall j, \forall s, \tag{4.10}$$

$$\sum_{s}\sum_{t} x_{jsdt} = 0, \qquad \forall j, \forall d \notin [ES_j, LS_j], \tag{4.11}$$

$$x_{jsdt} = 0, \qquad \forall j, \forall s, \forall d, \forall t | (t + p_j) > D^{M}_{sd}, \tag{4.12}$$

$$\sum_{j}\sum_{t} p_j\, x_{jsdt} \leq D^{M}_{sd}\; z_{sd}, \qquad \forall s, \forall d, \tag{4.13}$$

$$l_{sd} \geq \sum_{j} \sum_{\substack{t=D^{N}_{sd}-p_j+1\\ t>0}}^{T-p_j} \left(t + p_j - D^{N}_{sd}\right)\; x_{jsdt},\; \forall s, \forall d, \tag{4.14}$$

$$l_{sd} \in \mathbb{N}, \qquad \forall s, \forall d, \tag{4.15}$$

$$x_{jsdt},\; z_{sd} \in \{0, 1\}, \qquad \forall j, \forall s, \forall d, \forall t. \tag{4.16}$$

TAB. 4.1 – *Modélisation des contraintes liées au fonctionnement du quartier opératoire composé de salles dédiées.*

Les contraintes (4.2) assurent qu'une opération ne démarre qu'une et une seule fois dans un seul bloc opératoire. Les inégalités (4.3), (4.4) et (4.5) sont les contraintes de disponibilité des ressources. Les premières traitent des ressources renouvelables (infirmières, anesthésistes, etc.) et incluent les extra-consommations. Les contraintes (4.4) contrôlent la disponibilité des lits de réveil et assurent qu'un patient opéré n'attende pas avant de rejoindre la SSPI, tandis que les contraintes (4.5) traitent les ressources non-renouvelables (matériel chirurgical stérilisé, médicaments, etc.). Les contraintes (4.6) et (4.7) vérifient respectivement la présence des praticiens le matin et l'après-midi, en fixant x_{jsdt} à zéro lorsque le chirurgien respectif n'est pas disponible. Notons que, grâce à la restriction $\tau > T/2$ dans cette seconde équation, une intervention débutée le matin peut se terminer si nécessaire l'après-midi, même lorsque le chirurgien n'est pas censé être disponible dans l'après-midi. Les contraintes (4.8) préviennent le chevauchement de plusieurs opérations d'un même chirurgien, alors que les contraintes (4.9) proscrivent, elles, l'assignation de deux (ou plus) opérations au même instant dans la même salle. Les inégalités (4.10), connexes à l'accessibilité des blocs opératoires, vérifient que les interventions se déroulent dans des salles qui leur sont dédiées. Les équations (4.11) garantissent à chaque traitement de débuter entre les jours de début au plus tôt et au plus tard donnés, en fixant x_{jsdt} à zéro pour les jours n'appartenant pas à l'intervalle $[ES_j, LS_j]$ ainsi défini. Les contraintes (4.12) sont des contraintes de capacité fixant x_{jsdt} à zéro pour toutes les périodes de temps provoquant un dépassement de la disponibilité maximale du bloc opératoire D^M_{sd}. Les contraintes (4.13) sont des inégalités valides permettant de resserrer la formulation. Elles traitent le fait que, en dehors de toute contrainte de fonctionnement, une opération ne peut être programmée dans une salle si la somme des durées opératoires de toutes les interventions prenant place dans cette dernière excède D^M_{sd}. Les inégalités (4.14), de pair avec la minimisation de l'objectif et la positivité des variables l_{sd}, certifient que ces dernières représentent effectivement le nombre de périodes de temps utilisées en heures supplémentaires dans chaque salle. Pour finir, les contraintes (4.15) et (4.16) définissent d'une part la positivité et l'intégrité des variables l_{sd}, et d'autre part le caractère binaire des variables x_{jsdt} et z_{sd}. Notons que les contraintes (4.6), (4.7), (4.11) et (4.12) sont ici mentionnées par souci de précision mais qu'il est préférable de simplement ne pas créer les variables associées lors de l'implémentation du modèle.

Comme il dérive du RCPSP, le problème tel que modélisé dans ce chapitre est indubitablement NP-difficile. En raison de cette complexité et de la taille des instances à résoudre en pratique, il est peu probable que cette approche exacte produise rapidement des plannings optimums. Dans notre contexte, la priorité consiste à construire des programmes opératoires de qualité (sous-entendu non nécessairement optimums) en des temps raisonnables. En effet, ceux-ci seront sujets à perturbations dues à l'arrivée d'événements imprévus, comme des urgences par exemple, qui imposeront leur ajustement. Raison pour laquelle le gestionnaire requiert un bon calendrier opératoire initial, mais pas impérativement optimal. En fonction du temps que le décideur est disposé à accorder, l'on privilégie alors une méthode de résolution heuristique qui a la faculté de produire de bonnes solutions (c'est-à-dire supposées proches de l'optimum) sans aucune preuve d'optimalité. Ce type d'approche fait l'objet de la section 4.3.

4.2 Expérimentations numériques de l'approche exacte

Dans cette section, nous évaluons les performances de notre modèle MIP, en termes de valeur objectif et de temps de calcul, pour une résolution optimale du problème. Ces expériences numé-

riques mettront en exergue l'importance des ressources calculatoires nécessaires à cette résolution pour certains jeux de données, et l'utilité d'une approche de résolution plus facilement applicable en pratique. Le MIP est codé à l'aide du langage de modélisation AMPL et résolu par la version 10.0 du solveur MIP CPLEX. Celui-ci est paramétré de sorte à donner une priorité de branchement sur les variables z. Ces expériences numériques ont été conduites sous Windows XP équipé d'un processeur Pentium M 1.5 GHz et d'une mémoire RAM de 1.25 GB.

4.2.1 Données expérimentales

Nous traitons ici le cas de l'hôpital belge détaillé en section 3.2, dont le quartier opératoire se compose de sept blocs et onze lits de réveil. Le jeu de données complet présenté en section 3.2 a été subdivisé en neuf sous ensembles reprenant chacun l'activité opératoire élective observée sur une semaine du calendrier. Pour rappel, le quartier opératoire est ouvert cinq jours par semaine durant neuf heures par jour, auxquelles s'ajoutent les heures supplémentaires. Un jour se compose dès lors de 48 périodes de temps de 15 minutes chacune. Le nombre d'opérations par semaine varie de 60 à 100, dont 80 en moyenne, alors que le temps opératoire moyen est de 10 périodes de temps pour un écart-type de 5.9 périodes (voir la fonction de densité de probabilité, figure 3.1). Notons que les données dont nous disposons n'offrent aucune information sur les durées de réveil des patients. Néanmoins, connaissant leur taux de blocage, nous pouvons en déduire un temps de réveil moyen de 2 heures 25 minutes (ceci sera explicité plus loin, en section 7.5.2). Dès lors, connaissant la destination des patients opérés, nous appliquons une durée de réveil de 10 périodes pour chacun d'entre eux passant par la SSPI, et une durée nulle pour les autres (ces derniers sont alors directement dirigés vers l'unité de soins intensifs et n'influencent plus le système). Les ressources renouvelables sont au nombre de trois : anesthésistes, infirmières et instrumentalistes. Les coûts d'ouverture d'un bloc opératoire C^{open} sont estimés à 2040 euros, alors qu'une période de temps en heures supplémentaires est facturée $C^{over} = 50$ euros. Ces chiffres sont inspirés de données réelles fournies par notre partenaire hospitalier.

En raison de la nature complexe du problème, nous pouvons nous attendre à d'importants temps de calcul, tout du moins pour les instances de grande taille. En particulier, les contraintes (4.3) de disponibilité des ressources renouvelables sont très restrictives. Elles peuvent donc nécessiter un surplus de calculs. Dans la section 4.2.2, nous appliquons, dans un premier temps, le MIP sur un sous-ensemble des données représentant une unique journée d'utilisation du quartier opératoire, ce afin d'effectuer une analyse de sensibilité du modèle. Cette pratique revient donc à ne résoudre que l'étape d'ordonnancement du processus, le jour d'opération étant alors supposé connu. Cette analyse donne une première idée de l'applicabilité du modèle en pratique, nous positionnant de la sorte par rapport à une partie non négligeable de la littérature (voir le tableau 2.1). Par souci d'objectivité, les données utilisées ici ont été prélevées dans le sous-ensemble comportant 80 opérations, l'activité moyenne hebdomadaire. Le jeu de données comprend alors 19 interventions pratiquées par 12 chirurgiens sur un unique jour de travail, les autres chiffres restant identiques. Les durées opératoires varient alors de 4 à 21 périodes de temps, avec pour moyenne 10.9 périodes et 5.2 comme écart-type. Les performances de notre modèle sont ensuite évaluées sur la totalité des sous-ensembles de données afin d'évaluer la faisabilité de l'intégration des étapes de planification et d'ordonnancement qui composent le processus de programmation opératoire. Ces expériences nous permettent également, le cas échéant, de valider notre approche en testant sa robustesse lors de la construction de plannings hebdomadaires sur différentes semaines.

Nbr. Op.	Consommation Ressources	Objectif [€]	Temps [s]
16	R	12290*	1
	E	12290*	1
19	R	14430*	53
	E	15130 (1.62 %)	1800

TAB. 4.2 – *Performances du programme linéaire en nombres entiers* (4.1)–(4.16)*, pour 16 puis 19 interventions à ordonnancer sur un jour.*

4.2.2 Analyse de sensibilité

Le temps passé à l'élaboration du planning opératoire est relativement crucial pour le problème qui nous occupe. En effet, le gestionnaire de blocs opératoires ne peut pas se permettre de patienter durant des heures avant d'avoir son programme journalier à disposition. Dans cette section, nous effectuons une analyse de sensibilité du modèle mathématique proposé en section 4.1. Nous y évaluons comment objectif et temps de calcul sont affectés par le niveau de consommation de ressources (en termes de plus grandes demandes en ressources renouvelables ou de plus grand nombre d'interventions à réaliser). Nous ne considérons qu'un unique jour d'ouverture lors de cette analyse. Ceci nous permet, entre autres, d'évaluer l'utilité de l'approche lorsque, en cas d'imprévus, le gestionnaire se doit d'aménager le planning du jour, le matin même, pour tenir compte d'événements nouveaux. Etant donné que cette analyse ne se rapporte qu'à une utilisation quotidienne du quartier opératoire, les temps de calcul sont limités à 1800 secondes.

Nous évaluons les performances du modèle linéaire dans diverses situations en faisant varier la consommation de ressources et le nombre d'interventions à effectuer. Nous considérons deux niveaux de consommation de ressources : d'une part, un niveau régulier qui applique les consommations réellement observées et recensées par le partenaire hospitalier ; d'autre part, un niveau élevé augmentant les demandes réelles d'un facteur proportionnel à la capacité de chaque ressource. Nous débutons cette analyse de sensibilité par l'agencement d'un jour d'ouverture comportant 16 opérations à ordonnancer, ce qui correspond au nombre moyen d'interventions à réaliser journellement pour la base de données à disposition. Les performances du MIP sont identiques, quel que soit le niveau de consommation de ressources (voir tableau 4.2). La solution optimale est obtenue presque instantanément (moins d'une seconde) ; elle nécessite l'ouverture de six salles afin de programmer les seize interventions pour un coût total de 12290 euros. Si le nombre d'interventions à ordonnancer passe à dix-neuf, les sept salles sont nécessaires à leur accomplissement. A présent, une distinction entre niveau régulier et élevé de besoin en ressources s'impose. Si le cas d'une consommation régulière ne pose aucun problème avec une solution optimale de 14430 euros obtenue en 53 secondes, il n'en va plus de même lorsque celle-ci s'élève. Dans ce cas, le MIP ne peut prouver l'optimalité de sa solution endéans le temps imparti. Après une demi-heure de calcul, la solution proposée s'élève à 15130 euros, pour un écart avec la borne inférieure de 1.62%. On remarque ici l'impact de l'association des contraintes liées aux ressources renouvelables (4.3) et des contraintes d'accessibilité des salles (4.10) sur la complexité de résolution du modèle.

D'une manière générale, cette analyse de sensibilité nous apprend que pour les problèmes de petites tailles (consommation régulière des ressources, nombre moyen d'interventions), la formulation en nombres entiers se comporte bien. Elle atteint des solutions optimales presque ins-

tantanément. Ses performances restent totalement acceptables pour des problèmes de tailles plus importantes (consommation élevée de ressources, nombre élevé d'intervention) ne comprenant qu'un unique jour de programmation. Néanmoins, en raison de la complexité du problème et au vu des performances de l'approche appliquée sur une journée de travail chargée, nous pouvons attendre de longs temps de calcul afin d'optimiser la programmation hebdomadaire des blocs. Le niveau de consommation de ressources (en termes d'une demande plus importante en ressources ou d'un plus grand nombre d'interventions) engendre définitivement la contrainte la plus restrictive. Les temps de calculs résultant varient de quelques secondes pour une consommation régulière et 16 opérations, à plus de 1800 secondes pour une consommation élevée et 19 opérations. Notons toutefois que notre approche exacte reste entièrement applicable pour ordonnancer les opérations au sein des blocs puisqu'en trente minutes maximum elle produit des solutions (quasi-) optimales.

4.2.3 Programmation opératoire hebdomadaire des blocs dédiés

Dans cette section, nous appliquons la formulation mathématique proposée dans ce chapitre afin d'optimiser la programmation hebdomadaire du quartier opératoire lorsque les salles y sont dédiées. A titre comparatif, cela prend une journée complète au gestionnaire de l'hôpital partenaire pour établir manuellement le planning opératoire de la semaine suivante. Il réserve généralement son vendredi afin d'analyser les demandes de la semaine à venir, et de la sorte construire le planning opératoire en fonction des diverses contraintes de fonctionnement. Outre la difficulté de déterminer manuellement un programme réalisable, le résultat s'avère loin d'être optimal.

Les résultats de nos expériences sont présentés sous forme de tableau contenant des mesures de performance classiques du solveur CPLEX. Y figurent la valeur objectif, l'écart entre la solution courante et la meilleure borne inférieure (si le solveur est stoppé avant d'avoir prouvé l'optimalité de la solution), et le temps de calcul. L'écart, ou *gap*, est exprimé en pour-cent et apparaît entre parenthèses dans la colonne "*Objectif*". Un objectif marqué d'un astérisque signifie que la solution obtenue est optimale. L'écart résultant est par conséquent nul et ne paraît pas. Le tableau comporte deux colonnes "*Objectif*", l'une mentionnant la valeur de la première solution courante faisant passer le gap sous la barre de 1% (avec le temps y correspondant), et l'autre reprenant la solution obtenue après maximum 24 heures de calculs (ou la solution optimale et le temps s'y rapportant le cas échéant). Afin de rester cohérents avec la démarche du gestionnaire de blocs, nous limitons effectivement les temps de calcul à 24 heures. Notons que les expérimentations qui suivent ont été effectuées en appliquant uniquement les consommations de ressources réellement observées.

Globalement, le tableau 4.3 nous enseigne que le programme linéaire en nombres entiers mixtes peine à prouver l'optimalité de ses solutions lorsqu'il est appliqué sur des instances représentant une semaine d'activité opératoire. La figure 4.1 illustre l'évolution du temps de calcul nécessaire au solveur pour atteindre un certain niveau de qualité de la solution trouvée. Dans un tiers des situations, l'approche s'avère très efficace puisque les optimums sont atteints en moins de 25 minutes. Par contre, dans les deux tiers restants, il lui faut plus de six heures pour y parvenir alors que pour plus d'une instance sur deux la limite temporelle de 24 heures est atteinte sans preuve d'optimalité. Cependant, nous constatons également que limiter la recherche aux solutions ne dépassant pas la meilleure borne inférieure de plus d'un pour-cent permet d'obtenir des approximations de qualité nécessitant moins de ressources temporelles. Pour six de ces instances, le solveur prend moins d'une heure pour parvenir au seuil requis et fait de cette utilisation heuristique

Nbr.	Programme	MIP (Gap $< 1\%$)		MIP	
Op.	Effectif [€]	Obj. [€]	Temps	Obj. [€]	Temps
60	55140	37280 (0.7%)	*26 s*	37280*	*614 s*
70	56230	50130 (0.7%)	*68 s*	50030*	*152 s*
72	68630	55940 (0.1%)	*479 s*	55940 (0.0%)	*24 h*
80	66440	51780 (0.7%)	*5.9 h*	51580*	*6.5 h*
85	68530	58490 (0.9%)	*7064 s*	58140 (0.4%)	*24 h*
86	62050	46360 (0.8%)	*449 s*	46160 (0.3%)	*24 h*
88	62600	50780 (0.8%)	*917 s*	50380*	*1492 s*
91	63950	–	–	52320 (3.0%)	*24 h*
100	61160	57850 (0.3%)	*3583 s*	57700 (0.1%)	*24 h*

TAB. 4.3 – *Performances du programme linéaire en nombres entiers* (4.1)–(4.16) *appliqué à la programmation hebdomadaire d'un quartier opératoire composé de blocs dédiés. Comparaison des solutions obtenues en stoppant le solveur une fois le gap inférieur à 1% avec celles obtenues après maximum 24 heures de calcul.*

du modèle une bonne approche de résolution pour les données considérées. Il est effectivement peu probable qu'une quelconque heuristique soit en mesure de s'approcher autant de l'optimum tout en nécessitant moins de temps pour y parvenir. Néanmoins, pour les instances comportant 80, 85 et 91 opérations, cette façon de procéder n'est peu voire pas applicable du tout. Notamment, dans le cadre de cette dernière instance, il n'est tout bonnement pas possible d'obtenir de solution à moins de trois pour-cent de la borne inférieure en moins de 24 heures. Qui plus est, il faut parfois attendre plus d'une heure et demie avant que le solveur ne soit en mesure de proposer une solution réalisable, comme nous le montre la figure 4.2. Dans ces cas-là, l'application de techniques heuristiques, et plus précisément méta-heuristiques vu la complexité du problème abordé, se justifie amplement. Cette analyse révèle cependant tout l'intérêt de notre approche dans l'automatisation du processus de planification des blocs opératoires. Les solutions proposées ici permettent, en effet, une meilleure gestion de ceux-ci avec des coûts diminués, selon les cas, de 5% à 32% par rapport aux programmes opératoires effectifs réalisés manuellement pour les mêmes semaines. Ces chiffres nous renforcent également dans notre démarche d'intégration des deux étapes habituellement traitées séquentiellement du processus de programmation opératoire.

Les principaux enseignements tirés de ces expériences nous montrent que le MIP développé plus haut, en produisant des solutions optimales ou des solutions approchées de grande qualité, se révèle efficient pour 6 jeux de données sur neuf. A l'inverse, avec plusieurs heures de calcul nécessaires, celui-ci se montre nettement moins efficace pour les trois autres ensembles de tests. Ceci conditionne nettement une utilisation pratique de notre approche dans ce cas de figure et pourrait constituer un frein à son application dans des circonstances nécessitant une faculté de réactivité à des événements imprévus survenant tardivement. Notamment, si, pour une raison ou une autre, il est impératif de ré-agencer le planning de la semaine le lundi matin, le gestionnaire se doit d'avoir un outil efficient à disposition. Il est envisageable, par notre approche exacte, de modifier le programme opératoire afin d'insérer, retirer, avancer ou retarder certaines opérations, tout en perturbant le moins possible celles déjà programmées, en fixant certaines variables en conséquence dans le MIP. Cependant, le gestionnaire ne pourra se permettre d'attendre six heures (ou plus) pour faire débuter les interventions. Dans ce cas, nous privilégions l'utilisation d'approches moins

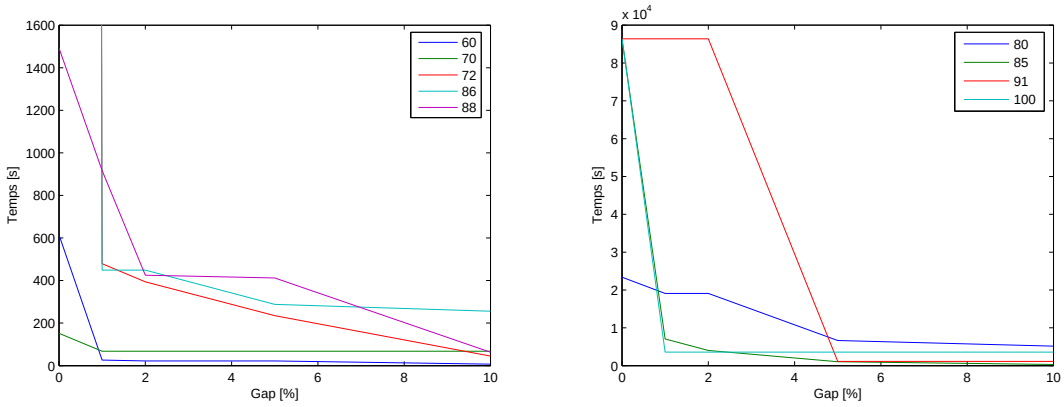

FIG. 4.1 – *Evolution du temps de calcul en fonction de l'écart (gap) entre la solution courante et la meilleure borne inférieure trouvée par l'algorithme de résolution. A gauche, les cinq instances les plus simples à résoudre, à droite les quatre les plus complexes.*

consommatrices de temps, quitte à obtenir un programme opératoire non optimal du moment qu'il offre un objectif de niveau satisfaisant. Cette description répond à la définition de techniques de résolution dite approchée, également appelées heuristiques. Plus particulièrement, dans la section suivante, nous allons nous intéresser à des méthodes qualifiées de méta-heuristiques, destinées à la résolution de problèmes d'optimisation réputés difficiles, comme l'est la planification et l'ordonnancement du quartier opératoire.

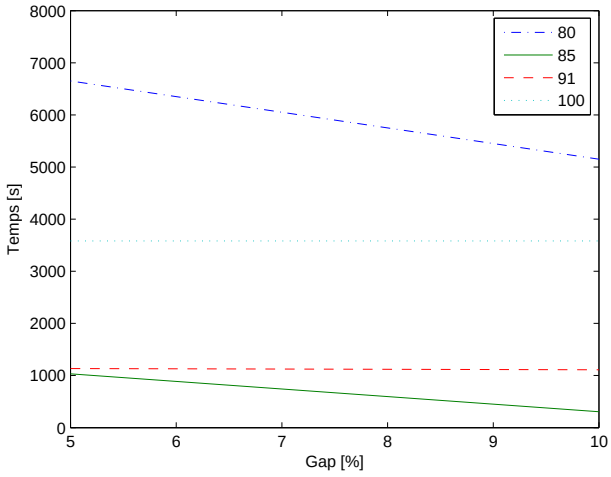

FIG. 4.2 – *Evolution du temps de calcul en fonction de l'écart (gap) entre la solution courante et la meilleure borne inférieure trouvée par l'algorithme de résolution. Illustration de la difficulté d'obtenir une solution réalisable pour les quatre instances les plus complexes.*

4.3 Approche méta-heuristique : algorithme génétique

Cette section développe une approche méta-heuristique supposée pourvoir des plannings opératoires de qualité en un temps de calcul raisonnable. En effet, le problème auquel nous nous attaquons est d'une grande complexité algorithmique. Les résultats et plus particulièrement les temps de calcul présentés en section 4.2 en témoignent. Une résolution exacte du problème envisagé semble peu appropriée pour certaines des instances de taille réelle portant sur un quartier opératoire équipé de salles d'intervention dédiées. C'est la raison pour laquelle nous avons entrepris de développer une méthode de résolution approchée, ici sous forme d'un algorithme génétique.

Les méta-heuristiques sont apparues dans les années 80 afin de résoudre les problèmes d'optimisation réputés difficiles, c'est-à-dire les problèmes pour lesquels la littérature ne recense aucun algorithme polynomial pouvant les résoudre à l'optimum, comme le sont les problèmes dits "NP-difficiles". Les méta-heuristiques s'appliquent autant aux problèmes discrets que continus et possèdent des caractéristiques communes (Dréo et al., 2003) :

- elles s'inspirent de systèmes réels tels que la physique (recuit simulé, etc.), la biologie (algorithmes génétiques, recherche Tabou, etc.) ou l'éthologie (colonies de fourmis, etc.) ;
- elles sont itératives et en partie stochastiques, pour faire face à l'explosion combinatoire des solutions ;
- elles sont directes et n'ont pas recours au calcul des gradients de la fonction objectif ;
- elles sont paramétrées et le réglage de ces paramètres peut s'avérer délicat.

Le principal atout des méta-heuristiques par rapport aux algorithmes itératifs classiques est leur capacité à s'extraire d'un optimum local. Pour ce faire, elles font usage d'un voisinage afin d'autoriser momentanément une dégradation de l'objectif (sans provoquer pour autant la divergence de l'algorithme) qui peut, à terme, mener à une meilleure solution. Les méta-heuristiques les plus répandues sont (Dréo et al., 2003) :

- *Le Recuit Simulé* (Kirkpatrick et al., 1983). Il transpose la technique métallurgique dite du "recuit" qui aboutit à un état solide bien ordonné et d'énergie minimale en portant le matériau à température élevée puis en abaissant progressivement celle-ci.
- *La Recherche Tabou* (Glover & Laguna, 1997). Elle met en pratique des mécanismes de la mémoire humaine en partant d'une situation courante et en l'actualisant au fil des itérations. Au cours d'une itération, la méthode effectue un mouvement dans le voisinage de la solution courante en tenant compte d'une liste de mouvements interdits qui sont les m derniers mouvements effectués (dénommés *tabous*), ceci pour éviter de boucler et de revenir à une solution déjà visitée précédemment.
- *Les Algorithmes Génétiques* (Goldberg, 1989). Ils tirent leur origine dans l'évolution biologique des espèces. Partant d'un ensemble de N points sélectionnés au hasard dans l'espace de recherche, l'algorithme va faire évoluer, par générations successives, l'ensemble initial de solutions (appelé population). Afin d'améliorer les solutions d'une population, celles-ci sont modifiées sur base de deux mécanismes de la théorie de Darwin : la sélection, qui favorise les individus les plus forts, et le croisement, qui combine les gênes des parents pour former la descendance.
- *Les Colonies de Fourmis* (Colorni et al., 1992). Cette approche simule la capacité qu'ont les fourmis à résoudre collectivement certains problèmes, notamment lorsqu'il s'agit d'aller chercher de la nourriture, alors même que la faculté d'un individu est très limitée. Chaque fourmi dépose une substance chimique (phéromone) le long du chemin qu'elle parcourt,

substance que tous les membres de la colonie perçoivent. Les individus s'orientent alors préférentiellement vers les chemins les plus "odorants", permettant de la sorte de trouver le plus court chemin.

Parce qu'ils ont pour aptitude de manipuler une population de solutions, les algorithmes génétiques sont particulièrement indiqués pour proposer un panel de solutions différentes quand la fonction objectif comporte plusieurs optima globaux. Ils produisent également des solutions offrant différents compromis lorsqu'ils sont dévolus à résoudre des problèmes comportant plusieurs objectifs potentiellement antagonistes. Pour cette raison, et afin de présenter un ensemble de solutions de qualité parmi lesquelles les personnes concernées pourront choisir la plus appropriée, nous avons opté pour cette catégorie d'algorithmes méta-heuristiques.

Comme leur nom l'indique, les algorithmes génétiques (ou AG) s'inspirent des principes d'évolution rencontrés en biologie. Ils parcourent globalement l'espace de solutions en recombinant des solutions déjà existantes pour en créer de nouvelles. Nous nous référons à (Goldberg, 1989; Michalewicz, 1998; Dréo et al., 2003) pour une introduction détaillée au vaste domaine des algorithmes génétiques, et à (Hartmann, 2001) pour une application de la programmation génétique au problème d'ordonnancement de projet multi-modes à ressources contraintes (MRCPSP).

4.3.1 Architecture basique des algorithmes génétiques

Les algorithmes génétiques forment une classe particulière d'algorithmes évolutionnaires, et appliquent des règles issues de la biologie évolutionnaire telles que le croisement, la mutation et la sélection. La puissance des AG vient de leur aptitude à traiter simultanément un ensemble de solutions du problème abordé et à les faire évoluer au fil des itérations vers des solutions de meilleure qualité (optima locaux ou globaux). Dans le vocabulaire génétique, il n'existe pas de solutions mais bien des individus ; un algorithme ne manipule pas un ensemble de solutions mais bien une population d'individus, de même qu'il ne se répète pas selon des itérations mais bien selon des générations. Les individus sont des représentations abstraites de solutions candidates du problème traité, qu'elles soient réalisables ou non. Un individu se présente sous forme de chromosomes reprenant les informations codées nécessaires à la reconstruction de la solution qu'il représente. Si, historiquement, les premiers codages de solutions en chromosomes étaient essentiellement binaires (un chromosome étant alors une chaîne de caractères binaires), bien d'autres codages sont à présent utilisés.

Un algorithme génétique débute par une phase d'initialisation durant laquelle une population contenant POP individus est créée, POP étant un nombre entier pair positif. Lors de chaque génération de l'algorithme, celui-ci associe à chaque individu une valeur *fitness* censée refléter la qualité de l'individu, et donc de la solution qu'il représente. C'est sur base de cette valeur que les individus pourront ensuite être comparés. Quant à l'évolution proprement dite, elle se produit en sélectionnant et en modifiant certains individus, sur base de leur valeur fitness. Les individus sélectionnés pour la phase évolutive de l'algorithme sont très naturellement appelés les *parents*. Ces derniers sont répartis aléatoirement par paires avant de subir l'opérateur de *croisement* (ou *crossover*). Cet opérateur va créer de nouveaux individus en croisant le patrimoine génétique des parents, ici sous forme de chromosomes. Les nouveaux individus ainsi créés, issus de l'évolution, sont logiquement nommés les *enfants*. L'algorithme applique ensuite aux seuls enfants l'opérateur de mutation qui modifie aléatoirement les chromosomes des enfants en produisant de nouveaux

gènes. Comme la taille de la population a ainsi été augmentée, l'algorithme applique finalement un opérateur de *sélection* destiné à ne préserver de la population que les POP individus qui constitueront la génération suivante. La sélection se fait également sur base de la valeur fitness des individus. S'agissant d'un algorithme itératif, cette procédure est répétée tant qu'un critère d'arrêt n'est pas rencontré, comme un nombre maximum GEN de générations ou une limite temporelle de traitement.

Cette procédure génétique se décrit par l'algorithme 1, où g est l'indice de génération courante, P_g représente la population courante, E_g l'ensemble des enfants issus de la population courante et T la limite de temps de calcul.

Algorithme 1 Schéma d'un algorithme génétique

1: $g := 1$;
2: Initialisation de la population P_1 contenant POP individus ;
3: Calcul de la valeur fitness des individus $I \in P_1$;
4: **while** $g \leq GEN$ et $t \leq T$ **do**
5: $E_g = \texttt{croisement}(P_g)$;
6: $E_g = \texttt{mutation}(E_g)$;
7: Calculer la valeur fitness des individus $I \in E_g$;
8: $P_g := P_g \cup E_g$;
9: $P_{g+1} = \texttt{sélection}(P_g)$;
10: $g := g + 1$;
11: **end while**

Voici décrit brièvement les principes de base qui régissent le fonctionnement d'un algorithme génétique. Les sections suivantes détaillent successivement quel codage nous appliquons pour la construction de plannings opératoires, comment nous initialisons la population de l'algorithme, quelle est la fonction qui permet d'évaluer la qualité des individus d'une génération, et enfin, comment nous définissons les opérateurs évolutionnaires (croisement et mutation) et l'opérateur de sélection.

4.3.2 Codage chromosomique des solutions

Un algorithme génétique est une procédure qui fait évoluer une population de solutions au moyen de divers opérateurs. Ces derniers sont définis de manière à faire converger au mieux les solutions vers des optima, globaux si possible. Afin d'appliquer ces opérateurs évolutionnaires, les solutions sont codées sous forme de chromosomes contenant toute l'information nécessaire à leur recouvrement. Pour le problème de programmation des blocs opératoires (voir chapitre 3), nous appliquons le codage suivant.

Chaque individu I se modélise par une paire de chromosomes non corrélés $I = (\delta, \sigma)$. Le chromosome δ représente une liste ordonnée de l'ensemble des opérations chirurgicales à planifier sur l'horizon de temps considéré. A chacune de ces interventions est associée la demi-journée durant laquelle celle-ci aura lieu. Un élément de δ se présente dès lors sous forme d'un couple (j_i, d_i) contenant l'identifiant d'une opération (j_i) et la demi-journée durant laquelle cette opération doit être réalisée (d_i). Le second chromosome σ alloue quant à lui une salle d'opération à

chacune des interventions de δ.

$$I = \begin{pmatrix} \delta \\ \sigma \end{pmatrix} = \begin{pmatrix} (j_1, d_1) & \dots & (j_J, d_J) \\ s_1 & \dots & s_J \end{pmatrix}.$$

En résumé, l'élément (j_i, d_i) de δ signifie que l'intervention j_i est planifiée la demi-journée d_i et sera traitée dans la salle d'opération représentée par s_i dans σ.

Le chromosome δ consiste en une liste ordonnée des opérations à effectuer. Or, a priori, il n'y a aucun lien de précédence entre les interventions et donc aucune véritable raison de les ordonner. Quel est donc l'intérêt qui nous pousse à donner un ordre aux opérations de cette liste ?

Bien qu'il ne soit pas inhérent au problème étudié, il existe bel et bien un lien de précédence entre les opérations se déroulant dans une même salle. Ce lien découle directement du décodage d'un individu, c'est-à-dire de la façon dont nous reconstruisons une solution à partir de sa représentation chromosomique. Pour récupérer le programme opératoire à partir d'un individu, nous appliquons un schéma de génération d'ordonnancements en série aux opérations devant être réalisées durant un même jour du planning. Ce dernier consiste à successivement prendre, dans la liste δ, la première opération non encore planifiée et à la programmer à la première période de temps satisfaisant les contraintes de ressources et de précédence temporelle. Cette manière de procéder nous permet d'appliquer un lien de précédence entre les interventions en place dans une même salle d'opération[1]. Dès lors, afin de considérer autant de solutions que possible, ces contraintes de précédence sont définies en fonction de l'ordre dans lequel les opérations sont reprises dans le chromosome δ. Cependant et de manière tout-à-fait évidente, les opérations planifiées le matin précèdent celles planifiées l'après-midi, même si ces dernières apparaissent en premier dans la liste.

Notons que deux individus différents peuvent représenter le même programme opératoire puisque permuter deux opérations consécutives de δ, programmées à des jours différents, n'aura aucun effet sur l'ordonnancement final. A l'inverse, chaque individu ne donne lieu qu'au mieux à un unique programme. En effet, étant donné que le codage appliqué peut ne pas satisfaire toutes les contraintes du problème (comme les contraintes de disponibilité maximale des salles ou les contraintes de ressources par exemple), le programme opératoire issu d'un individu peut s'avérer non réalisable.

4.3.3 Initialisation de la population

Ceci constitue la phase initiatrice de l'algorithme. Une première population d'individus y est constituée de manière aléatoire. Les POP individus de première génération sont générés comme suit. Les J interventions chirurgicales sont triées aléatoirement afin de former une séquence ordonnée dans δ. La demi-journée d_i associée à l'intervention j_i est sélectionnée aléatoirement parmi les disponibilités du chirurgien correspondant, pour autant que la date retenue satisfasse les contraintes de début d'opération. Pour des blocs opératoires dédiés, la salle d'opération s_i affectée à l'opération j_i est choisie aléatoirement parmi l'ensemble de salles accessibles à cette opération et représenté par $Q_{j_i,:}$, la $j_i^{\text{ème}}$ ligne de la matrice Q.

[1] Nous rejoignons de la sorte le problème classique d'ordonnancement de projet à ressources contraintes (RCPSP) dans lequel les différentes tâches constituant le projet sont généralement reliées entre-elles par des contraintes de précédence.

4.3.4 Evaluation de la qualité des individus

Les individus composant la population d'un algorithme génétique sont générés en codant des solutions du problème traité. Cette représentation chromosomique des solutions n'en facilite pas l'interprétation en termes de qualité, de valeur objectif. Dès lors, les algorithmes génétiques font usage d'une fonction fitness qui associe à chaque individu une valeur reflétant la qualité de la solution sous-jacente à cet individu. En ce qui nous concerne, la valeur fitness d'un individu est mesurée suivant la satisfaction de quatre contraintes non prises en compte dans le codage des individus. Il s'agit des contraintes de disponibilité des ressources non-renouvelables, de disponibilité des salles d'opération, de début d'intervention et de disponibilité des chirurgiens. La fonction fitness appliquée est partiellement décrite sur base du modèle mathématique en nombres entiers mixtes, détaillé en section 4.1. A partir du programme opératoire généré par un individu I, cette fonction mesure quatre quantités décrites ci-après.

Soit $L_k^\nu(d)$, la capacité journalière en surplus de la ressource non-renouvelable $k \in \mathcal{K}^\nu$:

$$L_k^\nu(d) = R_{kd}^\nu - \sum_j \sum_s \sum_t r_{jk}^\nu \, x_{jsdt}.$$

La négativité de cette quantité $L_k^\nu(d)$ implique la non-faisabilité du planning pour le jour d. La surconsommation totale de ressources non-renouvelables est donnée par :

$$L^\nu = \sum_{\substack{k \in \mathcal{K}^\nu \\ L_k^\nu(d) < 0}} \left| L_k^\nu(d) \right|.$$

Le codage des individus ne tient pas compte des ressources renouvelables ni de la capacité d'accueil de la salle de surveillance post-interventionnelle. Dès lors, à partir d'un individu, lorsque le programme opératoire est construit en considérant la disponibilité des ressources renouvelables et de même que celle des lits de réveil, les temps de fin d'opération peuvent excéder la capacité maximale D_{sd}^M allouée aux salles d'opération. Soit $L_s^\rho(d)$, le nombre quotidien de périodes de temps non utilisées dans la salle s :

$$L_s^\rho(d) = D_{sd}^M - \max_{j,t} \{(t + p_j) \, x_{jsdt}\}.$$

A nouveau, si $L_s^\rho < 0$ le programme opératoire ainsi généré n'est pas réalisable. Le nombre total de périodes de temps dépassant la capacité de la salle s'obtient par :

$$L^\rho = \sum_d \sum_{\substack{s \\ L_s^\rho(d) < 0}} \left| L_s^\rho(d) \right|.$$

Enfin, nous définissons deux autres quantités, L^{ST} et L^{CH}, traduisant respectivement le respect des contraintes de début d'intervention et le respect des contraintes de disponibilité des chirurgiens. Nous les fixons à zéro si leurs contraintes respectives sont satisfaites et à l'infini si ce

n'est pas le cas. Le choix d'une quantité infinie traduit simplement notre volonté de ne pas considérer des solutions violant ces deux types de contraintes. Sur base de ces quatre mesures, nous pouvons à présent décrire la fonction fitness $f(I)$ de l'algorithme.

$$f(I) = \begin{cases} C(I) & \text{si } I \text{ est réalisable} \\ P + L^{\nu} + L^{\rho} + L^{ST} + L^{CH} & \text{sinon,} \end{cases}$$

où $C(I)$ est la valeur objectif de la solution associée à l'individu I (fonction objectif représentée par l'équation (4.1) dans le programme mathématique décrit en section 4.1) et où P représente la borne supérieure de l'objectif. P est calculé en considérant le cas où chaque bloc opératoire est ouvert chaque jour et utilisé à hauteur de sa capacité maximale. Etant donné l'objectif de minimisation poursuivi, un individu sera d'autant meilleur que sa valeur fitness sera faible.

4.3.5 Opérateurs évolutionnaires

Les opérateurs évolutionnaires permettent aux différentes solutions représentées dans la population de converger, au fil des générations, vers des optima. Deux d'entre eux permettent la création de nouvelles solutions. L'opérateur de croisement détermine comment le matériel génétique est échangé entre les parents pour former la progéniture tandis que la mutation permet une altération aléatoire des gènes des enfants.

Opérateur de croisement

Cet opérateur génère deux nouveaux individus enfants, I^F et I^G, sur base du patrimoine génétique de deux individus parents, I^M et I^P, sélectionnés aléatoirement au sein de la population. Dans notre cas, la totalité des POP individus de la population courante sera sujette au croisement.

L'opérateur de croisement nécessite deux entiers q_1 et q_2 pris aléatoirement de sorte que $1 \leq q_1 , q_2 \leq J$. Considérant en premier lieu la création de l'individu Fille, nous procédons comme suit. Les q_1 premiers éléments de δ^F proviennent de l'individu Mère I^M :

$$(j_i^F, d_i^F) := (j_i^M, d_i^M) \quad \forall i = 1, \ldots, q_1.$$

Les $q_1 + 1, \ldots, J$ éléments restants sont obtenus du Père, en s'assurant toutefois qu'une activité héritée de la Mère n'est pas une nouvelle fois considérée :

$$(j_i^F, d_i^F) := (j_k^P, d_k^P) \quad \forall i = q_1 + 1, \ldots, J,$$

où k est le plus petit indice tel que $j_k^P \notin \{j_1^F, \ldots, j_{i-1}^F\}$. En ce qui concerne le second chromosome, les q_2 premières entrées de σ^F sont héritées de la Mère :

$$\sigma^F(j_i) = \sigma^M(j_i) \quad \forall i = 1, \ldots, q_2.$$

Ceci signifie que la salle d'opération s_i^F associée à l'intervention j_i^F est identique à celle de la Mère pour la même intervention j_i^F. La suite du chromosome σ^F est dérivée du Père.

La même démarche est appliquée afin d'obtenir l'individu Garçon I^G, en inversant Père et Mère. A titre illustratif, voici un exemple du fonctionnement de l'opérateur de croisement. Considérons un jeu de données fictives composé d'un jour de planification (c'est-à-dire deux demi-journées), de trois blocs opératoires et de quatre interventions. Supposant que $q_1 = 2$ et $q_2 = 3$, nous obtenons à partir des parents suivants :

$$I^M = \begin{pmatrix} (2,2) & (3,1) & (1,2) & (4,2) \\ 2 & 1 & 3 & 2 \end{pmatrix}, \quad I^P = \begin{pmatrix} (3,2) & (4,1) & (1,1) & (2,1) \\ 1 & 1 & 1 & 3 \end{pmatrix},$$

les enfants suivants :

$$I^F = \begin{pmatrix} (2,2) & (3,1) & (4,1) & (1,1) \\ 2 & 1 & 2 & 1 \end{pmatrix}, \quad I^G = \begin{pmatrix} (3,2) & (4,1) & (2,2) & (1,2) \\ 1 & 1 & 3 & 3 \end{pmatrix}.$$

Comme $q_1 = 2$, les opérations 2 et 3 de la Fille I^F sont effectivement héritées de la Mère I^M, tandis que les opérations 4 et 1 sont prises dans leur ordre d'apparition dans le chromosome δ^P du Père. Ensuite, q_2 valant trois, les salles d'opération des interventions 2, 3 et 4 sont les mêmes salles assignées à ces opérations dans le chromosome σ^M de la Mère ; la salle affectée à l'opération 1 est issue du Père I^P.

Opérateur de mutation

Cet opérateur produit des mutations aléatoires des gènes des enfants. Il s'applique à la totalité d'entre eux. Le chromosome δ d'un individu enfant $I = (\delta, \sigma)$ est modifié en deux étapes : tout d'abord, pour $i = 1, \ldots, J-1$, en intervertissant les opérations j_i et j_{i+1} avec une probabilité $p_{mut} = 0.05$. Notons que les opérations inversées gardent leurs affectations originales, c'est-à-dire leur jour de programmation et leur bloc opératoire. Ensuite, pour $i = 1, \ldots, J$, le chromosome δ est modifié en changeant aléatoirement le demi-jour d'opération d_i parmi les disponibilités du chirurgien qui satisfont les contraintes de début d'opération de l'intervention j_i, à nouveau avec une probabilité p_{mut}. Appliqué au second chromosome, l'opérateur de mutation remplace la salle d'opération s_i par une autre sélectionnée aléatoirement dans $Q_{j_i,:}$, l'ensemble d'unités de soins accessibles à l'intervention j_i, toujours selon la probabilité p_{mut}.

4.3.6 Opérateur de sélection

Il s'agit de l'opérateur sélectionnant les individus qui formeront la population de la génération suivante de l'algorithme. Ici, il consiste simplement à appliquer une méthode de tri élitiste qui ordonne les individus selon leur valeur fitness. L'opérateur ne garde que les POP meilleurs individus pour former la génération suivante. Dès lors, le meilleur programme opératoire s'obtient à partir du meilleur individu de la dernière génération de l'algorithme.

L'algorithme génétique décrit dans cette section s'inspire des travaux de Hartmann (2001) sur l'ordonnancement de projet à ressources contraintes multi-modes. Bien que pouvant paraître simpliste, l'algorithme génétique proposé par Hartmann produit des performances de qualité pour le problème MRCPSP. La programmation opératoire pouvant également s'exprimer comme un problème d'ordonnancement de projet à ressources contraintes, nous avons tout naturellement

orienté nos choix de modélisation génétique en lien avec ceux avancés par Hartmann. Ces choix concernent le codage des individus, la définition des opérateurs de croisement et de sélection, et la valeur de la probabilité de mutation.

4.4 Expérimentations numériques de l'approche méta-heuristique

La section précédente développe un algorithme génétique afin de pallier la complexité du processus de programmation opératoire envisagé, en proposant des plannings opératoires de qualité mais non prouvés optimums. A présent, nous en évaluons les performances sur base des données présentées en section 3.2, en se focalisant essentiellement sur les trois jeux de données pour lesquels l'approche exacte peinait à trouver une solution satisfaisante.

4.4.1 Validation de l'approche

Nous suivons ici la démarche appliquée pour l'évaluation des performances du modèle mathématique en nombres entiers, dans le cas de figure où les salles d'opération sont dédiées, présentée en section 4.2. A savoir, nous faisons varier le nombre d'interventions à programmer et, pour un horizon de temps limité à un unique jour (agencement de seize et dix-neuf opérations), le niveau de consommation de ressources renouvelables. Nous distinguons deux niveaux de consommation de ressources : l'un régulier (noté R) usant des besoins réellement observés, l'autre élevé (noté E) augmentant ces besoins réels d'un facteur proportionnel à la capacité de chaque ressource. Le tableau 4.4 permet la comparaison des performances méta-heuristiques en regard des (quasi-)optima produits par le programme mathématique MIP. Il reprend les valeurs objectifs et temps de calcul du MIP, la taille de la population et le nombre de générations de l'AG (paramètres POP et GEN), la valeur fitness du meilleur individu de la population, l'occurrence de ce meilleur individu durant les différentes réplications de l'expérience et le temps de calcul nécessaire pour atteindre la dernière génération de l'algorithme. Les performances de l'AG dépendent grandement de la valeur des paramètres POP et GEN ; ceux-ci sont fixés a priori de sorte que la population finale offre des solutions satisfaisantes.

Pour rappel, l'utilisation de blocs dédiés implique que chaque intervention ne peut s'effectuer que dans un nombre restreint d'entre eux, tous ces blocs ne présentant pas le même équipement. Dans la modélisation mathématique en nombres entiers, cela se traduit par l'ajout de la contrainte (4.10). A la lecture du tableau 4.4, il apparaît de suite que l'algorithme génétique ne peut rivaliser avec le MIP pour ces cas les plus aisés à résoudre. Cependant, ces cas simplistes que représente la programmation d'une unique journée d'ouverture du quartier opératoire (seize et dix-neuf interventions) nous permettent de valider notre approche génétique. En effet, ils mettent en exergue l'aptitude de l'algorithme génétique à s'approcher voire à atteindre une solution optimale en des temps certes plus long que ceux proposés par le MIP, mais toujours raisonnables.

A titre d'exemple, prenons le cas 19-R. Les paramètres de l'AG y sont fixés à cent individus pour cinquante générations. Sur les vingt réplications de l'expérience, les meilleurs individus produits par l'algorithme ne proposent jamais de solutions excédant l'optimum de plus d'un pourcent. Ces dernières se situent entre 14480 et 14580 euros. La meilleure solution approchée est produite dans une expérience sur quatre. De plus, si l'on élève le niveau de besoin en ressources, le MIP

Nbr.	Cons.	MIP (4.1)–(4.16)		AG			
Op.	Ress.	Obj.	Temps $[s]$	(POP, GEN)	Fit.	Occur. $[\%]$	Temps $[s]$
16	R	12290^*	1	(20,20)	12290	100	4
	E	12290^*	1	(20,20)	12290	100	4
19	R	14430^*	53	(100,50)	14480	25	80
	E	15130	1800	(100,100)	15180	10	162

TAB. 4.4 – *Comparaisons des performances de l'algorithme génétique pour 16 puis 19 interventions à ordonnancer sur un jour.*

ne peut prouver l'optimalité de sa solution endéans la demi-heure qui lui est impartie. La solution proposée par la méthode exacte de résolution s'élève à 15130 euros pour un écart de 1.6% avec la meilleure borne inférieure. Dans les mêmes conditions, l'algorithme génétique produit des solutions de qualité en quelques minutes à peine. Pour les vingt expériences réalisées, les meilleurs individus issus de ce dernier génèrent des solutions comprises entre 15180 et 15330 euros. A nouveau, l'écart entre solution exacte et approchée est très faible et ne dépasse pas 1.3% dans le pire des cas. La meilleure solution génétique est obtenue dans dix pourcents des cas alors que la solution la plus courante, avec près d'une occurrence sur deux, propose un programme opératoire à 15230 euros. L'algorithme génétique permet d'approcher à moins de 0.6% la solution issue du modèle mathématique dans plus d'une simulation sur deux pour le cas 19-E, ce qui nous conforte dans le choix de l'approche méta-heuristique adoptée. Le principal inconvénient de l'algorithme génétique, comme des autres grandes méta-heuristiques, vient du choix des valeurs à appliquer aux paramètres qui conditionnent ses performances (ici, essentiellement le nombre d'individus et le nombre de générations). La section suivante relate cet état de fait et se veut une aide à la paramétrisation de l'AG développé dans ce travail.

4.4.2 Paramétrisation de l'algorithme génétique

Les performances de l'algorithme génétique sont directement liées aux valeurs des paramètres qui le caractérisent. Les paramètres les plus influents sont la taille de la population POP et le nombre de générations GEN de l'algorithme. De la bonne paramétrisation de l'algorithme dépend la qualité des individus constituant la population finale. Or, sans connaissance a priori du problème étudié, il est délicat de déterminer quelles valeurs du couple (POP, GEN) satisferont aux exigences du gestionnaire de blocs. Nous poursuivons donc cette section de résultats par une étude paramétrique destinée à clarifier le choix des valeurs à assigner aux nombres POP et GEN. Cette étude se base sur les données relatives à la semaine d'exploitation comportant 80 opérations, la charge de travail hebdomadaire moyenne, et constitue un bon indicateur pour les autres instances à disposition.

Le tableau 4.5 illustre les performances obtenues par l'algorithme génétique appliqué sur une instance de taille réelle du problème traité. Il permet de se faire une idée de l'évolution de la qualité des individus lorsque leur nombre au sein de la population de même que le nombre de générations de l'algorithme varient. Le jeu de données utilisé comporte 80 opérations à planifier sur une semaine. Le calendrier opératoire effectivement appliqué lors de la semaine étudiée, et dressé manuellement, propose une solution à 66440 euros. Le tableau 4.5 reprend la valeur fitness du meilleur individu de la population finale de l'algorithme génétique, pour une unique expérience.

POP	GEN			
	500	1000	2000	5000
10	61820 (-6.9%)	59580 (-10.3%)	57580 (-13.3%)	55690 (-16.2%)
	305 s	*608 s*	*1232 s*	*3087 s*
20	59030 (-11.1%)	56000 (-15.7%)	56740 (-14.6%)	53310 (-19.8%)
	680 s	*1358 s*	*2739 s*	*1.8 h*
50	57240 (-13.8%)	55490 (-16.5%)	55890 (-15.9%)	55440 (-16.6%)
	2074 s	*1.15 h*	*2.3 h*	*5.75 h*
100	58880 (-11.4%)	56040 (-15.6%)	55840 (-15.9%)	55740 (-16.1%)
	1.6 h	*3.06 h*	*6.04 h*	*15.17 h*

TAB. 4.5 – *Illustration des performances de l'algorithme génétique en fonction des paramètres POP et GEN. L'algorithme est ici appliqué à la planification hebdomadaire de 80 interventions sur 7 blocs opératoires dédiés.*

Les chiffres entre parenthèses représentent l'écart entre le programme opératoire dressé manuellement et la solution issue du meilleur individu produit par l'algorithme. Les chiffres en italique mentionnent les temps de calcul nécessaires pour obtenir ces résultats.

De manière générale, nous observons que plus l'algorithme compte de générations, plus l'on peut s'attendre à ce que les solutions générées soient proches de l'optimum. Néanmoins, augmenter le nombre de générations tend à réduire la diversité au sein de la population. La taille de la population semble également avoir un impact bénéfique sur la convergence vers l'optimum. Cependant, cette tendance paraît comporter un seuil puisque passer de cinquante à cent individus n'influe pas sur la qualité de ces derniers. Selon le temps alloué à la procédure, une population comportant entre vingt et cinquante individus semble être un bon compromis. Bien entendu, plus les valeurs des paramètres POP et GEN seront élevées, plus l'application nécessitera de temps pour aboutir à la génération finale.

4.4.3 Evaluation des performances pour la programmation opératoire hebdomadaire

Lorsqu'il s'agit de programmer une semaine complète d'ouverture du quartier opératoire, l'approche méta-heuristique prend tout son sens. Ceci constitue bien entendu le type d'instance le plus complexe que nous ayons à traiter. Complexité qui se traduit au niveau du modèle en nombres entiers par des temps de calcul potentiellement trop élevés pour une utilisation en pratique. En effet, pour un tiers des instances le MIP nécessite au moins six heures pour obtenir une solution optimale ou quasi-optimale. Dans les cas extrêmes, à savoir pur 80 et 85 opérations, le modèle en nombres entiers réclame plus d'une heure pour déterminer une première solution réalisable. Ces temps excessifs ne permettent pas de réactivité en cas de changement de dernière minute.

Sur base du tableau 4.5, nous fixons les paramètres de l'algorithme génétique à 10 individus et 5000 générations. Le tableau 4.6 affiche les performances de l'AG en termes de meilleur individu issu de dix expérimentations de l'algorithme, et en termes de meilleur individu moyen reprenant les moyennes et écarts-types (présentés entre parenthèses) des valeurs objectifs et des temps de calcul du meilleur individu de chacun de ces dix tests. Il les compare ensuite aux performances

Nbr. Op.	MIP (4.1)–(4.16) (Gap $< 1\%$)		AG			
			Meilleur Ind.		Meilleur Ind. Moy.	
	Obj. [€]	Tps [s]	Obj. [€]	Tps [s]	Obj. [€]	Tps [s]
60	37280 (0.7%)	*26*	38720	*2354*	41415 (1302)	*2389 (18)*
70	50130 (0.7%)	*43*	51020	*2690*	53975 (1874)	*2684 (14)*
72	55940 (0.1%)	*479*	55940	*2941*	57687 (1271)	*2933 (9)*
80	51780 (0.7%)	*5.9 h*	53310	*2998*	56546 (2122)	*3024 (15)*
85	58490 (0.9%)	*7064*	59140	*3192*	61965 (2024)	*3203 (17)*
86	46360 (0.8%)	*449*	48670	*3240*	49693 (755)	*3237 (16)*
88	50780 (0.8%)	*917*	53050	*3285*	55154 (1252)	*3299 (18)*
91	52320 (3.0%)	*24 h*	50970	*3367*	52441 (1260)	*3376 (13)*
100	57850 (0.3%)	*3583*	60370	*3744*	61319 (899)	*3756 (17)*

TAB. 4.6 – *Comparaison des performances de l'algorithme génétique appliqué à la programmation hebdomadaire d'un quartier opératoire composé de blocs dédiés (pour 10 individus et 5000 générations), en regard des solutions produites par la formulation en nombres entiers* (4.1)–(4.16) *limitée à 1% de gap.*

du MIP dans son utilisation heuristique, c'est-à-dire pour un niveau de gap fixé à 1%. Ce tableau nous indique tout d'abord que, à l'inverse de l'approche exacte, l'approche méta-heuristique offre des performances calculatoires constantes dépendant de la valeur de ses paramètres. En moins d'une heure, et alors même que les branchements du solveur MIP peuvent ne pas avoir abouti à une seule solution réalisable, l'algorithme génétique produit des solutions de qualité. Globalement, l'individu moyen est distant de 6.5% des solutions quasi-optimales produites par le MIP. Cet écart tombe à 2% si l'on ne considère que les meilleurs individus. Qui plus est, l'AG égale le MIP pour 72 opérations et améliore la solution du MIP de plus de 2% pour 91 opérations. Si l'on ne considère que les trois instances justifiant pleinement l'utilisation d'une approche méta-heuristique (comportant respectivement 80, 85 et 91 interventions), l'écart moyen passe à 5% alors que la distance entre meilleur individu et MIP est inférieure à 1%, pour des temps de calcul deux à vingt-quatre fois moindres. Il va sans dire que, dans un contexte de salles dédiées, ces chiffres nous poussent à affirmer que l'approche génétique proposée est efficiente, proposant rapidement des solutions proches de l'optimum permettant un gain financier substantiel. En comparaison, rappelons que le gestionnaire de blocs réserve l'entièreté de son vendredi pour construire manuellement le planning opératoire de la semaine suivante.

Au regard de ces expériences numériques, nous pouvons affirmer que l'usage d'une approche méta-heuristique dans le cas de blocs opératoires dédiés se justifie pour des instances de grande taille contenant plusieurs jours à planifier. En effet, pour les problèmes de plus petites tailles, l'approche exacte se montre imbattable sur l'aspect temps de calcul. Elle met à disposition des programmes opératoires (quasi-) optimums en des temps très courts. Les limites de l'application pratique du MIP s'observent pour certaines instances prévoyant la construction du planning de la semaine à venir. Le modèle linéaire nécessite alors au minimum une heure de temps pour aboutir à une solution réalisable relativement distante de l'optimum. A l'inverse et pour autant que sa paramétrisation soit adéquate, l'algorithme génétique est toujours apte à produire des solutions de qualité pour un temps de calcul maîtrisable. Plus le problème traité est complexe, plus l'algorithme nécessite de générations pour rester efficient. Néanmoins, le tableau 4.6 nous prouve

que l'AG permet d'obtenir des programmes opératoires hebdomadaires de qualité en moins d'une heure de temps. Ceci offre la possibilité au gestionnaire de recommencer autant de procédures de planification que nécessaire, ce qu'il est impossible de faire manuellement. Enfin, contrairement au programme linéaire en nombres entiers qui n'offre qu'une unique solution, soit-elle optimale, l'algorithme génétique a l'avantage de produire un ensemble de solutions de qualité. Population de solutions au sein de laquelle le gestionnaire choisit d'appliquer celle qui lui semble la plus pertinente. Il est par contre délicat de déterminer a priori quels seront les jeux de données qui poseront problème au programme linéaire et qui nécessiteront l'emploi de l'algorithme génétique.

En conclusion, ce chapitre développe deux approches de résolution de philosophies différentes pour une même problématique : optimiser le processus de programmation du quartier opératoire lorsque celui-ci est composé de salles d'intervention dites dédiées. La première est une approche de résolution qualifiée d'exacte qui vise la détermination d'une solution optimale du problème étudié ; elle prend les traits d'un programme d'optimisation linéaire en nombres entiers mixtes. Les enseignements que nous avons pu tirer des expériences numériques de la section 4.2 privilégient l'utilisation de cette modélisation mathématique pour de petites instances du problème. Celle-ci s'est en effet révélée intraitable sur l'ordonnancement d'une unique journée d'ouverture du quartier opératoire, fût-elle une journée fort chargée. La seconde est une approche dite heuristique, ou approchée, qui s'attelle à produire un ensemble de solutions de qualité, mais non nécessairement optimales. Parmi les techniques méta-heuristiques dédiées à l'optimisation de problèmes complexes, nous avons opté pour un algorithme génétique pour sa faculté à produire une population de solutions. Ne pas prouver l'optimalité des solutions produites confère à cette approche heuristique un net avantage calculatoire sur les plus larges instances du problème. Les résultats présentés dans cette section mettent l'accent sur ses capacités à produire rapidement des plannings opératoires hebdomadaires à valeurs objectifs satisfaisantes. Quel que soit le cas de figure, nous avons donc à disposition des outils efficients de programmation opératoire de blocs dédiés. Cependant, qu'advient-il des ces techniques lorsque les blocs opératoires ne sont pas dédiés mais multidisciplinaires et donc polyvalents, comme l'on peut majoritairement l'observer dans la littérature ? Comme nous l'avons déjà mentionné, considérer des salles d'opération identiques rend le problème fortement symétrique. Certes, les deux approches proposées dans ce chapitres restent applicables dans de telles conditions, au prix d'un accroissement des temps de calcul, mais il est préférable de tirer profit de cette symétrie afin de limiter son impact sur les performances calculatoires nécessaires. Le chapitre suivant expose donc les techniques mises en œuvre afin de résoudre efficacement la programmation opératoire dans sa version symétrique.

Chapitre 5

Méthodes de résolution pour des blocs opératoires polyvalents

La revue de la littérature portant sur la programmation opératoire, présentée au chapitre 2, nous révèle que ce processus est presque exclusivement abordé en le scindant en deux étapes séquentielles, l'une déterminant le jour et éventuellement la salle de chaque intervention, l'autre assignant aux opérations devant être réalisées une même journée leur heure de début. La question sous-jacente, et l'une des motivations premières de cette recherche, est de savoir s'il est possible d'intégrer ces deux étapes composant le processus de programmation opératoire de façon efficiente sur des instances de taille réelle ? Le chapitre précédent nous a montré que le problème ainsi défini est complexe au sens algorithmique du terme et nécessite l'application d'une approche heuristique pour proposer des solutions de qualité aux instances les plus difficiles en un temps raisonnable. Ces conclusions ne valent évidemment que si l'on considère un quartier opératoire constitué de blocs dédiés, c'est-à-dire dont l'accessibilité est restreinte à certaines spécialités chirurgicales. Par contre, la nature pluridisciplinaire des salles d'opération polyvalentes considérées dans ce chapitre, alliée à une bonne gestion de la symétrie qui en découle, vont permettre à l'approche exacte décrite ici de répondre par l'affirmative à notre question de recherche. Cette approche exacte se compose ici de deux formulations en nombres entiers mixtes qui se différencient par la manière de traiter la symétrie inhérente au problème supposant des salles identiques et dont la seconde se montrera particulièrement efficiente sur l'ensemble des données à disposition.

La distinction entre blocs opératoires dédiés et polyvalents peut sembler anodine mais modifie en réalité radicalement le problème. En effet, l'alternative polyvalente apporte, comme mentionné ci-dessus, de la symétrie au problème. Dès lors, les approches de résolution utilisées vont différer selon qu'elles puissent tirer parti de ce caractère symétrique ou pas. Une méthode classique appliquée sur le problème symétrique risque fortement de se montrer moins efficiente que dans la version du problème comportant des salles dédiées, ce qui a effectivement été observé en pratique. Par contre, contrer cette symétrie par une approche spécifiquement développée dans ce but peut permettre à cette dernière d'être bien plus performante que son homologue classique. C'est pourquoi, dans la suite de ce document, nous distinguerons d'une part les approches de résolution à utiliser lorsque l'on considère des blocs opératoires dédiés, et d'autre part les techniques se montrant plus efficientes lorsque ces blocs sont polyvalents.

5.1 Salles polyvalentes et symétrie

L'objectif de ce travail est, jusqu'à présent, d'aider les gestionnaires hospitaliers à mieux programmer les interventions à réaliser au sein du quartier opératoire sur un court horizon de temps (typiquement une semaine). Comme mentionné précédemment, la décomposition de ce processus en deux sous-problèmes peut affecter la qualité des solutions produites étant donné l'interaction évidente qu'il existe entre ces deux étapes. Raison pour laquelle les approches développées dans le cadre de cette recherche s'évertuent à intégrer planification et ordonnancement des activités chirurgicales en une même procédure. Ce chapitre est consacré au développement d'une telle procédure lorsque les salles d'opération sont présumées identiques et donc polyvalentes. Cette similitude des blocs opératoires rend le problème fortement symétrique. L'application de l'approche exacte décrite en section 4.1 nécessite dès lors quelques ajustements afin de préserver son efficacité pour tenir compte de cette caractéristique.

5.1.1 Ajout de contraintes pour briser la symétrie

Un programme linéaire en nombres entiers est *symétrique* si ses variables peuvent être permutées sans changer la structure du problème. Ce type de programme symétrique est fréquent en combinatoire ou en optimisation et couvre des domaines aussi variés que la coloration de graphes, la construction de codes ou, en ce qui nous concerne, l'ordonnancement de tâches sur des machines parallèles. Ce genre de problème est généralement difficile à résoudre par les traditionnels algorithmes *branch-and-bound* dédiés à la résolution de modèles mathématiques en nombres entiers (voir Wolsey, 1998, pour une introduction à ces techniques). L'isomorphisme de nombreux sous-problèmes dans l'arbre d'énumération en est la cause et engendre des efforts de calcul inutiles.

Dans le problème qui nous occupe, relaxer les contraintes (4.10) d'accessibilité des blocs opératoires nous permet de planifier les opérations sans restriction sur ces blocs. En d'autres mots, ces derniers sont considérés comme polyvalents ; ils peuvent, par conséquent, traiter tout type de chirurgie. Le problème s'en trouve moins contraint, accroissant de la sorte l'espace de solutions. Le caractère polyvalent des salles d'opération offre donc plus de flexibilité dans la construction des programmes opératoires, mais potentiellement au prix d'une hausse des temps de calcul, ce malgré un problème résoluble plus facilement. Ce serait certainement le cas si nous n'avions porté une attention toute particulière à supprimer la symétrie induite par de telles salles polyvalentes. En effet, considérer la multifonctionnalité des salles d'intervention génère une grande symétrie dans le problème pour la raison suivante : comme les salles d'opération ne sont plus différenciées, intervertir les plannings opératoires de deux d'entre elles n'influe ni sur la faisabilité d'une solution, ni sur sa valeur objectif. La planification du quartier opératoire s'apparente dès lors à de l'ordonnancement sur machines parallèles identiques. Chaque programme opératoire existe donc à $S!$ permutations près, S étant le nombre de blocs opératoires. Si des mesures n'étaient pas prises en conséquence, nous observerions immanquablement une redondance des solutions et une perte d'information lors des branchements de l'algorithme de résolution du solveur (type *branch-and-bound*). Le nombre de branchements et de nœuds visités par cet algorithme s'en trouverait largement accru, ce qui ne manquerait pas de se répercuter sur les temps de calcul. Or, il existe toute une littérature sur le traitement de la symétrie dans la programmation linéaire en nombres entiers (voir Margot, 2008, pour une bonne introduction à cette problématique), et qui propose diverses approches pour en

limiter les effets. Parmi les techniques proposées pour exploiter la symétrie d'un problème, nous en avons retenu deux : fixer des variables et introduire des inégalités brisant cette symétrie.

Fixer des variables. Il s'agit d'une idée assez simple pour réduire la symétrie dans un programme en nombres entiers. Cela nécessite toutefois une bonne connaissance des propriétés du problème traité. Pour la planification du quartier opératoire, cette démarche est relativement implicite. En effet, le quartier opératoire est, par nature, une activité critique de la chaîne de création de soins de santé. De nombreuses interventions y sont réalisées quotidiennement. Or, si l'on considère des blocs opératoires polyvalents, leurs spécifications propres n'importent plus à l'accomplissement des chirurgies. Dès lors, en fonction du nombre moyen d'opérations à réaliser par jour, il est possible de fixer un certain nombre de salles immanquablement ouvertes puisque nécessaires à l'achèvement de ces opérations. Les salles étant identiques, il n'est plus nécessaire de les distinguer ; nous pouvons donc fixer les salles ouvertes en permanence en toute liberté. Nous supposons donc un certain nombre $\beta \leq S$ de salles ouvertes chaque jour de la semaine afin d'absorber l'activité journalière moyenne du quartier opératoire. De plus, imposer un nombre de blocs opératoires ouverts de façon récurrente chaque jour de la semaine permet de lisser la charge de travail de chacune de ces journées.

Introduire de nouvelles inégalités. Une façon naturelle d'éviter les désagréments liés à la symétrie d'un problème consiste à ajouter des inégalités qui vont briser cette symétrie. L'introduction de telles inégalités peut se faire de deux manières différentes. Une première approche génère ces inégalités durant le processus de résolution. Il s'agit d'inégalités contrant la symétrie dynamiquement. Ces inégalités peuvent ne pas être valides dans la formulation initiale du problème, mais s'avérer utiles au fil des itérations, sans pour autant empêcher l'aboutissement à une solution optimale. L'alternative à l'introduction dynamique d'inégalités insère celles-ci dans la formulation initiale du modèle. On parle alors d'inégalités contrant statiquement la symétrie. Elle permet de supprimer certaines des solutions symétriques tout en gardant au moins une solution optimale. C'est l'option que nous avons appliquée. Dans notre cas, la symétrie vient des S blocs opératoires polyvalents qui agissent en parallèle. Or, nous savons qu'un nombre β fixe les blocs ouverts chaque jour en permanence. Nous pouvons alors agir sur les $S-\beta$ blocs restant en restreignant leur ouverture. En effet, il n'est pas utile de considérer toutes les salles ouvertes lors de la construction du programme opératoire si toutes ces salles sont identiques. Par contre, il est plus intéressant de n'ouvrir une salle que si nécessaire, quitte à ré-optimiser le planning en considérant cette salle supplémentaire. Les salles étant identiques, seul leur identifiant les caractérise. Nous pouvons donc contrer la symétrie du problème en stipulant qu'une salle ne peut être utilisée que si la précédente l'est déjà. Par exemple, considérons la situation où les β blocs sont remplis et où le programme opératoire nécessite l'ouverture d'une autre unité chirurgicale. Ajouter de telles inégalités évite alors à l'algorithme de résolution de perdre du temps dans la comparaison de deux solutions totalement identiques dont la première agence les opérations supplémentaires dans la salle $\beta + 1$ et la seconde dans la salle $\beta + 2$.

Partant de la formulation pour blocs opératoires dédiés proposée en section 4.1, nous traitons le cas de quartiers opératoires composés de salles d'opérations polyvalentes en y ajoutant des contraintes spécifiques tenant compte de la symétrie que ces dernières engendrent. Dès lors, la programmation opératoire de blocs polyvalents telle qu'exposée au chapitre 3 se résout à l'aide

du programme en nombres entiers mixtes décrit par les équations (5.1) à (5.17) suivantes. Les contraintes (5.12) et (5.13) de symétrie y remplacent les contraintes (4.10) d'accessibilité des salles. Rappelons que les indices j, s, d et t symbolisent respectivement les interventions à pratiquer, les salles d'opération, les jours de planification et les périodes de temps opératoire. Les variables binaires x_{jsdt} spécifient quand une opération débute, les variables binaires z_{sd} indiquent si une salle est ouverte et les variables entières l_{sd} comptent le nombre de périodes de temps utilisées en heures supplémentaires. La fonction objectif (5.1) vise la minimisation des coûts opératoires :

$$\min_{z,l} \sum_{s} \sum_{d} [\, C^{open}\, z_{sd} + C^{over}\, l_{sd}\,], \tag{5.1}$$

sujette, pour des blocs opératoires polyvalents, aux contraintes (5.2)–(5.17) décrites dans le tableau 5.1.

Les contraintes de fonctionnement décrites ici étant scrupuleusement identiques aux contraintes (4.2)–(4.16) déjà présentées au chapitre précédent, nous nous contentons de ne commenter que les équations destinées à traiter la symétrie du problème. Les égalités (5.12) fixent z_{sd} à la valeur un pour les β blocs opératoires systématiquement ouverts quotidiennement. Les inégalités (5.13), quant à elles, ne s'appliquent qu'aux $S - \beta$ salles non concernées par les contraintes (5.12), et imposent que chacune d'entre elles ne puisse être utilisée que si la précédente l'est déjà.

5.1.2 Expérimentations numériques

Dans cette section, nous appliquons le programme mathématique détaillé ci-dessus sur les neuf jeux composant notre base de données (voir section 3.2 pour la description des données), retraçant chacun l'activité hebdomadaire réellement observée dans un quartier opératoire de Bruxelles. Rappelons que ce MIP est modélisé en AMPL et résolu à l'aide du solveur CPLEX. Les résultats de ces expérimentations vont nous donner un second élément de réponse, après celui du chapitre précédent, quant à la possibilité d'intégrer de façon efficiente les étapes du processus étudié.

Conformément aux données dont nous disposons, l'activité chirurgicale journalière est en moyenne de 16 interventions et nécessite un temps opératoire total moyen de 160 périodes de temps. Le nombre minimum de salles nécessaires à la réalisation de cette charge de travail est de quatre, de telle sorte que $\beta = 4$ dans les contraintes de symétrie (5.12) et (5.13). Les performances de cette première formulation mathématique, dans laquelle au moins quatre unités chirurgicales sont ouvertes chaque jour, sont présentées dans le tableau 5.2. Le coût du programme opératoire tel qu'effectivement construit par le gestionnaire du quartier opératoire apparaît, pour chaque instance, dans la colonne *Programme Effectif*. La différence entre les solutions optimales produites par le MIP (valeurs étoilées dans le tableau) et le programme effectif est présentée entre parenthèses dans la colonne *Valeur Objectif*. Tout d'abord, nous observons qu'une instance comportant une semaine d'activité peut maintenant être résolue à l'optimum en moins de deux heures en utilisant le modèle décrit dans cette section. En fonction de la taille de l'instance, les temps de calcul varient entre dix minutes et approximativement deux heures, ce qui représente une avancée considérable par rapport aux résultats du chapitre précédent. Ces temps sont considérés comme raisonnables étant donné la journée actuellement nécessaire au gestionnaire du quartier opératoire pour effectuer cette même tâche. En effet, ce dernier réserve généralement son vendredi afin d'établir le calendrier opératoire de la semaine à venir. Comparer les solutions produites par le MIP et

$$\sum_s \sum_d \sum_t x_{jsdt} \;=\; 1, \qquad \forall j, \tag{5.2}$$

$$\sum_j \sum_s \left[r^{\rho}_{jk} \sum_{\substack{\tau=t-p_j+1 \\ \tau>0}}^{t} x_{jsd\tau} + \Gamma^{\rho}_{jk} \sum_{\substack{\tau=t-\lambda_{jk}+1 \\ \tau>0}}^{t} x_{jsd\tau} \right] \;\leq\; R^{\rho}_{kd}, \qquad \forall d, \forall t, \forall k \in \mathcal{K}^{\rho}, \tag{5.3}$$

$$\sum_j \sum_s \sum_{\substack{\tau=t-(p_j+b_j)+1 \\ \tau>0}}^{t-p_j} x_{jsd\tau} \leq R^{B}, \qquad \forall d, \forall t, \tag{5.4}$$

$$\sum_j \sum_s \sum_t r^{\nu}_{jk}\, x_{jsdt} \;\leq\; R^{\nu}_{kd}, \qquad \forall k \in \mathcal{K}^{\nu}, \forall d, \tag{5.5}$$

$$x_{jsdt} = 0, \forall c, \forall j \in O(c), \forall s, \forall d | R^{C}_{c,2d-1} = 0, \forall t \leq \frac{T}{2}, \tag{5.6}$$

$$x_{jsdt} = 0, \;\forall c, \forall j \in O(c), \forall s, \forall d | R^{C}_{c,2d} = 0, \forall t > \frac{T}{2}, \tag{5.7}$$

$$\sum_s \sum_{j \in O(c)} \sum_{\substack{\tau=t-p_j+1 \\ \tau>0}}^{t} x_{jsd\tau} \;\leq\; 1, \qquad \forall t, \forall d, \forall c, \tag{5.8}$$

$$\sum_j \sum_{\substack{\tau=t-p_j+1 \\ \tau>0}}^{t} x_{jsd\tau} \;\leq\; z_{sd}, \qquad \forall s, \forall d, \forall t, \tag{5.9}$$

$$\sum_s \sum_t x_{jsdt} = 0, \qquad \forall j, \forall d \notin [ES_j, LS_j], \tag{5.10}$$

$$x_{jsdt} = 0, \qquad \forall j, \forall s, \forall d, \forall t | (t + p_j) > D^{M}_{sd}, \tag{5.11}$$

$$z_{sd} = 1, \qquad \forall d, \; s = 1, \ldots, \beta, \tag{5.12}$$

$$z_{sd} \geq z_{s+1,d}, \qquad \forall d, \; s = \beta, \ldots, S-1, \tag{5.13}$$

$$\sum_j \sum_t p_j\, x_{jsdt} \leq D^{M}_{sd}\; z_{sd}, \qquad \forall s, \forall d, \tag{5.14}$$

$$l_{sd} \geq \sum_j \sum_{\substack{t=D^{N}_{sd}-p_j+1 \\ t>0}}^{T-p_j} \left(t + p_j - D^{N}_{sd}\right)\; x_{jsdt}, \;\forall s, \forall d, \tag{5.15}$$

$$l_{sd} \in \mathbb{N}, \qquad \forall s, \forall d, \tag{5.16}$$

$$x_{jsdt}, \; z_{sd} \in \{0, 1\}, \qquad \forall j, \forall s, \forall d, \forall t. \tag{5.17}$$

TAB. 5.1 – *Modélisation des contraintes liées au fonctionnement du quartier opératoire composé de salles polyvalentes.*

Nbr	Programme	MIP (5.1)–(5.17)	
Op.	Effectif [€]	Objectif [€]	Temps[s]
60	55140	40800* (-26 %)	604
70	56230	45600* (-19 %)	1791
72	68630	44600* (-35 %)	2724
80	66440	49200* (-26 %)	1885
85	68530	50750* (-26 %)	3352
86	62050	45050* (-27 %)	1724
88	62600	45650* (-27 %)	6948
91	63950	48500* (-24 %)	3596
100	61160	49600* (-19 %)	6063

TAB. 5.2 – *Performances du MIP* (5.1)–(5.17) *conçu avec quatre blocs opératoires ouverts chaque jour ($\beta = 4$), appliqué à la programmation hebdomadaire d'un quartier opératoire composé de blocs polyvalents.*

les programmes établis par le gestionnaire suggère que notre approche est efficiente. Les frais de fonctionnement sont réduits d'au moins 19% et d'au plus 35%, ce qui peut représenter d'importantes économies pour l'hôpital. Ces premiers résultats nous permettent d'apporter un nouvel élément de réponse à la question : est-il possible d'intégrer de manière efficiente les étapes de planification et d'ordonnancement constituant le processus de programmation opératoire ? Le tableau 5.2 nous indique que, pour des blocs opératoires polyvalents, la réponse semble également être *oui* !

Le principal avantage du modèle composé des équations (5.1) to (5.17) est de lisser la charge de travail sur la semaine. En effet, contraindre l'approche à utiliser quotidiennement au moins un nombre β de salles d'opération assure que l'activité opératoire est distribuée équitablement dans la semaine, pour autant que β soit suffisamment grand (comme c'est le cas lorsqu'il est fixé en fonction de la charge de travail moyenne). Cependant, bien que nous ayons montré son aptitude à résoudre des instances réelles efficacement, cette approche n'est pas sans défaut. Son principal désavantage est précisément de contraindre le programme à fixer un nombre β de salles systématiquement ouvertes. La question qui s'ensuit est de savoir comment choisir la valeur adéquate pour β. En effet, le nombre β conditionne les performances du modèle mathématique en termes de valeur objectif et de temps de calcul. Dans notre cas, fixer $\beta = 4$ sur base de l'activité opératoire moyenne peut notamment ne pas être approprié pour les instances comportant moins d'interventions à programmer. Cela oblige le MIP à ouvrir plus de blocs opératoires que nécessaire, comme nous le verrons par la suite. Par conséquent, la valeur objectif obtenue pour un tel problème peut en pâtir et ne pas représenter l'optimum global réel. Plus précisément, cette solution sera bel et bien l'optimum mais d'un problème légèrement différent de celui présenté au chapitre 3. Par contre, fixer $\beta = 3$ n'affecte plus autant la qualité de la valeur objectif mais agit négativement sur les temps de calcul. A titre d'exemple, la plus petite instance ne comportant que 60 opérations requiert, dans ces conditions, plus de mille secondes pour être résolue à l'optimum (en lieu et place de six cents lorsque $\beta = 4$), optimum réduisant la valeur objectif présentée dans le tableau 5.2 de presque 10%. Ce constat empire si nous appliquons le modèle sur la seconde instance du même tableau puisqu'alors plus d'une journée complète est nécessaire pour trouver (ou prouver) la solution optimale.

Prendre en compte la symétrie engendrée par des salles polyvalentes permet donc de considérablement décroître les temps de calcul par rapport à l'approche exacte du chapitre traitant des salles dédiées. Néanmoins, les performances en termes de valeur objectif et de temps de calcul de ce premier modèle symétrique, décrit par les équations (5.1) à (5.17), paraissent trop sensibles à la valeur du paramètre β. Sachant cet inconvénient, cette manière de traiter la symétrie tend à être moins efficace qu'attendu. En effet, nous souhaitons proposer au gestionnaire un outil capable de l'aider dans la programmation du quartier opératoire. Cet outil se doit de fournir au décideur des plannings optimums ou quasi-optimums en des temps acceptables. Pour cette raison, nous considérons une autre manière de traiter la symétrie du problème et développons un nouveau modèle mathématique dans la section suivante.

5.2 Reformulation MIP du problème

Le principal défaut de la démarche ci-dessus vient de la nécessité de contraindre β blocs opératoires à être quotidiennement ouverts. Nous devons donc repenser notre approche face à la symétrie et trouver une manière plus efficace de limiter le rôle que cette dernière joue dans les performances calculatoires du MIP. Constatant que la plupart des contraintes (5.2) à (5.17) somment sur l'indice s, et étant donné que les salles ne peuvent être différenciées, l'idée ici est de supprimer de la formulation précédente l'indice s qui les identifie. Cela va diminuer de façon drastique le nombre de variables et de contraintes du modèle. Cette nouvelle approche de modélisation produira des solutions agrégées satisfaisant toutes les contraintes de fonctionnement mais ne déterminant que le jour et l'heure de chaque intervention. Par conséquent, nous serons forcés de développer un algorithme capable de désagréger les solutions produites par le programme linéaire en nombres entiers mixtes, et donc d'assigner une salle à chaque traitement chirurgical.

5.2.1 Redéfinition des variables de décision

Le problème que nous étudions dans cette section reste celui décrit au chapitre 3 mais les notations utilisées sont ajustées afin de tenir compte de la nouvelle approche de modélisation. La nouvelle formulation ne contient que les trois indices j, d et t décrits précédemment. Le modèle révisé inclut deux types de variables : les variables binaires x_{jdt} valant un si l'opération j débute à la période t du jour d, et les variables entières positives y_d représentant le nombre de salles d'opération utilisées durant le jour d. Outre la définition de nouvelles variables, cette nouvelle approche de modélisation impose d'autres remaniements. En effet, comme il ne considère plus les blocs opératoires, ce second modèle mathématique ne suscite plus l'introduction de contraintes brisant la symétrie du problème telles que les équations (5.12) and (5.13). En effet, cette caractéristique est à présent inhérente à la description des variables de décision puisque l'indice s n'y apparaît plus. La symétrie est tout simplement éradiquée du problème dès lors que l'on en ignore les salles. De même, les contraintes liées à l'utilisation des salles, comme interdire le chevauchement d'opérations dans celles-ci, modélisé par la contrainte (5.9), sont maintenant superflues. Ces contraintes seront en effet prises en compte par l'algorithme qui reconstruit le programme opératoire à partir de la solution agrégée du MIP. Par contre, il est à présent nécessaire de vérifier que le nombre de blocs opératoires utilisés chaque jour n'excède pas le nombre disponible. Le nouveau programme linéaire en nombres entiers mixtes est décrit par les équations (5.18) à (5.32).

$$\sum_d \sum_t x_{jdt} \;=\; 1, \qquad \forall j, \tag{5.19}$$

$$\sum_j \left[r^{\rho}_{jk} \sum_{\substack{\tau=t-p_j+1 \\ \tau>0}}^{t} x_{jd\tau} + \Gamma^{\rho}_{jk} \sum_{\substack{\tau=t-\lambda_{jk}+1 \\ \tau>0}}^{t} x_{jd\tau} \right] \;\leq\; R^{\rho}_{kd}, \qquad \forall d, \forall t, \forall k \in \mathcal{K}^{\rho}, \tag{5.20}$$

$$\sum_j \sum_{\substack{\tau=t-(p_j+b_j)+1 \\ \tau>0}}^{t-p_j} x_{jd\tau} \leq R^{B}, \qquad \forall d, \forall t, \tag{5.21}$$

$$\sum_j \sum_t r^{\nu}_{jk}\, x_{jdt} \;\leq\; R^{\nu}_{kd}, \qquad \forall k \in \mathcal{K}^{\nu}, \forall d, \tag{5.22}$$

$$x_{jdt} = 0, \quad \forall c, \forall j \in O(c), \forall d | R^{C}_{c,2d-1} = 0, \forall t \leq \frac{T}{2}, \tag{5.23}$$

$$x_{jdt} = 0, \quad \forall c, \forall j \in O(c), \forall d | R^{C}_{c,2d} = 0, \forall t > \frac{T}{2}, \tag{5.24}$$

$$\sum_{j \in O(c)} \sum_{\substack{\tau=t-p_j+1 \\ \tau>0}}^{t} x_{jd\tau} \;\leq\; 1, \qquad \forall t, \forall d, \forall c, \tag{5.25}$$

$$\sum_t x_{jdt} = 0, \qquad \forall j, \forall d \notin [ES_j, LS_j], \tag{5.26}$$

$$x_{jdt} = 0, \qquad \forall j, \forall d, \forall t | (t + p_j) > D^{M}_{d}, \tag{5.27}$$

$$\sum_j \sum_{\substack{\tau=t-p_j+1 \\ \tau>0}}^{t} x_{jd\tau} \;\leq\; y_d, \qquad \forall d, \forall t, \tag{5.28}$$

$$y_d \leq S, \qquad \forall d, \tag{5.29}$$

$$\sum_j \sum_t p_j\, x_{jdt} \;\leq\; D^{M}_{d}\, y_d, \qquad \forall d, \tag{5.30}$$

$$y_d \in \mathbb{N}, \qquad \forall d, \tag{5.31}$$

$$x_{jdt} \in \{0, 1\}, \qquad \forall j, \forall d, \forall t. \tag{5.32}$$

TAB. 5.3 – *Modélisation des contraintes liées au fonctionnement du quartier opératoire composé de salles polyvalentes après redéfinition des variables de décision.*

L'objectif d'optimisation (5.18) reste la minimisation des coûts du quartier opératoire :

$$\min_{x,y} \sum_d C^{open}\, y_d \;+ C^{over} \left(\sum_j \sum_d \sum_{\substack{t=D^{N}_{d}-p_j+1 \\ t>0}}^{T-p_j} (t + p_j - D^{N}_{d}) x_{jdt} \right), \tag{5.18}$$

fonction objectif sujette aux contraintes (5.19)–(5.32) reprises dans le tableau 5.3.

Les contraintes (5.19) à (5.25) correspondent à leurs homologues de la formulation précédente, à savoir (5.2) à (5.8). Cela englobe les contraintes de disponibilité des ressources (renouvelables, lits et non-renouvelables) de même que les disponibilités des chirurgiens. Les égalités (5.26) et (5.27) représentent les contraintes de début d'opération et de capacité des salles, et correspondent aux égalités (5.10) et (5.11) de la section précédente. Les contraintes (5.28) définissent y_d comme le nombre maximum d'unités chirurgicales utilisées simultanément par jour en évaluant, pour chaque période de temps, le nombre d'opérations se chevauchant (deux interventions se chevauchant devant impérativement être réalisées dans des blocs opératoires différents). Quant aux inégalités (5.29), celles-ci assurent que le nombre maximum S de blocs opératoires disponibles n'est pas outrepassé. Les inégalités valides (5.30) contrôlent que le temps total opératoire nécessaire par jour n'excède pas le temps total opératoire alloué à ce jour. Finalement, les contraintes (5.31) et (5.32) définissent, d'une part, la positivité et l'intégrité des variables y_d, et d'autre part, le caractère binaire des variables x_{jdt}. Notons que les modèles (5.1)–(5.17) et (5.18)–(5.32) dérivent tous deux du problème d'ordonnancement de projets à ressources contraintes (Brucker et al., 1999) et sont par conséquent assurément NP-difficiles.

5.2.2 Expérimentations numériques

Une première série d'expériences a montré que cette nouvelle formulation mathématique était en mesure de proposer de bonnes solutions très rapidement mais qu'elle peinait à prouver l'optimalité de celles-ci. Par exemple, pour l'instance comportant 80 opérations, moins de cinq minutes sont nécessaires pour obtenir une solution distante de moins d'1% de la meilleure borne inférieure, alors que prouver le programme opératoire optimal demande plusieurs heures. Excepté pour la planification de 70 opérations dont la solution optimale est trouvée après seulement 127 secondes, les autres jeux de données exigent beaucoup plus de temps pour être résolus à l'optimum. Bien entendu, nous pourrions nous contenter de bonnes approximations obtenues rapidement mais nous avons opté pour une autre alternative. Nous allons tirer avantage de la capacité qu'a le modèle de trouver des solutions quasi-optimales. L'approche consiste à faire tourner le solveur jusqu'à un niveau satisfaisant d'écart entre solution courante et meilleure borne inférieure (plus communément appelé *gap*). Ce niveau atteint, le solveur est alors stoppé pour récupérer les valeurs des variables y_d, c'est-à-dire le nombre de salles d'intervention ouvertes chaque jour. Ces valeurs sont ensuite figées de sorte que les y_d ne soient plus des variables mais deviennent des paramètres du modèle, et le solveur est relancé jusqu'à l'obtention d'un optimum, avec x_{jdt} pour seules variables. Il est évident que la qualité des solutions, de même que les performances calculatoires, dépendent du niveau de gap jugé satisfaisant, ce dernier déterminant l'arrêt du solveur. Dans la suite de ce chapitre, ce niveau sera fixé à 1% afin de s'assurer qu'aux variables y_d soient assignées des valeurs quasi-optimales avant de redémarrer CPLEX. Le tableau 5.4 compare les deux approches. Tout d'abord, la colonne "*Gap* $< 1\%$" expose les performances du modèle révisé lorsque celui-ci est simplement stoppé dès que l'écart entre la solution courante et la meilleure borne inférieure descend sous la barre de 1%. Ensuite, la colonne "*Var.* y_d *fixées*" montre comment ces solutions peuvent être améliorées si l'on relance le solveur jusqu'à l'optimum[1] avec les variables y_d fixées à leurs valeurs quasi-optimales, obtenues lors du premier jet du solveur et menant aux solutions de la colonne précédente. Dans ces deux colonnes, les nombres entre parenthèses représentent l'écart

[1]Les solutions présentées dans la colonne "*Var.* y_d *fixées*" sont dès lors optimales pour le problème considérant un nombre déterminé de salles d'opération utilisées par jour.

Nbr.	MIP (5.18)–(5.32) (Gap $< 1\%$)		MIP (5.18)–(5.32) (Var y_d fixées)	
Op.	Obj. [€]	Temps [s]	Obj. [€]	Temps [s]
60	33270 (0.81%)	30	33120 (0.36%)	168
70	44880*	127	44880*	127
72	42980 (0.95%)	3131	42780 (0.48%)	4911
80	49160 (0.65%)	303	48960 (0.24%)	415
85	51000 (0.87%)	4512	50750 (0.38%)	4814
86	44680 (0.88%)	1128	44330 (0.10%)	1481
88	45370 (0.95%)	243	45170 (0.51%)	439
91	48260 (0.36%)	676	48260 (0.36%)	680
100	49510 (0.45%)	210	49360 (0.15%)	520

TAB. 5.4 – *Performances du MIP* (5.18)–(5.32) *appliqué à la programmation hebdomadaire d'un quartier opératoire composé de blocs polyvalents. Comparaison des solutions obtenues en stoppant le solveur une fois le gap inférieur à 1% avec celles obtenues après avoir relancé le solveur avec les variables y_d fixées.*

entre la solution proposée et la même meilleure borne inférieure atteinte lors du premier jet du solveur (en d'autres termes, la véritable borne inférieure du problème étudié). Notons que les temps mentionnés dans cette seconde approche sont les temps globaux nécessaires à l'obtention des solutions proposées et pas uniquement les temps dévolus à la seconde application du solveur. Grâce au tableau 5.4, chacun peut évaluer l'utilité d'obtenir des approximations de meilleure qualité en relançant le solveur sur les variables x_{jdt} uniquement.

Sur base du tableau 5.4, il semble clair que les temps nécessaires à l'obtention de solutions quasi-optimales produites par ce nouveau modèle sont très courts. Pour les données à disposition, l'utilisateur doit au plus attendre une heure vingt minutes pour disposer de solutions de très bonne qualité, très proches de la meilleure borne inférieure trouvée par le solveur (ce qui ne signifie pas que ces solutions ne soient pas optimales). Relancer le solveur pour des valeurs de y_d déterminées, comme l'illustre la colonne "*Var. y_d fixées*", permet d'améliorer, à moindre coût, les solutions déjà très satisfaisantes acquises lors de la première application de celui-ci. Il revient alors au décideur de juger le profit qu'il peut tirer de ces gains, somme toute marginaux, en valeur objectif.

Le tableau 5.5 compare, lui, les performances des deux modèles de programmation opératoire présentés dans ce chapitre et traitant différemment la symétrie due au caractère polyvalent des blocs opératoires. Le second MIP, décrit par les équations (5.18)–(5.32) et appliqué en fixant les y_d puis en relançant le solveur, produit des solutions optimales améliorant celles que le modèle exposé en section 5.1 est en mesure de trouver. Seule l'instance contenant 85 interventions permet au premier modèle d'égaler le second. Ceci illustre clairement ce que nous soupçonnions précédemment, à savoir que fixer un nombre a priori de salles perpétuellement ouvertes a un impact sur la valeur objectif des solutions. Cet effet est assurément plus significatif pour les instances requérant moins d'unités chirurgicales.

Concernant les temps indispensables à l'obtention des solutions, la seconde formulation mathématique du problème améliore encore ceux-ci en ne consommant qu'au plus 25 minutes de calculs, excepté pour les instances composées de 72 et 85 opérations qui en réclament plus d'une

Nbr	MIP (5.1)–(5.17)		MIP (5.18)–(5.32)	
Op.	Obj. [€]	Temps [s]	Obj. [€]	Temps [s]
60	40800*	604	33120*	168
70	45600*	1791	44880*	127
72	44600*	2724	42780*	4911
80	49200*	1885	48960*	415
85	50750*	3352	50750*	4814
86	45050*	1724	44330*	1481
88	45650*	6948	45170*	439
91	48500*	3596	48260*	680
100	49600*	6063	49360*	520

TAB. 5.5 – *Comparaison des performances des formulations MIP* (5.1)–(5.17) *et* (5.18)–(5.32) *(relancé avec les variables y_d fixées), appliquées à la programmation hebdomadaire d'un quartier opératoire composé de blocs polyvalents.*

heure[2]. Ces temps plus élevés s'expliquent par la structure des données. Rappelons qu'il s'agit bien de données retraçant l'activité opératoire réelle de l'hôpital, ce qui explique la différence entre les neuf sous-ensembles de données. Analyser les deux instances comportant 72 et 85 traitements chirurgicaux nous enseigne qu'elles présentent les durées opératoires hebdomadaires moyennes les plus élevées. Ceci signifie qu'elles contiennent des cas longs et potentiellement critiques consommant beaucoup de ressources, les rendant plus complexes à résoudre. Cependant, en regard de la journée actuellement nécessaire au gestionnaire, ces temps de calcul apparaissent comme étant très courts. Si, comparativement à la formulation (5.1)–(5.17), le bénéfice de ce MIP révisé n'est pas éloquent en termes de valeur objectif sur les plus grandes instances, il est substantiel en termes de temps requis pour parvenir à la solution optimale. Pour les trois jeux composés de 88, 91 et 100 opérations, le solveur nécessite plus d'une heure pour aboutir à la solution optimale lorsqu'il est appliqué sur la formulation (5.1)–(5.17), alors qu'au plus onze minutes sont suffisantes s'il est appliqué sur la formulation (5.18)–(5.32). Ces résultats suggèrent que cette dernière formulation est encore plus efficiente que la première. De plus, le modèle (5.18)–(5.32) ne comporte pas de paramètre potentiellement difficile à manipuler, à savoir le nombre β de blocs opératoires systématiquement ouverts chaque jour. En revanche, ses performances dépendent d'un autre paramètre beaucoup plus intuitif, le niveau de gap satisfaisant, qui est, lui, directement relié à la qualité de la valeur objectif. Bien entendu, plus cet écart sera grand, moins les temps de calcul pour l'atteindre seront élevés. Cependant, comme nous l'avons fixé à une valeur très faible dans nos travaux, nous pouvons assurer que cette seconde approche est efficiente même pour de faibles niveaux de gap souhaités.

Le programme linéaire en nombres entiers mixtes développé ci-dessus apparaît comme un outil efficient pour créer des plannings opératoires hebdomadaires optimaux. Il tient compte de la plupart des contraintes de fonctionnement du quartier opératoire, comme, entre autres, la disponibilité du personnel médical. De plus, contrairement à la majeure partie des travaux de la littérature, il ne scinde pas le processus de programmation en deux étapes séquentielles. Il considère le processus dans sa totalité afin de proposer au décideur des solutions optimisées globalement.

[2]Notons cependant qu'en déterminant un niveau de gap satisfaisant à 1.1%, ces temps de calcul se réduisent à 3383 et 2591 secondes respectivement.

Cependant, il reste à développer un algorithme efficace capable de désagréger les solutions optimales obtenues par ce MIP pour déterminer l'affectation des salles d'opération. Un tel algorithme est présenté dans la section suivante.

5.2.3 Algorithme de reconstruction des solutions

Le programme en nombres entiers mixtes révisé détermine le nombre de salles à utiliser chaque jour de même que l'ensemble des opérations programmées pour y débuter au temps t, en s'assurant que le nombre d'opérations en cours à chaque instant n'excède jamais S, le nombre de salles disponibles. Cependant, il ne mentionne pas explicitement quelle opération est assignée à quel bloc opératoire. Le problème consistant à déterminer l'affectation des salles, c'est-à-dire l'assignation des opérations aux blocs opératoires, peut se résoudre exactement par l'algorithme glouton suivant.

L'algorithme de reconstruction est inspiré par le principe du problème de *bin packing*. Sur base d'un ensemble d'opérations dont les jours et heures de début sont connus, on souhaite remplir un minimum de blocs opératoires tout en assignant chaque opération à l'un d'entre eux. Le but de cet algorithme est donc d'allouer une unité chirurgicale à chaque intervention j, de sorte que le jour et l'heure d'opération de j, définis par le MIP, tout comme le nombre y_d^* de salles nécessaires chaque jour d, soient respectés. La philosophie de l'algorithme consiste, pour chaque jour d, à :

1. Remplir tour à tour chacune des y_d^* salles d'opération ouvertes $s = 1, \ldots, y_d^*$;
2. Ordonner les opérations j telles que $x_{jdt}^* = 1$ par ordre croissant d'heure de début t ;
3. Programmer la première opération disponible qui ne chevauche pas les opérations déjà assignées à la salle courante ;
4. Passer à la salle suivante si aucune opération ne peut plus être programmée dans la salle courante.

La procédure de reconstruction est présentée de manière plus détaillée par l'algorithme 2. Considérant chaque journée d séparément, l'algorithme commence par trier les opérations se déroulant le jour examiné, notées $\mathcal{O}_d$, par ordre croissant d'heure de début t_j. Alternativement, pour chaque salle $s = 1, \ldots, y_d^*$ ouverte ce jour-là, il initialise $T_s := 1$ comme étant la première période de temps à laquelle la salle s est disponible pour une opération non encore assignée. Ensuite, tant qu'il existe, dans l'ensemble des opérations sans bloc opératoire $\mathcal{O}_d$, une intervention dont le temps de début est supérieur ou égal à T_s, la salle s peut encore être allouée. Il s'agit alors de prendre la première opération j dont le temps opératoire $t_j \geq T_s$ est minimum (ligne 7). A cette opération j est ensuite assignée la salle courante s (ligne 8), la disponibilité T_s de cette salle étant mise à jour en concordance avec l'heure de fin de j (ligne 9). Une salle lui ayant été désignée, l'opération j est finalement supprimée de l'ensemble $\mathcal{O}_d$ des interventions devant encore être affectées à une salle (ligne 10). Lorsqu'aucune opération ne peut plus être assignée à la salle s (c'est-à-dire $t_j < T_s,\ \forall j \in \mathcal{O}_d$), l'algorithme passe au remplissage de la salle suivante.

Le fait que l'algorithme glouton utilise y_d^* blocs opératoires et produise bien un programme opératoire réalisable vient de la faculté à modéliser le problème comme un problème de coloration sur un graphe d'intervalles (chaque opération à réaliser une même journée peut être représentée par l'intervalle de temps durant lequel elle doit avoir lieu, deux intervalles étant reliés par une arête s'ils se chevauchent). Or, l'on sait (voir Golumbic, 1980) que pour ce type de graphes, le

Algorithme 2 Algorithme de reconstruction

for $d = 1$ à D **do**
  $\mathcal{O}_d := \{j \in J | d_j = d\}$;
  Trier $\mathcal{O}_d$ par ordre croissant d'heure de début t_j ;
  for $s = 1$ à y_d^* **do**
    $T_s := 1$;
    while $\exists\ j \in \mathcal{O}_d | t_j \geq T_s$ **do**
      Prendre $j = \min_{k \in \mathcal{O}_d,\ t_k \geq T_s} \{t_k\}$;
      $OR_j := s$;
      $T_s := t_j + p_j$;
      $\mathcal{O}_d = \mathcal{O}_d \setminus \{j\}$;
    end while
  end for
end for

nombre maximum d'opérations réalisées en parallèle (la clique maximum) $\omega(G) = y_d^* = \chi(G)$ le nombre minimum de blocs opératoires alloués pour traiter l'ensemble des opérations (le nombre chromatique). En voici une preuve plus formelle.

Proposition 5.1. *L'algorithme de reconstruction est optimal.*

Démonstration. D'abord, nous prouvons que le nombre de salles nécessaires à l'algorithme pour construire le programme opératoire est bien optimal et égal à y_d^*, déterminé par le MIP révisé. Ensuite, nous prouvons que la stratégie *earliest first* appliquée est également optimale. La solution issue du programme linéaire (5.18)–(5.32) peut être modélisée par un graphe d'intervalles, où chaque opération est représentée par l'intervalle de temps durant lequel elle est supposée se réaliser. Etant donné cet ensemble fini d'intervalles, le graphe associé se construit comme suit : à chaque intervalle correspond un point du graphe, et deux points sont connectés par une arête si et seulement si leurs intervalles respectifs se superposent au moins partiellement (comme l'illustre la figure 5.1). L'assignation des interventions aux blocs opératoires revient donc à colorer correctement le graphe qui s'y rapporte, où chaque couleur représente une salle d'intervention. Comme il s'agit d'un graphe parfait (Golumbic, 1980), nous savons que le nombre chromatique d'un graphe d'intervalles équivaut à la taille de sa clique maximum, définie par le nombre maximum d'intervalles se chevauchant, et donc d'opérations dans notre cas, ce qui définit effectivement y_d^*. Ce dernier est dès lors bien le nombre minimum de salles à ouvrir pour réaliser le programme opératoire.

Nous prouvons maintenant que la stratégie *earliest first* de l'algorithme est optimale. Au graphe d'intervalles peut être associée une matrice binaire A de taille $J \times T$ n'ayant pas de lignes nulles, où $J = |\mathcal{O}_d|$ est le nombre d'opérations à ordonnancer le jour d et où T représente le nombre de périodes de temps constituant cette journée. Chaque ligne $a_{j,:}$ de la matrice correspond à l'opération j et prend les valeurs 1 uniquement pour les périodes de temps constituant l'intervalle que représente cette opération. Soit $\sigma_t = \sum_j a_{jt}$, le nombre d'interventions (ou d'intervalles) en parallèle à la période t, et soit $\sigma^* = \max_t \sigma_t$ le nombre maximum d'interventions se chevauchant (initialement, $\sigma^* = y_d^*$, la taille de la clique maximum du graphe associé). σ^* peut également être perçu comme le nombre restant de blocs opératoires encore nécessaires pour réaliser les opérations non assignées à l'un d'eux. Une fois sa salle attribuée, la ligne correspondant à une opération

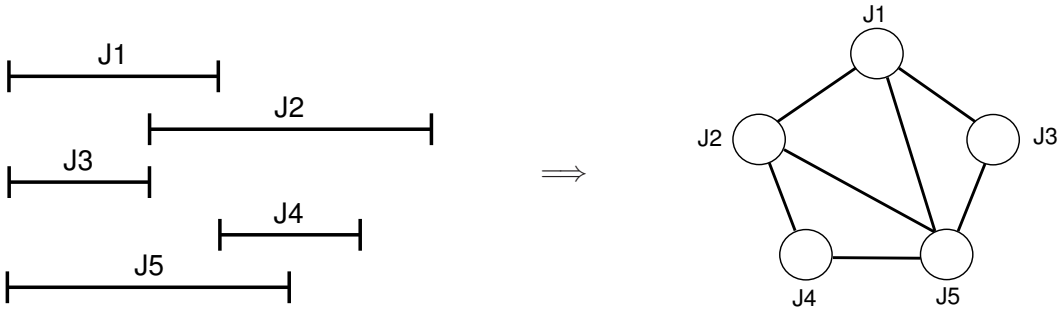

FIG. 5.1 – *Illustration d'un graphe d'intervalles où chaque opération est d'abord représentée par un intervalle ; ensuite un intervalle est modélisé par un nœud, deux nœuds étant reliés par une arête si leurs intervalles respectifs se chevauchent.*

est supprimée de la matrice A et l'ensemble des σ_t est actualisé, comme illustré par la figure 5.2. Supposons, sans perte de généralité, que nous attribuions des opérations à la première salle s_1. Les interventions éligibles pour la salle s_1 sont uniquement celles dont l'heure de début t est telle que $\sigma_t = \sigma^*$. En effet, si ce n'était pas le cas, et une fois s_1 remplie, σ^* resterait inchangé pour au moins une période de temps de telle sorte qu'au minimum $y_d^* + 1$ salles seraient utilisées au total, ce qui est en contradiction avec le nombre chromatique du graphe d'intervalles associé. De plus, ne pas assigner à s_1 la première opération éligible (c'est-à-dire celle minimisant t tel que $\sigma_t = \sigma^*$) signifie qu'il restera, après remplissage de la salle, au moins une période de temps pour laquelle $\sigma_t = y_d^*$, ce qui contredit à nouveau la première partie de cette preuve. □

Concernant les performances de l'algorithme de reconstruction, son temps de calcul est négligeable puisque, appliqué sur l'instance la plus grande, celui-ci s'élève à 0.2 seconde. Par conséquent, les performances calculatoires du modèle révisé développé dans cette section sont uniquement dictées par la procédure d'optimisation et non par la construction de la solution. Le programme en nombres entiers mixtes (5.18)–(5.32) semble donc être une approche tout-à-fait efficiente pour intégrer les étapes de planification et d'ordonnancement du quartier opératoire.

$$
A = \begin{array}{c} j_1 \\ . \\ . \\ . \\ j_5 \\ \\ \sigma_t \end{array}
\begin{array}{c}
\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \\
\\
\begin{array}{ccccccc} 3 & 3 & 3 & 3 & 2 & 1 & 0 \end{array}
\end{array}
\overset{(a)}{\Rightarrow}
\begin{array}{c}
\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \\
\\
\begin{array}{ccccccc} 2 & 2 & 2 & 2 & 1 & 1 & 0 \end{array}
\end{array}
\overset{(b)}{\Rightarrow}
\begin{array}{c}
\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \\
\\
\begin{array}{ccccccc} 1 & 1 & 1 & 1 & 0 & 0 & 0 \end{array}
\end{array}
$$

FIG. 5.2 – *Illustration de l'algorithme de reconstruction. La matrice A connexe au graphe d'intervalles représente la solution du MIP (5.18)–(5.32). L'étape (a) consiste à assigner les opérations j_1 et j_4 à la salle 1 en raison de leur temps opératoire t qui est tel que $\sigma_t = \sigma^* = 3$. Ensuite, l'étape (b) assigne les opérations j_3 et j_2 à la salle 2. A la fin de cette étape, seule l'opération j_5 n'est pas encore programmée dans une salle, et est donc assignée à la salle 3, qui est bien le nombre chromatique du graphe associé (voir figure 5.1).*

Dans les chapitres 4 et 5 nous avons présenté des approches de programmation opératoire intégrant les étapes usuellement séquentielles composant le processus de programmation opératoire. Les premières sont dévolues à la programmation de blocs opératoires dédiés, signifiant que chacun d'entre eux est équipé différemment, pour n'accueillir que certains types de chirurgie. Le problème envisagé étant d'une grande complexité algorithmique, la formulation mathématique en nombres entiers proposée peine à trouver des solutions de qualité sur plusieurs instances testées. Ceci justifiait pleinement le développement d'une approche méta-heuristique essentiellement destinée à traiter ces instances complexes. Utiliser l'une ou l'autre de ces approches permet d'obtenir des solutions quasi-optimales en des temps raisonnables, s'avérant dès lors efficientes pour la programmation de blocs opératoires dédiés. Supposer des salles d'intervention polyvalentes, et donc identiques, nous permet de prendre en considération un élément majeur de ce cas de figure, à savoir la symétrie générée par cette nature multifonctionnelle. Tenir compte de cette caractéristique procure un avantage indéniable aux deux modèles mathématiques proposés, qui diffèrent par la manière de contrer cette symétrie. La nature des blocs opératoires offre plus de flexibilité au gestionnaire et permet de réduire globalement les coûts. Les deux MIP développés dans ce chapitre sont largement plus efficients que le modèle proposé au chapitre précédent, de sorte que l'algorithme génétique y étant introduit ne trouve plus ici d'utilité. Un point sur la littérature nous enseigne que nos travaux, outre par l'intégration des étapes du processus étudié, se démarquent de celle-ci par les contraintes opératoires dont ils tiennent compte, notamment celles relatives à la disponibilité des ressources humaines. A ce stade, nous pouvons déjà souligner un apport significatif non seulement par rapport aux travaux de la littérature du domaine traité, mais également par rapport à la pratique de la programmation opératoire dans les hôpitaux. Nous pouvons donc répondre à certaines questions fondamentales que nous nous sommes posées en début de recherche. Oui, il semble possible d'intégrer planification et ordonnancement du quartier opératoire en une formulation unique, efficiente sur des données de taille réelle. Du moins, c'est ce que suggèrent les approches proposées dans les chapitres 4 et 5. Il est aujourd'hui possible, pour un gestionnaire de quartier opératoire, d'élaborer un planning optimisé dans un temps raisonnable là où cela lui coûtait une journée de travail avec les moyens dont il disposait. Grâce à notre outil, le décideur est en mesure de construire et de reconsidérer les programmes opératoires plusieurs fois pour parvenir à adhérer un maximum de praticiens à la solution qu'il leur propose. Cependant, dans leur état, ces dernières ne considèrent pas le facteur humain inhérent au monde hospitalier comme nous l'entendons. Bien entendu, avant d'envisager pouvoir impliquer davantage le personnel dans la prise de décision, il nous fallait un moteur efficace de programmation opératoire, en mesure de satisfaire un maximum de contraintes délimitant l'activité chirurgicale. Cet outil se devait également de pouvoir agir simultanément sur des niveaux tactique et opérationnel de décision, ce que ne proposait pas la littérature. Notre première contribution est donc d'avoir développé un tel moteur, sous les traits des méthodes de résolution présentées aux chapitres 4 et 5. Cependant, bien qu'à l'inverse de bon nombre d'approches de la littérature nous considérions les disponibilités des intervenants dans notre démarche, cela ne suffit pas. Les personnes concernées par la programmation du quartier opératoire sont en droit d'agir sur cette programmation, d'exprimer leurs avis et leurs préférences concernant leurs conditions de travail. Dès lors, il nous semble indispensable, dans le chapitre suivant, d'amener de la coopération dans ce processus de planification des blocs opératoires.

Chapitre 6

Coopération et programmation opératoire

Dans ce chapitre, nous tentons d'apporter au processus de programmation du quartier opératoire une certaine coopération qui interviendrait entre les acteurs dominants du processus de soins, à savoir les chirurgiens. Cette coopération concerne les plages horaires disponibles à l'activité opératoire et se met en œuvre par le biais de deux approches bien distinctes : considérer les préférences des acteurs pour ces périodes d'utilisation des salles ou mettre ces dernières aux enchères. La première de ces approches permet donc aux chirurgiens de déclarer des préférences quant aux différentes demi-journées de travail de la semaine. Confronter les desiderata des praticiens a pour but de maximiser l'utilité globale (ou, plus exactement, de minimiser la désutilité globale) et non l'utilité de chaque chirurgien individuellement. Cette nécessité de coopérer passe par la déclaration, par les médecins, de fonctions de désutilité associées aux différentes demi-journées constituant une semaine d'ouverture des blocs opératoires. Ces désutilités sont une manière de traduire la hauteur du désagrément encouru par un chirurgien lorsque ses préférences ne sont pas respectées. Un bon programme opératoire doit, en effet, considérer le bien être physique et moral du personnel médical, entre autres afin d'en améliorer l'efficacité et de réduire le risque d'absentéisme. Avec le même objectif poursuivi, la seconde approche consiste à placer aux enchères les différentes plages horaires constituant la semaine d'ouverture du quartier opératoire. Chaque praticien se voit alors remettre un certain nombre de jetons à répartir, selon ses desiderata, parmi les demi-journées de la semaine disponibles à l'activité opératoire programmée (rappelons que le quartier opératoire est typiquement ouvert cinq jours par semaine). Les enchérisseurs proposant les plus grandes mises remportent la plage horaire convoitée (sachant qu'une plage horaire concerne plusieurs salles occupées par plusieurs opérations, et donc par plusieurs chirurgiens). Ce mécanisme d'enchères s'accompagne d'un protocole de négociation offrant la possibilité aux éventuels chirurgiens mécontents de leurs attributions de négocier l'échange de deux interventions dans le planning proposé.

La première de ces approches ne sera implémentée que dans le cadre d'un quartier opératoire composé de salles dédiées alors que la seconde ne le sera que pour des salles polyvalentes. Bien entendu, chacune de ces approches est transposable à l'ensemble des méthodes de résolution développées dans les chapitres 4 et 5 mais ne le sera pas pour d'évidentes questions de redondance. Nous avons dès lors pris le parti de limiter l'intégration des préférences aux approches exacte

et heuristique vouées à la programmation de blocs opératoires dédiés, tandis que le mécanisme d'enchères, et plus particulièrement le protocole de négociation, mettront à profit l'efficience du modèle mathématique considérant des blocs opératoires identiques.

Dans ce chapitre, nous intégrons donc une part sociale à la fonction objectif afin de considérer les préférences des chirurgiens, qu'elles soient exprimées en tant que telles ou sous forme d'enchères. L'objectif est dès lors double : d'une part, minimiser les coûts de fonctionnement (C^{open} et C^{over}) et, d'autre part, minimiser la désutilité globale des chirurgiens ou allouer les plages horaires aux plus offrants. Cet apport de coopération dans notre démarche globale nécessitait une adaptation des approches de résolution proposées au chapitre 4 et 5, adaptation à laquelle répondent respectivement les sections 6.2 et 6.3 de ce chapitre. Au préalable, nous définissons les concepts de préférence et de désutilité liés à cette approche.

6.1 Vers une approche coopérative

Notre volonté, dans ce travail, est clairement d'intégrer le facteur humain dans l'établissement des programmes opératoires, en particulier en introduisant de la coopération dans ce processus de prise de décision. La littérature suggère essentiellement deux écoles abordant les aspects coopératifs : d'une part, la théorie des jeux, et, d'autre part, l'intelligence artificielle distribuée via les systèmes multi-agents.

La théorie des jeux est une branche des mathématiques, souvent appliquée dans des domaines tels que l'économie, la politique ou les sciences sociales, qui étudie les interactions entre des décideurs agissant dans leur propre intérêt. La plupart des modèles en théorie des jeux sont basés sur la théorie du choix rationnel qui stipule qu'un décideur sélectionne une action, parmi toutes les actions qui lui sont disponibles, en fonction de ses préférences. Ces dernières se traduisent au moyen d'une fonction d'utilité qui associe un "gain" ordinal, ou cardinal, à chaque action. Un jeu exprime alors une compétition entre différents décideurs dont l'utilité et donc la prise de décision dépendra des actions des autres décideurs (Osborne, 2004).

Les systèmes multi-agents, quant à eux, reposent sur une idée assez simple. Un agent est un système informatique capable d'actions indépendantes au nom de son utilisateur. Il peut déterminer seul les actions à entreprendre afin de satisfaire les objectifs qui lui sont désignés. Un système multi-agents est constitué de plusieurs agents intelligents qui interagissent entre eux, typiquement en s'échangeant des messages au moyen d'une infrastructure définie. Chaque agent a sa propre perception de l'environnement dans lequel il évolue et sélectionne ses actions sur base de ses connaissances. Les agents n'ayant pas forcément les mêmes buts poursuivis, leurs interactions seront basées sur leurs facultés à coopérer, à se coordonner et à négocier (Wooldridge, 2002).

A notre connaissance, il existe peu de travaux associant coopération et gestion du quartier opératoire. Par le biais de la programmation par buts (*goal programming*), certains travaux incluent des préférences des acteurs (Ozkarahan, 2000; Blake & Carter, 2002; Tan et al., 2007) mais l'on ne retrouve pas de réelle coopération comme la définissent la théorie des jeux ou l'intelligence artificielle distribuée. Dans ce cas précis de programmation par buts, il s'agit plus de se rapprocher d'un objectif (ou but) fixé au préalable. Pourtant, certaines phases du processus de programmation des blocs opératoires sont propices à l'introduction de collaboration. En effet, ce processus de planification est étroitement lié à la problématique de la planification de ressources

partagées. Les chirurgiens se partagent des salles et du personnel afin de procéder à des activités qui leur sont propres. Or, les besoins en ressources de chacun d'eux peuvent être compatibles ou, au contraire, s'avérer mutuellement exclusifs et engendrer le mécontentement du chirurgien lésé. C'est pourquoi il nous semble pertinent d'offrir la possibilité aux praticiens de négocier leur planning d'activité sur base de leurs préférences. La négociation consiste en l'obtention d'un accord entre parties. Elle peut se traiter sous différentes formes allant des approches théorie des jeux comme les enchères ou les protocoles de négociation, à l'argumentation propre aux systèmes multi-agents (Wooldridge, 2002). Par exemple, sur base d'une approche d'intelligence artificielle distribuée, Czap et al. (2005) utilisent une analyse conjointe (*conjoint analysis*) pour déterminer les fonctions d'utilité des chirurgiens, nécessaires pour déléguer le processus de négociation inhérent à un système d'agents intelligents. Bien que rare en programmation opératoire, cette nature coopérative se retrouve de plus en plus dans la littérature traitant de l'établissement de l'emploi du temps des infirmières, avec notamment des approches basées sur les mécanismes d'enchères (voir De Grano et al., 2009, par exemple) tels qu'imaginés ici.

Les deux méthodes que nous proposons afin d'inclure une dimension coopérative au processus de programmation opératoire s'inspirent de la théorie des jeux, et plus précisément de la théorie de la négociation de Nash (*Nash bargaining*) pour la première et des systèmes d'enchères pour la seconde. Bien qu'issues d'une philosophie identique, ces deux approches n'en sont pas moins complémentaires. En effet, d'un côté nous avons une approche qui permet au chirurgien de déterminer le nombre d'opérations qu'il souhaite effectuer par jour tout en pénalisant les périodes durant lesquelles il désire ne pas être présent. De l'autre, nous avons un mécanisme qui attribue les plages horaires sur base de mises effectuées par les médecins et établies à concurrence du poids que le chirurgien occupe dans le processus, poids qui peut varier au gré des négociations entre les acteurs. Quel que soit le mécanisme, en considérant un tel objectif d'ordre social, nous apportons un premier élément de réponse à un deuxième axe de recherche qui concerne l'intégration d'une certaine forme de coopération au problème, en maximisant l'utilité globale et non l'utilité individuelle de chaque chirurgien.

6.2 Intégration des préférences des acteurs

Cette première vision de la coopération prend ses racines dans la théorie de la négociation de Nash. Pour exprimer leurs préférences, nous allons associer à chaque chirurgien une fonction de désutilité qui lui est propre, où la désutilité est définie comme une mesure cardinale de mécontentement, de perte d'utilité, liée à la violation des préférences du praticien. Plus précisément, les chirurgiens spécifient des disponibilités par demi-journées. Ils précisent leurs préférences quant aux demi-journées de travail de deux façons différentes :

- soit en fixant le nombre d'opérations qu'ils souhaitent effectuer chaque demi-journée ;
- soit en stipulant pour chacune de leurs opérations la demi-journée durant laquelle ils souhaitent la réaliser.

Ces préférences seront pondérées par des mesures de désutilité particulières à chaque chirurgien, pénalisant plus ou moins fort le non respect des volontés exprimées par chacun. De même que l'utilité est une notion liée à l'idée de besoin, appréciant le bien-être émanant de l'obtention d'un bien ou d'un service, la désutilité est l'aptitude à contrarier un besoin (on parle en général de la désutilité du travail par rapport au temps libre, désutilité compensée alors par un salaire). En

reflétant les préférences individuelles, elle permet de traduire la hauteur du désagrément encouru par chaque chirurgien lorsque ses choix de prédilection ne sont pas respectés. A titre illustratif, on pourrait décrire les courbes d'indifférence de deux biens générant une désutilité (c.-à-d. deux biens procurant une utilité marginale négative) comme étant concaves, situées dans le troisième orthant en opposition aux biens engendrant une utilité situés, eux, dans le premier orthant, de telle sorte que la désutilité qu'ils suscitent diminue en s'approchant de l'origine de façon à éviter leur consommation.

Comme nous l'avons mentionné précédemment, cette première approche coopérative ne sera appliquée qu'aux méthodes de résolution destinées à la programmation de blocs opératoires dédiés. Il s'agit donc, dans cette section, d'adapter les programme linéaire et algorithme génétique développés respectivement en section 4.1 et 4.3 pour y incorporer une dimension coopérative.

6.2.1 Programme linéaire en nombres entiers mixtes

Dans cette section, nous adaptons le modèle mathématique décrit en section 4.1 en y incorporant une dimension de coopération. Le but est d'intégrer le facteur humain lié au quartier opératoire dans nos approches à caractère mathématique. En effet, la formulation du chapitre 4 peut être qualifiée de directive (en opposition à coopérative) car l'unique objectif poursuivi est la minimisation des coûts, sans préoccupation pour la coopération. En particulier, les disponibilités des chirurgiens y sont modélisées de façon rigide, sous forme de contraintes. Dans cette section, ces disponibilités ne transparaissent plus sous forme de contraintes mais au travers de préférences et de désutilités exprimées par les chirurgiens.

Le problème de programmation des blocs opératoires dédiés ne change pas pour autant et reste tel que décrit dans le chapitre 3. La seule différence consiste à ne plus modéliser les disponibilités des chirurgiens par l'intermédiaire d'une matrice binaire $R^C \in \{0,1\}^{C\times 2D}$, mais au moyen de préférences $\mathcal{P}$ et de désutilités $\mathcal{D}$ associées aux différentes journées de travail constituant une semaine d'ouverture du quartier opératoire. Le programme linéaire en nombres entiers comporte les mêmes variables :

- x_{jsdt} : dédiées aux activités chirurgicales, ces variables binaires valent 1 lorsque l'opération j débute dans le bloc s, le jour d au moment t ;
- z_{sd} : associées aux blocs opératoires, ces variables binaires valent 1 si l'on procède à l'ouverture du bloc s le jour d ;
- l_{sd} : ces variables entières positives représentent le nombre de périodes de temps en heures supplémentaires dans la salle s le jour d.

L'objectif d'ordonnancement se décrit de deux façons différentes suivant la manière dont sont modélisées les préférences des chirurgiens. Nous y retrouvons un premier terme commun, identique à la fonction objectif première (4.1), qui est la minimisation des coûts Δ inhérents à l'ouverture de blocs opératoires et à l'utilisation des salles en heures supplémentaires :

$$\Delta = \sum_{s} \sum_{d} \left[C^{open}\, z_{sd} + C^{over}\, l_{sd} \right].$$

Un terme variable minimise la différence entre les préférences des chirurgiens et le planning effectif, en fonction des désutilités déclarées. Pour rappel, les contraintes considérées sont présentées

$$\sum_s \sum_d \sum_t x_{jsdt} \;=\; 1, \qquad \forall j, \tag{4.2}$$

$$\sum_j \sum_s \left[r^{\rho}_{jk} \sum_{\tau=t-p_j+1}^{t} x_{jsd\tau} + \Gamma^{\rho}_{jk} \sum_{\tau=t-\lambda_{jk}+1}^{t} x_{jsd\tau} \right] \leq R^{\rho}_{kd}, \forall d, \forall t, \forall k \in \mathcal{K}^{\rho}, \tag{4.3}$$

$$\sum_j \sum_s \sum_{\tau=t-(p_j+b_j)+1}^{t-p_j} x_{jsd\tau} \leq R^{B}, \qquad \forall d, \forall t, \tag{4.4}$$

$$\sum_j \sum_s \sum_t r^{\nu}_{jk}\, x_{jsdt} \;\leq\; R^{\nu}_{kd}, \qquad \forall k \in \mathcal{K}^{\nu}, \forall d, \tag{4.5}$$

$$\sum_s \sum_{j \in O(c)} \sum_{\tau=t-p_j+1}^{t} x_{jsd\tau} \;\leq\; 1, \qquad \forall t, \forall d, \forall c, \tag{4.8}$$

$$\sum_j \sum_{\tau=t-p_j+1}^{t} x_{jsd\tau} \;\leq\; 1, \qquad \forall s, \forall d, \forall t, \tag{4.9}$$

$$\sum_d \sum_t x_{jsdt} \;\leq\; Q_{js}, \qquad \forall j, \forall s, \tag{4.10}$$

$$x_{jsdt} = 0, \qquad \forall j, \forall s, \forall t, \forall d \notin [ES_j, LS_j], \tag{4.11}$$

$$x_{jsdt} = 0, \qquad \forall j, \forall s, \forall d, \forall t | (t + p_j) > D^{M}_{sd}, \tag{4.12}$$

$$\sum_j \sum_t p_j\, x_{jsdt} \leq D^{M}_{sd}\; z_{sd}, \qquad \forall s, \forall d, \tag{4.13}$$

$$l_{sd} \geq \sum_j \sum_{t=D^{N}_{sd}-p_j+1}^{T-p_j} {}_{t>0} \left(t + p_j - D^{N}_{sd}\right)\; x_{jsdt},\; \forall s, \forall d, \tag{4.14}$$

$$l_{sd} \in \mathbb{N}, \qquad \forall s, \forall d, \tag{4.15}$$

$$x_{jsdt},\; z_{sd} \in \{0, 1\}, \qquad \forall j, \forall s, \forall d, \forall t. \tag{4.16}$$

TAB. 6.1 – *Contraintes liées au problème de planification du quartier opératoire avec coopération.*

dans le tableau 6.1. Seules les contraintes (4.6) et (4.7) associées aux disponibilités des chirurgiens n'y sont pas reprises, ces disponibilités n'apparaissant plus sous forme de contraintes dans le contexte coopératif. Notons que, dans ce tableau, les contraintes (4.10) ne s'appliquent qu'en raison de la nature non plurifonctionnelle des blocs opératoires.

Préférences sur le nombre d'opérations par demi-journée. Une première manière de modéliser les préférences des chirurgiens consiste à définir, pour chacun d'eux, le nombre d'opérations qu'il souhaite effectuer lors de chaque demi-journée de la semaine. A cette fin, nous définissons la matrice $\mathcal{P} \in \mathbb{Z}_{+}^{C \times 2D}$ dont chaque ligne correspond à un chirurgien particulier et contient le nombre d'interventions qu'il a l'intention d'effectuer chaque demi-journée. Nous définissons également une mesure de désutilité $\mathcal{D} \in \mathbb{Z}_{+}^{C \times 2D}$ par laquelle chaque chirurgien associe à chaque demi-journée un poids spécifique. Ce poids est supposé traduire le désagrément occasionné lorsque le nombre d'opérations planifiées diffère des préférences énoncées. Plus le poids $\mathcal{D}_{cd}$ associé par le chirurgien c à la demi-journée d est élevé, plus le désagrément encouru est grand, plus la valeur de la fonction objectif sera médiocre. Un poids nul reflète une indifférence totale quant au suivi des

préférences spécifiées pour la demi-journée correspondante. La fonction objectif devient :

$$\min_{z,l} \ \Delta + \sum_{c=1}^{C} \sum_{d=1}^{2D} [\, \mathcal{D}_{cd}\, m_{cd} \,], \tag{F1}$$

soumise aux contraintes du tableau 6.1, ainsi qu'aux contraintes définitionnelles suivantes :

$$m_{c,2d-1} \geq \left(\sum_{s} \sum_{j \in O(c)} \sum_{t=1}^{T/2} x_{jsdt} \right) - \mathcal{P}_{c,2d-1} \quad \forall c, \forall d$$

$$m_{c,2d} \geq \left(\sum_{s} \sum_{j \in O(c)} \sum_{t>T/2}^{T} x_{jsdt} \right) - \mathcal{P}_{c,2d} \quad \forall c, \forall d$$

$$m \in \mathbb{Z}_{+}^{C \times 2D}.$$

Comme il s'agit ici de minimiser la fonction objectif, et puisqu'elles sont définies positives, les variables m_{cd} représentent le nombre d'opérations réalisées la demi-journée d qui ne respectent pas les préférences du chirurgien c dans le programme opératoire effectif.

Préférences sur la demi-journée de chaque opération. Ensuite, une seconde approche consiste à décrire les préférences des chirurgiens au moyen d'une matrice binaire $\mathcal{P} \in \{0,1\}^{J \times 2D}$ dont l'entrée $\mathcal{P}_{jd} = 1$ si le chirurgien attitré souhaite réaliser l'opération j la demi-journée d. La mesure de désutilité s'exprime dès lors sous la forme d'une matrice $\mathcal{D} \in \mathbb{Z}_{+}^{J \times 2D}$ par laquelle chaque chirurgien exprime le désagrément qu'il aurait à effectuer une intervention particulière à une date différente de son choix de prédilection. Le principe est le même que précédemment, plus le poids attribué à une demi-journée est élevé, plus la contrariété est grande. La fonction objectif devient :

$$\min_{z,l} \ \Delta + \sum_{j=1}^{J} \sum_{d=1}^{2D} \mathcal{D}_{jd} \, [\, m_{jd} \, (1 - \mathcal{P}_{jd}) \,], \tag{F2}$$

soumise aux contraintes du tableau 6.1, ainsi qu'aux contraintes définitionnelles suivantes :

$$m_{j,2d-1} = \sum_{s} \sum_{t=1}^{T/2} x_{jsdt} \quad \forall j, \forall d$$

$$m_{j,2d} = \sum_{s} \sum_{t>T/2}^{T} x_{jsdt} \quad \forall j, \forall d$$

$$m \in \{0,1\}^{J \times 2D}.$$

Nous définissons de cette façon les variables binaires m_{jd} de telle sorte que $m_{jd} = 1$ si l'opération j est programmée pour la demi-journée d dans le planning effectif. L'objectif sera dès lors pénalisé d'un poids $\mathcal{D}_{jd}$ lorsque le jour effectif de l'intervention m_{jd} et le jour souhaité $\mathcal{P}_{jd}$ sont différents.

Bien entendu, prendre en compte l'avis des acteurs du processus opératoire lors de la construction des plannings nécessite la définition de nouvelles variables et l'ajout de contraintes supplémentaires. Ceci permet également plus de flexibilité dans l'établissement des emplois du temps.

Ce qui nous amène à penser que, d'une part, les approches exactes coopératives nécessiteront plus de ressources calculatoires que leur pendant non coopératif mais, d'autre part, qu'elles peuvent globalement diminuer les coûts opératoires.

6.2.2 Algorithme génétique multi-objectif

Dans cette section, nous modifions l'algorithme génétique proposé au chapitre 4.3 afin, à présent, de prendre en compte deux objectifs : l'un d'ordre économique et l'autre d'ordre social. Ces deux objectifs sont potentiellement conflictuels puisqu'il s'agit d'une part de minimiser les coûts de fonctionnement, et d'autre part de minimiser la désutilité globale des chirurgiens. Afin d'adapter le programme linéaire en nombres entiers, il est nécessaire d'agréger ces deux objectifs en un seul objectif multi-critère. Cela a pour désavantage de conditionner les solutions en fonction du poids attribué à chaque critère et de produire une valeur objectif difficilement interprétable. L'idéal, dans ce cas de figure, serait d'optimiser chacun des deux objectifs individuellement l'un de l'autre, et d'obtenir non plus une solution optimale mais bien un ensemble de solutions Pareto-optimales. Si notre approche exacte décrite plus haut se prête difficilement à ce type d'optimisation multi-objectif, ce n'est pas le cas de notre approche méta-heuristique (raison pour laquelle nous avons opté pour un algorithme génétique). Dès lors, comme nous entreprenons d'optimiser deux objectifs antagonistes, nous adaptons l'algorithme génétique décrit en section 4.3 sur base d'un schéma multi-objectif, en raison de son aptitude à produire une population de bonnes solutions en temps raisonnable. Les algorithmes évolutionnaires multi-objectifs (MOEA) ont, en effet, la capacité de converger vers la frontière Pareto, c.-à-d. de trouver plusieurs solutions Pareto-optimales à la fois, étant donné qu'ils génèrent une population de solutions.

Trois caractéristiques importantes que doit inclure ce type d'algorithmes multi-objectifs sont : (i) assigner une valeur fitness sur base de Pareto-dominance, afin de diriger la recherche de solutions vers l'ensemble Pareto-optimal ; (ii) maintenir une bonne diversité au sein des individus composant la population afin de diversifier l'information ; (iii) faire preuve d'élitisme pour préserver les meilleurs individus déjà trouvés et pour améliorer la convergence de l'algorithme. Parmi les MOEA présentant ces caractéristiques, nous avons opté pour le schéma NSGA-II, ou Non-dominated Sorting Genetic Algorithm II (Deb et al., 2002). Ce dernier semble en effet, et selon ses auteurs, être plus performant que ses concurrents, les méthodes élitistes PAES ou SPEA (Raghuwanshi & Kakde, 2004), que ce soit en termes de diversification de solutions ou en termes de convergence vers le front Pareto-optimal.

Architecture NSGA-II

NSGA-II est un algorithme génétique multi-objectif qui applique une procédure rapide de tri non-dominé, une approche élitiste et un opérateur de niche ne nécessitant pas de paramètres (Deb et al., 2002). Son principe de fonctionnement est identique à l'algorithme 1 présenté en section 4.3, à quelques ajustements près.

Comme nous travaillons à présent dans un environnement multi-objectif, la valeur fitness d'un individu ne peut plus représenter sa qualité uniquement qu'en termes de valeur objectif. En effet, la valeur fitness doit au contraire refléter le niveau de Pareto-dominance de chaque individu de la population. Pour un problème de minimisation de l'objectif, la relation de Pareto-dominance

peut se définir comme suit : soit $x, y \in \mathbb{R}^n$, x domine y (noté $x \prec y$) si et seulement si $x_i \leq y_i$ pour tout $i = 1, \ldots, n$ et il existe j tel que $x_j < y_j$. La procédure de Pareto-dominance appliquée dans NSGA-II trie les individus d'une population en différents fronts non-dominés. Le premier front contient tous les individus non-dominés de la population courante ; ces derniers se voient attribuer une valeur fitness égale à un. Pour déterminer les individus du deuxième front, la totalité des individus constituant le premier front est temporairement retirée de la population pour ensuite appliquer une seconde fois la procédure de tri non-dominé. Les individus non-dominés issus de cette seconde application constituent le deuxième front de non-dominance, et se voient attribuer la valeur fitness deux. Cette procédure est itérative et répétée tant que l'entièreté de la population courante n'est pas triée en fronts non-dominés.

Afin de préserver de la diversité parmi les solutions, le schéma NSGA-II utilise une stratégie de nichage basée sur une distance de masse (*crowding distance*). Celle-ci est appliquée par l'opérateur de sélection pour déterminer quels individus formeront la génération suivante de l'algorithme. La distance de masse est un paramètre qui mesure la proximité d'un individu par rapport à ses voisins le long de chaque objectif. Plus la distance de masse moyenne est grande, plus grande sera la diversité au sein de la population. Par conséquent, le processus de sélection est basé sur deux critères : la valeur fitness de front non-dominé et la mesure de distance de masse. Dès lors, entre deux individus appartenant à des fronts différents, seul celui faisant partie du front le moins dominé sera retenu ; entre deux solutions d'un même front, le processus sélectionnera celle présentant la plus grande distance de masse, c.-à-d. celle située dans une région moins peuplée de l'espace de solutions.

La procédure NSGA-II se décrit par l'algorithme 3, où g est l'indice de la génération courante, P_g est la population courante, E_g est l'ensemble des enfants issus de la population courante et T la limite de temps de calcul.

Algorithme 3 Schéma de l'algorithme multi-objectif NSGA-II

1: $g := 1$;
2: Initialisation de la population P_1 ;
3: **while** $g \leq GEN$ et $t \leq T$ **do**
4: $E_g = \texttt{croisement}(P_g)$;
5: $E_g = \texttt{mutation}(E_g)$;
6: $R_g := P_g \cup E_g$;
7: $\mathcal{F} = \texttt{tri non-dominé}(R_g)$;
8: $P_{g+1} = \emptyset$ and $i = 1$;
9: **while** $|P_{g+1}| + |\mathcal{F}_i| \leq POP$ **do**
10: $\texttt{mesure distance-masse}(\mathcal{F}_i)$;
11: $P_{g+1} = P_{g+1} \cup \mathcal{F}_i$;
12: $i = i + 1$;
13: **end while**
14: Trier $\mathcal{F}_i$ par ordre décroissant de distance de masse ;
15: $P_{g+1} = P_{g+1} \cup \mathcal{F}_i(1 : POP - |P_{g+1}|)$;
16: $g := g + 1$;
17: **end while**

La représentation chromosomique des individus, l'initialisation de la première population de l'algorithme et les opérateurs évolutionnaires restent identiques à ceux proposés en section 4.3. Rappelons qu'un individu se compose de deux chromosomes $I = (\delta, \sigma)$, le premier contenant une liste d'opérations ordonnées auxquelles sont associées des demi-journées d'opération ; le second alloue les blocs opératoires aux différentes interventions à programmer. Notons encore que, étant donné la définition par les chirurgiens de préférences et de désutilités, l'initialisation et la mutation ($p_{mut} = 0.05$) sont adaptées pour en tenir compte. En effet, les modifications de dates d'opération peuvent se faire en fonction d'un seuil de désutilité à ne pas dépasser. La sélection, quant à elle, applique à présent une technique qui trie les individus d'abord par fronts et ensuite par ordre décroissant de distance de masse. La principale différence entre l'algorithme génétique du chapitre 4 et celui-ci vient de l'évaluation de la qualité des individus. Les valeurs fitness sont ici calculées sur base du front de non-dominance auquel appartient chaque individu et non plus sur base de la valeur objectif atteinte par la solution que représente cet individu. Cependant, le schéma NSGA-II est une procédure d'optimisation non contrainte. Or, la planification du quartier opératoire est un problème hautement contraint, dont certaines de ces restrictions ne peuvent être prises en compte par le codage des individus. Il nous faut donc mettre en œuvre une façon d'incorporer ces contraintes dans la procédure NSGA-II. Pour ce faire, nous allons pénaliser les valeurs objectifs de la solution émanant d'un individu lorsque celle-ci n'est pas réalisable. La manière de procéder est décrite dans la section suivante.

Manipuler les contraintes

Les algorithmes génétiques utilisent une fonction fitness pour décrire la qualité de la solution générée par un individu. Ici, la valeur fitness d'un individu I représente son niveau de non-dominance. Puisque la représentation chromosomique ne peut satisfaire toutes les contraintes du problème étudié, nous prenons en compte les contraintes non satisfaites d'une autre façon. Etant donné que la procédure de tri non-dominé fonctionne sur base des valeurs objectifs afin d'attribuer le front d'un individu, nous allons modifier celles-ci afin de leur faire refléter la satisfaction ou non de ces contraintes. Les valeurs objectifs d'un individu sont donc modifiées si l'une des trois contraintes suivantes n'est pas respectée : la disponibilité des ressources non-renouvelables, la capacité maximale des salles d'opération et le respect des dates de débuts au plus tôt et au plus tard de chaque intervention.

Tout comme dans la fonction fitness de la page 51, nous définissons $L_k^\nu(d)$ comme étant le surplus journalier de consommation de la ressource non-renouvelable $k \in \mathcal{K}^\nu$. La négativité de cette mesure implique un programme opératoire non réalisable pour le jour incriminé. La surconsommation totale de ressources non-renouvelables est définie par :

$$L^\nu_{\text{nsga}} = \sum_{\substack{k \in \mathcal{K}^\nu \\ L_k^\nu(d) < 0}} \left| L_k^\nu(d) \right| .$$

De même, la reconstruction du planning opératoire à partir d'un individu, sur base des disponibilités des ressources renouvelables et de celles des lits de réveil, peut nécessiter un dépassement de la disponibilité maximale qu'offre une salle. Si $L_s^\rho(d)$ représente le nombre journalier de périodes de temps encore accessibles dans la salle s, le total de périodes de temps en dépassement de capacité

de salle est décrit par :

$$L^{\rho}_{\text{nsga}} = \sum_{d} \sum_{\substack{s \\ L^{\rho}_{s}(d)<0}} |L^{\rho}_{s}(d)| \,.$$

Finalement, soit L^{ST}_{nsga}, quantité reflétant la satisfaction des dates de début au plus tôt et au plus tard d'une opération, et valant zéro en cas de respect de ces contraintes et l'infini en cas de non respect. Nous définissons les valeurs objectifs modifiées d'un individu I comme suit :

$$f_{\text{nsga}}(I) = \begin{cases} C(I) & \text{si } I \text{ est réalisable} \\ P + L^{\nu}_{\text{nsga}} + L^{\rho}_{\text{nsga}} + L^{ST}_{\text{nsga}} & \text{sinon,} \end{cases}$$

où $C(I) \in \mathbb{R}^2_+$ est le vecteur des valeurs objectifs associées à l'individu I, et où $P \in \mathbb{R}^2_+$ est le vecteur des bornes supérieures des deux objectifs poursuivis. $P(1)$ est calculé en considérant le cas où chaque bloc est ouvert chaque jour et utilisé à hauteur de sa capacité maximale, tandis que $P(2)$ est calculé en maximisant la désutilité de l'ensemble des chirurgiens.

De cette manière, nous espérons que les individus constituant le premier front soient réalisables et offrent de bonnes valeurs objectifs.

6.2.3 Expérimentations numériques

A présent, nous allons évaluer le comportement des approches exacte et heuristique décrites ci-dessus dans un contexte coopératif. Nous analysons l'implication que peut avoir la modélisation des préférences sur les performances de ces deux approches. Rappelons que nous supposons deux façons pour les chirurgiens de déclarer leurs agendas de disponibilités :

1. soit ils précisent le nombre d'opérations qu'ils souhaitent effectuer lors de chacune des demi-journées de la semaine ;
2. soit ils spécifient, pour chacune de leurs opérations, la demi-journée durant laquelle ils préfèrent la réaliser.

Pour rappel, le quartier opératoire dont nous souhaitons programmer l'activité se compose de sept blocs opératoires supposés dédiés, chacun étant donc inaccessible à certaines pathologies. Le jeu de données utilisé est décrit en page 32. Les données à disposition ne mentionnent ni préférences, ni désutilités comme nous l'envisageons. Par conséquent, les préférences et désutilités appliquées dans ces expériences ne modélisent pas réellement le sentiment des acteurs. En effet, nous n'avons pas eu l'occasion d'interviewer les chirurgiens impliqués, et les données en notre possession ont été collectées une année avant le début de nos travaux. Dès lors, ces préférences ont été fixées de manière à correspondre aux disponibilités des chirurgiens réellement observées pour la semaine étudiée. Les désutilités sont, quant à elles, fixées à zéro pour les demi-journées de présence dans le planning effectif et à cent en cas d'absence. Bien entendu, il s'agit d'une modélisation arbitraire simplement destinée à permettre la comparaison des solutions optimisées avec le véritable programme opératoire tel qu'appliqué *in situ*. De plus, en modélisant de la sorte, les deux formulations en nombres entiers décrites dans ce chapitre ont la possibilité d'aboutir à des solutions de même valeur objectif que le programme linéaire proposé en section 4.1, ce qui en facilite la comparaison.

Form. Math.	Cons. Ress.	16 Opérations Objectif	16 Opérations Temps [s]	19 Opérations Objectif	19 Opérations Temps [s]
F0	R	12290*	1	14430*	53
	E	12290*	1	15130 (1.62 %)	1800
F1	R	12290*	3	14430*	121
	E	12290*	4	15280 (2.75 %)	1800
F2	R	12290*	1	14430*	80
	E	12290*	1	15230 (2.30 %)	1800

TAB. 6.2 – *Performances des deux formulations linéaires en nombres entiers F1 et F2 pour 16 et 19 interventions à ordonnancer sur un jour.*

Performances des programmes linéaires en nombres entiers

Les diverses formulations introduites en section 6.2.1 sont ici comparées quantitativement, sur base de leurs performances calculatoires. Le tableau 6.2 présente les résultats des formulations F1 et F2 pour une journée d'ouverture contenant seize puis dix-neuf opérations. Les conclusions émises en section 4.1 sont également applicables ici. Pour une unique journée d'ouverture des blocs, les programmes linéaires coopératifs proposent des performances tout à fait semblables à celles de la formulation directive du chapitre 4. A nouveau, seul le cas considérant des blocs dédiés pour la programmation de dix-neuf opérations, avec niveau élevé de consommation de ressources, peut poser problème. Les autres solutions sont toutes optimales et obtenues endéans la limite de temps fixée. On remarque que la formulation F2 se distingue quelque peu et semble mieux adaptée aux instances à blocs opératoires dédiés que F1.

Le tableau 6.3 résume les performances des deux formulations coopératrices F1 et F2 appliquées à la programmation hebdomadaire du quartier opératoire. Afin d'évaluer l'impact de la coopération sur le processus étudié, figurent également dans ce tableau les performances de la formulation directive proposée au chapitre 4 et dénotée ici, par souci de clarté, F0. Cette dernière s'étant déjà montrée largement plus efficiente que la procédure manuelle, les coûts des plannings effectifs ne sont plus reproduits ici. L'on constate d'emblée que les deux approches coopératrices souffrent du même mal que leur pendant directif et peinent à prouver l'optimalité des solutions qu'elles produisent. Seule F2 y parvient pour les instances comportant 60 et 70 interventions, F1 n'atteignant, quant à elle, l'optimum que pour 70 opérations. Cependant, toujours à l'instar de leur prédécesseur, elles proposent en moins d'une heure des solutions distantes de moins d'un pour-cent de la meilleure borne inférieure, déterminée par l'algorithme de résolution, dans quatre expérimentations sur les neuf pour F2 et trois pour F1, comme l'illustre la figure 6.1. Les autres instances utilisées pour valider nos approches réclament, elles, au minimum deux heures de calculs pour atteindre ce niveau de qualité. Il est cependant délicat de comparer ces deux formulations sur base de leur besoin en ressource temporelle. En effet, nulle ne se distingue clairement de l'autre, F1 semblant plus apte à se rapprocher à moins d'un pour-cent de la borne inférieure et F2 à prouver l'optimalité de ses solutions. Globalement, ces deux formulations ne sont applicables en pratique, dans leur utilisation heuristique, que pour la moitié des tests réalisés ici. Qui plus est, dans un tiers des situations, F1 et F2 requièrent plus de 15 heures pour se rapprocher à moins de 1% de la borne inférieure, alors que dans deux de ces cas, ce temps passe à plus de 24 heures. La figure 6.1 nous montre encore que pour les instances comportant 80, 85 et 91 (resp. 80 et 85) interventions, F2 (resp. F1) ne propose aucune solution réalisable avant minimum trois heures de temps,

Nbr.	Form.	Gap $< 1\%$			
Op.	Math.	Objectif	Temps	Objectif	Temps
60	F0	37280 (0.7%)	*26 s*	37280*	*614 s*
	F1	35990 (0.9%)	*88 s*	35940 (0.1 %)	*24 h*
	F2	36790 (0.9%)	*130 s*	36590*	*2538 s*
70	F0	50130 (0.7%)	*68 s*	50030*	*152 s*
	F1	49250 (0.7%)	*266 s*	49100*	*11393 s*
	F2	51090 (0.9%)	*2143 s*	50800*	*6193 s*
72	F0	55940 (0.1%)	*479 s*	55940 (0.0 %)	*24 h*
	F1	46210 (0.8%)	*9950 s*	46060 (0.2 %)	*24 h*
	F2	46460 (0.8%)	*579 s*	46260 (0.1 %)	*24 h*
80	F0	51780 (0.7%)	*5.9 h*	51580*	*6.5 h*
	F1	50300 (0.9%)	*15.3 h*	50100 (0.5 %)	*24 h*
	F2	49800 (0.9%)	*22.1 h*	49650 (0.6 %)	*24 h*
85	F0	58490 (0.9%)	*7064 s*	58140 (0.4%)	*24 h*
	F1	–	–	54530 (4.5 %)	*24 h*
	F2	–	–	–	–
86	F0	46360 (0.8%)	*449 s*	46160 (0.3%)	*24 h*
	F1	44770 (0.6%)	*1870 s*	44670 (0.4 %)	*24 h*
	F2	45770 (0.9%)	*1892 s*	45560 (0.2 %)	*24 h*
88	F0	50780 (0.8%)	*917 s*	50380*	*1492 s*
	F1	47500 (0.9%)	*10.5 h*	47400 (0.7 %)	*24 h*
	F2	49350 (0.9%)	*2.9 h*	49300 (0.7 %)	*24 h*
91	F0	–	–	52320 (3.0 %)	*24 h*
	F1	–	–	49400 (1.0 %)	*24 h*
	F2	–	–	50540 (1.9 %)	*24 h*
100	F0	57850 (0.3%)	*3583 s*	57700 (0.1 %)	*24 h*
	F1	53230 (0.9%)	*6732 s*	53080 (0.3 %)	*24 h*
	F2	53380 (0.3%)	*9593 s*	53330 (0.1 %)	*24 h*

TAB. 6.3 – *Performances des trois formulations linéaires en nombres entiers appliquées à la programmation hebdomadaire de blocs opératoires dédiés. Comparaison des solutions obtenues en stoppant le solveur une fois le gap inférieur à 1% avec celles obtenues après maximum 24 heures de calcul. Ici, F0 correspond à la formulation directive présentée en section 4.1, page 38.*

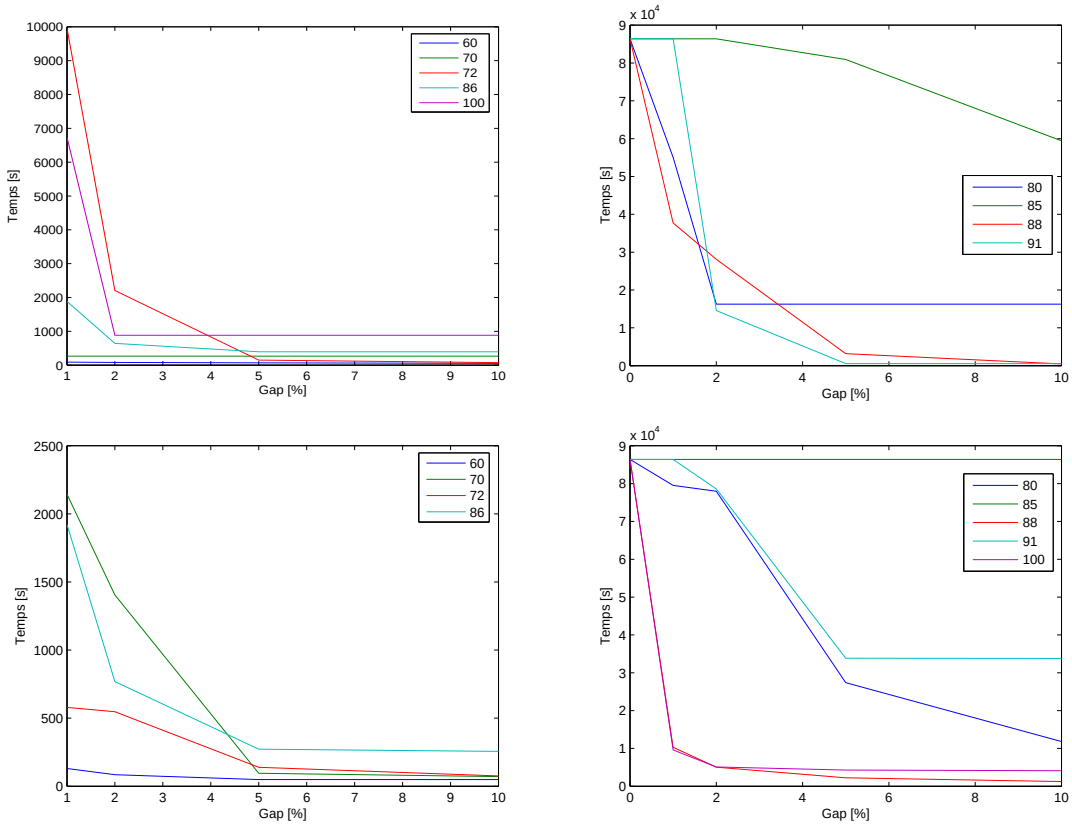

FIG. 6.1 – *Evolution du temps de calcul des formulations F1 (haut) et F2 (bas) en fonction de l'écart (gap) entre la solution courante et la meilleure borne inférieure trouvée par l'algorithme de résolution. A gauche, les instances les plus simples à résoudre, à droite les plus complexes.*

cette dernière ne trouvant pas la moindre solution réalisable endéans les 24 heures imparties pour la programmation de 85 opérations. Dès lors, plus encore que pour leur équivalent directif, ces approches coopératrices nécessitent l'utilisation de techniques méta-heuristiques plus spécifiques afin d'obtenir des programmes de qualité en des temps plus courts.

L'apport de ces deux formulations coopératrices est cependant clairement visible au niveau de la valeur objectif obtenue. En effet, bien qu'elles ne soient pas nécessairement prouvées optimales, les solutions coopératrices offrent systématiquement des valeurs objectifs inférieures comparativement à la formulation directive initiale (lorsque les salles d'opération sont présumées dédiées). Ceci met en relief la possibilité qu'offre la coopération de réduire le coût global, dans ce cas de figure. En effet, si chacun est enclin à faire un effort et à revoir ses préférences, la performance globale peut être améliorée, à l'avantage de tout le monde. Coopérer autorise dès lors plus de flexibilité dans la construction des plannings opératoires et permet d'outrepasser, en partie, la rigidité induite par l'accessibilité restreinte des salles. On se rapproche alors des performances des méthodes appliquées lorsque ces dernières sont identiques et polyvalentes. Le tableau 6.4 décompose les valeurs objectifs des solutions produites par les MIPs coopératifs en un terme économique et un terme social. Si la formulation F2 présente des valeurs objectifs plus élevées que F1 (voir tableau 6.3), cette différence tend à s'atténuer si l'on ne considère que la partie économique de la

Nbr.	Form. F0	Form. F1		Form. F2	
Op.	Coûts [€]	Coûts [€]	Désutilité	Coûts [€]	Désutilité
60	37280	35640 (-4.4%)	300 (2%)	35590 (-4.5%)	1000 (3.7%)
70	50030	48500 (-3.1%)	600 (3.8%)	48700 (- 2.7%)	2100 (3.3%)
72	55940	45160 (-19.3%)	900 (3.9%)	45060 (-19.5%)	1200 (3.7%)
80	51580	49200 (-4.6%)	900 (3.4%)	49250 (-4.5%)	400 (3.1%)
85	58140	53330 (-8.3%)	1200 (5%)	–	–
86	46160	44570 (-3.4%)	100 (0.5%)	44860 (-2.8%)	700 (0.9%)
88	50380	46700 (-7.3%)	700 (3.2%)	47000 (-6.7%)	2300 (2.9%)
91	52320	48700 (-6.9%)	700 (3.2%)	49140 (-6.1%)	1400 (3.4%)
100	57700	52280 (-9.4%)	800 (2.8%)	52030 (-9.8%)	1300 (2.9%)
Moy.		**-7.4**%	**689 (3.1%)**	**-7.1**%	**1300 (3.0%)**

TAB. 6.4 – *Comparaison des coûts et désutilités générés par les formulations coopératrices F1 et F2 en regard de la formulation directive F0.*

fonction objectif. En effet, il est clair, d'après le tableau 6.4, que ces deux formulations génèrent des frais de fonctionnement similaires et permettent, en moyenne, une économie de 7.4% et 7.1% respectivement sur le coût global proposé par la version directive du MIP. C'est là une manière de chiffrer l'apport de la coopération dans le processus de programmation opératoire. Cet écart s'explique par la plus grande flexibilité qu'offre la coopération en permettant d'outrepasser les exigences de certains praticiens pour obtenir des plannings opératoires globalement moins onéreux. Par contre, F1 et F2 se distinguent par la désutilité engendrée. En effet, F2 transgresse en moyenne treize préférences exprimées par les chirurgiens (désutilité moyenne de 1300) alors que F1 n'en enfreint que sept (désutilité moyenne de 689). Cette différence est imputable à la modélisation des préférences puisque F2 compte J expressions des desiderata des praticiens (rappelons que pour F2 les préférences sont exprimées pour chaque opération) alors que F1 n'en dénombre que C, le nombre de chirurgiens. Comme ici $C < J$, il est dès lors logique que F2 nécessite de contrevenir plus souvent aux préférences pour atteindre un objectif économique équivalent. Par contre, si nous normalisons ces données (pourcentages entre parenthèses dans la colonne *Désutilité*), nous remarquons que cette différence s'atténue puisque les deux formulations coopératrices génèrent, en moyenne, environ 3% de la totalité des désutilités exprimées. Bien entendu, les résultats présentés ici dépendent des poids associés à chaque objectif. La partie sociale de la fonction objectif est, ici, moins pondérée que son équivalent économique, ce qui explique comment réduire les coûts grâce aux formulations coopératrices. A charge du gestionnaire de faire le compromis entre les parties économique et sociale de la fonction objectif.

Même si elles sont globalement plus lentes que la formulation directive, les formulations coopératrices offrent plus de flexibilité tout en tenant compte du bien-être social des principaux acteurs du quartier opératoire, les chirurgiens. En effet, la formulation directive du chapitre 4 ne considère pas les préférences des chirurgiens mais impose leurs disponibilités à l'aide de contraintes. Par conséquent, cette formulation initiale peut ne pas être résoluble en raison de telles contraintes, ces dernières pouvant s'avérer trop exigeantes. Problème que ne rencontreront pas les deux formulations coopératrices, celles-ci ne contraignant pas les solutions aux disponibilités des chirurgiens mais proposant des solutions minimisant leur désagrément. De plus, les deux formulations coopératrices présentées dans ce chapitre affichent des performances comparables sur la programmation hebdomadaire de blocs opératoires dédiés, que ce soit en termes de coût écono-

POP	GEN			
	500	1000	2000	5000
	59230–62570	56940–57240	55550–56690	53560–58480
10	0–200	0–200	100–300	0–300
	342 s	*691 s*	*1390 s*	*3409 s*
	59580–61770	62410–63510	54450–56840	56340–56790
20	0–300	0–300	0–100	0–100
	771 s	*1358 s*	*2813 s*	*7188 s*
	59870–63910	55740–56290	53650–55350	54200–55840
50	200–800	0–200	100–300	0–100
	2261 s	*4223 s*	*8909 s*	*21776 s*

TAB. 6.5 – *Illustration des performances de l'algorithme génétique multi-objectif pour la modélisation 1 des préférences, en fonction des paramètres POP et GEN. L'algorithme est ici appliqué à la planification hebdomadaire de 80 interventions sur 7 blocs opératoires dédiés.*

mique ou d'applicabilité. Cependant, les mêmes réserves quant à leur exploitation en pratique sur certaines données peuvent être émises. Si d'aventure une grande capacité de réactivité était nécessaire, ces formulations, tout comme la formulation directive, sont difficilement applicables pour l'ensemble des instances.

Performances de l'algorithme génétique multi-objectif

Nous évaluons à présent le comportement de l'algorithme génétique multi-objectif face aux deux modélisations des préférences des chirurgiens. Dans la section précédente, nous avons pu observer que le solveur CPLEX était sensible à la formulation mathématique d'un problème. Certes la taille des données y joue un rôle prédominant, mais, pour diverses raisons, l'écriture même des contraintes influence les performances du solveur. Ceci explique les différences observées entre les trois formulations en nombres entiers mixtes. Il n'en va pas de même pour l'algorithme génétique. Celui-ci n'est pas sensible à la formulation mathématique mais uniquement à la taille des données. Or, si les deux modélisations des préférences proposées nécessitent des données de tailles différentes, cette différence est minime. Comme le montrent les tableaux 6.5 et 6.6 suivants, la modélisation des préférences n'a pas de répercussion observée sur les performances de l'algorithme génétique multi-objectif. Ces performances seront, par contre, dictées par les paramètres constituant cette approche méta-heuristique, à savoir le nombre d'individus et le nombre de générations.

Le tableau 6.5 illustre les performances, pour l'équivalent génétique de F1, obtenues pour une unique expérience sur l'instance moyenne comportant 80 opérations. Ceci donne un aperçu de l'évolution des valeurs objectifs en fonction des paramètres de l'algorithme. Les deux paramètres testés sont le nombre d'individus composant la population (POP) et le nombre de générations (GEN) de l'algorithme. Le tableau 6.6 fait de même pour la deuxième proposition de modélisation des préférences, spécifiées sur la demi-journée de chacune des interventions du chirurgien. Pour chaque valeur du paramètre POP, la première ligne de ces tableaux représente la gamme

POP	GEN			
	500	1000	2000	5000
	61330–63220	58240–59680	58680–59080	53950
10	200–400	0–200	0–100	0
	339 s	*699 s*	*1359*	*3427 s*
	59380–61320	59080–61620	55800–57840	54750–56640
20	0–200	0–100	0–300	0–300
	777 s	*1443 s*	*2906 s*	*7139 s*
	58940–65960	55100–60530	55100–56490	54000–55000
50	100–500	0–500	0–200	0–200
	2260 s	*4377 s*	*8910 s*	*21549 s*

TAB. 6.6 – *Illustration des performances de l'algorithme génétique multi-objectif pour la modélisation 2 des préférences, en fonction des paramètres POP et GEN. L'algorithme est ici appliqué à la planification hebdomadaire de 80 interventions sur 7 blocs opératoires dédiés.*

minimum/maximum de valeurs de l'objectif économique pour l'ensemble de la population finale de l'algorithme. La deuxième ligne indique, elle, les différents niveaux de l'objectif social et la dernière ligne affiche les temps de calcul (en secondes) nécessaires à l'obtention de ces résultats.

Au vu de ces deux tableaux, il n'y a pas de raison pour privilégier l'une ou l'autre modélisation des préférences ; aucune ne se distingue. Elles suivent les mêmes tendances sans réellement modifier les valeurs objectifs (économique ou sociale). En règle générale, et tout comme pour la version mono-objectif de l'algorithme, plus l'on augmente le nombre de générations, plus l'on peut s'attendre à être proche de la frontière de Pareto. Néanmoins, accroître le nombre de générations de l'algorithme tend à réduire la diversité au sein de la population. A l'inverse, le nombre d'individus dans la population ne semble pas avoir d'impact majeur sur la qualité des solutions produites, mais bien sur la diversité au sein de ces dernières. De toute évidence, plus les paramètres POP et GEN seront élevés, plus le seront les temps de calcul.

Le choix de la modélisation des préférences important peu ici, nous limitons les expérimentations au seul cas de desiderata exprimés en termes de nombre d'opérations par demi-journée, à savoir le pendant heuristique de F1 (celle-ci semblant la plus apte à trouver rapidement des solutions quasi-optimales). Sur base des performances affichées dans le tableau 6.5, les paramètres de cette version multi-objectif de l'algorithme génétique, décrite en sous-section 6.2.2, sont fixés à 10 individus pour 5000 générations. Le schéma NSGA-II appliqué à notre recherche est testé sur l'ensemble des instances à disposition. Toutefois, celui-ci ne devrait rivaliser avec la version coopérative du MIP et montrer tout l'intérêt d'une approche heuristique que pour les instances les plus complexes (typiquement celles comportant 80, 85, 88 et 91 interventions).

Le tableau 6.7 semble nous donner raison. Ce dernier reprend respectivement les performances du MIP dans sa formulation F1 (l'application de celle-ci étant limitée à l'obtention de solutions à moins d'un pour-cent de la meilleure borne inférieure), les valeurs objectifs du meilleur individu issu de dix réplications de l'algorithme génétique multi-objectif, ainsi que les valeurs objectifs et écarts-types (mentionnés entre parenthèses) du meilleur individu moyen observé sur ces dix ex-

Nbr. Op.	MIP form. F1 (Gap $< 1\%$)			NSGA-II Meilleur Ind.			Meilleur Ind. Moy.		
	Coût	Dés.	Tps	Coût	Dés.	Tps [s]	Coût	Dés.	Tps [s]
60	35690	300	*88 s*	36580	300	*2663*	38075 (791)	300 (141)	*2596 (28)*
70	48750	500	*266 s*	52260	0	*2976*	53820 (662)	100 (53)	*3010 (31)*
72	45410	800	*9950 s*	51610	500	*3064*	53173 (1638)	370 (156)	*3063 (15)*
80	49500	800	*15.3 h*	52320	400	*3314*	54462 (1177)	280 (154)	*3333 (40)*
85	53330	1200	*24 h*	56250	100	*3563*	59435 (1491)	190 (119)	*3540 (38)*
86	44670	100	*1870 s*	46230	100	*3590*	49258 (1422)	100 (66)	*3616 (38)*
88	46800	700	*10.5 h*	51760	200	*3697*	53359 (858)	230 (141)	*3667 (36)*
91	48700	700	*24 h*	51310	300	*3693*	52646 (1164)	190 (128)	*3713 (36)*
100	52430	800	*6732 s*	57140	200	*4132*	59884 (2016)	240 (150)	*4171 (28)*

TAB. 6.7 – *Comparaison des performances de l'algorithme génétique multi-objectif en regard des solutions produites par la formulation en nombres entiers F1 limitée à 1% de gap.*

périmentations. Les temps nécessaires pour obtenir ces solutions y sont également mentionnés. Comme la partie économique de la fonction objectif de F1 possède un poids plus important, les meilleurs individus retenus pour la méta-heuristique sont ceux présentant les moindres coûts financiers. Comme attendu, l'AG ne peut produire des individus aussi efficients que les solutions de l'approche exacte pour les instances les plus simples à traiter (60, 70, 86 et 100 opérations). Par contre, à l'inverse de la formulation F1, la méta-heuristique produit systématiquement des solutions réalisables de qualité pour les paramètres fixés. Notamment, rappelons que pour les instances comportant 80 et 85 interventions, cette même formulation réclame plus de trois heures avant de proposer une solution réalisable, alors que l'algorithme génétique est en mesure de le faire en une heure environ. Ce dernier semble également plus efficace que l'approche exacte du point de vue social. En effet, bien que les coûts des solutions génétiques soient plus importants, la désutilité générée par ces solutions est, elle, en moyenne inférieure de 66% à la désutilité engendrée par les solutions exactes. Ceci s'explique par la nature multi-objectif de l'AG, alors que le MIP est, lui, multi-critère. La désutilité, dans l'approche génétique, est alors traitée comme un objectif indépendant à optimiser. Sa prise en compte dans la performance globale ne dépend alors plus du poids qui lui est accordé dans la fonction objectif agrégée du MIP. Comparer les deux approches sur le plan économique nous apprend en revanche que les meilleures solutions génétiques sont distantes, en moyenne, de 7% des solutions exactes, cet écart variant de 2% à 13% ; l'écart moyen entre individu moyen et F1 est, quant à lui, de 11%. Si l'écart relatif entre les deux approches peut paraître conséquent, notons toutefois que l'approche génétique permet d'améliorer les solutions issues de la formulation directive du problème (F0) dans cinq cas sur neuf. Ce gain est en moyenne de 3% (variant de 1% à 8%) alors que les meilleurs individus génétiques ne sont distants des quasi-optima directifs que de 2% en moyenne (variant de 0.1% à 4%). Ceci suggère donc que l'algorithme génétique tel que décrit en sous-section 6.2.2 est une alternative de qualité aux approches exactes directive et coopératrices, en mesure de proposer des solutions proches de l'optimum. De plus, les performances génétiques présentent sans conteste une constance dans la consommation de ressources calculatoires, quelles que soient les instances traitées, ce que ne sont pas en mesure d'assurer les différentes formulations en nombres entiers.

La figure 6.2 reprend un exemple des solutions produites par notre procédure génétique multi-objectif, pour vingt individus et cinq mille générations. Nous pouvons y observer des solutions

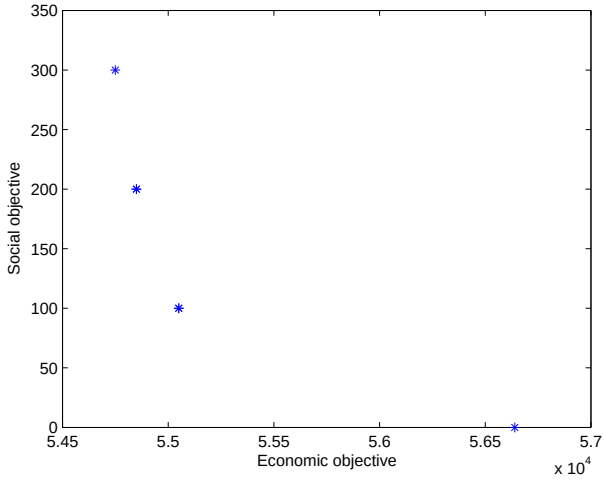

FIG. 6.2 – *Exemple d'estimation de la frontière de Pareto proposée par l'algorithme génétique multi-objectif. Les paramètres sont fixés à 20 individus et 5000 générations.*

variant de 54750 € à 56640 € pour l'objectif financier, et de 0 à 300 pour la désutilité globale des chirurgiens. Nous remarquons également qu'à partir des vingt individus constituant la population, seules quatre solutions différentes sont générées. Ceci illustre la perte de diversité des solutions engendrée par la multiplication des générations de l'algorithme. D'un point de vue théorique, nous n'avons aucune information sur la frontière Pareto-optimale du problème, celui-ci présentant une grande complexité de résolution. En effet, pour des espaces de décision de grande taille, il devient rapidement difficile d'obtenir le front Pareto-optimal en temps raisonnable ; l'on se contente alors, comme ici, d'une bonne approximation de cette frontière. Les solutions proposées dans cette figure améliorent le planning opératoire effectif de plus de 14.75% en termes de frais de fonctionnement, alors que le coût social reste nul. Si le gestionnaire est enclin à augmenter ce dernier, la participation financière peut diminuer à 54750 euros, ce qui représente une réduction de 18% des coûts du planning établi manuellement. Ces solutions sont obtenues en seulement deux heures de temps.

Bien entendu, ces résultats sont conditionnés par les désutilités fixées par les chirurgiens. Ici, le coût social augmente de 100 unités pour toute préférence non respectée. Cependant, le désagrément encouru peut différer selon les situations. Il est raisonnable de penser que l'incommodité perçue est différente selon que l'on programme une intervention l'après-midi alors que le chirurgien opère déjà le matin, ou que l'on programme une intervention durant le jour de repos du chirurgien. Ces différences sur l'importance du désagrément occasionné se traitent facilement par la définition de désutilités appropriées.

6.3 Mécanisme d'enchères et négociation

Dans cette section, la coopération est vue au travers d'un mécanisme d'enchères offrant aux chirurgiens la possibilité de spécifier des disponibilités par demi-journées et plus particulièrement leurs préférences quant à ces différentes plages horaires constituant la semaine. Le poids d'un chirurgien dans ce processus de décision se traduit par un certain nombre de jetons à placer sur les

demi-journées ayant sa préférence. Ce nombre peut différer selon les chirurgiens ou au contraire être uniforme (auquel cas les chirurgiens sont représentés de manière équitable).

Comme précisé en début de chapitre, cette seconde approche coopérative ne sera utilisée que dans le cadre d'un quartier opératoire comportant des salles d'intervention polyvalentes. En effet, le protocole de négociation proposé plus loin se doit de reposer sur un modèle mathématique résoluble rapidement si l'on souhaite qu'il soit utilisable pratiquement. Dès lors, dans cette section il s'agit de modifier la formulation en nombres entiers mixtes développée en section 5.2 et l'adapter au contexte coopératif qui nous préoccupe ici.

6.3.1 A propos des enchères...

Une enchère est un mécanisme comportant deux parties, le(s) vendeur(s) et le(s) acheteur(s), mis en concurrence pour l'obtention d'un bien ou d'un service octroyé au plus offrant. Un agent unique y met en présence un nombre indéfini d'autres agents afin de vendre ou d'acheter l'objet convoité. Il existe donc deux configurations pour un mécanisme d'enchères : la première considère un vendeur aux prises avec un ensemble d'acheteurs potentiels, la seconde confronte un acheteur à divers vendeurs potentiels. Dans ce cadre, la théorie des jeux permet d'analyser le comportement des agents économiques et de définir un équilibre stratégique du jeu indiquant à chaque joueur la stratégie optimale à adopter compte tenu des stratégies utilisées par les autres joueurs. Notons que le type de configuration du mécanisme d'enchères considéré n'influence pas l'analyse de l'attitude à suivre par les enchérisseurs ni le revenu de l'enchère. Il existe différents types d'enchères dont les plus connus sont :

- *Les enchères anglaises*, ou ascendantes, sont certainement les plus connues et nécessitent la présence d'un commissaire-priseur. Ce dernier commence par annoncer un prix de départ. Ensuite, chaque personne intéressée propose successivement une surenchère, c'est-à-dire un prix plus élevé que le dernier prix annoncé. Lorsque plus personne n'enchérit, le bien est alors attribué au plus offrant, en d'autres termes le dernier enchérisseur. Ce mécanisme se rencontre typiquement dans les salles de ventes.
- *Les enchères hollandaises*, ou descendantes, se déroulent à l'inverse des enchères anglaises. Le commissaire-priseur annonce un prix de départ supérieur à la demande et le diminue successivement au fil du temps jusqu'à ce qu'un candidat se manifeste, le bien lui étant alors attribué au prix en cours. Ce type d'enchères est essentiellement utilisé pour la vente de denrées périssables comme c'est le cas pour le marché aux fleurs d'Aalsmeer, ville hollandaise qui leur donne leur nom.
- *Les enchères scellées au premier prix*, ne comportent, elles, qu'un seul tour. Chaque participant remet une enchère sous enveloppe et détermine son prix sans connaissance préalable des offres des autres candidats. Les enchérisseurs ne s'échangent pas d'information, et le bien est attribué au plus offrant qui paie alors le montant de son offre. On rencontre ce type d'enchères notamment dans l'attribution des marchés publics.
- *Les enchères de Vickrey*, ou enchères scellées au second prix, sont semblables aux enchères scellées. La procédure est identique, mais l'enchérisseur remportant l'enchère paie le prix proposé par le second meilleur participant, sans partage de l'information. Ces enchères sont, entre autres, utilisées par Google pour l'attribution de ses espaces publicitaires ou, dans une version dynamique, par Ebay.

Le théorème d'équivalence des revenus

Notre but dans ce travail étant de maximiser le bien-être social (défini par les préférences des chirurgiens), nous sommes en droit de nous demander quel mécanisme d'enchères sera le plus propice pour atteindre cet objectif, c'est-à-dire la satisfaction des praticiens. C'est exactement à cette question que répond le théorème d'équivalence des revenus. En substance, ce dernier stipule (Klemperer, 1999) :

> *Pour autant que les agents soient neutres face au risque et que leurs estimations de l'objet à pourvoir (en d'autres mots leurs intentions de payer) soient privées et tirées indépendamment et aléatoirement d'une même fonction de distribution cumulative croissante et continue, alors chacun des mécanismes décrits ci-dessus offre le même revenu.*

Dans notre cas, nous prenons l'hypothèse que les acteurs du système sont neutres face au risque. Qui plus est, ils ne sont pas supposés s'échanger d'information quant à leurs préférences, considérant de plus que ces préférences individuelles ne sont pas influencées par celles d'autrui, et l'hypothèse d'estimations privées tient également. Le gestionnaire n'a donc aucun intérêt à privilégier un système par rapport à un autre. Ce théorème nous indique également qu'enchères anglaises et scellées au second prix sont stratégiquement équivalentes, de même pour les enchères hollandaises et scellées au premier prix. Ces quatre mécanismes ne se distinguant pas sur le revenu escompté qu'ils assurent, nous utilisons la simplicité de mise en œuvre comme critère de sélection. Comme nous privilégions une enchère statique ne comportant qu'un unique tour à une enchère dynamique comprenant plusieurs phases de mises consécutives, et vu sa facilité d'interprétation, nous optons pour un système d'enchères scellées (sans prix de réserve). De plus, nous envisageons un mécanisme dans lequel les chirurgiens miseraient un nombre prédéterminé de jetons sur les différentes plages horaires de leur choix, participant de la sorte à des enchères multiples (chaque enchère correspond à une plage horaire). Ces enchères se déroulant simultanément, en parallèle, les mises de l'une ne peuvent être réutilisées pour une autre. D'une certaine manière, toute mise est donc due. Ce schéma caractérise une enchère scellée au premier prix et non au second prix.

Enchères scellées au premier prix : information (im)parfaite

En situation d'information parfaite, chaque enchérisseur connaît l'estimation que chaque autre enchérisseur fait de la valeur du bien à acheter. Chaque joueur a à sa disposition un ensemble d'actions correspondant à l'ensemble des mises possibles. Ses préférences sont traduites par une fonction *payoff* qui définit le bénéfice du joueur i sur base de son évaluation v_i du bien. Le joueur i obtient un bénéfice de $v_i - b_i$ si b_i est la mise qui remporte l'enchère, et n'obtient rien sinon (bénéfice nul).

Dans une enchère scellée au second prix, il est dans l'intérêt de chaque joueur de miser à hauteur de son évaluation v_i du bien, le gagnant retirant un profit de $v_1 - v_2$ puisqu'il paie le prix de la seconde mise[1]. Dans une enchère scellée au premier prix, miser $b_i < v_i$ est une stratégie

[1] Les enchérisseurs étant généralement triés par ordre décroissant de mise, le vainqueur d'une enchère scellée au second prix retire donc un profit de $v_1 - p$, p étant le montant déboursé. Or, par définition, le vainqueur paie le prix de la seconde meilleure enchère. La stratégie optimale pour les acteurs de ce type de mécanisme étant de miser à hauteur de leur évaluation du bien, p vaut tout logiquement v_2.

dominant faiblement celle consistant à miser $b_i = v_i$ ou toute autre stratégie. En effet, miser v_i domine faiblement toute mise plus grande, mais est à son tour dominée par une mise plus petite. Miser v_i revient à obtenir un profit nul quelle que soit l'issue de l'enchère. Par contre, miser $b_i < v_i$ procure un bénéfice nul si le joueur i perd l'enchère mais un bénéfice positif s'il la remporte.

Le théorème d'équivalence des revenus vaut également lorsque l'information est imparfaite, pour autant que les valeurs respectives des joueurs associées au bien soient issues d'une même fonction de distribution cumulative, continue et croissante. Dans ce cas de figure, les participants ne sont pas parfaitement informés des estimations de la valeur du bien des autres, et l'on se retrouve dans une situation de jeu Bayésien. En cas d'information imparfaite, le concept approprié d'équilibre est un équilibre de Nash Bayésien. En d'autres termes, la stratégie de chaque joueur dépend de sa propre information (ici l'évaluation du bien) et maximise son payoff attendu étant donné la stratégie des autres joueurs et ses croyances sur l'information des autres joueurs. On peut montrer (Osborne, 2004) qu'en cas d'information imparfaite, dans les conditions requises, une enchère scellée au premier prix possède un équilibre de Nash Bayésien dans lequel chaque joueur mise $E[X|X < v_i]$, la valeur attendue de la plus grande mise des autres enchérisseurs sachant que v_i est l'évaluation la plus grande du bien.

Enchères scellées au premier prix : information asymétrique

Associée à la neutralité face au risque et au caractère indépendant et privé de l'information à disposition de chaque joueur, la troisième grande hypothèse du théorème d'équivalence des revenus concerne la symétrie des valeurs associées au bien. Le théorème suppose effectivement que ces valeurs soient issues d'une même fonction de distribution. Que se passe-t-il dès lors lorsque ce n'est pas le cas et que les joueurs sont en présence d'une asymétrie de l'information ?

Considérer des acteurs hétérogènes implique que leurs estimations du bien sont issues soit de fonctions de distribution différentes, soit d'une même fonction de distribution mais avec une répartition différente de la masse de probabilité (les joueurs forts ayant plus de masse sur les valeurs élevées que les joueurs faibles). Cependant, l'asymétrie entre les enchérisseurs n'affecte pas leur comportement dans un mécanisme d'enchères scellées au second prix. Miser la valeur associée au bien reste une stratégie faiblement dominante. Par contre, dans une enchère au premier prix, on peut prouver (Krishna, 2002) qu'un joueur plus faible (en d'autres termes, la fonction de distribution des valeurs qu'il associe au bien est dominée par celles d'autres joueurs, ses valeurs étant statistiquement plus faibles que celles de certains concurrents) aura tendance à miser plus agressivement qu'un joueur fort, dans le sens où, pour une même valeur v fixée, la mise du joueur faible sera plus élevée et donc plus proche de la valeur v qu'il associe au bien. Ce type d'enchères en condition d'asymétrie peut donc favoriser les joueurs faibles au détriment du joueur ayant l'évaluation la plus élevée. Il est dès lors plausible qu'une enchère scellée au premier prix engendre un revenu plus important, en moyenne, qu'une enchère de second ordre, mais soit moins efficiente sur l'allocation du bien (le bien n'étant pas assigné au joueur en ayant la plus grande utilité) en cas d'asymétrie. Toujours dans ces conditions, il est difficile d'exprimer l'équilibre du jeu sous une forme arrêtée tel qu'il est possible de le faire en cas d'information (im)parfaite et le problème consiste plus à déterminer si un tel équilibre existe. Cependant, si l'on suppose des mises discrètes (par exemple de l'argent ou des jetons, comme ce sera le cas dans notre mécanisme) l'on peut prouver qu'il existe bel et bien un équilibre à l'enchère scellée de premier prix dans lequel chaque

enchérisseur suit une stratégie qui est une fonction non-décroissante de son évaluation du bien (Krishna, 2002).

Contrainte budgétaire

Une autre raison pour laquelle le théorème d'équivalence des revenus peut être invalidé survient lorsque les joueurs sont soumis à une contrainte budgétaire. Dans ce cas, le revenu attendu d'une enchère au premier prix est plus élevé que celui d'une enchère au second prix (Klemperer, 1999). Intuitivement, ceci se justifie par le caractère moins restrictif de la contrainte de budget dans la première, la stratégie d'équilibre consistant à miser moins que la valeur de l'objet. Supposons une situation dans laquelle le théorème s'applique et les acteurs ont des valeurs indépendantes et privées v_i, et supposons que le joueur i ne puisse dépasser un budget b_i. Si, dans une enchère au second prix, i évalue l'objet à $e_i = \min\{v_i, b_i\}$ sans contrainte de budget, par l'équivalence des revenus, le profit attendu $\mathcal{R}$ est identique à celui issu d'une enchère au premier prix dans la même configuration. Maintenant, reconsidérant la contrainte de budget, le revenu de l'enchère au second prix est également $\mathcal{R}$ étant donné que les mises dans les deux situations sont égales. Le bénéfice dans une enchère au premier prix est, quant à lui, plus grand que $\mathcal{R}$ car on compare alors une situation dans laquelle les acteurs ont des valeurs $v_i \geq e_i$ et des budgets $b_i \geq e_i$ avec une situation dans laquelle ils ont des valeurs e_i.

Dans le mécanisme que nous envisageons, les acteurs se voient attribuer un budget global qu'ils répartissent sur les différentes enchères attribuant les plages horaires du planning. Les jetons n'ayant aucune valeur marchande en soi, et ceux éventuellement épargnés ne pouvant être reportés d'une semaine à l'autre, il n'est pas dans l'intérêt des praticiens de miser moins que le nombre total de jetons dont ils disposent chacun[2]. Ceci constitue bien une contrainte de budget, justifiant par ailleurs notre choix pour un schéma d'enchères au premier prix (notre objectif étant de maximiser l'utilité globale des chirurgiens, et donc le revenu, en termes de jetons, provenant des enchères).

Enchères multiples

Il existe différents types d'enchères multiples. Il peut s'agir de biens identiques que l'on vend lors d'une unique enchère à un ou plusieurs acheteurs, de biens identiques vendus unitairement lors d'enchères séquentielles ou encore de biens différents mais substituts l'un de l'autre vendus en une enchère unique. La situation telle que nous l'envisageons est, quant à elle, une combinaison de deux types d'enchères multiples. En effet, il s'agit d'allouer simultanément des plages horaires différentes, chacune pouvant être substituée par une autre, mais où une plage horaire peut elle-même être attribuée à un ensemble de praticiens dépendant du nombre d'interventions déjà assignées à cette dernière. Autrement dit, plus le plus gros enchérisseur a d'opérations à réaliser, moins il reste de capacité pour les plus petits joueurs. Le nombre d'actes chirurgicaux qu'un petit enchérisseur peut espérer placer sur une plage horaire particulière dépend donc du nombre de traitements à réaliser que compte déjà celle-ci (tout ceci étant déterminé simultanément par l'approche de résolution décrite plus loin). La littérature relative à la vente d'objets multiples traite essentiellement des situations où chaque enchérisseur ne souhaite qu'une seule unité du bien (Klemperer, 1999). Il y a donc peu à dire sur les cas qui nous intéressent vraiment.

[2] Et ce quelles que soient les valeurs qu'ils attribuent à chacune des plages horaires.

Objets identiques. Il y a essentiellement deux formats d'enchères scellées multiples pour vendre W objets identiques : les enchères discriminatoires où chaque vainqueur paie le prix qu'il a annoncé (les prix pour un même bien sont donc différents) et les enchères uniformes (ou concurrentielles) où les gagnants déboursent le prix proposé par le dernier acquéreur (chacun paie alors le même prix). Chaque enchérisseur propose donc un prix pour chaque unité w supplémentaire du bien à pourvoir, représentant la valeur marginale associée à l'obtention de ce $w^{\text{ième}}$ exemplaire de l'objet. Les enchères discriminatoires étant l'équivalent multiple des enchères au premier prix, c'est à elles que nous nous intéressons. Cependant, à l'inverse des dernières, même lorsque les joueurs sont supposés symétriques, il n'est pas possible de dériver une stratégie dominante sous une forme déterminée pour des enchères multiples discriminatoires. De manière générale, celles-ci se montrent inefficaces dans l'allocation des biens, tout comme les enchères concurrentielles d'ailleurs. Le théorème d'équivalence des revenus peut, quant à lui, s'étendre au cas multiple de manière assez directe lorsqu'un enchérisseur souhaite au plus une unité du bien vendu. Dans ce cas, enchères concurrentielles et discriminatoires sont équivalentes, sinon rien ne peut être inféré sur leur revenu respectif (Klemperer, 1999; Krishna, 2002).

Objets différents. La littérature est encore moins abondante sur le sujet. Nous pouvons toutefois mentionner qu'il existe une équivalence des revenus au sens faible (à une constante additive près) lorsque, pour les mécanismes en relation, miser la valeur de l'objet est un équilibre. Parmi l'ensemble des mécanismes allouant différents objets, celui de Vickrey-Clarke-Groves, dérivé d'une enchère au second prix, maximise le profit attendu de chaque agent (Krishna, 2002).

6.3.2 Méthodologie pour la programmation opératoire

Après avoir succinctement présenté les différents mécanismes d'enchères et les stratégies optimales à appliquer par les acteurs, nous développons à présent notre vision des enchères au sein du processus étudié. Le processus de programmation opératoire tel qu'envisagé comporte trois étapes : tout d'abord, une phase d'enchères dans laquelle les chirurgiens concernés font des offres pour les plages horaires de leurs choix ; ensuite, la phase de programmation opératoire proprement dite qui détermine le planning optimal de la semaine tenant compte de ces choix ; enfin, une étape facultative de négociation permet aux éventuels chirurgiens mécontents de solliciter un changement de plage horaire avec un confrère.

Mécanisme d'enchères. Comme nous l'avons vu, la théorie des jeux recense bon nombre de mécanismes d'enchères mais pour sa simplicité de mise en œuvre, nous avons retenu le principe d'enchères scellées. Ce dernier se caractérise donc par un unique tour d'enchères durant lequel chaque participant remet une offre sans connaître celles émises par ses adversaires et sans pouvoir surenchérir une fois l'offre déterminée. L'enchère la plus élevée remporte alors l'objet convoité. Pratiquement, chaque chirurgien se voit attribué un certain nombre de jetons sur base du poids qu'il prend dans le processus de décision et selon les concessions issues de négociations antérieures. Ce poids peut varier selon l'ancienneté du chirurgien, selon le nombre d'interventions qu'il a à réaliser durant la période considérée, selon son enclin à coopérer, etc. Dans ce travail, nous supposons que chaque chirurgien est représenté équitablement dans le système d'enchères. En fonction de leurs préférences, les chirurgiens placent ensuite leurs jetons sur les plages horaires (ici des demi-

journées) durant lesquelles ils désirent opérer. Le nombre de jetons à disposition d'un chirurgien peut bien entendu être distribué sur plusieurs plages horaires. Par hypothèse, les jetons non utilisés sont perdus, de même que les mises n'ayant pas remporté d'enchères. Il n'est effectivement pas permis de reporter des jetons d'une semaine à l'autre (la programmation opératoire déterminant ici l'activité hebdomadaire des chirurgiens). A la différence d'une enchère scellée traditionnelle, il y aura ici plusieurs vainqueurs par plage horaire. Une plage horaire correspond, en effet, à la mise à disposition des différentes salles en parallèle, chacune pouvant accueillir plusieurs opérations et dès lors plusieurs chirurgiens. De même, comme dix plages horaires sont à allouer (rappelons qu'une plage horaire correspond à une demi-journée du planning, celui-ci comportant cinq jours d'ouverture), les dix enchères associées se déroulent simultanément. Les vainqueurs, à savoir les chirurgiens ayant placé les plus grosses mises, sont sélectionnés par la phase suivante du processus, soit la conception du planning opératoire optimal pour la semaine considérée. En effet, l'attribution des plages horaires ne dépend pas uniquement des enchères, mais également de la satisfaction des contraintes associées au fonctionnement du quartier opératoire. Cette approche d'optimisation aura notamment pour objectif de maximiser le montant des mises remportant une enchère, assurant de la sorte aux plus gros enchérisseurs d'obtenir la plage horaire de leur choix.

Programmation opératoire. La construction des programmes opératoires hebdomadaires se fait sur base d'une approche d'optimisation. Un modèle mathématique en nombres entiers mixtes permet de définir les plannings optimaux tout en tenant compte d'une série de contraintes liées au fonctionnement du quartier opératoire. L'objectif poursuivi est double : d'une part minimiser les coûts liés à l'utilisation des salles d'opération, et d'autre part, maximiser les mises gagnantes associées aux plages horaires attribuées. Dans ce dernier objectif, nous avons pris le parti de pondérer les mises des chirurgiens par le nombre d'opérations assignées aux plages horaires correspondantes. Les chirurgiens possédant initialement le même nombre de jetons, cela permet de différencier ceux devant intervenir plus de fois au cours de la semaine. Nous aurions pu procéder autrement et ne pas tenir compte du nombre d'interventions réalisées par chacun, l'objectif ne consistant alors qu'à maximiser le nombre de mises gagnantes. Cependant, cette dernière modélisation de l'objectif a logiquement pour inconvénient de répartir la charge d'un praticien sur un maximum de plages horaires, même dans les situations où il est possible de concentrer celle-ci sur une ou deux demi-journées. Pondérer les mises par le nombre d'opérations programmées, et donc la démarche que nous avons adoptée, évite de scinder l'activité opératoire plus que nécessaire, et permet d'épargner au chirurgien un désagrément certain en lui évitant, par exemple, de se présenter pour une unique opération. Lorsque le calendrier opératoire est dressé, il peut arriver que, pour des questions de conflits ou de contraintes fonctionnelles, certains chirurgiens soient mécontents des plages horaires qui leur sont attribuées. Vient alors une phase de négociation permettant à un chirurgien de proposer un échange de plage horaire.

Protocole de négociation. Le chirurgien insatisfait de son horaire peut encore essayer de modifier celui-ci. Dans un premier temps, il contacte les chirurgiens assignés à la plage horaire qu'il convoite (ph_{conv}) afin de procéder à un échange des demi-journées. Pour ce faire, il s'adresse prioritairement aux chirurgiens devant y opérer et n'ayant pas misé sur celle-ci. Ce sont en effet les acteurs n'ayant rien à perdre dans un échange de plages horaires et par conséquent certainement les plus aisés à convaincre. Le cas échéant, si tous les praticiens assignés à la plage ph_{conv} ont une préférence pour celle-ci, le chirurgien mécontent classe alors les confrères à contacter selon

un ordre bien établi : il les sélectionne par ordre décroissant de différence entre les mises que ces derniers ont placées sur la plage convoitée ph_{conv} et les mises placées sur sa plage horaire initiale (ph_{init}). Le premier chirurgien joint est donc préférentiellement un praticien sans mise sur ph_{conv}, sinon celui ayant les enchères les plus proches sur ph_{init} et ph_{conv}. Le chirurgien approché évalue alors l'offre proposée. Il va de soi que, si l'échange des plages horaires occasionne une perte d'utilité pour ce dernier (en d'autres mots, il lui sera assigné une plage horaire sur laquelle il avait placé une moindre mise), il sera peu enclin à accepter l'inversion des deux interventions. Le chirurgien initiant la négociation peut alors s'engager à lui payer, en contrepartie, une quantité de jetons égale à la différence entre l'utilité que son confrère obtiendra après modification et celle qu'il possède actuellement (utilité ici traduite en nombre de jetons), ce afin d'assurer le bon déroulement de la négociation. Si ce montant est estimé trop important par le chirurgien demandeur, ce dernier met fin à la négociation et chacun garde son horaire tel qu'établi à l'étape précédente. En revanche, si le paiement est jugé acceptable, le chirurgien désireux de changer de plage horaire accepte la transaction et cédera, lors de la semaine suivante (ou lors d'une semaine ultérieure à déterminer entre les intervenants), une partie de ses jetons au chirurgien avec lequel il vient de négocier. L'avantage pour ce dernier est donc d'avoir à disposition un plus grand nombre de jetons lors de la prochaine phase d'enchères, et dès lors un poids plus important durant les mises. Si le programme opératoire devait ne pas satisfaire plusieurs chirurgiens, leurs demandes respectives seraient alors traitées de manière séquentielle, dans l'ordre de manifestation de leur mécontentement, c'est-à-dire selon la règle du premier arrivé, premier servi.

Cette étape de négociation s'effectue à l'aide du modèle mathématique utilisé dans l'étape précédente du processus. Ce dernier est appliqué séparément à chacun des jours impliqués dans la négociation (à savoir les jours comportant les plages horaires ph_{conv} et ph_{init} des deux chirurgiens en négociation), moyennant des contraintes supplémentaires fixant les plages horaires des chirurgiens présents durant ces jours et ne participant pas à la négociation. Notons que seule la demi-journée de ces interventions est fixée, l'heure de début quant à elle pouvant varier au sein de cette demi-journée générant de la sorte une perturbation du calendrier opératoire initial. La fonction objectif est, par conséquent, également adaptée et comporte à présent un terme minimisant la perturbation du planning initialement construit lors de la phase deux du processus de programmation opératoire.

6.3.3 Modèle mathématique

Pour rappel, la formulation initiale sur laquelle se base la suite de cette section avait pour objet d'éradiquer la symétrie qui caractérise la programmation opératoire de salles identiques. Ce modèle, présenté en page 65, est ici adapté pour correspondre aux préoccupations sociales qui nous animent et s'ajuster au mécanisme d'enchères proposé, ainsi qu'au protocole de négociation qui l'accompagne. A cette fin, le système d'enchères scellées est modélisé par la matrice $E \in \mathbb{N}^{C \times 2D}$, dont l'entrée $E_{c,2d-1}$ désigne la mise, c'est-à-dire le nombre de jetons que le chirurgien c place sur la matinée du jour d, et $E_{c,2d}$ est celle placée sur l'après-midi du même jour.

Le problème de programmation des blocs opératoires, tel qu'envisagé ici, peut se formuler mathématiquement au moyen du programme linéaire suivant. Celui-ci est en grande partie équivalent au modèle décrit en section 5.2 dont nous rappelons ici les grandes lignes. Les indices j, d et t représentent respectivement les interventions chirurgicales à réaliser, les jours de planifica-

tion et les périodes de temps opératoire. Cette formulation comporte deux variables : les variables binaires x_{jdt}, dédiées aux activités chirurgicales, prennent la valeur 1 uniquement lorsque l'opération j débute le jour d au moment t, et valent zéro sinon ; les variables entières y_d, quant à elles, représentent le nombre de blocs opératoires utilisés durant la journée d. Les contraintes de fonctionnement du quartier opératoire sont décrites par les équations (5.19) à (5.32) explicitées en page 66. Nous considérons deux fonctions objectifs associées à ces contraintes, selon que le modèle soit utilisé pour la phase de programmation opératoire hebdomadaire ou pour la phase de négociation du processus.

Programmation opératoire hebdomadaire. La fonction objectif pour cette phase du processus consiste à minimiser les coûts en termes d'ouverture de salles et en termes de travail en heures supplémentaires, tout en maximisant les mises liées aux enchères satisfaites (c'est-à-dire maximiser le nombre total de jetons associés aux plages horaires attribuées). Cette première fonction objectif s'écrit comme suit :

$$\min_{x,y} \sum_{d} C^{open}\, y_d \;-\; \epsilon \sum_{c} \sum_{h=1}^{2D} E_{ch}\, a_{ch} \tag{6.1}$$
$$+\, C^{over} \left(\sum_{j} \sum_{d} \sum_{\substack{t=D_d^N-p_j+1 \\ t>0}}^{T-p_j} (t + p_j - D_d^N) x_{jdt} \right),$$

où ϵ, le poids des enchères dans la fonction objectif, est à mettre en balance avec l'objectif économique de celle-ci. La fonction (6.2) requiert l'introduction des variables a définies par les contraintes (6.2)-(6.4) suivantes :

$$a_{c,2d-1} = \sum_{j \in O(c)} \sum_{t=1}^{T/2} x_{jdt} \qquad \forall c, \forall d, \tag{6.2}$$

$$a_{c,2d} = \sum_{j \in O(c)} \sum_{t>T/2}^{T} x_{jdt} \qquad \forall c, \forall d, \tag{6.3}$$

$$a \in \mathbb{N}^{C \times 2D}. \tag{6.4}$$

Les variables a_{ch} représentent alors le nombre d'opérations que le chirurgien c réalise effectivement la demi-journée h. Il en découle que l'objectif économique de la programmation opératoire est diminué du nombre d'opérations effectuées chaque demi-journée pondéré par la mise associée à cette demi-journée par le chirurgien attitré. Plus cette mise est importante, plus ce dernier aura de chance de se voir attribué la demi-journée de son choix. Rappelons que, pour chaque plage horaire, nous pondérons les mises par le nombre d'opérations réalisées afin d'éviter que l'activité opératoire d'un chirurgien soient trop morcelée dans le planning.

Négociation. A l'inverse, la phase de négociation du processus ne nécessite plus de tenir compte des enchères placées par chacun des chirurgiens. En effet, cette phase n'implique directement que

deux chirurgiens : l'un, mécontent de la plage horaire ph_{init} assignée à l'une de ses interventions j^{neg1} et désireux de la changer, et l'autre, disposé à céder la tranche horaire occupée par son intervention j^{neg2} dans la demi-journée ph_{conv} convoitée par le premier (moyennant rétribution éventuelle sous forme de jetons). Cette dernière étape du processus ne remet en question qu'au plus deux journées du programme opératoire initial, les autres restant bien évidemment inchangées. L'approche proposée consiste à appliquer le MIP séparément sur les deux journées en question, en partant de la solution initiale et en y substituant, dans l'ensemble J_d des interventions assignées au jour d qu'il faut modifier, les opérations adéquates (par exemple, $J_d = \left(J_d \setminus \{j^{neg2}\}\right) \cup \{j^{neg1}\}$ pour la plage horaire ph_{conv}, et inversement pour ph_{init}). Il est évident que les opérations de la journée d n'intervenant pas dans le protocole de négociation ne peuvent changer de plage horaire (c'est-à-dire de demi-journée). Il s'agit donc d'insérer une opération j^{neg1} en remplacement d'une autre j^{neg2}, en minimisant les coûts de fonctionnement du quartier opératoire, tout en minimisant la perturbation du planning initial de la journée considérée. Ceci se fait par l'entremise de la fonction objectif suivante, où l'indice d des jours n'est préservé que pour une question de consistance des notations[3] :

$$\min_{x,y} \sum_{d} C^{open}\, y_d \;+\; \delta_{jdt} \;+\; C^{over} \left(\sum_{j \in J_d} \sum_{d} \sum_{\substack{t=D_d^N - p_j + 1 \\ t>0}}^{T-p_j} (t + p_j - D_d^N) x_{jdt} \right), \qquad (6.5)$$

où δ_{jdt} sont des variables entières positives mesurant la différence entre les plannings initial et final, définies par les équations suivantes :

$$\delta_{jdt} \geq x_{jdt} - X_{jdt}, \qquad \forall j \in J_d \setminus \{j^{neg}\}, \forall d, \forall t, \qquad (6.6)$$

$$\delta_{jdt} \in \mathbb{N}, \qquad \forall j \in J_d \setminus \{j^{neg}\}, \forall d, \forall t, \qquad (6.7)$$

où X_{jdt} est le programme opératoire initial de la journée d sujette à modification. Les contraintes (6.6) et (6.7) diffèrent selon que l'on modifie la demi-journée convoitée ph_{conv} ou initiale ph_{init} ; j^{neg} y représente alors respectivement j^{neg1} et j^{neg2}. Cette phase nécessite encore l'introduction de contraintes supplémentaires fixant les demi-journées des interventions n'étant pas impliquées dans le protocole de négociation, ceci afin de ne pas léser les chirurgiens satisfaits du calendrier opératoire initial. Soit J^{am} et J^{pm}, respectivement l'ensemble des opérations devant se dérouler le matin et l'ensemble des opérations devant se réaliser l'après-midi du jour d, et tels que $J^{am} \cup J^{pm} = J_d$, ces contraintes s'expriment comme suit :

$$\sum_{d} \sum_{t=1}^{T/2} x_{jdt} = 1, \qquad \forall j \in J^{am}, \qquad (6.8)$$

$$\sum_{d} \sum_{t>T/2}^{T} x_{jdt} = 1, \qquad \forall j \in J^{pm}. \qquad (6.9)$$

[3] L'indice d est en effet superflu puisque la phase de négociation considère chaque jour séparément, d ne peut dès lors y prendre qu'une unique valeur représentant le jour considéré.

Mise en œuvre. Les modèles mathématiques utilisés dans notre approche sont décrits ci-dessus. Voici à présent brièvement exposée la marche à suivre afin d'optimiser le processus de programmation opératoire dans une optique de coopération. Après la détermination des mises individuelles sur les différentes plages horaires, l'on procède à la construction du planning opératoire de la semaine en résolvant le modèle MIP constitué des contraintes (5.19)–(5.32) et de la fonction objectif (6.2)–(6.4). Ensuite, s'il n'y a aucune réclamation de la part des chirurgiens, le planning est appliqué tel quel. Par contre, si un ou plusieurs chirurgiens sont mécontents, l'on passe alors par une phase de négociation. Celle-ci se réalise en appliquant la formulation composée des contraintes (5.19)–(5.32) et de la fonction objectif (6.5)–(6.9). Si l'inversion de deux opérations est réalisable et si les chirurgiens concernés se sont mis d'accord sur une éventuelle rétribution (sous forme de jetons), alors le programme opératoire est adapté (pour autant que le gestionnaire de blocs ait remis un avis favorable). Sinon, la négociation a échoué ou elle ne satisfait pas à toutes les contraintes, le planning reste inchangé et l'on traite l'éventuelle négociation suivante. Notons encore que, à l'instar de la formulation (5.18) à (5.32), les modèles utilisés ici ne déterminent pas directement la salle d'opération assignée à chaque intervention. Ce choix de modélisation nous permet en effet de contrer la symétrie du problème et n'affecte en rien la solution proposée étant donné la polyvalence supposée des salles. Pour affecter les opérations aux différents blocs opératoires, il suffit d'appliquer, sur la solution optimale, l'algorithme 2 de type bin packing décrit en section 5.2.3.

6.3.4 Expérimentations numériques

Les programmes mathématiques présentés ci-dessus modélisent le processus de programmation du quartier opératoire. Ils ont pour objectif d'aider le gestionnaire à mieux prendre en compte les contraintes liées au fonctionnement de ce dernier, et par conséquent d'améliorer ce processus de programmation en en minimisant les coûts. Afin de jauger l'applicabilité d'une telle approche, nous l'avons appliquée sur un cas d'étude réel présenté en section 3.2. Le MIP est toujours codé à l'aide du langage de modélisation AMPL et résolu par la version 10.0 du solveur CPLEX.

A l'instar de la formulation non coopérative décrite au chapitre 5, de premières expérimentations nous ont montré que la formulation destinée à optimiser la programmation hebdomadaire du quartier opératoire (composée des contraintes (5.19)–(5.32) et de la fonction objectif (6.2)–(6.4)) est en mesure de produire rapidement de bonnes solutions sans pouvoir prouver suffisamment vite leur optimalité. Dans le cas de l'instance comportant 80 interventions (la charge hebdomadaire moyenne), seules six minutes sont nécessaires pour obtenir une solution à moins d'un pour-cent de la meilleure borne inférieure, alors qu'il faut plusieurs heures pour prouver que le programme opératoire ainsi créé est optimal. Par conséquent, nous appliquons la même stratégie que pour la formulation initiale (5.18)–(5.32), à savoir résoudre le problème jusqu'à un certain niveau prédéfini satisfaisant de gap pour déterminer les valeurs quasi optimales des variables y_d. Ces valeurs sont ensuite fixées de sorte que les variables y_d deviennent des paramètres du modèle. Le solveur est alors relancé jusqu'à l'obtention d'un optimum, avec x_{jdt} pour uniques variables. Bien entendu, la qualité de la solution finale de même que le temps de calcul dépendent du niveau de gap fixé. Ici, ce niveau est fixé à 2% afin d'être assuré que les variables y_d soient fixées à des valeurs quasi-optimales avant de relancer le solveur.

Programmation opératoire. Le tableau 6.8 reprend les performances de notre modèle appliqué à la phase de programmation opératoire hebdomadaire du processus. Nous y comparons les versions coopérative et directive du MIP (pour reprendre les termes déjà utilisés en sous-section 6.2.3), cette dernière ayant déjà prouvé sa supériorité face aux solutions observées en pratique, issues des jeux de données dont nous disposons. Tout d'abord, nous y observons que toutes les instances testées sont résolues en moins d'une demi-heure. En fonction de la complexité de l'instance, le temps de calcul nécessaire à sa résolution varie effectivement d'une à trente minutes. Ensuite, dans la section précédente, nous avons vu que dans le cadre d'un quartier opératoire constitué de blocs dédiés, la coopération permettait non seulement une satisfaction accrue du personnel, mais également une diminution moyenne de 7% des coûts du planning opératoire. Qu'en est-il lorsque ces mêmes salles sont identiques et polyvalentes ? Manifestement, dans ce contexte, la coopération ne permet plus d'économie sur le processus. Ceci s'explique aisément. Lorsque les unités opératoires sont dédiées, la coopération, au prix d'une certaine désutilité, offre la possibilité de diminuer le nombre total de salles utilisées en déplaçant des interventions, contrevenant dès lors les préférences des chirurgiens. N'ayant pas de restriction sur l'accessibilité des salles, le MIP, dans la situation présente, minimise déjà le nombre total de blocs opératoires utilisés, ceux-ci étant identiques. Y greffer un terme social ne lui permet tout logiquement pas de diminuer ce nombre, ce qui explique pourquoi la coopération n'améliore pas l'objectif économique. Par contre, elle ne génère pas non plus de surplus financier (ou de façon très marginale), ce qui en fait une approche tout aussi efficiente que son pendant directif. Notons cependant que la formulation coopérative est ici appliquée avec un niveau de gap de 2% alors que la version directive est, elle, appliquée avec un niveau de 1%. Ceci explique dès lors aussi bien les différences de valeurs objectifs que les différences de temps de calcul. Le seuil de 2% pratiqué ici permet d'obtenir des solutions quasi-optimales en des temps raisonnables pour toutes les instances, ce que ne permettait pas le seuil de 1%. Comparativement aux calendriers opératoires conçus manuellement, les solutions coopératrices optimisées réduisent donc les coûts opératoires de 19% à 40%, ce qui représente à nouveau des gains non négligeables pour l'hôpital. Ces résultats nous renforcent encore dans notre démarche d'intégration des étapes de planification et d'ordonnancement du processus de programmation opératoire.

Concernant l'aspect humain du processus, chaque praticien s'est vu accorder un poids de 100 jetons dans la procédure d'enchères. N'ayant pas à disposition les préférences réelles des chirurgiens, nous avons équitablement réparti les jetons de chacun parmi les demi-journées durant lesquelles chacun a réalisé des interventions dans le planning effectif construit par le gestionnaire. Si la valeur associée à l'objectif social est délicate à interpréter telle quelle, l'analyser plus en détails nous apprend qu'en moyenne 56% des mises remportent une enchère, que 58% des jetons misés sont valorisés par le programme et que la totalité des interventions programmées le sont sur des plages horaires comportant une mise du chirurgien attitré. Le tableau 6.8 nous révèle donc que le modèle est efficient économiquement, mais qu'il l'est également socialement puisque les mises du mécanisme d'enchères sont bien respectées.

Négociation. Afin d'illustrer la phase de négociation, nous nous concentrons sur l'instance comportant 88 opérations à programmer. Dans le programme opératoire optimal, nous constatons que le chirurgien 8 se voit programmer la plupart de ses opérations le mardi, excepté l'opération 17 qui, elle, est prévue le jeudi. Désireux de modifier cet état de fait, il contacte alors le chirurgien 48 qui a placé des enchères équivalentes sur les journées de mardi et jeudi et qui possède une plage

Nbr.	Form. Directive		Form. Coopératrice		
Op.	Coûts [€]	Tps [s]	Coûts [€]	Enchères	Tps [s]
60	33120*	*168*	33120*	1642	*79*
70	44880*	*127*	45120*	1845	*162*
72	42780*	*4911*	42780*	2331	*670*
80	48960*	*415*	48960*	2247	*383*
85	50750*	*4814*	50750*	2409	*705*
86	44330*	*1481*	44570*	2353	*767*
88	45170*	*439*	45170*	2805	*625*
91	48260*	*680*	48260*	2789	*1883*
100	49360*	*520*	49600*	2872	*462*

TAB. 6.8 – *Performances des formulations directive* (5.18) *et coopérative* (6.2)–(6.4)*, toutes deux associées aux contraintes* (5.19)–(5.32)*, sur neuf jeux de données contenant chacun l'activité opératoire d'une semaine (*$\epsilon = 1$ *dans la fonction objectif* (6.2)*). Le niveau satisfaisant de gap est fixé à 1% pour la formulation directive et à 2% pour la formulation coopérative, avant de relancer le solveur avec les variables* y_d *fixées.*

horaire le mardi matin (tous les chirurgiens ayant exprimé des préférences sur la plage horaire qu'il convoite, le demandeur contacte donc le collègue ayant la plus grande mise sur la plage horaire qu'il souhaite céder). Supposant que ces deux praticiens se soient accordés sur une potentielle rétribution, l'on tente alors d'inverser ces deux opérations à l'aide du modèle composé des équations (5.19)–(5.32) et (6.5)–(6.9). Le remplacement de l'opération 24 par la 17 le mardi matin se fait en quatre secondes, tandis que l'insertion de l'opération 24 dans la journée du jeudi nécessite 174 secondes, des temps somme toute raisonnables. La différence de temps nécessaire à l'aménagement de ces deux journées s'explique par la différence de charge de travail entre celles-ci. En effet, la journée du jeudi est de loin la plus chargée avec 23 opérations planifiées dans cinq salles. En revanche, la journée de mardi ne comprend que 14 interventions chirurgicales réparties dans quatre salles. La figure 6.3 illustre l'aménagement qu'engendre l'insertion de l'opération 17 dans la journée du mardi et l'opération 24 dans celle du jeudi. La durée opératoire de l'intervention 17 étant bien supérieure à celle de l'intervention 24, son intégration au planning du mardi requiert l'ouverture d'une salle supplémentaire. Lorsque les modifications souhaitées par les médecins génèrent une augmentation substantielle des coûts, comme c'est le cas ici, c'est au gestionnaire de blocs que revient la décision d'entériner, ou non, ces modifications. D'autres cas de figure problématiques rencontrés sont l'impossibilité d'échanger de plage horaire par manque de ressources disponibles ou en raison d'une surcharge de travail pour le chirurgien attitré, violant de la sorte les contraintes (5.25)[4]. Cet outil permet donc d'évaluer facilement et instantanément la possibilité d'échanger les plages horaires de deux opérations et, le cas échéant, et moyennant un temps de calcul raisonnable, de modifier le planning opératoire initial pour en tenir compte.

Ce chapitre démontre qu'il est possible d'inclure une dimension coopérative au processus de programmation opératoire, tout en préservant l'efficience des approches de résolution dévelop-

[4]En d'autres mots, il est impossible de placer l'opération à insérer dans la demi-journée convoitée car cela provoquerait un chevauchement de plusieurs opérations pour le chirurgien concerné. Toutefois, cette problématique peut être contournée en supprimant l'opération concernée de l'ensemble J^{am} dans les contraintes (6.8) ou, le cas échéant, de l'ensemble J^{pm} dans les contraintes (6.9), selon que cette opération soit à placer le matin ou l'après-midi.

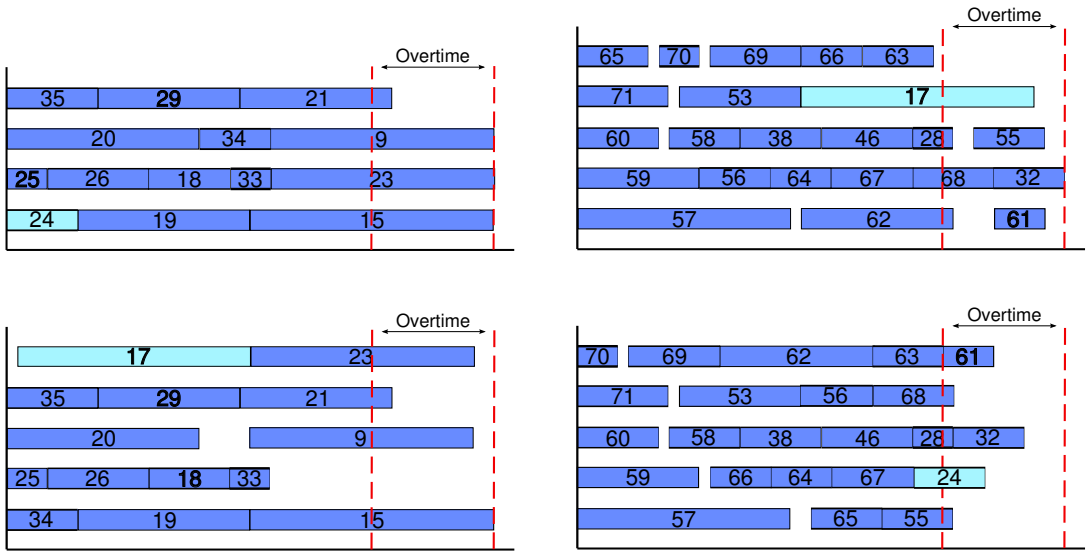

FIG. 6.3 – *Modification des journées du mardi (à gauche) et du jeudi (à droite) après l'échange des opérations 24 et 17.*

pées dans les chapitres précédents. Dans le cas de blocs opératoires dédiés, la coopération permet notamment de réduire les coûts tout en maximisant l'utilité globale des chirurgiens. Dans le cas de blocs opératoires polyvalents, la coopération, telle qu'envisagée dans cette partie, n'offre plus d'avantage économique mais ne dégrade pas pour autant les solutions issues de la formulation mathématique directive du problème. Les temps de calcul restent tout aussi raisonnables alors que la distribution des plages horaires est, elle, plus équitable. Un protocole de négociation basé sur une formulation mathématique en nombres entiers permet en outre l'échange de plages horaires entre d'éventuels acteurs mécontents du planning établi. Capitaliser sur notre moteur de programmation opératoire, incarné par les approches de résolution des deux chapitres précédents, en y incorporant une dimension coopérative marque une nouvelle différence de taille entre nos travaux et ceux de la littérature.

Jusqu'à présent, nous avons toujours travaillé dans un contexte déterministe, ce qui peut soulever la question de la validité de nos approches. En effet, les solutions proposées par celles-ci feront face, en pratique, à certains aléas comme des urgences venant s'intercaler entre deux interventions électives, ou des durées opératoires exceptionnellement longues. Par conséquent, ces solutions doivent être vues comme des plannings de base ayant pour principale fonction d'allouer les ressources nécessaires aux opérations. Cependant les méthodes développées jusqu'ici peuvent prendre en compte certaines incertitudes. En effet, minimiser la quantité de salles ouvertes permet de réserver, autant que possible, des blocs opératoires pour le traitement des aléas, qu'il s'agisse de cas urgents ou de grands retards. Nous pouvons également considérer des durées opératoires, bien que déterministes, incluant des *time buffers* afin d'absorber de petits retards. Bien que les gestionnaires hospitaliers appliquent, eux-mêmes, des procédures déterministes, il serait toutefois opportun d'évaluer quel impact peut avoir la nature stochastique du quartier opératoire sur les performances de celui-ci, entre autres afin d'estimer dans quelle mesure nos approches doivent être adaptées pour en tenir compte, et accroître de la sorte leur robustesse. Cette étude de la variabilité intrinsèque du milieu hospitalier chirurgical fait l'objet du chapitre suivant.

Chapitre 7

Stochasticité et programmation opératoire

Notre recherche, motivée par une application réelle en milieu hospitalier, a été articulée autour de deux perspectives peu abordées dans la littérature sur le sujet : d'une part la possibilité d'intégrer les étapes usuellement séquentielles du processus de programmation opératoire, d'autre part la prise en considération d'une dimension humaine caractéristique du milieu hospitalier. Dans le chapitre 4 nous avons développé un modèle linéaire en nombres entiers mixtes de même qu'une méta-heuristique afin de répondre à la première de ces questions dans un contexte où les salles d'opération sont dédiées. Les expérimentations qui y ont été conduites permettent de répondre par l'affirmative à notre première interrogation. Le chapitre 5, par le biais d'un modèle linéaire en nombres entiers mixtes adapté à la symétrie du problème, nous apporte la même réponse lorsque ces mêmes salles sont identiques et polyvalentes. Ensuite, le chapitre 6 adapte ces mêmes approches au contexte coopératif. Celui-ci passe soit par la déclaration de préférences, soit par un mécanisme d'enchères, afin d'allouer les plages horaires disponibles aux différents chirurgiens, en maximisant l'utilité globale de ces derniers. Cette démarche nous distingue de la littérature classique du domaine et répond à la seconde question initiatrice de ce travail. Nous disposons donc, à ce stade, d'approches de programmation opératoire efficientes, capables de prendre en compte une partie du facteur humain inhérent à l'activité chirurgicale. Il reste néanmoins un point important sur lequel ces méthodes de résolution restent lacunaires : la gestion des événements aléatoires. En effet, l'ensemble de nos approches supposent un contexte déterministe dans lequel il n'y a pas de place pour l'incertitude. Notamment, la durée des opérations est connue avec certitude, les urgences n'ont aucun impact sur le processus, etc. Or, il paraît évident que la réalité est toute autre et que ces événements conditionneront le bon déroulement du calendrier opératoire. Toutefois, ceci ne remet pas en cause la pertinence de nos approches, bien au contraire, puisque les techniques de programmation utilisées *in situ* par les gestionnaires hospitaliers sont elles-mêmes déterministes. Par contre, il s'agit d'évaluer l'impact de la variabilité inhérente au milieu chirurgical sur les performances du quartier opératoire, afin de prendre les dispositions qui s'imposent pour renforcer la robustesse de nos approches déterministes. Principalement, estimer l'importance de la perturbation engendrée par ces aléas permet de déterminer le pourcentage de capacité des blocs opératoires à préserver lors de la construction des plannings afin d'absorber les imprévus sans trop mettre à mal ces programmes opératoires. C'est tout l'objet de ce dernier chapitre qui, par l'intermédiaire de la théorie Markovienne, va nous permettre de rationaliser la prise en compte de l'aléatoire dans le processus étudié.

7.1 Introduction

Il semble évident que l'activité chirurgicale se déroule dans un environnement hautement aléatoire. En effet, la réalisation du calendrier opératoire ne peut se prédire parfaitement en pratique car la réalité est incertaine. Notamment, la durée d'une opération varie en fonction du chirurgien qui la pratique, du patient, des circonstances dans lesquelles se déroule l'intervention (complications ou autres), etc., et ne peut donc qu'être estimée a priori. En outre, le planning prévisionnel peut se voir perturber par des opérations urgentes non prévues survenant inopinément. Cette nature imprévisible des urgences fait d'elles des interventions difficiles à planifier qui peuvent troubler le bon déroulement du programme opératoire. Pour éviter ce genre de désagrément, mais également pour minimiser leur temps d'attente avant prise en charge, certains hôpitaux vont même jusqu'à dédier une ou plusieurs salles d'opération pour le seul traitement des cas urgents. Une autre source majeure de perturbation du planning chirurgical est due au nombre limité de lits de réveil disponibles en salle de surveillance post-interventionnelle (SSPI). Dans le cours normal d'une opération, un patient, une fois opéré, est ensuite conduit en salle de réveil pour y occuper un lit jusqu'à ce qu'il récupère de l'intervention qu'il vient de subir. La durée de cette phase de réveil, et donc le temps d'occupation d'un lit, est évidemment une donnée qui ne peut être déterminée avec certitude. Cependant, il peut arriver qu'à la fin d'une opération, aucun lit ne soit disponible. Dans une telle situation, le patient opéré ne pouvant attendre qu'un lit se libère avant de se voir prodiguer les soins post-interventionnels, celui-ci débute sa phase de réveil dans le bloc opératoire où s'est déroulée son intervention chirurgicale. Ce dernier est dès lors bloqué, ce qui a pour conséquence de retarder, de manière totalement inattendue, les opérations suivantes prévues dans cette même salle. Ce travail ayant pour objectif d'améliorer la gestion du quartier opératoire, nous ne pouvions occulter sa nature imprévisible. Il s'agit donc, dans ce chapitre, d'évaluer dans quelle mesure il est possible de mieux tenir compte des aléas de sorte à proposer des plannings opératoires robustes.

En pratique, afin de pallier la stochasticité de leur environnement, bon nombre de gestionnaires de blocs opératoires construisent des plannings déterministes qui remplissent les salles pour une fraction donnée de leur capacité, c'est-à-dire du temps opératoire disponible (le taux d'occupation), et préservent donc un pourcentage prédéterminé de ce dernier (souvent aux alentours de 15%, voir MEAH, 2006, par exemple). Autrement dit, ils conservent une marge de sécurité qui devrait être suffisante, selon eux, pour absorber l'ensemble des événements non prévus. Ils espèrent, de la sorte, que le temps de travail non alloué sera suffisant pour traiter les opérations étonnamment longues, pour soigner les urgences et pour supporter le possible manque de lits de réveil. L'optique de cette démarche est, bien entendu, d'éviter que l'activité réelle n'engendre trop d'heures supplémentaires tout en assignant de façon efficiente les salles d'intervention. La plupart du temps, le taux d'occupation des blocs est fixé subjectivement, sur base de l'expérience du staff médical ou à l'aide de benchmarks.

Dans ce chapitre, notre objectif est de proposer un outil capable d'aider les gestionnaires à traiter la stochasticité inhérente au quartier opératoire, et en fin de compte de les aider à gérer plus efficacement ce dernier à des niveaux de décision stratégique et tactique. Nous considérons trois sources d'incertitude que sont les temps opératoires, les durées de réveil et l'arrivée inattendue d'urgences. En particulier, notre outil devrait proposer des mesures de performance permettant aux gestionnaires de choisir de manière rationnelle le nombre de lits de réveil à pourvoir ou le nombre de salles à dédier au traitement unique des urgences (niveau de décision stratégique), ou encore la marge de sécurité à prendre sur la capacité régulière, dénommée taux d'occupation ici, pour absorber les aléas (niveau de décision tactique).

Dans cette optique, nous proposons un modèle analytique du quartier opératoire, basé sur un processus de Markov duquel seront également dérivées des distributions *phase-type*, dont les grandes lignes sont explicitées plus loin (voir section 7.3). Notre approche fournit des estimations de diverses mesures de performance comme le temps de perturbation causé par les urgences, le temps d'attente moyen des urgences, le temps de blocage dû à la saturation de la salle de réveil, le temps de travail dans les salles ou encore le taux d'occupation correspondant à un niveau donné d'heures supplémentaires. A des degrés divers, ces mesures donnent une vue d'ensemble de l'efficacité du quartier opératoire en termes de qualité de service aux patients ou de surcoût généré. En pratique, un gestionnaire peut utiliser cet outil pour choisir, parmi un ensemble de décisions, celle qui sera la plus appropriée en évaluant l'impact qu'a chacune sur le comportement du système. Cette partie de notre travail ne comporte pas de volet financier. Les implications que suggère notre modèle sont donc à mettre en correspondance avec le surplus de coûts qu'elles pourraient générer.

Ce dernier chapitre s'articule comme suit. La section suivante est dédiée à un bref survol de la littérature sur le sujet. Nous décrivons ensuite le quartier opératoire tel que nous le modélisons ainsi que les différentes sources de stochasticité que nous considérons. Ensuite, après un bref rappel de la théorie Markovienne, nous détaillons le processus de Markov appliqué pour modéliser analytiquement l'évolution du système, nous offrant la possibilité de calculer une série de mesures de performance. Enfin, nous appliquons notre approche à l'hôpital bruxellois qui nous sert d'étude de cas tout au long de cette recherche et pour lequel nous disposons de données réelles, illustrant ainsi son utilisation pratique.

7.2 Etat de l'art sur la programmation opératoire stochastique

Notre recherche s'articule autour de la gestion des quartiers opératoires et tente d'en rationaliser les prises de décision. Un état de l'art substantiel sur le processus de programmation du quartier opératoire est proposé au chapitre 2. Ici nous nous concentrons sur les travaux qui proposent des approches stochastiques dans le domaine plus large de la gestion hospitalière. Ces approches sont au nombre de trois : la théorie des files d'attente, la théorie Markovienne et la simulation.

Tout d'abord, la **théorie des files d'attente** fut appliquée à de nombreuses reprises en gestion hospitalière au travers de diverses études. Dans ce champ spécifique d'investigation, Green (2006) et Creemers & Lambrecht (2007) présentent de bonnes revues de la littérature. La majorité des études appliquant cette théorie traite du dimensionnement de ressources critiques (comme les lits de réveil par exemple, voir Kao & Tung, 1981) ou de l'allocation de ressources (Gorunescu et al., 2004). A cette fin, Kim et al. (1999) se concentrent sur l'évaluation du temps d'attente nécessaire avant d'entrer dans une unité de soins intensifs. Pour ce qui est de la gestion du flux d'urgences, de Bruin et al. (2007) ont utilisé la théorie des files d'attente pour modéliser le flux de patients cardiaques et ont étudié dans quelle mesure pouvait survenir un phénomène de *bottleneck*, c'est-à-dire de goulot d'étranglement. Les files d'attente forment également une approche classique appliquée au problème de prise de rendez-vous en milieu hospitalier (voir Cayirli & Veral, 2003, pour une revue de la littérature s'y rapportant).

Ensuite, certains travaux ont montré toute l'utilité des **chaînes de Markov** dans la modélisation des flux patients (voir Hall, 2006; McClean et al., 2009). Notamment, pour le dimensionnement de ressources critiques, Harrison et al. (2005) ont montré que les chaînes de Markov of-

fraient une modélisation plus détaillée que les simples files d'attente, et permettaient de calculer le dépassement de capacité. McClean et al. (2006) ont appliqué la théorie Markovienne et utilisé des distributions phase-type afin d'évaluer la durée de séjour de patients en unité de soins gériatriques. L'utilisation de ces distributions phase-type en gestion hospitalière a été passée en revue par Fackrell (2007).

Enfin, les approches de **simulation** ont été largement utilisées afin d'analyser des cas particuliers, mettant en exergue leur principal atout : leur habileté à modéliser en détail des systèmes bien spécifiques. En particulier, les urgences, qui sont stochastiques par nature, ont été abordées par différents auteurs. Kolker (2008) étudie les caractéristiques du flux de patients dans une unité de soins dédiée aux urgences. Komashie & Mousavi (2005) examinent le comportement du système quant aux causes cachées de temps d'attente excessifs. Ruohonen et al. (2006) évaluent la possibilité de mettre en place une équipe de triage des malades afin de mieux allouer les ressources disponibles. Guinet et son équipe de recherche ont développé plusieurs modèles de simulation pour améliorer la gestion de diverses activités dans le quartier opératoire ou dans le service d'urgences (voir en particulier Wang et al., 2008). Dans un autre registre, Hammami et al. (2003) étudient, à l'aide d'une approche par goal programming, l'opportunité d'insérer une urgence dans le programme opératoire tout en minimisant et en équilibrant la charge de travail.

Dans notre étude de la stochasticité, nous considérons deux flux de patients différents convergeant vers le quartier opératoire : les opérations électives et donc planifiées, et les urgences. Dans la littérature dédiée à la programmation opératoire, la place des urgences est assez restreinte (voir sous-section 2.2.2). A notre connaissance, peu de papiers considèrent les deux flux patients et incorporent les urgences dans la planification du quartier opératoire. Gerchak et al. (1996) proposent un modèle stochastique de programmation dynamique afin de déterminer le nombre de demandes additionnelles d'opérations électives que l'on peut accepter par jour, en fonction des urgences et des interventions déjà planifiées. Ils maximisent une fonction profit tout en pénalisant les heures supplémentaires et les reports d'intervention. Un second exemple nous vient de Lamiri et al. (2008). Les auteurs développent une méthode de construction de plannings opératoires tenant compte des urgences. Cependant, ils considèrent les urgences au moyen d'une durée aléatoire requise dans le programme, évitant des mesures telles que le temps d'attente des urgences par exemple. De plus, les temps opératoires des interventions électives y sont déterministes. Aucune de ces deux approches n'envisage la possibilité de dédier des salles au traitement exclusif des urgences. Wullink et al. (2007) étudient bien le moyen le plus adéquat de préserver du temps opératoire afin de traiter les opérations urgentes (dédier des salles ou répartir la capacité requise sur l'ensemble des salles), mais ils ne considèrent pas l'impact de la salle de réveil sur le système ni ne permettent de déterminer quel taux d'occupation appliquer. Nous proposons un modèle qui combine flux électif et urgent, et qui prend en compte l'effet de blocage induit par la saturation de la salle de réveil, celle-ci influençant les performances du système étudié au même titre que les flux de patients.

La problématique des patients entamant la phase de réveil dans les blocs opératoires, bloquant par la sorte ces derniers, a été très peu abordée. Nous pouvons citer Augusto et al. (2010) qui proposent une approche de relaxation Lagrangienne pour déterminer des programmes opératoires quasi-optimaux considérant, si nécessaire, le possible rétablissement dans les salles d'opération. Ils prennent en compte trois types de ressources : les transporteurs, les salles d'opération et les lits de réveil. Fei et al. (2007) traitent le problème de programmation hebdomadaire en le décompo-

sant en deux étapes. Dans l'étape d'ordonnancement journalier, ils envisagent la capacité d'accueil de la SSPI comme une contrainte et autorisent le blocage des salles d'opération. Cependant, ils ne mentionnent pas la présence d'urgences et abordent le quartier opératoire comme une entité déterministe. Finalement, nous n'avons trouvé aucune recherche qui permette à la fois de fixer de façon rationnelle le taux d'occupation idéal qui permettrait d'absorber les événement imprévisibles, de déterminer combien de lits de réveil sont nécessaires pour préserver une faible probabilité de blocage, ou de mesurer l'impact d'une salle dédiée sur le temps d'attente des urgences.

7.3 Théorie de Markov

Pour étudier l'impact de la stochasticité du quartier opératoire sur ses performances, afin d'en tenir compte dans nos approches déterministes et de construire des programmes opératoires robustes, nous allons modéliser le comportement de ce dernier à l'aide d'un processus de Markov. Les processus de Markov ont été largement utilisés en pratique pour déterminer des mesures de performance et de fiabilité. Quelques-unes de leurs applications sont la quantification du débit de lignes de production, la détermination du temps moyen entre deux défaillances d'un système de production ou encore l'identification de goulots d'étranglement dans des réseaux de communication. Avant d'entrer dans le vif du sujet, et sans verser dans les détails, nous allons présenter les grandes lignes de la théorie que l'on doit au mathématicien russe A. A. Markov. Ses travaux portaient sur la théorie des probabilités, et particulièrement sur la mise au point d'un processus stochastique dont le résultat d'une expérience affecte le résultat de l'expérience suivante. Ce type de processus prend le nom de *chaîne de Markov*. Cette section s'entend donc comme un bref rappel de la théorie Markovienne.

7.3.1 Chaîne de Markov à temps discret

Les chaînes de Markov représentent une classe importante de processus stochastiques à temps discret, pour lesquels le comportement futur ne dépend que du présent, et non du passé. Rappelons qu'un processus stochastique $\{X_t\}$ est une collection de variables aléatoires indexées par un indice $t \in T$ (représentant généralement le temps) et définies sur un même espace de probabilités. La variable X_t représente alors l'état du processus au temps t, l'ensemble des valeurs que peut prendre cette variable étant appelé l'espace des états du processus et noté S. Si l'espace S est fini (ou dénombrable), le processus est appelé chaîne ; si l'espace T est fini (ou dénombrable), le processus est dit à temps discret. Un état est généralement représenté par un identifiant, comme un vecteur par exemple, dont les composants décrivent les différentes parties du système analysé. Cet identifiant doit être suffisamment détaillé pour caractériser entièrement l'objet d'étude, tout en restant aussi succinct que possible pour éviter l'explosion de l'espace des états et les importants efforts de calcul qui vont de pair.

Une chaîne de Markov est donc un processus stochastique à temps discret $\{X_n, n \geq 0\}$, défini sur un ensemble fini d'états S, vérifiant la propriété de Markov suivante :

$$P[X_n = i | X_{n-1}, \ldots, X_0] \;=\; P[X_n = i | X_{n-1}], \qquad \forall i \in S,\; n \geq 1.$$

Autrement dit, l'état courant comporte toute l'information nécessaire permettant d'accéder à l'état suivant du système, et le chemin parcouru dans le passé jusqu'à cet état n'influence en rien sur

le futur du système. Une chaîne de Markov démarre dès lors dans un état particulier du système et évolue à intervalles de temps réguliers, passant successivement d'un état à un autre, au gré des probabilités de transition $p_{ij} = P[X_n = j|X_{n-1} = i]$ entre états. Dès lors, p_{ij} représente la probabilité conditionnelle que le système évolue vers l'état j sachant qu'il est pour l'instant dans l'état i. La matrice P contenant l'ensemble des probabilités p_{ij} est appelée la matrice de transition de la chaîne de Markov. Par définition, la somme des éléments sur chacune de ses lignes vaut nécessairement 1.

Deux types de probabilités sont calculables à partir d'un tel processus stochastique : les probabilités transitoires, considérant alors le système à un moment donné n, ou stationnaires, déterminant le système lorsqu'un équilibre est atteint. Initialement, la chaîne se trouve dans un état du système selon une probabilité distribuée par π^0. Son comportement transitoire à l'instant n dépend de sa situation de départ et se représente par un vecteur de probabilités π^n qui caractérise l'état du système après n transitions. Ce vecteur se définit comme suit :

$$\pi^n = \pi^{n-1}P = \pi^{n-2}P^2 = \cdots = \pi^0 P^n. \tag{7.1}$$

Sur le long terme, ces probabilités transitoires convergent vers un équilibre : les probabilités stationnaires. Si la chaîne est irréductible et apériodique (autrement dit, tous les états communiquent en un nombre fini et non périodique de transitions), les probabilités stationnaires $\pi = \lim_{n\to\infty} \pi^n$ existent et sont déterminées en résolvant l'équation suivante :

$$\pi = \pi P, \tag{7.2}$$

où $\sum_{i\in S} \pi_i = 1$. Les probabilités stationnaires π s'interprètent alors comme la proportion du temps durant laquelle le système se trouve dans un état particulier, en moyenne. A l'inverse des probabilités transitoires, les probabilités stationnaires sont indépendantes de la distribution initiale π^0.

A titre illustratif, voici un petit exemple classique de chaîne de Markov. Modélisons l'évolution de la météo d'une île paradisiaque où il ne pleut jamais s'il a fait beau la veille, et s'il pleut, ce n'est jamais plus de deux jours de suite. Nous supposons que le temps d'aujourd'hui suffit à déterminer les probabilités du temps de demain. S'il y fait beau, il y a quatre fois plus de chances qu'il fasse beau le lendemain qu'il ne fasse nuageux. S'il fait nuageux, la moitié du temps il fera toujours nuageux le lendemain, alors que dans 30% des cas il fera beau, sinon il pleuvra. S'il pleut pour la première fois, alors dans un cas sur deux il fera beau le jour suivant, dans 30% des cas il fera nuageux, sinon il pleuvra pour la deuxième fois consécutive. Enfin, si la pluie persiste depuis deux jours, l'on peut être sûr que le lendemain il fera beau. Il s'agit bel et bien d'une île paradisiaque et définitivement pas de la Belgique ! Sur base de ces informations, nous pouvons créer une chaîne de Markov en prenant comme états les types de temps : B (pour beau), N (pour nuageux), P_1 (pour le premier jour de pluie) et P_2 (pour le deuxième jour de pluie consécutif), comme l'illustre la figure 7.1. La matrice de transition définissant cette chaîne de Markov est la suivante :

$$P = \begin{pmatrix} 0.8 & 0.2 & 0 & 0 \\ 0.3 & 0.5 & 0.2 & 0 \\ 0.5 & 0.3 & 0 & 0.2 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Supposons que, initialement, il pleuve depuis deux jours, de sorte que $\pi^0 = (0\ 0\ 0\ 1)$. Après trois

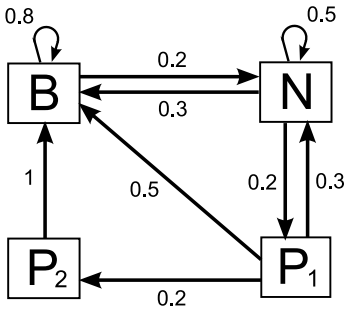

FIG. 7.1 – *Exemple illustratif d'une chaîne de Markov modélisant l'évolution de la météo dans une île paradisiaque.*

jours, les probabilités transitoires des états sont obtenues en appliquant l'équation $\pi^3 = \pi^0 P^3$, où :

$$P^3 = \begin{pmatrix} 0.66 & 0.28 & 0.05 & 0.01 \\ 0.59 & 0.31 & 0.08 & 0.02 \\ 0.66 & 0.28 & 0.05 & 0.01 \\ 0.70 & 0.26 & 0.04 & 0 \end{pmatrix}.$$

Donc $\pi^3 = (0.7 \;\; 0.26 \;\; 0.04 \;\; 0)$. Autrement dit, il y a 70% de chance qu'il fasse beau dans trois jours et seulement 4% qu'il pleuve. De plus, les probabilités stationnaires peuvent être calculées en appliquant la formule (7.2) ; ce qui nous donne $\pi = (0.64 \;\; 0.29 \;\; 0.06 \;\; 0.01)$ à l'équilibre. Globalement, sur cette île lointaine, il ne pleut que sept jours sur cent en moyenne.

7.3.2 Chaîne de Markov à temps continu

Ce que traditionnellement l'on nomme processus de Markov n'est autre que la version en temps continu de la chaîne de Markov. Un processus de Markov se caractérise toujours par un ensemble fini d'états S, mais les transitions entre états se passent différemment. Comme le temps est continu, il n'y a plus de pas de temps à proprement parler qui caractérise le passage d'un état du système à un autre. Par contre, à partir d'un état particulier, ce sont à présent des événements propres au système étudié qui vont provoquer les transitions. L'événement qui fait évoluer le système d'un état i à un état j (avec nécessairement $j \neq i$) survient après un temps distribué exponentiellement, de paramètre q_{ij} ; q_{ij} est donc le taux de transition de i vers j. Dès lors, le temps que le système passe dans l'état i suit une distribution exponentielle de paramètre $\sum_{j \neq i} q_{ij}$, alors que la probabilité de transition vers l'état j se calcule par $p_{ij} = q_{ij} / \sum_{k \neq i} q_{ik}$. La distribution exponentielle joue un rôle essentiel dans un processus de Markov. En effet, dans le cas continu, la propriété de Markov consiste en l'absence de mémoire du système étudié ; le comportement de celui-ci ne dépend donc pas du passé mais uniquement du présent. Dès lors, le temps passé entre deux transitions dans un processus de Markov doit également satisfaire cette propriété d'absence de mémoire ; le temps passé dans un état du système n'influence pas son futur. La distribution du temps restant avant la prochaine transition dépend uniquement de l'état dans lequel le système se trouve, et non du temps déjà passé dans cet état. Et la seule distribution satisfaisant cette propriété d'absence de mémoire est la distribution exponentielle.

Comme dans le cas discret, nous pouvons construire une matrice de taux de transition Q d'éléments q_{ij} qui caractérise le processus de Markov. Par contre, à l'inverse du cas discret, il n'y a pas de transition d'un état vers lui-même. Par conséquent, nous définissons pour l'état i, $q_{ii} = -\sum_{j \neq i} q_{ij}$, de sorte que la somme des lignes de Q soit nulle (la masse de probabilité sortant de l'état i vers un autre état est conservée). Le comportement transitoire du système à l'instant t, défini par le vecteur de probabilités des états $\pi(t)$, est donné par :

$$\pi(t) \; = \; \pi(0)\, e^{Qt},$$

où $\pi(0)$ est la probabilité initiale d'être dans l'un des états du système. A nouveau, si tous les états du processus de Markov communiquent, nous pouvons déterminer les probabilités stationnaires $\pi = \lim_{t \to \infty} \pi(t)$ des états du système. Ces probabilités s'obtiennent à l'aide de la formule :

$$0 \; = \; \pi\, Q, \tag{7.3}$$

où $\sum_{i \in S} \pi_i = 1$. A nouveau, ces probabilités représentent la proportion moyenne du temps que le système passe dans un état, et ne sont dépendantes ni de la probabilité initiale ni du temps.

Distributions phase-type. Les distributions *phase-type* représentent une classe polyvalente de distributions, théoriquement capables d'approximer toute distribution sur des nombres réels positifs, avec autant de précision que voulue, moyennant une complexité suffisante. Une telle distribution phase-type peut se modéliser comme la distribution du temps avant absorption dans une chaîne de Markov à temps continu comportant un état absorbant (c'est-à-dire un état i dans lequel le système est bloqué, tel que $\sum_{j \neq i} q_{ij} = 0$ et $\lim_{t \to \infty} \pi_i(t) = 1$). Elle est entièrement caractérisée par la matrice de taux de transition du processus de Markov et les probabilités initiales de démarrer dans un état particulier du système (Neuts, 1981). La matrice de taux de transition d'un tel processus comportant $m + 1$ états, dont les m premiers sont transitoires (c'est-à-dire non absorbants), s'écrit comme suit :

$$\Theta = \begin{pmatrix} Q & -Q\mathbf{1} \\ \mathbf{0} & 0 \end{pmatrix},$$

où Q est la matrice carrée de dimension m des taux de transition des états transitoires, $\mathbf{1}$ est un vecteur colonne de dimension m dont les entrées valent un, et $\mathbf{0}$ est un vecteur ligne nul. Si Z est la variable aléatoire modélisant le temps jusqu'à absorption dans l'état $m + 1$, alors la fonction de distribution de cette variable est donnée par :

$$\begin{aligned} F(t) \; &:= \; P[Z \leq t] \\ &= \; 1 - \alpha \, \exp(Qt)\, \mathbf{1}, \end{aligned}$$

où α est le vecteur des probabilités initiales associées aux états transitoires du processus de Markov. Une distribution phase-type continue est donc une composition complexe de distributions exponentielles, représentées par un processus de Markov. La théorie relative à ces distributions est mûre et permet, à partir de la matrice Q et du vecteur α, de calculer bon nombre de mesures comme la moyenne, la variance, la fonction génératrice des moments, etc. Des exemples typiques de distributions pouvant se modéliser via phase-type sont les distributions Erlang (distributions exponentielles en série), hyper-exponentielle (distributions exponentielles en parallèle), Cox (distribution Erlang avec sortie préemptive) ou exponentielle généralisée.

7.3.3 Résolution des chaînes de Markov

Une chaîne de Markov se caractérise par un ensemble d'états et par les transitions qui s'opèrent entre ceux-ci. Une fois spécifiée, la chaîne de Markov peut ensuite être résolue afin de calculer quelles sont les probabilités transitoires et stationnaires des états, pour pouvoir ensuite en dériver des mesures de performance du système étudié. Une vaste littérature existe à ce sujet et propose diverses approches analytiques de résolution (Stewart, 2000). Nous nous intéressons ici uniquement à celles dévolues au calcul des probabilités stationnaires.

Pour déterminer la distribution stationnaire π des états de la chaîne, il est nécessaire de résoudre une équation matricielle qui diffère selon que le temps soit discret ou continu, représentée respectivement par (7.2) et (7.3). Il s'agit donc de déterminer la solution d'un système d'équations linéaires. La résolution directe de ce type d'équation matricielle par les méthodes classiques de résolution de systèmes linéaires, basées sur l'élimination Gaussienne (en appliquant une décomposition LU de la matrice de transition), s'avère généralement délicate lorsque le nombre d'états |S| est grand. C'est pourquoi des techniques spécifiques ont été développées, dont nous en mentionnons les principales ci-dessous. La résolution des chaînes de Markov n'étant pas l'objet de nos travaux, cette section se veut uniquement introductive. Nous nous référons à Stewart (1994), la référence dans le domaine, pour plus de détails.

Méthodes itératives basiques. Certaines techniques comme la *méthode des puissances* basée sur l'équation (7.1) – consistant à calculer les différentes puissances de P afin de résoudre l'équation (7.2) – et les méthodes de *Jacobi* ou de *Gauss-Seidel* qui en dérivent – pour la résolution de (7.3) – déterminent π de façon itérative au départ d'une estimation de ce vecteur solution (voir Stewart, 1994, chapitre 3). Malgré une certaine facilité de mise en œuvre, elles présentent néanmoins l'inconvénient d'un taux de convergence assez lent.

Méthodes par blocs. Lorsque l'espace d'états peut être partitionné avantageusement en sous-ensembles, il est possible de scinder la matrice de transition en concordance afin d'élaborer une résolution itérative basée sur cette partition (voir Stewart, 1994, section 3.3). Ces méthodes de résolution par "blocs" nécessitent généralement plus de calculs par itération mais présentent une convergence plus rapide que les seules techniques exposées précédemment. Si S est subdivisé en N blocs, le problème consiste alors à la résolution de N systèmes d'équations de plus petites tailles. Du choix de la partition dépendra l'efficience de la méthode.

Méthodes de projection. Alors que les méthodes itératives modifient une solution approximée du problème à chaque itération, convergeant (supposément) vers la solution du système linéaire (à titre illustratif, considérons le système $\pi Q = 0$), les méthodes de projection créent des sous-espaces vectoriels et y cherchent la meilleure approximation possible. A partir d'un sous-espace particulier, il est possible de construire une combinaison linéaire $\hat{\pi}$ des vecteurs de base de cet espace qui minimise $\|\hat{\pi}Q\|$ selon une norme vectorielle choisie. Ce vecteur $\hat{\pi}$ représente alors une approximation de la solution de l'équation (7.3). Ceci constitue, par exemple, la base de l'algorithme GMRES (*Generalized Minimal Residual*) qui peut également se décliner sous une forme itérative (voir Stewart, 1994, chapitre 4).

Matrix Analytic Methods. Lorsque la matrice de transition d'une chaîne de Markov présente une structure bien définie et affiche les propriétés de "*skip-free*" (impossibilité de passer d'un état n à un état $n \pm 2$ en une transition, dans au moins une direction de la chaîne) et d'"*homogénéité spatiale*" (la structure de la matrice de transition est répétée à chaque ligne avec un décalage) la distribution π peut se caractériser récursivement. C'est la démarche prônée par les techniques analytiques matricielles (*Matrix Analytic Methods*, voir Neuts, 1981; Latouche & Ramaswamy, 1999), largement utilisées pour l'analyse exacte d'une classe fréquente de modèles de file d'attente. Dans ces modèles, les chaînes de Markov imbriquées sont une généralisation bi-dimensionnelle des files d'attente élémentaires $M/G/1$ et $GI/M/1$, et du processus *Quasi-Birth-Death* qui peut être vu comme leur intersection (les probabilités ou taux de transition deviennent des matrices). Ces techniques mettent à profit la structure de ces problèmes spécifiques afin de restreindre les aspects calculatoires (voir également Stewart, 1994, chapitre 5). Nous retrouverons notamment ce type de structures dans les distributions phase-type détaillées précédemment.

Voilà qui clôture ce petit rappel sur la théorie de Markov et sur les distributions phase-type. Rappelons que la résolution analytique des chaînes de Markov n'est pas abordée dans ce travail et n'est donc mentionnée qu'à titre indicatif. Elle constitue toutefois une suite logique de nos travaux, notamment via l'application de Matrix Analytic Methods. La section suivante est consacrée à la description du problème que nous envisageons de résoudre.

7.4 Description du problème

Dans cette section, nous commençons par détailler la modélisation du quartier opératoire, en focalisant sur les sources de stochasticité et leur impact, et relions ensuite notre modèle au cas d'étude utilisé pour valider nos approches et qui a inspiré nos travaux.

7.4.1 Modélisation du quartier opératoire

Nous nous proposons d'étudier la stochasticité émanant du fonctionnement du quartier opératoire afin de mieux l'appréhender. Le caractère aléatoire du quartier opératoire a essentiellement trois origines.

Les durées opératoires. Les temps nécessaires à l'accomplissement des actes chirurgicaux diffèrent d'un cas à l'autre et ne peuvent être prédits avec certitude. En particulier, pour diverses raisons médicales, une opération peut durer plus longtemps que prévu, retardant alors les interventions programmées dans la même salle.

Les urgences. Ces cas urgents surviennent inopinément et sont supposés être traités dans la journée. Les urgences peuvent provenir de l'extérieur mais également de l'intérieur de l'hôpital si l'état de santé d'un patient hospitalisé décline subitement. Comme elles requièrent des ressources et retardent les opérations électives, elles perturbent le programme opératoire. Leur impact sur ce dernier peut cependant être amoindri en leur dédiant une ou plusieurs salles d'opération, comme il est parfois possible de l'observer en pratique (Trilling, 2006; Augusto et al., 2010). Cette stratégie

permet d'améliorer le soin apporté aux urgences en réduisant leur temps d'attente moyen, tout en diminuant la perturbation et le retard des cas planifiés. Cependant, dédier une salle d'opération réduit le temps opératoire disponible pour réaliser les interventions programmées puisqu'une salle ne leur est dorénavant plus accessible. Cette manière de procéder constitue toutefois une option de notre modèle. Notre modèle offre en effet la possibilité de différencier des salles dédiées aux urgences et des salles polyvalentes (qui permettent de traiter des opérations électives et, si nécessaire, également des urgences). Notons encore qu'un quartier opératoire ne comportant pas de salle dédiée peut aisément être modélisé, chaque intervention urgente étant alors réalisée au sein des salles polyvalentes.

Les durées de réveil. La troisième source d'imprévu provient de la salle de surveillance post-interventionnelle : le nombre de lits de réveil y est limité par des contraintes légales, physiques ou budgétaires, et leurs temps d'occupation, c'est-à-dire le temps nécessaire au patient pour émerger de son opération, sont variables et donc stochastiques. Or, si la totalité des lits disponibles est occupée lorsqu'une opération en cours se termine, le patient opéré ne pouvant dès lors accéder à la SSPI, celui-ci entame sa phase de réveil dans le bloc opératoire dans lequel s'est pratiquée sa chirurgie. Ce bloc opératoire ne peut conséquemment plus être utilisé pour réaliser les interventions à venir y étant programmées, et reste bloqué tant que le patient en attente n'a pas récupéré ou qu'un lit de réveil ne s'est pas libéré. Ce phénomène de blocage génère donc une nouvelle perturbation du programme opératoire, perturbation dont l'ampleur dépend fortement du nombre de lits composant la salle de réveil. Une règle générale consiste à pourvoir cette dernière d'un nombre de lits proche d'une fois et demie le nombre de blocs composant le quartier opératoire (voir notamment Dexter & Tinker, 1995, qui mentionnent une proportion allant de 1.5 à 2).

Le modèle que nous décrivons par la suite est en mesure de tenir compte de ces trois sources d'aléas et permet de mesurer leur effet sur le fonctionnement du quartier opératoire et sur sa programmation. Ceci nous amène au modèle illustré en figure 7.2, où trois processus peuvent y être distingués : l'arrivée des opérations, la réalisation des opérations chirurgicales dans les salles d'intervention et la phase de réveil dans la SSPI.

Les opérations planifiées sont pratiquées dans les salles polyvalentes supposées identiques. Nous considérons que ces dernières sont continuellement approvisionnées par un flux d'interventions électives dont le processus d'arrivée est dit saturé. En d'autres mots, comme ces opérations sont prévues dans le planning prévisionnel, nous supposons que l'opération élective à venir est toujours prête à débuter lorsque celle qui la précède se termine. A l'inverse, les urgences arrivent quant à elles aléatoirement, selon un processus stochastique, et sont considérées, ici, comme des événements prioritaires à traiter au plus vite. Lorsqu'une urgence survient, elle est directement admise dans une salle dédiée, pour autant que la configuration du quartier opératoire en comporte et que l'une d'entre elles soit libre. Le cas échéant, elle rejoint une file d'attente prioritaire jusqu'à ce qu'une unité chirurgicale, quelle qu'elle soit (dédiée aux urgences ou polyvalente), se libère. Si la première opération qui s'achève libère une salle polyvalente, comme elle a priorité sur les cas électifs, l'urgence entre dans cette salle polyvalente et retarde les opérations qui y sont encore à réaliser, perturbant de la sorte le planning établi.

Installés dans les blocs opératoires, qu'ils soient polyvalents (pouvant accueillir cas programmés et urgents) ou dédiés (n'accueillant que les urgences), les patients sont opérés durant un temps stochastique. Les durées opératoires sont des variables aléatoires indépendantes et identiquement

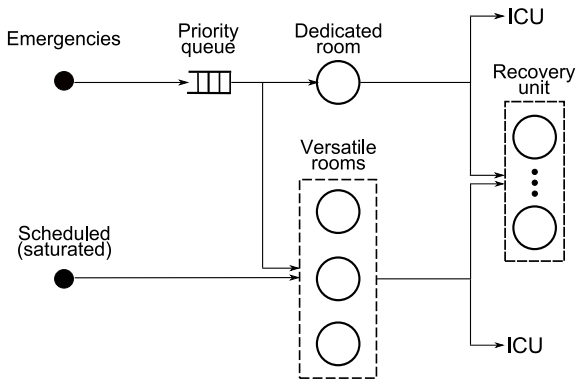

FIG. 7.2 – *Illustration d'un quartier opératoire tel que modélisé ici, composé de quatre blocs opératoires dont trois sont polyvalents et un est dédié aux urgences, et de cinq lits de réveil (configuration* $[3|1|5]$*).*

distribuées, dont les distributions diffèrent selon que le patient soit électif ou urgent. Nous supposons en outre que le temps de réalisation d'une opération urgente ne dépend pas significativement du type de salle (dédiée ou polyvalente) dans laquelle elle prend place. Notons toutefois que cette hypothèse pourrait facilement être levée.

A la fin de leur opération chirurgicale, la plupart des patients sont immédiatement admis en salle de réveil, les autres étant dirigés vers l'unité de soins intensifs (USI) si leur état est critique, ou pire, vers la morgue si l'issue de l'intervention est fatale. Les patients ne passant pas par la SSPI quittent alors le quartier opératoire sans avoir occupé de lit de réveil. En effet, l'unité de soins intensifs est indépendante du quartier opératoire et gérée de façon autonome ; elle n'est par conséquent pas prise en compte dans notre modèle. Dès lors, avec une probabilité P_p (respectivement P_e), son état de santé permet au patient électif (respectivement au patient urgent) de rejoindre la salle de réveil, nécessitant la disponibilité d'un lit. Avec une probabilité $(1-P_p)$ (respectivement $(1 - P_e)$), ce même état de santé mène le patient programmé (respectivement le patient urgent) directement en soins intensifs, celui-ci quitte alors le quartier opératoire et n'influence plus son fonctionnement. Dans le cas d'un sujet joignant la SSPI, ce dernier occupe un lit pendant une durée de réveil aléatoire avant de quitter le quartier opératoire et de rejoindre sa chambre d'hospitalisation ou, selon son état, l'USI, ou encore de rentrer chez lui s'il s'agit d'un patient ambulatoire. Par contre, si aucun lit n'est disponible en SSPI, comme le patient ne peut attendre avant de recevoir les soins post-interventionnels, il débute sa phase de réveil dans la salle d'opération qu'il occupe, cette dernière étant alors bloquée. Si le quartier opératoire comporte plusieurs types de salles et que certaines d'entre elles sont dédiées aux urgences, une stratégie de déblocage doit être mise sur pied. En effet, si les deux sortes de blocs opératoires sont immobilisées simultanément, et qu'un lit de réveil se libère, la salle qui se décharge en premier lieu dépendra de cette stratégie. Nous en envisageons trois : la stratégie S_1 qui donne la priorité aux salles polyvalentes, la stratégie S_2 qui privilégie les salles dédiées et la stratégie S_3 qui priorise les salles polyvalentes excepté si une urgence patiente dans la file d'attente, auquel cas elle donne priorité aux salles dédiées.

Le modèle du quartier opératoire que nous considérons dans ce chapitre est à présent défini. Par la suite, un quartier opératoire composé de n_v salles polyvalentes, n_d salles dédiées aux urgences et n_b lits de réveil sera dit avoir une configuration $[n_v|n_d|n_b]$. L'objectif de notre approche est

à présent d'évaluer l'effet de la stochasticité sur un tel système, proposant diverses mesures de performance. La démarche de résolution analytique, basée sur la théorie de Markov, est présentée en section 7.5. Mais avant cela, nous rappelons brièvement le cas d'étude qui sert de corps à ce travail et qui sera utilisé afin d'illustrer notre outil.

7.4.2 Rappel du cas d'étude

L'étude de la nature imprévisible du quartier opératoire réalisée dans cette recherche trouve son origine dans notre collaboration avec un hôpital bruxellois qui, d'une part, nous donne un cadre pratique d'investigation, et d'autre part nous offre des données réelles sur lesquelles nous basons nos expérimentations pour évaluer la pertinence de nos travaux.

Rappelons que le quartier opératoire de cet hôpital est ouvert cinq journées par semaine durant huit heures par jour, et qu'à ces heures doivent être ajoutées les heures supplémentaires souvent nécessaires pour achever la réalisation du planning quotidien. La réduction de l'utilisation des salles en heures supplémentaires, vecteurs de coûts importants, est un souci permanent pour les gestionnaires, et la stochasticité du système joue un rôle majeur dans la génération de l'activité en dépassement horaire. On chiffre en moyenne sept heures supplémentaires à prester chaque jour, réparties sur l'entièreté du quartier opératoire, dans les données à disposition. Ce dernier comporte sept salles polyvalentes (en mesure d'accepter tout type de chirurgie), aucune n'est donc à proprement parler réservée pour le traitement des urgences. Cependant, analyser plus en détail les données nous apprend qu'en réalité seulement six de ces salles peuvent être utilisées simultanément, et ce en raison d'un cruel manque de personnel, en particulier d'anesthésistes et d'infirmières. En outre, il ressort que 38% des urgences arrivant au quartier opératoire sont traitées dans la même unité chirurgicale. Il serait dès lors intéressant d'évaluer l'opportunité de dédier l'une de ces six salles aux urgences, voire la septième. Comme seuls six blocs opératoires sont utilisés en pratique, nous allons considérer dans la suite que la SSPI est équipée de neuf lits, appliquant de la sorte la proportion de 1.5 lits par bloc opératoire comme le suggère la règle communément admise en Belgique. En effet, considérer le nombre maximum de lits disponibles équipant la salle de réveil diminuerait immanquablement l'impact du blocage sur les performances du système.

Rappelons également que ce jeu de données contient toute l'activité opératoire observée sur deux mois, ce qui correspond à 632 opérations électives et 125 urgences traitées sur 42 jours ouvrables. Le nombre moyen d'opérations programmées par jour s'élève à 15 tandis que son équivalent pour les urgences est légèrement inférieur à 3 (notons qu'ici les interventions reprises sous le terme d'opérations ajoutées dans les données sont considérées comme des urgences en raison de la perturbation qu'elles engendrent dans le planning). Une opération prévue au planning prend en moyenne 153 minutes pour être réalisée alors qu'une urgence, elle, nécessite seulement 87 minutes en moyenne. Notons que ces temps opératoires représentent la durée d'occupation globale de la salle (préparation, anesthésie et chirurgie). Les fonctions de densité empiriques de ces deux types d'intervention, issues de nos données, sont illustrées dans la figure 7.3. Le détail de ces données reflète donc très clairement la nature stochastique du quartier opératoire.

Une fois opéré, bon nombre de patients rejoignent la salle de réveil où ils se rétablissent de leur intervention. Dans nos données, près de 90% (respectivement 87%) des opérations planifiées (respectivement urgentes) devraient aboutir en SSPI. Cependant, il peut s'avérer que les lits de réveil soient tous occupés, bloquant alors l'opération qui viendrait à terminer et perturbant le programme

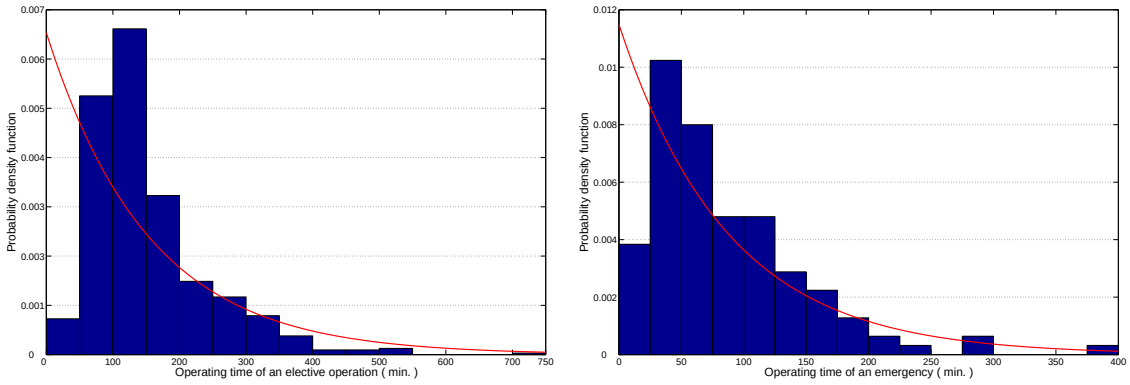

FIG. 7.3 – *Fonctions de densité de probabilité des durées opératoires des opérations électives (figure de gauche) et urgentes (figure de droite). Les distributions empiriques (issues des données) et théoriques (distributions exponentielles de taux* $\lambda_p = 153$ *et* $\lambda_e = 87$*) y sont tracées.*

établi. Toujours selon nos données, 8.1% des patients se retrouvent bloqués dans leur salle d'intervention et y débutent la phase de réveil. Le blocage apparaît donc comme un événement loin d'être anodin qui justifie sa prise en compte dans notre modèle. Les patients qui ne rejoignent pas la salle de réveil, autrement dit les cas les plus critiques (qu'ils soient urgents ou électifs), quittent le quartier opératoire. Soit ils sont directement admis dans l'unité de soins intensifs (10.3% de l'ensemble des patients), soit, malheureusement, ils ne survivent pas à leur intervention (0.3%).

La section suivante décrit l'approche analytique que nous envisageons pour évaluer l'impact des aléas sur les performances du quartier opératoire. Les chiffres présentés ici seront utilisés par défaut pour illustrer le fonctionnement du modèle et pour le tester.

7.5 Modèle Markovien, extensions et mesures

L'approche que nous proposons dans cette section a pour but de rationaliser la considération de la variabilité inexorablement présente lors de la réalisation du programme opératoire. A cette fin, la méthodologie présentée mesure l'impact de cette stochasticité, issue des trois grandes sources détaillées précédemment (urgences, durées opératoires et de réveil), sur différentes mesures de performance du quartier opératoire comme le temps d'attente avant traitement des urgences, la perturbation du programme, le temps de blocage ou le nombre d'heures supplémentaires nécessaires. Sur base de ces indications, nous pouvons déduire des actions à entreprendre sur un plan stratégique et tactique afin de mieux gérer la stochasticité d'un quartier opératoire particulier. Il est en effet essentiel que l'outil développé soit générique et puisse être appliqué à des cas particuliers sur base uniquement de certaines caractéristiques comme les temps opératoires moyens, le flux d'urgences, la répartition des blocs opératoires et d'autres.

Du point de vu management, cet outil devrait permettre de calculer le taux d'urgences perturbant le planning opératoire (c'est à dire le nombre moyen d'opérations urgentes devant être réalisées, par jour, dans les salles polyvalentes), le temps nécessaire pour réaliser ces opérations

urgentes ou encore la probabilité qu'une salle d'intervention soit bloquée en raison de la saturation de la SSPI. Concernant la qualité de soins aux urgences, il serait intéressant de pouvoir mesurer leur temps d'attente moyen afin de déterminer si des actions sont à entreprendre en vue d'améliorer la réponse apportée à de tels événements. Ce type d'actions concerne, entre autres, la réservation de blocs opératoires pour la réalisation exclusive d'interventions urgentes. Il faudrait donc pouvoir apprécier dans quelle mesure dédier une salle aux urgences peut être profitable au comportement du système. De même, l'outil pourrait déterminer, autrement que sur base d'une règle empirique, le nombre de lits à pourvoir dans la salle de réveil en fonction d'une série d'indicateurs sur les performances du quartier opératoire. La finalité ultime de notre modèle serait de pouvoir estimer objectivement le taux d'occupation à appliquer dans les salles lors de la construction des calendriers opératoires (pour rappel, le taux d'occupation correspond ici au pourcentage de capacité à préserver lors de la construction des plannings afin d'absorber les aléas).

Dans ce chapitre, nous proposons et appliquons un modèle stochastique analytique du quartier opératoire, et ce en opposition aux techniques de simulation. Modèles analytiques et simulation ont leurs propres avantages et peuvent être perçus comme complémentaires ; ils sont souvent utilisés pour valider l'un l'autre. Dans notre cas, l'adéquation de l'approche mathématique peut se justifier. Notre travail se positionne, dans sa partie sur l'analyse de la stochasticité du milieu, à un niveau de décision stratégique/tactique. Nous souhaitons pouvoir fournir des indications sur le comportement d'un système et sur les mesures à entreprendre pour en améliorer la gestion à partir de peu de données le caractérisant. Dans ce contexte, le principal atout de la simulation, c'est-à-dire son aptitude à modéliser de manière plus détaillée l'objet d'étude, ne se justifie pas, ou du moins n'est pas très fructueux vu que cette technique réclame une connaissance du système plus approfondie que nous ne l'envisageons. Pour le problème qui nous préoccupe ici, la démarche analytique a l'avantage d'offrir rapidement des mesures de performance constantes et fiables, nécessitant relativement peu de données. Par ailleurs, même si la définition du problème peut faire référence à la théorie des files d'attente, la structure étudiée s'avère trop compliquée pour être modélisée par une telle approche. En conséquence, nous avons développé un modèle basé sur la théorie Markovienne, reposant plus précisément sur un processus de Markov continu. Ce dernier nous permet de prendre en compte les diverses sources d'aléas pour obtenir une modélisation globale de l'effet de la variabilité sur le quartier opératoire. De plus, l'approche Markovienne offre l'opportunité d'estimer la distribution de différentes mesures de performance sous forme de distributions phase-type dérivées du processus de Markov. C'est notamment sur base de la distribution du temps d'occupation global du quartier opératoire que nous espérons pouvoir déterminer un taux d'occupation des salles adéquat en fonction des exigences du gestionnaire.

Afin d'appliquer la théorie de Markov, nous supposons l'ensemble des durées distribuées exponentiellement. La démarche eût été tout aussi valide en modélisant ces mêmes durées par des distributions phase-type, offrant un degré de précision plus élevé mais impliquant un modèle plus complexe nécessitant plus de données. La suite décrit la liste des quatre hypothèses nécessaires à l'utilisation de la théorie Markovienne.

– La durée d'une opération programmée est supposée être distribuée exponentiellement avec un taux λ_p, exprimé en opérations par jour, et où un jour est censé compter huit heures. Nous prenons le parti de ne pas distinguer les interventions électives. Nous pourrions différencier celles-ci par salle d'opération, c'est-à-dire accorder aux différents blocs opératoires des temps de traitement différents permettant de modéliser plus en détails les diverses spécialités

chirurgicales, mais cela mènerait à une chaîne de Markov de plus grande dimension et donc plus difficile encore à traiter. De plus, il n'est pas absurde de considérer que les différences entre les opérations sont prises en compte par leur nature stochastique.

– Les urgences arrivent selon un processus de Poisson, avec un taux d'arrivée λ_i. Il s'agit là d'une hypothèse classique de la littérature, qui a déjà été validée dans de nombreux contextes (voir Green, 2006, par exemple).
– Le temps d'intervention d'une urgence est également distribué exponentiellement, avec un taux λ_e. Notons à ce propos que les temps des différents traitements, électifs ou urgents, peuvent inclure les soins préopératoires et postopératoires.
– La durée de réveil d'un patient, qu'il soit programmé ou urgent, est supposée suivre une distribution exponentielle de taux λ_b. En d'autres termes, si $\lambda_b = 3.3$ cela signifie que, par jour, 3.3 patients occupent un même lit de réveil en moyenne (le temps de réveil moyen est donc de 2h25). A nouveau, nous pouvons considérer que la différence entre les durées de réveil des deux types de patients (électif ou urgent) est modélisée par leur caractère stochastique.

L'hypothèse portant sur la nature exponentielle des distributions des durées opératoires est faite en premier lieu pour des raisons de modélisation, afin de pouvoir appliquer la théorie de Markov. Cependant, sur base des données provenant de l'hôpital bruxellois, nous pouvons voir que cette hypothèse n'est pas exagérément trompeuse. Dans la figure 7.3, nous représentons les distributions empiriques et théoriques (c'est-à-dire exponentielles) des temps opératoires des interventions électives et des urgences. Pour une comparaison équitable, les surfaces sous les courbes sont ajustées à un. Bien que des distributions offrant de meilleures approximations puissent être utilisées (en particulier pour les durées les plus courtes), les distributions empiriques et théoriques sont relativement similaires. De plus, les coefficients de variation observés (0.58 et 0.72) ne sont pas trop éloignés de l'unité qui caractérise la distribution exponentielle. Dans de futures recherches, il serait envisageable d'améliorer l'adéquation aux temps opératoires observés par l'intermédiaire de distributions phase-type, notre outil restant totalement applicable dans de telles conditions.

Dans la suite de cette section, nous introduisons progressivement notre approche Markovienne. Nous commençons par décrire un modèle assez simpliste afin de clarifier les idées principales de celle-ci, et l'étendons ensuite pour tenir compte des lits de réveil et estimer des mesures de performance comme le temps d'attente des urgences ou le nombre d'heures supplémentaires générées.

7.5.1 Modèle initial

Nous présentons ici un modèle simplifié qui a pour principal objectif d'illustrer les idées essentielles de notre approche de modélisation, de telle sorte que le processus de Markov qui en découle reste aisé à illustrer. Pour cela, nous nous focalisons sur les salles d'opération afin de mesurer la perturbation du planning opératoire due aux urgences, et ne considérons pas, à ce stade, la salle de réveil (voir figure 7.4).

Le processus de Markov modélisant le quartier opératoire sans SSPI est représenté en figure 7.5, pour un exemple comportant trois salles polyvalentes et une dédiée. Nous décrivons à présent ce que représentent les états du processus Markovien ainsi que les différentes transitions qui les unissent. Chaque état dvq est défini par le nombre d'urgences présentes dans chaque partie du système : les salles dédiées aux urgences ($\underline{d}vq$), les salles polyvalentes ($d\underline{v}q$) et la file d'attente

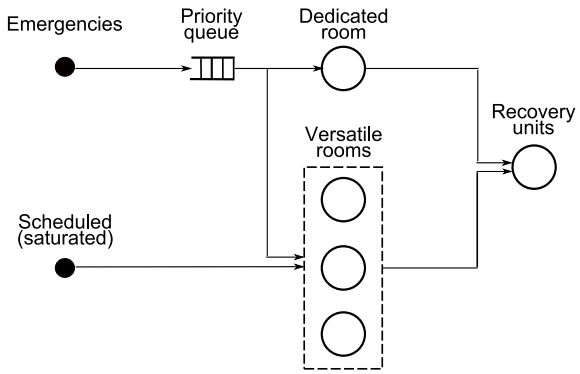

FIG. 7.4 – *Quartier opératoire (sans SSPI) composé de quatre salles, dont trois sont polyvalentes et une est dédiée aux urgences (configuration* $[3|1]$*).*

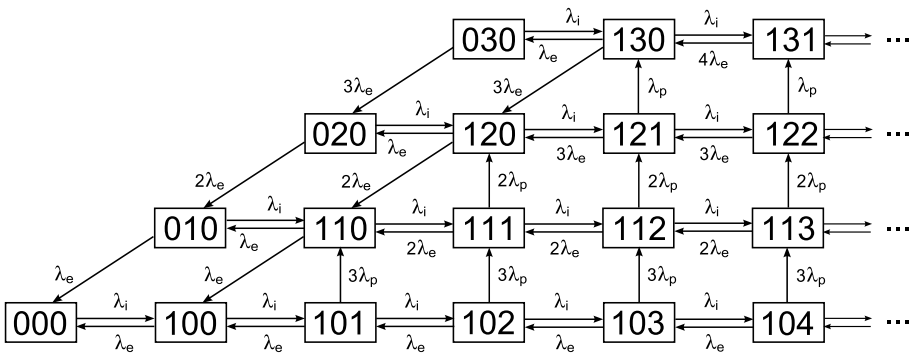

FIG. 7.5 – *Processus de Markov modélisant un quartier opératoire (sans SSPI) composé de quatre salles, dont trois sont polyvalentes et une est dédiée aux urgences (configuration* $[3|1]$*).*

prioritaire destinée aux urgences ($dv\underline{q}$). Dans le cas des salles polyvalentes, le nombre d'opérations électives en cours de traitement peut facilement se retrouver en retranchant du nombre n_v de ces salles le nombre v d'urgences qui y sont actuellement réalisées. A titre d'exemple, l'état $dvq = 112$ caractérise un quartier opératoire dans lequel une urgence est traitée dans une salle qui lui est dédiée, une urgence est insérée dans les salles polyvalentes (ce qui signifie donc que deux interventions programmées sont en cours puisque $n_v = 3$), et deux urgences attendent qu'un bloc se libère pour être prises en charge.

Les transitions entre ces états correspondent aux différents événements qui peuvent se produire à partir d'un état donné du système. Ces événements sont de quatre types.

- Lorsqu'une urgence arrive dans le système (avec un taux λ_i), elle est, si possible, directement admise dans une salle dédiée (par exemple, la transition de 020 vers 120 de la figure 7.5), autrement elle rejoint la file d'attente (de 112 vers 113).
- Quand une urgence s'achève dans une salle dédiée (avec taux $d\lambda_e$), le patient quitte la salle correspondante. Si une autre urgence attend sa prise en charge, elle remplace immédiatement celle qui vient de se finir, dans la même salle d'opération (de 112 vers 111). Sinon, la file d'attente reste vide et la salle dédiée devient accessible (de 120 vers 020).
- Lorsqu'une urgence libère une salle polyvalente (avec un taux $v\lambda_e$), elle est remplacée par

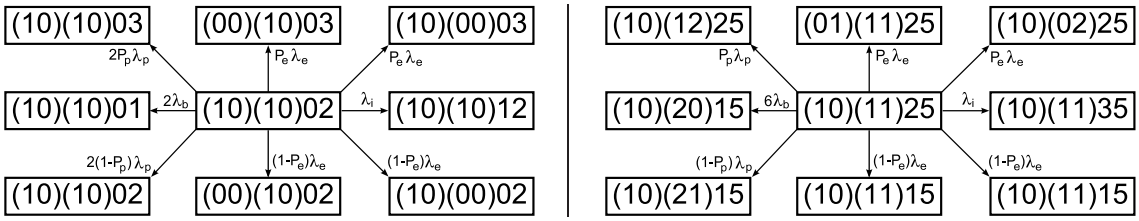

FIG. 7.6 – *Exemples de transitions du processus de Markov pour un quartier opératoire présentant une configuration* $[3|1|5]$. *La figure de gauche représente chaque transition possible à partir d'un état initial sans blocage ; la figure de droite représente les transitions au départ d'un état avec blocage.*

une autre urgence si un tel patient est en attente de traitement (de 112 vers 111). Par contre, si la file d'attente est vide, c'est une opération programmée qui peut prendre place dans la salle libérée (de 120 vers 110).

– Le dernier cas correspond à une opération élective se terminant (avec un taux $(n_v - v)\lambda_p$). Si la file d'attente n'est pas vide, la première urgence qui patiente rejoint la salle polyvalente tout juste libérée (de 112 vers 121), sinon une nouvelle opération planifiée entre dans la salle et l'état du système ne change pas (120).

La figure 7.5 montre que la structure du processus de Markov est relativement régulière. Elle diffère légèrement selon que la file d'attente contienne des urgences ou non, ce qui peut se déduire de la description des différentes transitions traduisant le système. La structure homogène de la chaîne, et en particulier la faible densité de la matrice de transition qui la caractérise, facilitent le calcul des probabilités stationnaires des états.

Le modèle de base présenté ici illustre comment modéliser le quartier opératoire et l'impact de sa variabilité à l'aide d'un processus de Markov. Cependant, il ne tient pas compte de l'unité de réveil et ne permet pas d'obtenir d'indications plus détaillées à propos des distributions de certaines mesures de performance ; les résultats obtenus ici valent uniquement en moyenne. Nous allons à présent proposer des extensions à ce modèle initial afin d'en surmonter les limites.

7.5.2 Inclure la salle de réveil

Jusqu'à présent, nous avons toujours défini la salle de réveil comme une partie intégrante du quartier opératoire, ce qui se justifie notamment par l'impact que celle-ci peut avoir sur ce dernier en termes de blocage (rappelons que plus de 8% des opérations sont au moins partiellement bloquées dans notre cas d'étude belge). Le terme "blocage" signifie ici qu'un ou plusieurs patients ont débuté leur phase de réveil dans un bloc opératoire en raison de l'absence de lits de réveil disponibles en SSPI. Les durées de réveil des patients ne sont pas prévisibles avec exactitude, ce qui en font une source de stochasticité à part entière. Comme nous les supposons distribuées exponentiellement, l'évolution du quartier opératoire complet (incluant la SSPI) peut également se modéliser à l'aide d'un processus de Markov. Ce dernier est une version élaborée du modèle initial présenté précédemment, qui peut difficilement s'illustrer en totalité. Nous présentons deux fragments de ce nouveau processus Markovien dans la figure 7.6 et décrivons la signification des états et transitions le caractérisant.

Prendre en compte la salle de réveil induit une complexification de la description des états composant le processus de Markov, comparativement au modèle initial. Les indices correspondant aux deux types de salles sont à présent chacun divisé en deux identifiants afin de compter le nombre de salles d'intervention bloquées, de même qu'un indice est ajouté pour comptabiliser le nombre de patients en salle de réveil. Chaque état du processus est dès lors composé de six indices et dénoté $[(d_e d_b)(v_e v_b)qb]$. En voici la signification :

- $(d_e d_b)$ sont associés aux salles dédiées. d_e représente le nombre d'urgences en cours dans ce type de salles alors que d_b représente le nombre de salles dédiées bloquées par un patient attendant un lit de réveil.
- $(v_e v_b)$ sont liées aux blocs opératoires polyvalents. v_e est le nombre d'urgences en cours dans ce type de blocs tandis que v_b reprend le nombre de salles polyvalentes bloquées par un patient quel qu'il soit. Le nombre d'interventions planifiées en cours est donc $n_v - v_e - v_b$.
- q est le nombre d'urgences en attente de traitement dans la file prioritaire.
- b est le nombre de lits de réveil occupés par des patients opérés (urgents ou non).

Par exemple, supposant une configuration $[3|1|5]$ du système comme celle de la figure 7.6, l'état $[(d_e d_b)(v_e v_b)qb] = [(10)(11)25]$ caractérise un quartier opératoire dans lequel une urgence est traitée dans la salle dédiée ; aucune salle dédiée n'est bloquée ; une urgence est insérée dans les salles polyvalentes ; une salle polyvalente est bloquée ; deux urgences patientent dans la file et l'ensemble des cinq lits de réveil disponibles sont occupés (ce qui explique pourquoi une salle polyvalente est bloquée). Les transitions du processus de Markov sont également rendues plus compliquées en raison de la présence des lits de réveil et du blocage potentiel qu'ils apportent. Elles correspondent aux cinq types d'événements qui peuvent se produire à partir d'un état particulier, événements qui sont explicités ci-après.

- Une urgence arrive (avec un taux λ_i). Elle entre, si possible, dans une salle dédiée, sinon elle rejoint la file d'attente (gauche de la figure 7.6, de $[(10)(10)02]$ vers $[(10)(10)12]$).

- Une opération élective s'achève dans une salle polyvalente (avec un taux $(n_v - v_e - v_b)\lambda_p$). Avec une probabilité $(1 - P_p)$, celui-ci sort du système pour rejoindre l'unité de soins intensifs. Par contre, avec une probabilité P_p le patient accède à la salle de réveil, pour autant qu'un lit y soit encore disponible, le nombre de lits de réveil occupés augmentant alors d'une unité. Si une urgence attend d'être traitée, elle rejoint directement la salle polyvalente ainsi libérée (de $[(10)(11)25]$ vers $[(10)(21)15]$), sinon une nouvelle opération élective y prend place. Finalement, si le patient doit passer par la salle de réveil mais qu'aucun lit n'y est inoccupé, la salle polyvalente dans laquelle se trouve ce patient devient bloquée (de $[(10)(11)25]$ vers $[(10)(12)25]$).

- Une urgence se termine dans une salle dédiée (avec un taux $d_e\lambda_e$). Avec une probabilité $(1 - P_e)$, le patient urgent quitte le quartier opératoire pour l'USI, et aucun lit de réveil n'est nécessaire (de $[(10)(10)02]$ vers $[(00)(10)02]$). A l'inverse, avec une probabilité P_e, il est dirigé vers la salle de réveil (si $b < n_b$) et y occupe un lit supplémentaire (de $[(10)(10)02]$ vers $[(00)(10)03]$). Dans les deux situations, si une urgence est en attente de traitement, elle remplace celle qui s'est achevée dans la même salle dédiée. Enfin, si le patient urgent doit joindre la SSPI mais qu'aucun lit n'y est libre ($b = n_b$), la salle dédiée devient bloquée (de $[(10)(11)25]$ vers $[(01)(11)25]$).

– Une urgence se termine dans une salle polyvalente (avec un taux $v_e\lambda_e$). Le patient rejoint la SSPI ou quitte le système avec les mêmes probabilités qu'énoncées plus haut. Si l'urgence traitée doit accéder à la salle de réveil mais ne peut le faire, la salle polyvalente devient bloquée (de $[(10)(11)25]$ vers $[(10)(02)25]$). Si au contraire elle peut quitter la salle, elle est instantanément remplacée par une autre urgence si la file d'attente en compte encore (de $[(10)(11)25]$ vers $[(10)(11)15]$), et par une opération planifiée sinon.

– La dernière transition correspond au réveil d'un patient (avec un taux $(b + d_b + v_b)\lambda_b$). Si aucune salle d'intervention n'est bloquée, le patient quitte le lit de réveil, ce dernier redevenant libre. Par contre, si un bloc opératoire est monopolisé ($d_b > 0$ or $v_b > 0$), deux cas sont alors envisageables. Si un patient termine sa phase de réveil en SSPI, le patient bloqué accède à celle-ci et y occupe le lit nouvellement disponible. La durée de réveil qu'il lui faut encore accomplir reste distribuée exponentiellement avec le même taux λ_b, par la propriété d'absence de mémoire de ces distributions. Autrement, si un patient achève de se réveiller dans la salle dans laquelle il fut opéré, il quitte directement le système, sans passer par la SSPI. Ces deux cas mènent au même état (droite de la figure 7.6, de $[(10)(11)25]$ vers $[(10)(20)15]$). Quand les deux types de salles sont bloquées simultanément ($d_b, v_b > 0$), les transitions entre les états dépendront de la stratégie de blocage adoptée (voir section 7.4.1).

Sur base des définitions de ces états et transitions, nous pouvons à présent construire le processus de Markov continu modélisant le quartier opératoire complet. Celui-ci nous offre un outil flexible dont les différents paramètres sont à déterminer : les différents taux (λ_i, λ_e, λ_p et λ_b), le nombre de salles d'opération, leur répartition en blocs dédiés et polyvalents (n_v et n_d), le nombre de lits de réveil n_b, les probabilités de rejoindre la salle de réveil P_e et P_p, et la stratégie de blocage S. L'outil ainsi implémenté permet de calculer les probabilités stationnaires des états de la chaîne de Markov. De manière informelle, la probabilité stationnaire d'un état peut être vue comme la proportion du temps durant laquelle le système, c'est-à-dire le quartier opératoire dans notre cas, est dans cet état particulier. Les probabilités stationnaires π_i des états du processus de Markov fini continu sont facilement obtenues à l'aide de la formule classique $\pi Q = 0$, avec $\sum_i \pi_i = 1$ et où Q est la matrice de transition du processus. Notons que le processus de Markov est tronqué, les états présentant une probabilité d'occurrence négligeable étant supprimés afin de garder un nombre fini d'états sans affecter les mesures de performance.

A partir des probabilités stationnaires, nous pouvons dériver différentes mesures de performance en fonction des paramètres particuliers au quartier opératoire étudié. Nous en présentons les plus pertinentes dans ce qui suit. Nous analysons et illustrons comment ces mesures de performance sont affectées par divers facteurs caractérisant le quartier opératoire. A cette fin, nous basons toutes nos expérimentations sur le cas d'étude de l'hôpital belge présenté en sections 3.2 et 7.4. Pour faciliter la lecture de ces résultats, les principaux paramètres relatifs au quartier opératoire bruxellois sont résumés dans le tableau 7.1. Par la suite, la stratégie de blocage appliquée est

n_v	n_d	n_b	λ_p	λ_i	λ_e	λ_b	P_p	P_e
6	0	9	3.14	2.98	5.52	3.32	0.9	0.87

TAB. 7.1 – *Valeurs par défaut des paramètres du processus de Markov, déterminées sur base de l'analyse statistique du cas d'étude belge.*

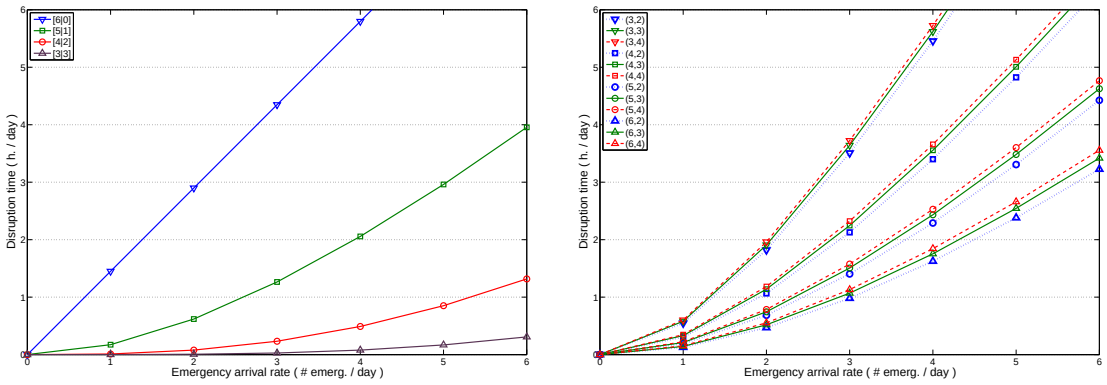

FIG. 7.7 – *Temps opératoire total des urgences perturbatrices dans les salles polyvalentes, en fonction du taux d'arrivée des urgences dans le système (λ_i). A gauche, la configuration du quartier opératoire change, la légende mentionnant les valeurs du couple (n_v,n_d). A droite, supposant une configuration $[5|1|9]$, ce sont les temps opératoires moyens des opérations urgentes et électives qui varient, c'est-à-dire le couple (λ_e,λ_p).*

supposée être S_3, celle qui nous semble la plus pertinente (voir sous-section 7.4.1). Pour rappel, nous ne disposons pas d'information sur les durées de réveil des patients mais pouvons estimer leur temps de réveil moyen sur base du taux de blocage observé de ces derniers (après leur intervention, 8.1% des patients débutent leur phase de réveil en salle d'opération). En effet, à partir des probabilités stationnaires des états menant à du blocage, calculées par notre outil, nous avons pu déterminer que la durée de réveil moyenne justifiant le taux de blocage observé était de 2 heures 25 minutes ($\lambda_b = 3.32$). Cette valeur se trouve dans la lignée des observations de Dexter & Tinker (1995), ce qui conforte notre estimation. A partir de ce peu de paramètres, nous sommes maintenant en mesure d'estimer l'impact de la stochasticité inhérente au milieu chirurgical, et ce sur différents aspects du quartier opératoire. En fonction des mesures de performance analysées, nous faisons varier un ou plusieurs paramètres afin d'évaluer leur effet sur celles-ci, les autres paramètres étant fixés à leurs valeurs par défaut reprises dans le tableau 7.1.

Perturbation du programme opératoire par les urgences. Pour estimer l'effet des urgences sur le planning prévisionnel, nous en calculons le taux de perturbation. Il s'agit du taux d'urgences entrant dans les salles polyvalentes et retardant de la sorte le calendrier établi, et est noté λ_{ip}. Il est défini comme la somme des taux des transitions menant à une telle perturbation, pondérée par les probabilités stationnaires des états initiaux correspondants. Notons que si aucune salle n'est dédiée au traitement exclusif des cas urgents, alors $\lambda_{ip} = \lambda_i$. Le taux de perturbation peut être directement traduit en un nombre moyen d'heures nécessaires à l'achèvement des urgences réalisées dans les salles polyvalentes. Le temps de perturbation par les urgences, ainsi défini, donne un premier aperçu du temps moyen qu'il sera utile de réserver pour le traitement des urgences lors de la construction des programmes opératoires.

La figure 7.7 illustre l'évolution du temps de perturbation par les urgences en fonction de divers facteurs. Bien entendu, ce temps de perturbation augmente avec le taux d'arrivée des urgences. La figure de gauche nous indique toutefois que dédier des salles aux urgences diminue significativement leur impact sur le planning opératoire, justifiant une telle pratique

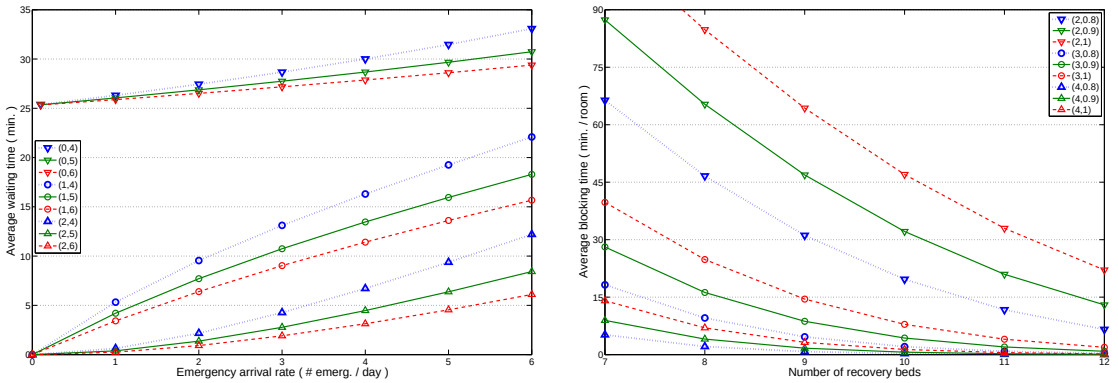

FIG. 7.8 – *A gauche, évolution du temps d'attente moyen en fonction du taux d'arrivée des urgences λ_i, du nombre de salles dédiées n_d et du taux d'opération des urgences λ_e. La légende y donne le couple (n_d, λ_e). A droite, évolution du temps moyen de blocage en fonction du nombre de lits de réveil n_b, du taux de réveil λ_b et de la probabilité de rejoindre la SSPI (ici $P_p = P_e$). La légende y donne le couple (λ_b, P_p).*

lorsque le flux des cas urgents est suffisamment élevé. Dans la figure de droite, nous pouvons observer que, intuitivement, le temps opératoire moyen des urgences influence directement la perturbation qu'elles causent dans les salles polyvalentes. A l'inverse, la durée moyenne des opérations planifiées ne modifie que marginalement ces mêmes courbes.

Temps d'attente des urgences. Afin d'évaluer la qualité des soins offerts aux cas les plus urgents, nous pouvons en calculer la probabilité d'attente. Celle-ci revient à mesurer la probabilité qu'une urgence arrive lorsque toutes les salles dédiées sont occupées, en d'autres termes la somme des probabilités stationnaires des états portant cette caractéristique. Dans le même but, nous pouvons estimer le temps d'attente moyen d'une urgence dans la file prioritaire, qui dérive de la mesure précédente en appliquant la loi de Little.

La partie gauche de la figure 7.8 reprend l'évolution du temps d'attente moyen en fonction de divers facteurs. Le temps d'attente augmente avec le nombre d'urgences à traiter par jour et avec la durée de ces interventions. Afin d'améliorer la qualité de service aux urgences et de réduire leur temps d'attente moyen, une option efficace consiste à leur dédier une ou plusieurs salles d'intervention, comme l'illustre la figure 7.8. Notre outil permet donc de mesurer l'impact d'une telle décision.

Blocage. Lorsqu'une opération se termine et qu'aucun lit ne peut accueillir ce patient, les soins postopératoires sont réalisés dans la salle d'opération qui se retrouve ainsi bloquée. Le blocage peut être évalué via la probabilité qu'une salle d'opération soit bloquée, c'est-à-dire la somme des probabilités stationnaires des états dans lesquels cette salle est bloquée. A partir de cette mesure, nous pouvons aisément déduire le temps moyen durant lequel un bloc opératoire est bloqué.

La droite de la figure 7.8 illustre comment le temps de blocage moyen par salle décroît lorsque le nombre de lits de réveil augmente. Cette figure montre également comment le blocage augmente lorsque la durée de réveil augmente ou lorsque la probabilité qu'un patient rejoigne la SSPI augmente.

Autres. D'autres mesures de performance peuvent être calculées sur base des probabilités stationnaires, telles que le taux d'occupation moyen des lits de réveil, etc.

Le processus de Markov offre donc une image fidèle du quartier opératoire ainsi qu'une façon rationnelle de mesurer l'impact de la variabilité de son milieu, au moyen d'un panel d'indicateurs de performance. Cependant, l'ensemble des mesures décrites jusqu'à présent sont des moyennes tirées des probabilités stationnaires des états du système. Afin d'améliorer la connaissance du comportement du quartier opératoire que l'on peut extraire de notre approche, nous allons désormais calculer des mesures basées sur les distributions du temps d'attente des urgences et du temps de travail global.

7.5.3 Distribution du temps d'attente des urgences

Le principal critère d'évaluation de la qualité de service prodigué aux patients urgents est leur temps d'attente avant traitement. Comme expliqué précédemment, à partir des probabilités stationnaires du processus Markovien il est possible de déduire la probabilité qu'un tel patient doive attendre avant d'être pris en charge, de même que la durée moyenne de cette attente. Cependant, d'autres types de mesures semblent pertinents dans le cas d'une chirurgie urgente. Notamment, il serait intéressant de savoir quelle est la probabilité qu'une urgence attende plus d'une demi-heure par exemple ? En effet, au-delà d'un certain délai, la vie du patient urgent peut être mise en jeu. Pour répondre à cette question, nous devons connaître la distribution du temps d'attente d'une urgence.

La distribution du temps d'attente d'une urgence arrivant au quartier opératoire peut être obtenue en modifiant le processus de Markov modélisant le quartier opératoire pour obtenir une distribution phase-type. Une distribution phase-type continue est une composition complexe de distributions exponentielles qui peut se modéliser comme le temps nécessaire avant absorption dans un processus Markovien contenant un unique état absorbant. Elle est complètement définie par la matrice de transition du processus de Markov et par les probabilités initiales de démarrer dans l'un des états de la chaîne (voir Neuts, 1981). La théorie sur les distributions phase-type est bien établie. La fonction de densité et la fonction de distribution, de même que les moments comme la moyenne ou la variance, sont connus et déterminés sur base de la matrice de transition et des probabilités initiales. Afin d'expliciter comment calculer la distribution phase-type du temps d'attente, et par souci de simplicité, nous nous référons au modèle initial, sans tenir compte des lits de réveil. La généralisation au cas comportant la SSPI s'ensuit logiquement. Pour faciliter la compréhension du raisonnement effectué, le processus de Markov initial modélisant le quartier opératoire, décrit par la figure 7.5, est ici dénoté MP1. Celui-ci est modifié pour obtenir le processus de Markov caractérisant la distribution phase-type, dépeint en figure 7.9, et nommé MP2.

Supposons qu'un patient urgent se présente au quartier opératoire. Nous souhaitons calculer son temps d'attente. Si une salle dédiée est disponible, le patient y est directement admis et opéré ; son temps d'attente est donc nul. Par contre, si aucune salle dédiée aux urgences n'est vacante, il doit patienter dans une file d'attente prioritaire tant qu'un bloc opératoire, dédié ou polyvalent, ne lui est pas accessible. Son temps d'attente dépend alors du nombre d'urgences déjà présentes dans la file et des événements qui s'ensuivent dans le quartier opératoire (événements qui vont conduire à la libération des salles). Le nombre de personnes patientant déjà dans la file à l'arrivée du patient peut être déduit des probabilités stationnaires du processus de Markov original MP1. Celles-ci

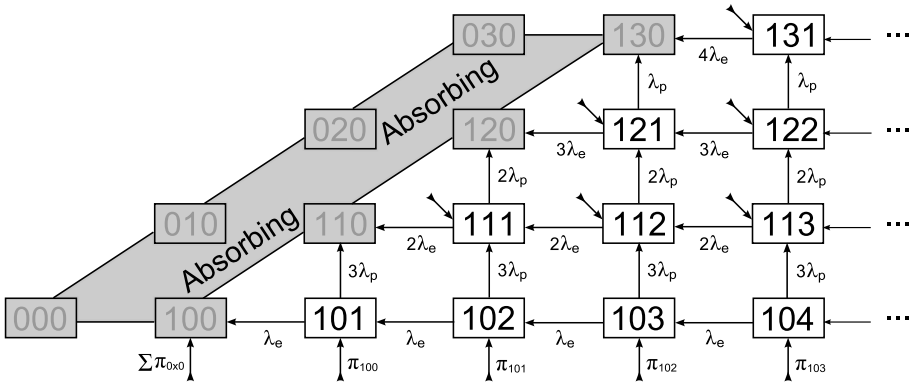

FIG. 7.9 – *Processus de Markov associé à la distribution phase-type du temps d'attente d'une urgence, dans un quartier opératoire composé de quatre salles, dont une dédiée aux urgences, et ne comportant pas de SSPI.*

nous donnent les probabilités initiales associées à la distribution phase-type. Les événements qui se produisent dans le quartier opératoire alors que le patient urgent attend, et qui vont finalement lui permettre d'accéder à une salle, ont été modélisés précédemment dans le processus Markovien MP1. Le processus de Markov MP2 caractérisant la distribution phase-type du temps d'attente est donc similaire à MP1 (voir figures 7.5 et 7.9), à deux exceptions près. Premièrement, les transitions associées à l'arrivée d'urgences (taux λ_i) sont supprimées. En effet, les cas critiques se présentant après notre patient (celui dont nous souhaitons établir le temps d'attente) n'ont aucune influence sur le temps d'attente de ce dernier. Nous ne tenons compte, dès lors, que des autres événements, c'est-à-dire ceux liés à la fin d'interventions programmées ou urgentes (taux λ_p et λ_e). Deuxièmement, les états dvq tels que $q = 0$ sont regroupés en un état absorbant unique dans MP2. En effet, $q = 0$ correspond à l'ensemble des cas où la file d'attente est vide, signifiant donc que notre patient a rejoint une salle d'opération et, par conséquent, la fin du calcul du temps d'attente. En d'autres mots, conformément à la définition de la distribution phase-type, l'état absorbant est atteint.

Afin de clarifier nos explications, prenons l'exemple d'un patient urgent se présentant au quartier opératoire alors qu'une urgence patiente déjà dans la file, une autre est en cours de traitement dans la salle dédiée et trois opérations électives sont réalisées dans les salles polyvalentes. La probabilité que le patient arrive alors que le système est dans cet état est donné par la probabilité stationnaire π_{101} (voir la sous-section précédente pour le calcul). A son arrivée, notre patient est donc le second dans la file d'attente, ce qui nous place dans l'état 102. Il faut donc que deux opérations se terminent dans le quartier opératoire, toutes salles confondues, pour que l'intervention de notre patient débute et que ce dernier cesse d'attendre. Dans le processus MP2, décrit par la figure 7.9, la probabilité de démarrer à l'état 102 est donc π_{101}, et quatre combinaisons différentes de deux transitions mènent à l'état absorbant. Ces combinaisons dépendent du type d'opérations qui s'achèvent, mettant ainsi à disposition des salles de types différents. Par exemple, si une opération planifiée arrive à son terme en premier lieu (menant à l'état 111, avec un taux $3\lambda_p$) et que l'une des deux urgences en cours se termine ensuite (l'opération planifiée ayant laissé place à une urgence, menant à l'état absorbant avec un taux $2\lambda_e$), le temps d'attente de notre patient est donné par la succession de deux distributions exponentielles (de taux $3\lambda_p$ et $2\lambda_e$). L'un dans l'autre, la

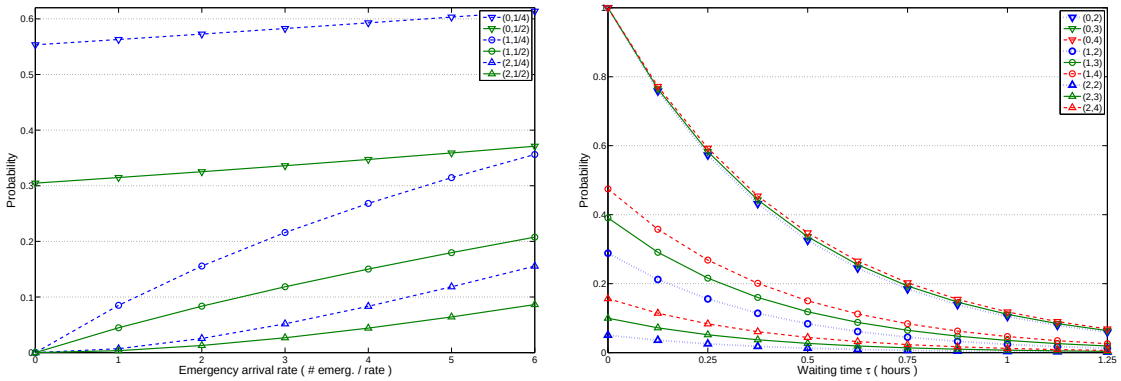

FIG. 7.10 – *Probabilité qu'une urgence attende plus qu'un seuil fixé τ (exprimé en heures). A gauche, le taux d'arrivée des urgences (λ_i) varie ; la légende donne le couple (n_d,τ). A droite, le seuil τ évolue ; la légende donne le couple (n_d,λ_i).*

distribution phase-type représentée par la figure 7.9 rassemble toutes les combinaisons possibles et donne la distribution du temps d'attente d'une urgence se présentant au quartier opératoire lorsque celui-ci est supposé ne pas être équipé de salle de réveil.

De la même manière, une distribution phase-type peut être construite afin de déterminer la distribution du temps d'attente d'une urgence tout en tenant compte de la salle de surveillance post-interventionnelle. Les probabilités initiales sont à nouveau déduites des probabilités stationnaires de se trouver dans un état spécifique du processus de Markov de départ, illustré en figure 7.6. Le processus Markovien relatif à la distribution phase-type s'élabore à partir du processus modélisant le quartier opératoire, en supprimant l'arrivée d'urgences et en regroupant les états affichant une file d'attente vide en un unique état absorbant. A l'aide de cette distribution phase-type, caractérisée par un processus de Markov et les probabilités initiales d'accéder à l'un de ses états, nous sommes à présent en mesure de calculer la probabilité, pour une urgence, d'attendre plus qu'un certain seuil τ. Cette probabilité est obtenue en appliquant la formule :

$$1 - F(\tau) \;=\; \alpha \;\exp(Q\tau)\;\mathbf{1},$$

où $F(t)$ est la fonction de distribution du temps d'attente, α est le vecteur des probabilités initiales associées aux états du processus de Markov, Q est la matrice de transition associée à la distribution phase-type, et $\mathbf{1}$ est un vecteur comportant uniquement la valeur un.

La figure 7.10 illustre deux mesures calculées sur base de la distribution du temps d'attente d'une urgence. La partie gauche décrit comment la probabilité d'attendre plus qu'un seuil déterminé τ, exprimé en heures, augmente en fonction du nombre d'urgences à réaliser, et dans quelle proportion la valeur de ce seuil affecte cette probabilité. Cette figure nous enseigne également que dédier une ou plusieurs salles au traitement exclusif des urgences augmente sensiblement la qualité du service qui leur est proposé, en diminuant leur probabilité d'attendre trop longtemps. Le graphe de droite représente, quant à lui, l'évolution de cette dernière en fonction de la valeur du seuil τ, de la configuration du quartier opératoire et du taux d'arrivée des urgences. Notons que, assez implicitement, dédier des salles aux urgences permet à certaines d'entre elles de ne pas avoir à attendre du tout. Ces figures illustrent également comment notre outil peut aider le gestionnaire

en lui permettant de prendre des décisions rationalisées, par exemple dans l'optique de réserver un bloc opératoire à l'usage unique des urgences.

Nous venons donc de montrer comment déterminer la distribution phase-type du temps d'attente des urgences. D'une manière toute similaire, l'estimation du nombre d'heures supplémentaires générées nécessite la même démarche, démarche présentée dans le point suivant.

7.5.4 Heures supplémentaires

Les heures supplémentaires sont le temps durant lequel les ressources opératoires (salles, personnel médical) sont utilisées au-delà des heures normales d'ouverture du quartier opératoire (un jour comprenant ici huit heures). Elles représentent un vecteur de coûts majeur et doivent être évitées autant que faire se peut (entre autres car le personnel médical travaillant en heures supplémentaires est plus rétribué). Cependant, le dépassement horaire ne peut virtuellement pas être supprimé, en raison de la probabilité d'occurrence non nulle de jours exceptionnels comportant des interventions étonnamment longues ou de nombreuses urgences. Le gestionnaire se doit donc de trouver le bon équilibre entre les coûts dus aux heures supplémentaires et les coûts d'opportunité dus à la non affectation des ressources durant la marge de sécurité qu'il préserve lors de la construction des plannings (destinée à absorber les aléas). Dans ce contexte, nous souhaitons aider le gestionnaire en répondant à l'une de nos questions de recherche, à savoir quel taux d'occupation appliquer lors de la programmation opératoire, ou autrement dit combien d'opérations programmer, afin d'atteindre un nombre d'heures supplémentaires acceptable. Pour ce faire, le gestionnaire a besoin d'une mesure fiable du temps de travail effectué en heures supplémentaires.

Pour disposer de mesures de performance telles que le nombre moyen d'heures supplémentaires générées ou la probabilité de travailler en heures supplémentaires, nous devons déterminer la distribution du temps de travail global du quartier opératoire, pour une journée. Par temps de travail global nous entendons la somme des temps de travail dans chaque salle polyvalente, puisque le but est de trouver le nombre d'opérations à planifier rendant nos approches déterministes robustes. Cette distribution est une nouvelle fois phase-type : il s'agit d'une succession de distributions exponentielles correspondant aux durées opératoires des opérations électives (taux λ_p) et urgentes (taux λ_e), et au temps de blocage des salles (taux $n_b\lambda_b$). Nous supposons ici que le nombre d'opérations programmées par jour, no_p, est fixé a priori, comme un paramètre du système, et correspond au taux d'occupation décidé par le gestionnaire. A l'inverse, le nombre quotidien de distributions exponentielles relatives aux urgences, no_e, et son équivalent pour les temps de blocage, no_b, varient aléatoirement (chacun de ces nombres ayant une certaine probabilité d'occurrence). Notons que, dans le calcul de la distribution du temps de travail global, nous négligeons la possibilité qu'un patient bloqué dans une salle d'intervention achève sa phase de réveil sans avoir rejoint la SSPI. La raison en est simple : ajouter cet état de fait augmenterait considérablement la complexité de calcul car il nous faudrait adjoindre deux nouveaux indices à la modélisation des états du processus de Markov, et ce afin de compter le nombre de blocages, sans pour autant améliorer significativement la précision de l'estimation de la distribution du temps de travail global. Ne pas tenir compte des patients se réveillant dans les blocs opératoires nous amène à légèrement surestimer le temps de blocage, cette hypothèse étant toutefois acceptable puisque ce dernier ne constitue qu'une faible partie du temps de travail total. En effet, dans le cas de l'hôpital belge que nous étudions, notre modèle estime que 13% des patients bloqués terminent leur phase

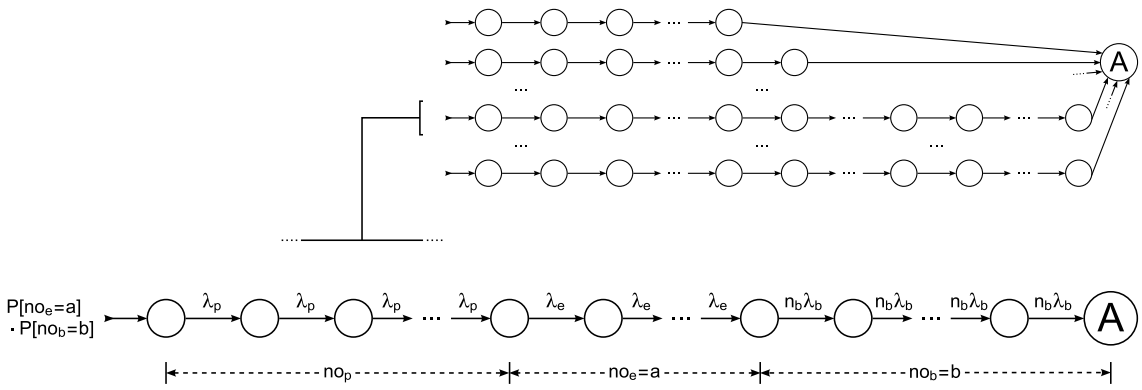

FIG. 7.11 – *Processus de Markov associé à la distribution phase-type du temps de travail global (partie supérieure), et détail d'une des branches de cette distribution (partie inférieure).*

de réveil en salle d'opération. Ce pourcentage se déduit des transitions du processus de Markov décrit en sous-section 7.5.2. Il suffit, pour ce faire, de multiplier les probabilités conditionnelles qu'une, deux, etc. salles polyvalentes soient bloquées sachant qu'il y a du blocage, par la probabilité qu'un patient bloqué se réveille avant qu'un lit ne se libère. Mathématiquement, cela se traduit par la formule $\sum_{i=1}^{n_v} P[v_b = i | v_b > 0] \cdot \frac{i}{(n_b+i)}$.

La distribution phase-type qui en résulte est présentée en figure 7.11. Chacune de ses branches est donc une suite de no_p, no_e et no_b distributions exponentielles de taux λ_p, λ_e et $n_b\lambda_b$ respectivement, ayant une probabilité initiale égale au produit des probabilités d'avoir no_e et no_b distributions exponentielles (rappelons que no_p est fixé par le gestionnaire). Par conséquent, pour définir entièrement la distribution du temps de travail global, nous devons connaître les distributions des nombres no_e et no_b. Nous expliquons comment y parvenir dans les paragraphes suivants.

Urgences. Concernant le temps dévolu au traitement des urgences, nous devons déterminer la distribution du nombre d'urgences perturbant le programme opératoire par jour, c'est-à-dire la probabilité d'avoir zéro, une, deux, etc. urgences opérées dans les salles polyvalentes. Pour calculer le nombre d'urgences perturbatrices se produisant par jour, nous ne pouvons faire usage des probabilités stationnaires. Nous devons nous appuyer sur une mesure transitoire et compter le nombre de cas urgents s'insérant dans le planning. Pour ce faire, le processus de Markov modélisant le fonctionnement du quartier opératoire, et décrit en section 7.5.2, doit être étendu. Un nouvel indice est greffé à chacun des états afin de compter le nombre d'urgences intercalées parmi les interventions planifiées dans les salles polyvalentes ; en d'autres termes, $[(d_e d_b)(v_e v_b) q b]$ devient $[(d_e d_b)(v_e v_b) q b \underline{c}]$. Les transitions entre états sont, bien entendu, ajustées en conséquence. Pour chaque transition correspondant à l'insertion d'un cas urgent dans les salles polyvalentes, le compteur c est augmenté d'une unité. Par exemple, dans la figure 7.6, à partir de l'état $[(10)(11)25\underline{1}]$, si la salle polyvalente bloquée devient accessible, une urgence y est admise, menant à l'état $[(10)(20)15\underline{2}]$ avec un taux $5\lambda_b$[1]. La chaîne de Markov destinée à calculer la distribution du nombre no_e est donc similaire à celle décrite en section 7.5.2. De manière informelle, celle-ci

[1]Rappelons que pour déterminer la distribution du temps de travail global, nous omettons la possibilité que les patients achèvent leur phase de réveil dans les blocs opératoires. Chaque patient ne quittant pas le système pour l'USI passe immanquablement par la SSPI, ce qui justifie le $5\lambda_b$ dans cet exemple.

peut être vue comme ayant une dimensionnalité additionnelle, les transitions se produisant d'un niveau vers le niveau supérieur (de c vers $c+1$) si une urgence accède à une salle polyvalente. Notons que, comme pour le cas classique, le processus est tronqué afin de garder un nombre fini d'états. Le compteur c de même que la taille de la file d'attente q sont limités de telle façon que les états négligés présentent une probabilité d'occurrence très faible.

A partir de cette chaîne de Markov étendue, nous pouvons calculer les probabilités transitoires $\pi(t)$ des états à l'aide de la formule classique $\pi(t) = \pi(0)e^{Qt}$ (voir Stewart, 1994, par exemple), où Q est la matrice de transition et $\pi(0)$ est la probabilité initiale, déterminée en considérant qu'il n'y a pas d'urgence dans le quartier opératoire au début de la journée. Comme nous considérons huit heures ouvrables par jour, les probabilités transitoires $\pi(8\,\text{heures})$ indiquent la probabilité d'être dans chacun des états à la fin de la journée de travail. Or, par construction du processus de Markov étendu, nous cherchons la probabilité $P[no_e = a]$ que a urgences perturbent le planning durant une journée. Pour la trouver, il suffit simplement de sommer les probabilités transitoires des états dont l'indice c prend la valeur a.

Blocage. Nous procédons de façon similaire pour calculer le temps, par jour, durant lequel une salle d'opération est bloquée par un patient y entamant sa phase de réveil, ainsi que le nombre de distributions exponentielles qui y sont associées. Le procédé est cependant quelque peu plus compliqué que précédemment. Pour mesurer l'impact temporel des urgences perturbatrices, il nous suffisait de compter leur nombre, qui correspondait lui-même au nombre de distributions exponentielles à inclure dans la branche de la distribution phase-type, étant donné que le temps opératoire des urgences est distribué exponentiellement. Ceci n'est plus vrai pour le temps de blocage. Quand une salle d'intervention est bloquée en raison de l'engorgement de la salle de réveil, le temps durant lequel le patient opéré monopolise cette salle n'est plus nécessairement distribué exponentiellement. Si aucune autre salle n'est bloquée lorsqu'il débute sa phase de réveil dans son bloc opératoire, un patient devra attendre durant un temps distribué exponentiellement, de taux $n_b\lambda_b$, que l'un des n_b lits de réveil devienne disponible avant de libérer son bloc[2]. Par contre, si une autre salle est déjà bloquée lorsque son intervention est terminée, un patient devra tout d'abord attendre que la personne en attente qui le précède accède à la SSPI (distribution exponentielle de taux $n_b\lambda_b$), et ensuite qu'un nouveau lit de réveil devienne inoccupé pour qu'il rallie lui-même la SSPI (distribution exponentielle de taux $n_b\lambda_b$, par propriété d'absence de mémoire). Dans ce cas, notre patient bloquera donc sa salle d'opération pour une durée suivant une distribution Erlang(2,$n_b\lambda_b$). De manière générale, le temps de blocage d'une salle suit une distribution Erlang ($z+1$,$n_b\lambda_b$), c'est-à-dire $z+1$ distributions exponentielles successives, où z est le nombre de salles déjà bloquées à la fin de l'intervention[3]. Dès lors, il faut ajouter le bon nombre de distributions exponentielles, représentant chaque blocage, dans la distribution phase-type du temps de travail global détaillée en figure 7.11.

Calculer la distribution du nombre total no_b de distributions exponentielles relatives au temps de blocage nécessite d'adapter une nouvelle fois le processus de Markov modélisant le fonctionnement du quartier opératoire. Tout comme pour le cas des urgences perturbatrices, un nouvel indice comptant le nombre de distributions exponentielles requises est ajouté à chaque état :

[2] Rappelons à nouveau que pour calculer la distribution du temps de travail global du quartier opératoire, nous faisons l'hypothèse qu'aucun patient bloqué ne se réveille avant d'avoir rejoint la SSPI.

[3] La distribution du temps de blocage peut être vue comme la distribution du temps d'attente dans une file de type $GI/M/s$, qui a été démontrée Erlang (voir Asmussen, 2003, par exemple).

$[(d_e d_b)(v_e v_b)qb]$ devient alors $[(d_e d_b)(v_e v_b)qb\underline{f}]$. Celui-ci est incrémenté comme suit : si z salles sont bloquées lorsqu'une opération s'achève, bloquant de la même manière la salle dans laquelle elle s'est déroulée, l'indice f est augmenté de $z+1$ (pour les $z+1$ distributions exponentielles successives composant la distribution Erlang($z+1$,$n_b\lambda_b$) associée au temps de blocage). A titre illustratif, partant de la figure 7.6 et supposant un état $[(10)(11)25\underline{1}]$, si une opération élective s'achève, comme les lits de réveil sont tous occupés, le patient opéré reste bloqué dans sa salle d'opération. Etant donné qu'une autre salle polyvalente est déjà immobilisée, ce dernier doit donc attendre que deux lits se libèrent avant d'être admis en SSPI, incrémentant le nouveau compteur de deux et menant à l'état $[(10)(11)25\underline{3}]$. Sur base de cette chaîne de Markov étendue, nous pouvons à présent calculer les probabilités transitoires $\pi(8\,\text{heures})$ et dès lors la probabilité $P[no_b = b]$ que le temps de blocage total soit composé de b distributions exponentielles successives sur un jour. Ce compteur f est également borné afin de garder un processus de Markov fini.

Nous venons d'expliquer comment établir la distribution des nombres no_e et no_b nécessaires pour obtenir davantage d'informations sur le nombre d'heures supplémentaires générées. Nous sommes donc en mesure de calculer la distribution phase-type du temps de travail global du quartier opératoire, comme décrite en figure 7.11. La probabilité de travailler en heures supplémentaires est intimement liée au temps de travail global puisqu'elle correspond à la probabilité que ce dernier dépasse huit heures par salle, c'est-à-dire $1 - F(n_v\,8\,\text{heures})$, où $F(t)$ est la fonction de distribution cumulative du temps de travail global dont nous venons d'expliciter le calcul. Nous disposons donc de l'outil que nous souhaitions développer : nous pouvons à présent lier le nombre d'opérations programmées par le gestionnaire (no_p) et la probabilité d'avoir recours aux heures supplémentaires. Ceci nous permet de répondre à des questions telles que : quel taux d'occupation doit être pratiqué, en d'autres mots quelle charge de travail doit contenir le programme opératoire, afin de n'avoir pas plus d'une journée présentant des heures supplémentaires par semaine, en moyenne ?

La figure 7.12 illustre les mesures que l'on peut tirer de la distribution du temps de travail total du quartier opératoire. Le graphique de gauche présente la probabilité de générer des heures supplémentaires en fonction du nombre d'interventions que planifie le gestionnaire par jour, autrement dit en fonction du taux d'occupation des salles appliqué lors de leur programmation. Sur base des courbes contenues dans ce graphique, il est possible de déterminer combien d'opérations peuvent être planifiées quotidiennement, en moyenne, pour garder une probabilité acceptable d'achever le planning prévisionnel en heures supplémentaires. Nous pouvons également y observer l'impact du taux d'arrivée des urgences de même que celui du nombre de blocs opératoires dédiés à ces dernières sur l'allure des courbes. Notamment, la perturbation engendrée par l'admission d'urgences dans les salles polyvalentes, et dès lors le nombre d'heures supplémentaires indispensables pour réaliser l'ensemble des activités chirurgicales planifiées, diminue lorsque l'on dédie certains blocs au seul traitement des cas urgents. Cette figure apprend au gestionnaire d'un quartier opératoire dont un bloc est dédié qu'il peut programmer douze interventions par jour pour ne dépasser l'horaire qu'une fois par semaine, en moyenne. La droite de la figure 7.12 décrit, quant à elle, la distribution du temps de travail global lorsque le niveau acceptable d'apparition des heures supplémentaires est fixé (ici, une fois toutes les deux semaines, une puis deux fois par semaine en moyenne). Ces courbes suggèrent également que la distribution du nombre d'heures supplémentaires générées, c'est-à-dire la durée du dépassement horaire moyen, peut être calculée à l'aide de notre approche.

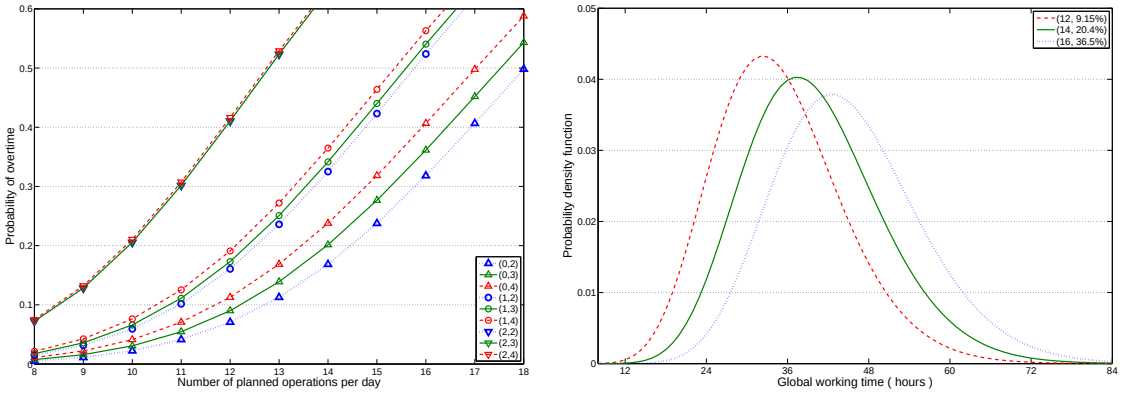

FIG. 7.12 – *A gauche, probabilité de générer des heures supplémentaires en fonction du nombre d'opérations planifiées ; la légende donne le couple (n_d,λ_i). A droite, fonctions de densité de probabilité du temps de travail global du quartier opératoire ; la légende précise le nombre d'opérations programmées et la probabilité de générer des heures supplémentaires qui en découle.*

Définir la fraction de capacité à préserver lors de l'élaboration des programmes opératoires, ou autrement dit le taux d'occupation, n'est pas chose aisée. En effet, cette décision conditionne le rendement du quartier opératoire et donc l'activité en elle-même. Si le processus de Markov développé ici représente un outil précieux pour appréhender la variabilité de l'environnement chirurgical, il ne détermine pas quelle est la meilleure décision à prendre. Cette tâche incombe au gestionnaire du quartier opératoire qui, seul, a les informations nécessaires pour rationaliser ce choix. En l'occurrence, lorsqu'il s'agit de construire le planning opératoire prévisionnel, est-il préférable de programmer quotidiennement 14 interventions, en moyenne, pour n'appliquer qu'un taux d'occupation de 80%, ou de prévoir 15 opérations par jour et n'épargner que 15% de la capacité totale des salles ? Choisir l'une de ces options se fait en lien avec des impératifs de profit, de coût, de prise en charge des patients, et surtout en fonction de l'importance de la variabilité caractérisant l'activité du quartier opératoire concerné. Il est dès lors envisageable d'adjoindre une fonction objectif au processus Markovien, fonction objectif qui n'influencerait pas les mesures qu'il propose mais permettrait de mieux les décrypter. Dans le cadre du remplissage des blocs opératoires lors de leur programmation, cette fonction objectif comporterait essentiellement deux termes : un terme de profit principalement associé au nombre d'opérations planifiées (et donc à l'utilisation des ressources chirurgicales) et un terme de coût combinant heures supplémentaires, report d'intervention, bien-être du patient ou d'autres. Cette fonction objectif se devrait donc de mettre en relation le taux d'occupation, c'est-à-dire le nombre d'opérations à programmer, avec l'utilité du décideur, utilité probablement exprimée monétairement. N'ayant pas à disposition les informations nécessaires pour détailler une telle fonction objectif, nous nous contenterons de mentionner la possibilité d'en élaborer une.

Lorsque le taux d'occupation optimal a pu être fixé, se pose alors la question de la répartition de la capacité réservée sur les différentes salles, au cours de la journée. En effet, la mesure proposée par notre outil concerne le temps de travail global, toute salle confondue. Encore faut-il pouvoir désagréger cette capacité pour effectivement partager celle-ci au sein du quartier opératoire. Si des études (Wullink et al., 2007, notamment) préconisent de répartir la capacité retenue sur l'en-

semble des blocs opératoires plutôt que de dédier une salle à l'absorption des aléas (les urgences en l'occurrence dans cette étude), cette répartition ne doit pas nécessairement être équitable. Elle peut, en effet, tenir compte de la variabilité des opérations devant se réaliser dans chacun des blocs opératoires (variabilité en termes de durée d'intervention), et accorder une plus grande marge de sécurité dans les blocs dont les opérations présentent un plus haut degré d'incertitude. Quant à savoir où placer la capacité préservée dans la journée, il paraît censé de le faire dynamiquement, en fonction du déroulement de l'activité chirurgicale. En effet, pourquoi retarder une intervention qui peut débuter en temps voulu, voire plus tôt, sous prétexte qu'une tranche horaire disposée en pleine journée est destinée à absorber les événements aléatoires ?

Heure de fermeture maximum. Comme elle provient de la distribution du temps de travail global du quartier opératoire, c'est-à-dire la somme des temps de travail de chaque salle, notre estimation du nombre d'heures supplémentaires générées ne donne qu'une mesure moyenne, partagée par toutes les salles. Une autre mesure intéressante à propos des heures supplémentaires consiste à déterminer, le cas échéant, le dépassement horaire maximum au sein des salles. Ceci revient à calculer l'heure de fermeture de la dernière salle encore occupée, de sorte que les no_p patients électifs soient traités et que toutes les salles d'opération polyvalentes soient libérées de leurs patients (y compris les éventuels patients bloqués toujours présents après la fermeture de la dernière salle). La distribution du temps de fermeture maximum d'une salle peut également être modélisée à l'aide d'une distribution phase-type. Pour ce faire, le processus de Markov, décrit en section 7.5.2 et illustré par la figure 7.6, est modifié par l'ajout d'un indice supplémentaire comptabilisant le nombre de patients électifs entrant dans une salle polyvalente ; $[(d_e d_b)(v_e v_b)qb]$ évolue en $[(d_e d_b)(v_e v_b)qb\underline{h}]$. L'heure de fermeture maximum est donc une succession de distributions exponentielles correspondant aux transitions dans ce nouveau processus de Markov. La probabilité initiale est de un pour l'état $[(00)(00)00\underline{n_v}]$ comme, au début de la journée, les n_v salles polyvalentes du quartier opératoire sont occupées par des patients électifs et ne comportent ni urgences ni blocage, la salle de réveil étant elle-même vide. Ensuite, lorsque de nouveaux patients programmés entrent dans le quartier opératoire au fil de la journée, le compteur h est progressivement incrémenté (comme dans les cas précédents, les transitions se font d'un niveau au niveau supérieur). Quand ce compteur atteint la valeur no_p, donc au niveau no_p du processus de Markov, aucun autre patient électif ne peut plus accéder au quartier opératoire (à ce moment, les salles polyvalentes sont artificiellement remplacées par des salles dédiées, pour permettre aux opérations en cours et aux éventuelles urgences à venir avant la fermeture du quartier opératoire d'être réalisées). L'état absorbant, représentant la fin du calcul du temps de fermeture de la dernière salle restée ouverte, est atteint lorsque tous les patients électifs ont quitté les salles d'opération. Cette description informelle nous permet de caractériser entièrement la distribution phase-type (processus de Markov, probabilités initiales et état absorbant) et de la mettre en œuvre[4] ; les résultats concernant cette nouvelle mesure de performance sont exposés dans la prochaine section portant sur le cas d'étude de l'hôpital belge. De toute évidence, le nombre maximum d'heures supplémentaires générées dans une salle est considérablement plus grand que le nombre moyen, parmi toutes les salles, calculé précédemment.

[4]Notons que, à l'inverse du calcul du temps de travail global, le modèle utilisé pour estimer l'heure de fermeture maximale d'une salle autorise le réveil de patients en salle d'opération. Les patients ne sont donc plus automatiquement dirigés vers la SSPI s'ils ne sont pas admis en unité de soins intensifs.

Cette section a décrit notre approche Markovienne, qui offre une vue relativement complète du quartier opératoire et de l'effet des événements stochastiques qui s'y produisent. Cet outil permet de mesurer les performances du quartier opératoire de diverses manières et permet de guider le gestionnaire dans sa prise de décision. La section suivante illustre comment cet outil peut s'utiliser en pratique, ici sur base du cas d'étude récurrent tout au long de ce travail.

7.6 Illustration pratique

Le cas illustratif de l'hôpital bruxellois, résumé en section 7.4.2 et dont les principaux paramètres sont repris dans le tableau 7.1, nous sert ici pour démontrer l'utilité de notre approche de gestion des aléas.

Tout d'abord, le gestionnaire hospitalier peut dresser un portrait complet du fonctionnement du quartier opératoire et de l'effet de la stochasticité sur ses performances. Les mesures de performances calculées dans ce contexte sont présentées dans le tableau 7.2. Comme aucune salle n'est actuellement dédiée aux urgences, l'ensemble de celles-ci perturbe le programme opératoire en s'intercalant entre deux patients électifs dans les salles polyvalentes, durant 4 heures et 19 minutes en moyenne. Le temps d'attente moyen d'une urgence est de 27 minutes alors que la probabilité d'attendre plus d'une demi-heure avoisine les 34%. En moyenne, une salle est bloquée durant 1% de la journée, ce qui engendre la perte d'une demi-heure de temps opératoire répartie sur la totalité des salles polyvalentes pour raison de blocage. Planifier 15 opérations revient à appliquer un taux d'occupation estimé à 88.5%, urgences et blocage inclus. Par ailleurs, le calcul de la distribution du temps de travail global (somme des temps opératoires de l'ensemble des blocs polyvalents), pour les 15 opérations planifiées par jour, permet d'estimer la probabilité de travailler en heures supplémentaires à 28%, ce qui revient à un dépassement horaire se répétant trois jours toutes les deux semaines. Si maintenant nous nous focalisons sur le temps de travail maximum au sein des salles (voir sous-section 7.5.4), le dernier bloc opératoire à fermer est utilisé en moyenne pendant 10 heures et 24 minutes par jour. La probabilité que l'heure de fermeture la plus tardive d'une salle dépasse les huit heures réglementaires, c'est-à-dire la probabilité de dépassement horaire "dans le pire des cas", est estimée à 75%.

Malheureusement, les évaluations fournies par notre outil sur le temps d'attente des urgences et sur le blocage des salles ne peuvent être entérinées sur base du jeu de données disponible. Cependant, certains recoupements peuvent être faits concernant l'estimation de la distribution du temps de travail global du quartier opératoire. La figure 7.13 compare les distributions théorique, calculée par notre outil, et empirique, observée dans les données (sur 42 jours ouvrables). Nous y remarquons que la distribution phase-type établie par notre approche est très similaire à la distribution réelle. Ceci suggère notamment que l'hypothèse bannissant le réveil des patients en salle d'opération n'est pas trop restrictive. La principale différence s'observe pour les valeurs extrêmes, où empiriquement nous observons plus de jours comprenant entre 43 et 51 heures de travail global. Ceci peut s'interpréter de la manière suivante : nous pouvons supposer qu'en pratique, lorsque le gestionnaire est face à des journées se prolongeant plus que la normale, le planning est adapté et certaines opérations sont postposées (ce qui engendre toutefois un coût, ne serait-ce qu'humainement parlant pour le patient). Les jours dépassant en principe les 51 heures repris dans la courbe théorique sont en réalité tronqués, et ramenés à des jours contenant entre 43 et 51 heures de travail, ce qui explique pourquoi notre approche en sous-estime le nombre. En d'autres mots, notre

Configuration	[6|0|9]	[5|1|9]	[6|1|9]	[6|0|10]	[5|1|10]	[6|1|10]
Taux urg. perturb. *(nb. op. / jour)*	2.98	0.86	0.89	2.98	0.86	0.88
Temps urg. perturb. *(heures)*	4.32	1.25	1.29	4.32	1.24	1.28
Temps attente moy. *(minutes)*	27.4	9.74	8.35	27.5	9.70	8.26
Proba. attente > 30 min. *(%)*	33.6	11.8	9.61	33.6	11.7	9.52
Proba. blocage *(%, QO)*	1.03	0.60	1.22	0.47	0.24	0.63
Temps blocage moy. *(minutes / QO)*	29.6	17.2	44.8	13.4	6.96	21.1
Proba. heures supp. (15 op.) *(%)*	27.9	44.0	19.3	27.5	43.7	18.8
Nbr. op. à programmer *(nb. op. / jour)*	14 (20.4)	12 (17.3)	15 (19.3)	14 (20.0)	12 (17.1)	15 (18.8)

TAB. 7.2 – *Estimation de plusieurs mesures de performance pour le cas de l'hôpital belge, pour différentes configurations du quartier opératoire (QO). Les heures supplémentaires sont ici représentées par leur nombre moyen pour toutes les salles (et pas par le nombre maximum pour une salle, voir sous-section 7.5.4). La probabilité d'heures supplémentaires est calculée pour 15 interventions planifiées comme c'est le cas dans les données. Le nombre d'opérations à programmer est calculé sur base d'un dépassement horaire généré moins d'une fois par semaine (la probabilité d'heures supplémentaires associée est indiquée en pour-cent en dessous).*

prédiction théorique semble relativement précise et s'avère une base objective pour comparer les implications des différentes décisions que peut prendre le gestionnaire, même si en pratique la réalisation du planning prévisionnel peut être sujette à caution (pour éviter des journées anormalement longues par exemple). Afin de tenir compte de cet état de chose, la distribution estimée pourrait être tronquée.

Afin de convaincre davantage le lecteur de la justesse de notre estimation de la distribution du temps global de travail, nous appliquons un test statistique pour comparer l'échantillon fourni par le cas d'étude belge et la distribution de référence, calculée par notre outil. L'hypothèse nulle que nous testons concerne l'égalité des distributions empirique et théorique estimée. Les deux tests les plus classiques dans ce domaine sont le test du chi carré et celui de Kolmogorov-Smirnov (voir Saporta, 2006; Shao, 2003). Le test du chi carré amène à ne pas rejeter l'hypothèse nulle (même à 20%), avec sept ($D^2 = 8.34$) comme avec treize classes ($D^2 = 14.02$). De même, le test de Kolmogorov-Smirnov ne permet pas non plus de rejeter l'hypothèse nulle, même à 20% ($D_n = 0.1557$). Finalement, le calcul du nombre maximum d'heures supplémentaires générées dans une salle peut être recoupé avec les données. La probabilité estimée que le temps de travail maximum dans une salle excède les huit heures (75%) est proche de la probabilité équivalente calculée à partir des données : 73%. Cette différence peut à nouveau être partiellement justifiée par le fait que, en pratique, les jours trop longs sont évités en postposant des interventions programmées.

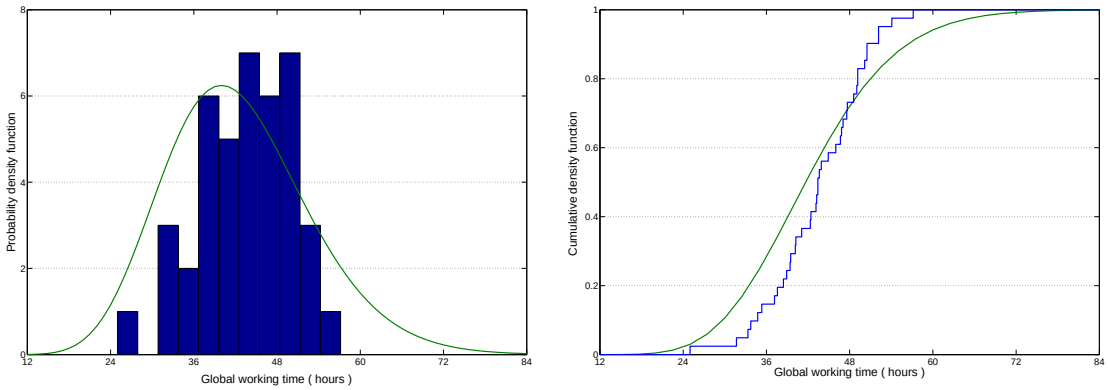

FIG. 7.13 – *Comparaison des distributions empirique et théorique du temps global de travail, pour le cas de l'hôpital belge. A gauche, la fonction de densité de probabilité et à droite la fonction de distribution cumulative.*

Outre offrir un aperçu rationnel et quantitatif du comportement d'un quartier opératoire face à la variabilité de l'environnement dans lequel il évolue, notre approche permet au gestionnaire d'évaluer le résultat de changements qu'il entrevoit pour améliorer les performances des unités chirurgicales. Un souci majeur dans ce domaine concerne le travail en heures supplémentaires qui engendre d'importants coûts additionnels. Dans la configuration actuelle du bloc opératoire belge (à savoir $[6|0|9]$), avec 15 opérations programmées chaque jour, ce qui revient à appliquer un taux d'occupation de 88% (urgences et blocage compris), on estime à trois jours toutes les deux semaines, en moyenne, la nécessité de travailler en heures supplémentaires. Pour réduire cette proportion, par exemple aux alentours d'une fois par semaine, le nombre d'opérations à programmer décroît à 14, le taux d'occupation des salles atteignant alors 83%. Grâce à cet outil, le taux d'occupation à appliquer lors de la construction des programmes opératoires peut donc être fixé sur base du travail en heures supplémentaires qu'il est susceptible de générer.

A un niveau de décision stratégique, le gestionnaire peut évaluer l'impact d'un changement de configuration du quartier opératoire. Actuellement, il peut observer que la qualité de prise en charge des urgences est assez pauvre : un patient urgent attend en moyenne 27 minutes avant d'être admis dans une salle d'opération. De plus, ces urgences perturbent significativement le planning opératoire, et pénalisent de la sorte la qualité de service offerte aux patients électifs (qui sont alors potentiellement retardés). Afin de rectifier cet état de fait, le gestionnaire peut envisager de dédier un bloc opératoire à l'unique réalisation des interventions urgentes. Comme mentionné dans le tableau 7.2, le temps d'attente moyen est facilement réduit à moins de 10 minutes (avec seulement 12% de chances d'attendre plus d'une demi-heure), de même que l'est la perturbation du planning prévisionnel par l'insertion d'urgences dans les salles polyvalentes. Cependant, cette stratégie a un coût : seules cinq salles polyvalentes sont encore disponibles pour traiter les cas électifs, avec pour conséquence une augmentation du travail en heures supplémentaires (44%) ou une diminution du nombre d'interventions à programmer (en préservant alors un niveau acceptable d'heures supplémentaires). Dans la description du quartier opératoire belge étudié dans ce document, résumée en sous-section 7.4.2, un septième bloc opératoire existe mais n'est pas utilisé par manque de personnel. Dédier cette septième salle aux urgences permettrait de maintenir le nombre d'opérations

à programmer par jour à 15 unités tout en gardant une faible probabilité d'heures supplémentaires et en améliorant significativement la qualité de service, en particulier aux urgences dont le temps d'attente moyen serait de seulement 8 minutes.

Dans la configuration présente du quartier opératoire, le gestionnaire peut encore vouloir réduire la perturbation du planning due au blocage des salles d'opération. A cette fin, l'option la plus directe est d'augmenter la capacité d'accueil de la salle de réveil. A nouveau, nous pouvons estimer l'effet d'une telle décision à partir des chiffres exposés dans le tableau 7.2. Ajouter un lit de réveil permet de réduire le temps opératoire perdu lors du blocage des salles à environ 13 minutes. Cette stratégie semble également avoir un léger impact positif sur le nombre d'heures supplémentaires générées. La configuration idéale pour le cas sur lequel repose notre étude est certainement la configuration $[6|1|10]$ qui permet, en moyenne, d'améliorer la qualité de service aux patients tout en programmant 15 opérations par jour avec des dépassements horaires prévus une fois par semaine (probabilité de pratiquement 19%). A la charge du gestionnaire d'ensuite prendre les décisions qui s'imposent, à la lumière des coûts qu'elles engendreront.

La validation de notre approche Markovienne par le cas d'étude dont nous disposons clôture ce chapitre. Nous avons donc développé un outil qui permet d'appréhender l'incertitude inhérente à l'activité opératoire et, au travers d'un panel d'indicateurs de performance, d'évaluer l'impact de décisions stratégiques ou tactiques sur le comportement du quartier opératoire. Cet outil, combiné aux techniques de programmation opératoire développées précédemment, constitue une aide à la décision précieuse soutenant les décideurs dans la délicate tâche d'améliorer la gestion des ressources opératoires. Le processus de Markov détaillé ici s'inscrit directement dans la suite logique des approches des chapitres précédents en déterminant le pourcentage de capacité à préserver avant leur application, ce dans le but d'accroître leur robustesse aux aléas. Cet outil répond aux dernières de nos perspectives de recherche, mises en exergue par la revue de la littérature, et marque la fin de nos travaux qui, nous l'espérons, apportent une contribution de qualité à cette dernière.

Chapitre 8

Conclusion générale et perspectives

Les centres hospitaliers sont des systèmes complexes qui regroupent, autour d'une même entité, à savoir le patient, une multitude de métiers aux missions bien distinctes, utilisant pour la plupart des ressources mutualisées et opérant dans un environnement en perpétuel changement. Au delà de cette complexité inhérente au milieu, la gestion de tels établissements est d'autant plus difficile que les pressions économiques forcent les hôpitaux à rationaliser les dépenses de soins de santé. Dans ce contexte financièrement contraint, l'étude du quartier opératoire semble à propos. En effet, il s'avère que ce dernier soit la plus importante source de dépenses de l'hôpital et consomme entre dix et quinze pourcents du budget global de celui-ci. De plus, les frais engendrés proviennent essentiellement de la gestion des ressources chirurgicales. Rendre la gestion des blocs opératoires plus efficace passe notamment par une utilisation plus rationnelle de ces ressources, et donc par une meilleure programmation des interventions au sein des blocs. Le processus de programmation opératoire conditionne, en effet, les performances du quartier opératoire, entre autres en termes d'utilisation des salles d'opération et d'heures supplémentaires générées. Ces deux critères sont principalement appliqués pour estimer la qualité d'un calendrier opératoire, le premier étant un vecteur de profit et le second un vecteur de coûts. Construire de meilleurs plannings opératoires est une des solutions pour améliorer la gestion du quartier opératoire, en réduisant les coûts liés aux ressources tout en maintenant la qualité des soins prodigués.

La programmation opératoire est, en pratique, une tâche complexe et fastidieuse si le décideur ne dispose pas d'outils adaptés. Outre le fait que de tels outils n'existent pas dans le commerce[1], ce processus est encore effectué manuellement dans certains établissements. C'est notamment le cas pour le gestionnaire du quartier opératoire de l'hôpital dont est issue notre étude appliquée. Chaque vendredi, ce dernier est monopolisé pour la journée entière afin d'établir le planning de la semaine à venir. Qui plus est, ce planning est loin d'être optimum : le gestionnaire se contente le plus souvent d'un programme qui satisfasse au mieux les contraintes, et ne dispose pas du temps nécessaire pour refaire cette tâche si celle-ci s'avérait devoir être aménagée. Enfin, même si le gestionnaire de blocs fait de son mieux pour contenter les desiderata de tout le monde, il est inévitable que certains chirurgiens se montreront mécontents des plages horaires qui leur seront attribuées. Or, le planning établi, le gestionnaire n'est plus en mesure de rediscuter les décisions prises, par

[1] En réalité des outils d'aide à la planification existent bel et bien mais ils sont généralement dérivés de produits destinés au monde industriel et sont rarement adaptés aux spécificités des centres hospitaliers. Les gestionnaires de blocs que nous avons pu rencontrer et qui étaient équipés de tels logiciels s'en sont le plus souvent montrés insatisfaits.

manque de temps. Ceci souligne évidemment tout l'intérêt des outils que nous souhaitions développer. Non seulement, nos méthodes de résolution facilitent la tâche du décideur en proposant des solutions (quasi) optimales en des temps raisonnables. Mais aussi, nos approches ont la faculté de pouvoir être relancées, pour envisager différents cas de figure, en fixant certaines variables et donc certaines plages horaires pour pouvoir pallier d'éventuels changements de dernière minute, etc. Un tel travail est presque impossible pour un être humain sans l'aide d'approches telles que celles que nous développons. Enfin, nos approches sont en mesure de tenir compte des préférences des acteurs principaux de l'activité opératoire, à savoir les chirurgiens, permettant ainsi une plus grande coopération au sein du processus de programmation opératoire.

8.1 Résultats de la recherche

En parcourant la littérature ayant trait au domaine, nous avons mis en évidence trois points qui ont motivé notre recherche. Tout d'abord, il nous est apparu que la majorité des travaux portant sur le processus de programmation opératoire scinde systématiquement celui-ci en deux étapes séquentielles, l'une portant sur l'attribution d'une date et d'une salle d'opération, l'autre sur l'heure théorique de début d'intervention. Ensuite, nous avons constaté que les ressources opératoires indispensables au bon déroulement de l'activité chirurgicale, qu'elles soient humaines ou matérielles, sont rarement évoquées. Or, dans un environnement centré sur l'humain, les ressources essentielles sont les membres du corps médical, qui ont les compétences requises pour mener à bien le processus de soin. Enfin, l'activité opératoire se déroule dans un environnement incertain qui influence fortement la réalisation des plannings prévisionnels établis. Or, la majorité des travaux ne traite pas cet aspect imprévisible du quartier opératoire, se contentant de développer des approches déterministes sans aide à la décision sur la prise en compte de la stochasticité du milieu.

Notre premier objectif était donc d'intégrer les étapes d'affectation des interventions et d'agencement des blocs opératoires en une unique procédure optimisant ces deux étapes globalement, et non plus localement comme le fait la littérature. De plus, cette procédure se devait de prendre en compte les disponibilités des ressources opératoires telles que les chirurgiens et les anesthésistes, mais également de considérer celles des équipements spécifiques comme certains matériels stériles. En outre, la littérature recense deux grands types de salles d'opération : celles dédiées et celles polyvalentes (ou identiques) ; il fallait donc que cette procédure puisse également prendre en compte cette spécificité. Dans un premier temps, nous nous sommes concentrés sur le cas d'un quartier opératoire composé de blocs dédiés, équipés différemment et ne pouvant accueillir chacun que certaines chirurgies (chapitre 4). Nous avons tout d'abord formulé mathématiquement le problème afin de le résoudre de façon exacte, à l'aide d'un modèle en nombres entiers mixtes. L'ensemble de nos approches ont été testées sur des données de taille réelle, regroupant l'activité opératoire observée en pratique dans un hôpital belge de la région bruxelloise. Cette activité a été subdivisée en neuf jeux de données représentant chacun une semaine d'ouverture du quartier opératoire. Sur ces neuf instances, l'approche exacte s'est révélée efficiente à six reprises, en produisant des solutions optimales ou quasi-optimales de grande qualité. Cependant, cette efficience ne se vérifie pas sur les trois autres instances, le modèle réclamant plusieurs heures de calcul pour déterminer des solutions satisfaisantes. Ceci pouvant conditionner une application en pratique de ce type d'approche exacte, nous avons alors proposé un algorithme génétique, autrement dit une méta-heuristique, capable de contrer la complexité du problème traité. L'utilisation d'une telle

méthode de résolution approchée ne se justifie que pour les trois ensembles de test sur lesquels le programme en nombres entiers mixtes peine à produire des solutions de qualité en des temps raisonnables. Toutefois, l'algorithme génétique est toujours en mesure d'offrir une population de solutions proches de l'optimum, en des temps maîtrisables déterminés par les valeurs des paramètres caractérisant celui-ci.

Nos travaux traitant d'un quartier opératoire composé de salles dédiées ont été validés par diverses publications : dans une revue internationale avec comité de lecture (Roland et al., 2010), lors d'un congrès national (Roland, DiMartinelly & Riane, 2007) et lors de conférences internationales avec comité de lecture (Roland et al., 2006; Roland, DiMartinelly, Riane & Pochet, 2007).

- Roland, B., Di Martinelly, C. & Riane, F. (2006). Operating theatre optimization : A resource constrained-based solving approach, International Conference on Service Systems and Service Management, IEEE/ICSSSM'06, Troyes, France.
- Roland, B., Di Martinelly, C. & Riane, F. (2007). On the integration of resources' preferences in the operating theatre optimization, Annual conference of the Belgian Operations Research society, ORBEL'21, Luxembourg.
- Roland, B., Di Martinelly, C., Riane, F. & Pochet, Y. (2007). Scheduling operating theatre under human resources constraints, International Conference on Industrial Engineering and Systems Management, IESM'07, Beijing, China.
- Roland, B., Di Martinelly, C., Riane, F. & Pochet, Y. (2010). Scheduling an operating theatre under human resource constraints, *Computers & Industrial Engineering* **58** : 212–220.

Nous avons ensuite analysé le processus de programmation d'un quartier opératoire dont les blocs sont identiques, et donc polyvalents (chapitre 5). Ce problème, bien qu'apparemment très semblable au premier, comporte toutefois une particularité qui le distingue nettement du précédent : la symétrie. En effet, nous avons observé que la nature polyvalente des salles d'intervention génère une grande symétrie qui, si elle n'est pas traitée de façon appropriée, peut s'avérer très néfaste sur les performances de l'algorithme de résolution. Le modèle en nombres entiers mixtes proposé au chapitre 4 a donc été adapté en conséquence, avant d'évaluer la nécessité d'une nouvelle approche heuristique. Deux formulations mathématiques de philosophies différentes sont proposées afin de briser la symétrie du problème : la première fixe des contraintes et ajoute des inégalités, la seconde redéfinit les variables de décision et propose des solutions agrégées nécessitant l'application d'un algorithme d'allocation des salles d'opération. Les deux approches se sont montrées efficientes sur les ensembles de test dont nous disposons. La première a toutefois l'inconvénient de comporter une contrainte fixant un nombre a priori de blocs opératoires systématiquement ouverts, chaque jour, afin de traiter une activité opératoire moyenne. Les performances du modèle, en termes de temps de calcul ou de valeur objectif, sont sensibles à la valeur de ce paramètre. Son choix est donc critique, ce qui le rend particulièrement délicat à manipuler. Ceci justifie le développement d'une seconde approche initiant un traitement différent de la symétrie. Les résultats de cette dernière ne sont, quant à eux, pas conditionnés par un paramètre du problème. Ce second modèle a démontré une aptitude certaine à produire rapidement des solutions quasi-optimales, peinant toutefois à les prouver optimales. C'est pourquoi nous avons également proposé une stratégie d'utilisation de ce dernier, consistant à le résoudre jusqu'à atteindre un certain niveau d'écart avec la meilleure borne inférieure, avant de le relancer, en ayant au préalable fixé certaines variables à des valeurs quasi-optimales. Ces deux modèles en nombres entiers mixtes surpassent leur prédécesseur, permettant une résolution du problème en des temps plus que satis-

faisants. Les performances proposées par ces deux formulations ont montré qu'aucune approche heuristique n'était plus nécessaire pour aborder la gestion de blocs opératoires polyvalents. Les approches de ces deux chapitres ont démontré que la dissociation du processus étudié, que prône la littérature, ne semble plus justifiée.

Le contenu de ce chapitre a donné lieu à publication dans une conférence internationale avec comité de lecture (Roland & Riane, 2010c, papier ayant obtenu la mention honorifique du jury dans la compétition étudiante) et dans un congrès national (Roland & Riane, 2009).

- Roland, B. & Riane, F. (2009). Symétrie et préférences des ressources dans la planification des blocs opératoires, Congrès de la société française de Recherche Opérationnelle et d'Aide à la Décision, ROADEF'09, Nancy, France.
- Roland, B. & Riane, F. (2010c). When OR techniques help ORs management, International Conference on Information Systems, Logistics and Supply Chain, ILS'10, Casablanca, Morocco. *Professor Salah E. Elmaghraby's Honorable Mention award*.

Notre second objectif était d'induire une forme de coopération dans l'établissement des programmes opératoires, notamment entre les chirurgiens qui constituent la ressource primordiale du processus. Nous avons alors adapté les approches précitées en permettant à ces praticiens de spécifier des préférences quant à l'attribution des plages horaires de travail (chapitre 6). Ces préférences sont tout d'abord traduites au moyen d'une fonction de désutilité propre à chaque chirurgien. Cette fonction associe à chacune des demi-journées de l'horizon de temps considéré, un poids, c'est-à-dire une mesure cardinale traduisant le désagrément subi par chacun lorsque ses préférences, exprimées pour la plage horaire en question, ne sont pas respectées. Cette première approche coopérative n'est appliquée qu'aux seules méthodes de résolution du chapitre 4. Le respect des préférences est alors intégré à la fonction objectif des approches exacte et heuristique pour obtenir des plannings minimisant les coûts d'utilisation des salles et la désutilité globale des chirurgiens, et non la désutilité de chaque chirurgien individuellement. Le programme linéaire en nombres entiers et l'algorithme génétique, tous deux adaptés au contexte coopératif et appliqués à des données réelles, montrent à nouveau des performances satisfaisantes. Ces expérimentations ont mis en évidence le fait que, dans la gestion de blocs opératoires dédiés, la coopération permettait de surmonter l'accessibilité limitée des salles, et donc de diminuer ostensiblement les coûts d'utilisation de ces dernières, au prix d'une certaine désutilité. En minimisant la désutilité globale des chirurgiens, nous avons donc incorporé une dimension coopérative à la programmation des interventions dans les blocs opératoires. Cependant, nos approches n'offraient pas la possibilité aux acteurs d'émettre leur avis sur le planning proposé. Les calendriers opératoires ainsi établis étaient construits sans avoir de retour des personnes concernées. Or, en pratique, le programme opératoire résulte bien souvent de négociations entre certains des acteurs, raison pour laquelle nous avons proposé une deuxième approche coopérative, basée, elle, sur un mécanisme d'enchères. Les chirurgiens y misent des jetons virtuels sur leurs plages horaires de prédilection, l'attribution de celles-ci étant déterminée par un programme en nombres entiers mixtes dérivant du modèle révisé présenté au chapitre 5. Outre, la minimisation des coûts, cette formulation comporte un objectif supplémentaire visant à assigner les demi-journées aux praticiens ayant soumis les mises les plus importantes. Si, à l'issue de cette phase d'enchères, certains chirurgiens se montrent insatisfaits d'une ou plusieurs demi-journées leur étant allouées, une nouvelle extension du modèle mathématique leur permettra alors de solliciter un échange de plages horaires avec leurs confrères. La négociation qui en découle met en jeu les jetons virtuels que le demandeur est prêt à céder au

confrère sollicité afin d'assurer la bonne fin de la transaction. Si l'intérêt du premier est assez évident dans cette démarche, celui du second est d'obtenir plus de jetons lors d'une phase d'enchères ultérieure, et dès lors d'avoir plus de poids dans le mécanisme et de s'assurer une meilleure représentation de ses choix. Cette seconde approche coopérative, appliquée sur les données réelles, s'est également montrée efficiente. Ce chapitre présente donc deux techniques efficaces permettant d'impliquer les acteurs principaux du quartier opératoire, les chirurgiens, dans le processus de programmation de ce dernier, intégrant par la sorte à nos approches une nouvelle dimension humaine rarement prise en compte dans la littérature.

Ces techniques coopératrices ont fait l'objet de publications dans une revue internationale avec comité de lecture (Roland & Riane, 2010b) et dans des conférences internationales avec comité de lecture (Roland et al., 2008; Roland & Riane, 2008, 2010a).

- Roland, B., Di Martinelly, C. & Riane, F. (2008). Planification du quartier opératoire intégrant les préférences des ressources humaines, Conférence Internationale de Modélisation et Simulation, MOSIM'08, Paris, France.
- Roland, B. & Riane, F. (2008). A constrained NSGA-II to consider surgeons' preferences while scheduling the operating theatre, International Conference on Information Systems, Logistics and Supply Chain, ILS'08, Madison (WI), USA.
- Roland, B. & Riane, F. (2010a). Enchères et coopération dans la programmation opératoire, Conférence Internationale de Modélisation et Simulation, MOSIM'10, Hammamet, Tunisia.
- Roland, B. & Riane, F. (2010b). Integrating surgeons' preferences in the operating theatre planning, *European Journal of Industrial Engineering* . Accepted, to appear.

Une autre perspective importante que nous souhaitions aborder dans ce travail de recherche concerne le caractère aléatoire du problème. En effet, le quartier opératoire évolue dans un environnement incertain qui va perturber l'application en pratique du programme prévisionnel. Notamment, ce dernier est construit à l'aide d'approches déterministes, comme celles utilisées par les gestionnaires hospitaliers, sur base d'estimations des temps d'opération. Or ces durées sont incertaines et difficiles à estimer avec exactitude. Les plannings que nos approches établissent seront donc soumis à modifications, en raison de cette nature stochastique du problème. L'objectif du chapitre 7 était donc de déterminer dans quelle mesure nous pouvions tenir compte de ces incertitudes dans nos méthodes de résolution afin de produire des programmes opératoires robustes, c'est-à-dire pouvant absorber une grande partie des aléas sans devoir être reconsidérés. La littérature classique aborde assez peu les aspects stochastiques. Les travaux traitant de la variabilité du quartier opératoire considèrent généralement l'une des deux grandes sources d'aléas, à savoir l'inconstance des durées opératoires et l'arrivée aléatoire d'urgences prioritaires devant être intercalées dans le planning. Or, il existe d'autres phénomènes qui s'observent en pratique et perturbent le programme établi. Notamment, la saturation des unités de réveil provoque un blocage des salles d'opération par des patients opérés qui ne peuvent attendre, par manque d'autonomie respiratoire, avant de rejoindre un lit en salle de surveillance post-interventionnelle. Ceux-ci entament dès lors leur phase de réveil dans la salle qui a vu le déroulement de leur intervention chirurgicale, empêchant alors les opérations programmées dans cette salle, encore à réaliser, de débuter. La plupart du temps, ces différentes sources de stochasticité vont perturber le programme opératoire et imposer le recourt aux heures supplémentaires ou le report d'interventions, le premier générant un important surplus de coûts et le second engendrant l'insatisfaction des patients. Concrètement, les gestionnaires hospitaliers se prémunissent de tels désagréments en préservant une partie du

temps opératoire total destinée à assimiler un maximum de ces événements imprévisibles. Cette marge de sécurité est définie subjectivement, par expérience, et rien dans la littérature ne permet de rationaliser ce choix. C'est pourquoi, dans ce chapitre, nous avons proposé une approche permettant aux gestionnaires de blocs opératoires de mesurer l'impact de la stochasticité de leur environnement, et, sur base des mesures de performance qu'offre cette approche, de rationaliser les décisions stratégiques et tactiques qu'ils ont à prendre. Les trois principales sources d'aléas y sont combinées : l'incertitude des durées opératoires, l'arrivée imprévisible d'urgences et les durées aléatoires de la phase de réveil engendrant potentiellement le blocage des salles d'intervention en raison de l'engorgement de la salle de surveillance post-interventionnelle. Notre approche consiste en un processus de Markov modélisant, au travers de ses différents états, l'évolution du système. De ce processus, nous avons ensuite dérivé des distributions phase-type du temps d'attente des urgences, ainsi que du temps d'utilisation globale du quartier opératoire ou du temps de fermeture maximum de celui-ci. En outre, cette chaîne de Markov à temps continu permet aisément de tester différentes configurations du quartier opératoire, offrant même la possibilité de dédier des salles au traitement exclusif des urgences. Les mesures de performance que nous en avons tirées concernent, par exemple, la perturbation du calendrier opératoire par les urgences ou par le blocage de salles d'intervention, les heures supplémentaires susceptibles d'être générées, ou encore le temps d'attente des urgences. De surcroît, cet outil permet d'estimer le nombre d'opérations à programmer relatif à un certain niveau d'heures supplémentaires, et donc le taux d'occupation des blocs opératoires à appliquer lors de l'élaboration de ces calendriers opératoires. Il évalue encore l'impact du nombre de lits de réveil équipant la salle de surveillance post-interventionnelle sur les performances du quartier opératoire. L'application de notre approche a finalement été illustrée sur le cas d'étude de l'hôpital belge et s'est avérée traduire fidèlement le comportement observé du quartier opératoire.

L'approche Markovienne présentée dans ce chapitre est le fruit de la collaboration avec Messieurs Jean-Sébastien Tancrez et Jean-Philippe Cordier. Cette collaboration a donné lieu à plusieurs publications comme chapitre de livre avec comité de lecture (Tancrez et al., 2009) ou lors de conférences internationales avec comité de lecture (Tancrez et al., 2008; Roland et al., 2009, le premier papier ayant obtenu le prix de la meilleure communication étudiante).

- Roland, B., Cordier, J.-P., Tancrez, J.-S. & Riane, F. (2009). Recovery beds and blocking in stochastic operating theatres, International Conference on Industrial Engineering and Systems Management, IESM'09, Montreal, Canada.
- Tancrez, J.-S., Roland, B., Cordier, J.-P. & Riane, F. (2008). Etude de la perturbation par les urgences du planning opératoire, Conférence de Gestion et Ingénierie des SystèmEs Hospitaliers, GISEH'08, Lausanne, Suisse. *Meilleure communication d'étudiant-Docteur*. Egalement publié dans *Logistique & Management* (à paraître).
- Tancrez, J.-S., Roland, B., Cordier, J.-P. & Riane, F. (2009). How stochasticity and emergencies disrupt the surgical schedule, *in* S. McClean, P. Millard, E. El-Darzi & C. Nugent (eds), *Intelligent Patient Management*, Vol. 189 of *Studies in Computational Intelligence*, Springer, pp. 221–239.

8.2 Perspectives de recherches futures

Les travaux exposés dans ce document, du fait des hypothèses émises ou des méthodologies appliquées, comportent autant de limites que d'opportunités de recherches ultérieures.

Les modèles mathématiques en nombres entiers mixtes, qu'ils s'appliquent aux cas de blocs opératoires dédiés ou polyvalents, offrent actuellement des performances satisfaisantes. Notamment, les poids économiques de la fonction objectif associés à l'ouverture des salles d'une part, et à l'utilisation des périodes de temps en heures supplémentaires d'autre part, y sont fixés en accord avec les données dont nous disposons. Il serait cependant intéressant d'analyser la sensibilité des formulations mathématiques aux valeurs que prennent ces poids, en particulier si la minimisation du nombre d'heures supplémentaires nécessaires prévaut. Par ailleurs, afin d'évaluer l'exportabilité de ces approches exactes, dont la résolution par le logiciel CPLEX est efficiente, il serait nécessaire de leur appliquer des solveurs d'optimisation tels que XPRESS OPTIMIZATION ou GUROBI, ou encore GLPK pour une utilisation libre par le plus grand nombre.

L'algorithme génétique, par son codage chromosomique actuel, peut générer des individus soit identiques, perdant alors de l'information, soit irréalisables lors de l'initialisation des populations. Ceci réclame indéniablement plus de générations avant de produire des solutions satisfaisantes, notamment en raison d'une redondance de l'information présente au sein des populations. Cet écueil pourrait toutefois être contourné par une manipulation somme toute assez simple des chromosomes définissant un individu. Il s'agirait de trier les opérations constituant ces chromosomes par jour et par salle d'opération, rendant les individus aisément comparables, et d'appliquer un nombre limité de modifications locales afin d'espérer obtenir un individu différent de ceux déjà existants. Dans le même ordre d'idée, la recherche globale de solutions qu'effectue l'algorithme génétique pourrait être hybridée à une technique de recherche locale, afin de tenter d'améliorer chaque individu nouvellement créé par des variations de certains gènes prises aléatoirement dans un voisinage direct. Il semble en effet que, de nos jours, les approches méta-heuristiques les plus aptes à résoudre les problèmes d'optimisation sont le fruit d'une hybridation entre recherche locale et algorithmes évolutionnaires. Elles combinent les qualités de différentes méta-heuristiques pour proposer généralement des solutions supérieures aux résultats obtenus par les méthodes sur lesquelles elles se basent. En particulier, les algorithmes mémétiques, issus de l'hybridation entre une recherche tabou et un algorithme génétique sont particulièrement en vogue dans le domaine de l'optimisation multi-objectif (pour un état de l'art, voir Knowles & Corn, 2004). Parmi ceux-ci, nous pouvons distinguer les approches suivantes : IMMOGLS (*Ishibuchi Murata MultiObjective Genetic Local Search*, Ishibuchi & Murata, 1996), MOGLS (*MultiObjective Genetic Local Search*, Jaszkiewicz, 1998), PMA (*Pareto Memetic Algorithm*, Jaszkiewicz, 2001) et M-PAES (*Memetic Pareto Archived Evolution Strategy*, Knowles & Corne, 2000). Nous pouvons également pointer les travaux de Lust et Teghem qui proposent deux nouvelles méta-heuristiques particulièrement bien adaptées aux problèmes de *knapsack* multi-dimensionnel et du voyageur de commerce multi-objectif, à savoir MEMOTS (*Memetic MultiObjective Tabu Search*, Lust & Teghem, 2008) et 2PPLS (*Two-Phase Pareto Local Search*, Lust & Teghem, 2009). Ces schémas hybridés constituent sans nul doute l'avenir des méta-heuristiques et dès lors une voie naturelle d'évolution de nos travaux.

Concernant le traitement de la symétrie induite par la multifonctionnalité des blocs opératoires, nous savons que fixer un nombre a priori d'entre eux systématiquement ouverts affecte les perfor-

mances du modèle. Par contre, nous ne connaissons pas l'impact du niveau de gap satisfaisant à partir duquel le solveur doit être arrêté afin de fixer le nombre de salles d'intervention utilisées par jour, avant de relancer le solveur jusqu'à l'optimum. Il serait instructif d'estimer l'éventail de valeurs de cet écart avec la borne inférieure pour lequel la stratégie que nous suggérons reste applicable et continue à produire des solutions quasi-optimales. Ce traitement de la symétrie pourrait, en outre, être étendu au cas de blocs opératoires dédiés. En effet, dans ce contexte, bien qu'une opération ne puisse accéder à l'ensemble des salles d'intervention, elle peut cependant, en général, être réalisée dans un nombre plus restreint de salles, souvent deux ou trois. Le modèle développé dans ce cadre fait donc face à une symétrie partielle, qui pénalise immanquablement la résolution de celui-ci. Ce modèle gagnerait très certainement à se voir également appliquer une technique brisant cette symétrie partielle.

La démarche coopérative que nous prônons est également sujette à approfondissements. Tout d'abord, il nous semble à propos de mesurer l'impact des poids associés à ce nouvel objectif social sur la construction des programmes opératoires et sur l'attribution des plages horaires. Ensuite, la désutilité endurée, que traduisent ces poids, reste identique quel que soit le nombre de préférences du chirurgien déjà non respectées ; en quelque sorte, la désutilité est linéaire en ce nombre. Ceci résulte d'un comportement que nous qualifions d'*indifférent*. Cependant, nous pourrions imaginer des comportements sujets à des effets de *lassitude* ou de *résignation*. Un comportement de lassitude signifierait que le chirurgien est enclin à supporter les premiers inconvénients, mais trop de dérangements provoqueraient la lassitude de ce dernier et sa désutilité augmenterait non linéairement en conséquence. A l'inverse, la résignation représenterait un comportement où le chirurgien serait très affecté par les premiers non-respects de ses préférences, mais où la désutilité diminuerait graduellement en fonction du nombre de préférences ignorées par le planning. Ces deux comportements constituent des perspectives d'amélioration de la modélisation des préférences des chirurgiens. Concernant les mécanismes d'enchères, les mises censées représenter les préférences des chirurgiens ont été ici réparties subjectivement, sur base de leurs journées de présence observées en pratique. Il serait bon d'évaluer le comportement de notre approche exacte lorsque ces mêmes mises sont distribuées aléatoirement, par exemple selon une loi uniforme. A propos du protocole de négociation, nous supposons actuellement que les demandes des chirurgiens sont traitées dans leur ordre d'arrivée ; la négociation ne concerne par conséquent que deux parties. Cependant, nous pourrions étendre notre approche de deux manières : d'une part, considérer comment un chirurgien sollicité pour un échange de plages horaires pourrait mettre en concurrence plusieurs praticiens demandeurs, afin d'obtenir un profit maximum de la négociation qui s'en suit. D'autre part, nous pourrions envisager permettre à un chirurgien demandeur de mettre en compétition plusieurs confrères enclins à céder leur plage horaire contre rétribution, afin pour le premier de minimiser le prix assurant l'accord des deux parties.

Enfin, le processus de Markov permettant d'appréhender la variabilité du milieu chirurgical suppose bon nombre d'hypothèses, notamment sur la distribution exponentielle des durées opératoires. Afin de mieux représenter la réalité, ces distributions exponentielles pourraient être écartées au profit de distributions phase-type plus complexes, permettant une modélisation plus fidèle de la distribution réelle des temps de réalisation des interventions, au prix d'une complexité accrue du processus Markovien. Ceci justifierait davantage l'application de méthodes de résolution analytique (Matrix Analytic Methods) des chaînes de Markov afin d'améliorer les performances calculatoires de notre modèle, autant en l'état qu'étendu en appliquant des distributions phase-type des durées opératoires. Dans le même registre, les durées de traitement des opérations électives suivent

ici une distribution exponentielle de même paramètre. Bien que la variabilité de cette distribution peut modéliser les différences entre les diverses chirurgies, nous pourrions toutefois envisager de différencier les salles polyvalentes sur base de temps opératoires disparates. Une autre perspective qui nous semble pertinente serait d'évaluer dans quelle mesure notre approche pourrait être liée à une méthodologie de programmation plus détaillée, pour atteindre un niveau opérationnel de décision. L'idée sous-jacente serait de pouvoir greffer un planning tel que dressé par le gestionnaire et, appliquant un outil proche du nôtre, évaluer la robustesse de ce planning en le confrontant aux sources stochastiques qui caractérisent le quartier opératoire, identifiées par l'utilisateur.

Autant de pistes, autant de questions ouvertes, autant de sujets de recherche... La programmation du quartier opératoire n'a pas cessé de faire couler beaucoup d'encre !

Bibliographie

Arenas, M., Bilbao, A., Caballero, R., Gomez, T., Rodriguez, M. & Ruiz, F. (2002). Analysis via goal programming of the minimum achievable stay in surgical waiting lists, *Journal of the Operational Research Society* **53** : 387–396.

Asmussen, S. (2003). *Applied Probability and Queues*, Springer-Verlag, New York.

Augusto, V., Xie, X. & Perdomo, V. (2008). Operating theatre scheduling using lagrangian relaxation, *European Journal of Industrial Engineering* **2**(2) : 172–189.

Augusto, V., Xie, X. & Perdomo, V. (2010). Operating theatre scheduling with patient recovery in both operating rooms and recovery beds, *Computers & Industrial Engineering* **58** : 231–238.

Baubeau, D. & Thomson, E. (2002). Les plateaux techniques liés aux interventions sous anesthésie entre 1992 et 2000, *Etudes et Résultats* **189**.

Beliën, J. & Demeulemeester, E. (2007). Building cyclic master surgery schedules with leveled resulting bed occupancy, *European Journal of Operational Research* **176** : 1185–1204.

Beliën, J. & Demeulemeester, E. (2008). A branch-and-price approach for integrating nurse and surgery scheduling, *European Journal of Operational Research* **189** : 652–668.

Beliën, J., Demeulemeester, E. & Cardoen, B. (2006). Visualizing the demand for various resources as a function of the master surgery schedule : A case study, *Journal of Medical Systems* **30**(5) : 343–350.

Blake, J. & Carter, M. (2002). A goal programming approach to strategic resource allocation in acute care hospitals, *European Journal of Operational Research* **140** : 541–561.

Blake, J., Dexter, F. & Donald, J. (2002). Operating room managers' use of integer programming for assigning block time to surgical groups : A case study, *Anesthesia and Analgesia* **94** : 143–148.

Blake, J. & Donald, J. (2002). Mount Sinai hospital uses integer programming to allocate operating room time, *Interfaces* **32** : 63–73.

Bowers, J. & Mould, G. (2004). Managing uncertainty in orthopaedic trauma theatres, *European Journal of Operational Research* **154**.

Bowers, J. & Mould, G. (2005). Ambulatory care and orthopaedic capacity planning, *Health Care Management Science* **28**(8) : 41–47.

Brucker, P., Drexl, A., Möhring, R., Neumann, K. & Pesch, E. (1999). Resource-constrained project scheduling : Notation, classification, models, and methods, *European Journal of Operational Research* **112** : 3–41.

Calichman, M. (2005). Creating an optimal operating room schedule, *AORN Journal* **81** : 580–588.

Cardoen, B., Demeulemeester, E. & Beliën, J. (2010). Operating room planning and scheduling : A literature review, *European Journal of Operational Research* **201** : 921–932.

Cardoen, B., Demeulemeester, E. & Beliën, J. (2006a). Optimizing a multiple objective surgical case scheduling problem, *Technical report*, Katholieke Universiteit Leuven, Belgium.

Cardoen, B., Demeulemeester, E. & Beliën, J. (2006b). Scheduling surgical cases in a day-care environment : A branch-and-price approach, *Technical report*, Katholieke Universiteit Leuven, Belgium.

Cayirli, T. & Veral, E. (2003). Outpatient scheduling in health care : A review of literature, *Production & Operations Management* **12**(4) : 519–549.

Chaabane, S. (2004). *Gestion prédictive des blocs opératoires*, PhD thesis, Institut National des Sciences Appliquées de Lyon.

Chaabane, S., Meskens, N., Guinet, A. & Laurent, M. (2006). Comparison of two methods of operating theatre planning : Application in belgian hospitals, International Conference on Service Systems and Service Management.

Chaabane, S., Meskens, N., Guinet, A. & Laurent, M. (2007). Comparaison des performances des politiques de programmation opératoire, *Logistique et Management* **15** : 17–26.

Colorni, A., Dorigo, M. & Maniezzo, V. (1992). Distributed optimization by ant colonies, ECAL'91 - First European Conference on Artificial Life.

Creemers, S. & Lambrecht, M. (2007). Modeling a healthcare system as a queueing network : The case of a belgian hospital, *Technical report*, K.U.Leuven.

Czap, H., Becker, M., Poppensieker, M. & Stotz, A. (2005). Utility-based agents for hospital scheduling : A conjoint-analytical approach, *AAMAS Workshop*, Utrecht.

de Bruin, A. M., van Rossum, A. C., Visser, M. C. & Koole, G. M. (2007). Modeling the emergency cardiac in-patient flow : an application of queuing theory, *Health Care Management Science* **10 (2)** : 125–137.

De Grano, M., Medeiros, D. & Eitel, D. (2009). Accomodating individual preferences in nurse scheduling via auctions and optimization, *Health Care Management Science* **12** : 228–242.

Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T. (2002). A fast elitist multi-objective genetic algorithm : NSGA-II, *IEEE Transactions on Evolutionary Computation* **6** : 182–197.

Demeulemeester, E. & Herroelen, W. (2002). *Project Scheduling : A Research Handbook*, Kluwer Academic Publishers.

Denton, B., Viapiano, J. & Vogl, A. (2007). Optimization of surgery sequencing and scheduling decisions under uncertainty, *Health Care Management Science* **10** : 13–24.

Dexter, F. (2000). A strategy to decide whether to move the last case of the day in an operating room to another empty operating room to decrease overtime labor costs, *Anesthesia and Analgesia* **91** : 925–928.

Dexter, F., Blake, J., Penning, D., Sloan, B., Chung, P. & Lubarsky, D. (2002). Use of linear programming to estimate impact of changes in a hospital's operating room time allocation on perioperative variable costs, *Anesthesiology* **96** : 718–724.

Dexter, F. & Ledolter, J. (2003). Managing risk and expected financial return from selective expansion of operating room capacity : Mean-variance analysis of a hospital's portfolio of surgeons, *Anesthesia and Analgesia* **97** : 190–195.

Dexter, F., Ledolter, J. & Wachtel, R. (2005). Tactical decision making for selective expansion of operating room resources incorporating financial criteria and uncertainty in subspecialties' future workloads, *Anesthesia and Analgesia* **100** : 1425–1432.

Dexter, F., Macario, A. & Lubarsky, D. (2001). The impact on revenue of increasing patient volume at surgical suites with relatively high operating room utilization, *Anesthesia and Analgesia* **92** : 1215–1221.

Dexter, F., Macario, A. & O'Neill, L. (2000). Scheduling surgical cases into overflow block time - computer simulation of the effects of scheduling strategies on operating room labor costs, *Anesthesia and Analgesia* **90** : 980–988.

Dexter, F., Macario, A., Traub, R. & Lubarsky, D. (2003). Operating room utilization alone is not an accurate metric for the allocation of operating room block time to individual surgeons with low caseloads, *Anesthesiology* **98**(5) : 1243–1249.

Dexter, F. & O'Neill, L. (2001). Weekend operating room on-call staffing requirements, *Association of periOperative Registered Nurses Journal (AORN J)* **74** : 666–671.

Dexter, F. & Tinker, J. H. (1995). Analysis of strategies to decrease postanesthesia care unit costs, *Anesthesiology* **82(1)** : 94–1001.

Dexter, F. & Traub, R. (2002). How to schedule elective surgical cases into specific operating rooms to maximize the efficiency of use of operating room time, *Anesthesia and Analgesia* **94** : 933–942.

Dréo, J., Pétrowski, A., Siarry, P. & Taillard, E. (2003). *Métaheuristiques pour l'optimisation difficile*, Eyrolles.

Everett, J. (2002). A decision support simulation model for the management of an elective surgery waiting list, *HCMS* **5** : 89–95.

Fackrell, M. (2007). Modelling healthcare systems with phase-type distributions, *Health Care Management Science* **12 (1)** : 11–26.

Fei, H. (2006). *Vers un outil d'aide à la planification et à l'ordonnancement des blocs opératoires*, PhD thesis, Facultés Universitaires Catholiques de Mons.

Fei, H., Chu, C., Meskens, N. & Artiba, A. (2008). Solving surgical cases assignment problem by a branch-and-price approach, *International Journal of Production Economics* **112** : 96–108.

Fei, H., Meskens, N. & Chu, C. (2006). An operating theatre planning and scheduling problem in the case of a block scheduling strategy, International Conference on Service Systems and Service Management.

Fei, H., Meskens, N. & Chu, C. (2007). An operating theatre planning and scheduling problem in the case of an open scheduling strategy, International Conference on Industrial Engineering and Systems Management.

Gerchak, Y., Gupta, D. & Henig, M. (1996). Reservation planning for elective surgery under uncertain demand for emergency surgery, *Management Science* **42** : 321–334.

Glover, F. & Laguna, M. (1997). *Tabu search*, Kluwer Academic Publishers.

Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley.

Golumbic, M. (1980). *Algorithmic Graph Theory and Perfect Graphs*, Computer Science and Applied Mathematics, Academic Press.

Gordon, T., Lyles, A. & Fountain, J. (1988). Surgical unit time review : resource utilization and management implications, *Journal of Medical Systems* **12** : 169–179.

Gorunescu, F., McClean, S. I. & Millard, P. H. (2004). Using a queueing model to help plan bed allocation in a department of geriatric medicine, *Health Care Management Science* **5** : 307–312.

Green, L. (2006). Queueing analysis in health care, *in* R. Hall (ed.), *Patient Flow : Reducing Delay in Healthcare Delivery*, Springer US, chapter 10, pp. 281–307.

Guinet, A. (2001). A linear programming approach to define the operating room opening hours, *International Conference on Integrated Design and Production*.

Guinet, A. & Chaabane, S. (2003). Operating theatre planning, *International Journal of Production Economics* **85** : 69–81.

Hall, R. W. (ed.) (2006). *Patient Flow : Reducing Delay in Healthcare Delivery*, International Series in Operations Research & Management Science, Springer US.

Hammami, S. (2006). *Aide à la décision dans le pilotage des flux matériels et patients d'un plateau d'un plateau médico-technique*, PhD thesis, Institut National Polytechnique de Grenoble.

Hammami, S., Jebali, A., Ladet, P. & Hadj Alouane, A. (2003). Approche multi-objectifs pour l'introduction de l'urgence dans le programme opératoire, 5ème Congrès International de Génie Industriel (CIGI'03), Canada.

Hans, E., Wullink, G., VanHoudenhoven, M. & Kazemier, G. (2008). Robust surgery loading, *European Journal of Operational Research* **185** : 1038–1050.

Harper, P. (2002). A framework for operational modelling of hospital resources, *Health Care Management Science* **5** : 165–173.

Harrison, G. W., Shafer, A. & Mackay, M. (2005). Modelling variability in hospital bed occupancy, *Health Care Management Science* pp. 325–334.

Hartmann, S. (2001). Project scheduling with multiple modes : a genetic algorithm, *Annals of Operations Research* **102** : 111–135.

Hsu, V., de Matta, R. & Lee, C.-Y. (2003). Scheduling patients in an ambulatory surgical center, *Naval Research Logistics* **50** : 218–238.

Ishibuchi, H. & Murata, T. (1996). Multi-objective genetic local search algorithm, IEEE International Conference on Evolutionary Computation, Nagoya, Japan.

Jaszkiewicz, A. (1998). Genetic local search for multiple objective combinatorial optimization, *Technical report*, Institute of Computing Science, Poznan University of Technology.

Jaszkiewicz, A. (2001). A comparative study of multiple-objective metaheuristics on the biobjective set covering problem and the pareto memetic algorithm, *Technical report*, Institute of Computing Science, Poznan University of Technology.

Jebali, A., Hadj Alouane, A. & Ladet, P. (2006). Operating rooms scheduling, *International Journal of Production Economics* **99** : 52–62.

Kao, E. & Tung, G. (1981). Bed allocation in a public health care delivery system, *Management Science* **27** : 507–520.

Kharraja, S. (2003). *Outils d'aide à la planification et l'ordonnancement des plateaux médico-techniques*, PhD thesis, Université Jean Monnet.

Kharraja, S., Albert, P. & Chaabane, S. (2006). Block scheduling : Toward a master surgical schedule, International Conference on Service Systems and Service Management.

Kim, S.-C. & Horowitz, I. (2002). Scheduling hospital services : the efficacy of elective-surgery quotas, *Omega - The International Journal of Management Science* **30** : 335–346.

Kim, S.-C., Horowitz, I., Young, K. & Buckley, T. (1999). Analysis of capacity management of the intensive care unit in a hospital, *European Journal of Operational Research* **115** : 36–46.

Kirkpatrick, S., Gelatt, C. & Vecchi, M. (1983). Optimization by simulated annealing, *Science* **4398** : 671–680.

Klein, R. (2000). *Scheduling of resource constrained projects*, Kluwer Academic Publishers.

Klemperer, P. (1999). Auction theory : A guide to the litterature, *Journal of Economic Surveys* **13** : 227–286.

Knowles, J. & Corn, D. (2004). Memetic algorithms for multiobjective optimization : issues, methods and prospects, *in* N. Krasnogor, J. Smith & W. Hart (eds), *Recent Advances in Memetic Algorithms*, Springer, p. 313–352.

Knowles, J. & Corne, D. (2000). M-PAES : A memetic algorithm for multiobjective optimization, Congress on Evolutionary Computation, New Jersey, USA.

Kolisch, R. & Hartmann, S. (2005). Experimental investigation of heuristics for resource-constrained project scheduling : An update, *European Journal of Operational Research* .

Kolker, A. (2008). Process modeling of emergency department patient flow : Effect of patient length of stay on ED diversion, *Journal of Medical Systems* **32(5)** : 389–401.

Komashie, A. & Mousavi, A. (2005). Modeling emergency departments using discrete event simulation techniques, *Proceedings of the Winter Simulation Conference*.

Krempels, K.-H. & Panchenko, A. (2006). An approach for automated surgery scheduling, 6th International Conference on the Practice and Theory of Automated Timetabling.

Krishna, V. (2002). *Auction Theory*, Academic Press.

Lamiri, M., Xie, X., Dolgui, A. & Grimaud, F. (2008). A stochastic model for operating room planning with elective and emergency demand for surgery, *European Journal of Operational Research* **185** : 1026–1037.

Landry, S. & Beaulieu, M. (2002). Logistique hospitalière : un remède aux maux du secteur de la santé ?, *Gestion* **26** : 34–41.

Latouche, G. & Ramaswamy, V. (1999). *Introduction to Matrix Analytic Methods in Stochastic Model*, ASA/SIAM Series on Statistics and Applied Probability.

Lebowitz, P. (2003). Schedule the short procedure first to improve OR efficiency, *AORN Journal* **78**(4) : 651–659.

Lovejoy, W. & Li, Y. (2002). Hospital operating room capacity expansion, *Management Science* **48**(11) : 1369–1387.

Lust, T. & Teghem, J. (2008). MEMOTS : a memetic algorithm integrating tabu search for combinatorial multiobjective optimization, *RAIRO Oper. Res.* **42** : 3–33.

Lust, T. & Teghem, J. (2009). Two-phase pareto local search for the biobjective traveling salesman problem, *Journal of Heuristics* . accepted, in press.

Macario, A., Vitez, T., Dunn, B. & McDonald, T. (1995). Where are the costs in perioperative care ? analysis of hospital costs and charges for inpatient surgical care, *Anesthesiology* **83** : 1138–1144.

Magerlein, J. & Martin, J. (1978). Surgical demand scheduling : A review, *Health Services Research* **13** : 418–433.

Marcon, E. (2004). Dimensionnement des ressources des plateaux médico-techniques des établissements hospitaliers, *Journal Européen des Systèmes Automatisés RS-JESA, Logistique Hospitalière* **38** : 631–656.

Marcon, E. & Dexter, F. (2006). Impact of surgical sequencing on post anesthesia care unit staffing, *Health Care Management Science* **9** : 87–98.

Marcon, E. & Dexter, F. (2007). An observational study of surgeons' sequencing of cases and its impact on postanesthesia care unit and holding area staffing requirements at hospitals, *Anesthesia and Analgesia* **105** : 119–126.

Marcon, E., Kharraja, S. & Simonet, G. (2003). The operating theatre planning by the follow-up of the risk of no realization, *International Journal of Production Economics* **85** : 83–90.

Margot, F. (2008). Symmetry in integer linear programming, *Technical report*, Tepper School of Business.

McClean, S., Faddy, M. & Millard, P. (2006). Using markov models to assess the performance of a health and community care system, *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*.

McClean, S., Millard, P., El-Darzi, E. & Nugent, C. (2009). *Intelligent Patient Management*, Vol. 189 of *Studies in Computational Intelligence*, Springer Berlin / Heidelberg.

MEAH (2006). Gestion et organisation des blocs opératoires dans les hôpitaux et cliniques : Recueil des bonnes pratiques organisationnelles observées, *Technical report*, Mission nationale d'Expertise et d'Audit Hospitalier, Paris.

Michalewicz, Z. (1998). *Genetic Algorithms + Data Structure = Evolution Programs*, Springer.

Moisdon, J. & Tonneau, D. (1999). *La démarche gestionnaire à l'hôpital*, Selan Arslan Edition.

Nemhauser, G. & Wolsey, L. (1999). *Integer and Combinatorial Optimization*, Wiley-Interscience.

Neuts, M. (1981). *Matrix-Geometric Solutions in Stochastic Models : an Algorithmic Approach*, Dover Publications, New York.

Ogulata, S. & Erol, R. (2003). A hierarchical multiple criteria mathematical programming approach for scheduling general surgery operations in large hospitals, *Journal of Medical Systems* **27**(3) : 259–270.

Osborne, M. (2004). *an introduction to Game Theory*, Oxford University Press.

Ozkarahan, I. (2000). Allocation of surgeries to operating rooms by goal programming, *Journal of Medical Systems* **24** : 339–378.

Pham, D.-N. & Klinkert, A. (2008). Surgical case scheduling as a generalized job shop scheduling problem, *European Journal of Operational Research* **185** : 1011–1025.

Raghuwanshi, M. & Kakde, O. (2004). Survey on multiobjective evolutionary and real coded genetic algorithms, 8th Asia Pacific Symposium on Intelligent and Evolutionary Systems.

Rakotondranaivo, A. (2006). *Contribution de la modélisation à l'évolution des performances des organisations de santé*, PhD thesis, Institut National Polytechnique de Lorraine.

Rohleder, T., Sabapathy, D. & Schorn, R. (2005). An operating room block allocation model to improve hospital patient flow, *Clinical and Investigative Medicine* **28**(6) : 353–355.

Roland, B., Cordier, J.-P., Tancrez, J.-S. & Riane, F. (2009). Recovery beds and blocking in stochastic operating theatres, International Conference on Industrial Engineering and Systems Management, IESM'09, Montreal, Canada.

Roland, B., DiMartinelly, C. & Riane, F. (2006). Operating theatre optimization : A resource-constrained based solving approach, International Conference on Service Systems and Service Management, IEEE/ICSSSM'06, Troyes, France.

Roland, B., DiMartinelly, C. & Riane, F. (2007). On the integration of resources' preferences in the operating theatre optimization, Annual conference of the Belgian Operations Research society, ORBEL'21, Luxembourg.

Roland, B., DiMartinelly, C. & Riane, F. (2008). Planification du quartier opératoire intégrant les préférences des ressources humaines, Conférence Internationale de Modélisation et Simulation, MOSIM'08, Paris, France.

Roland, B., DiMartinelly, C., Riane, F. & Pochet, Y. (2007). Scheduling operating theatre under human resources constraints, International Conference on Industrial Engineering and Systems Management, IESM'07, Beijing, China.

Roland, B., DiMartinelly, C., Riane, F. & Pochet, Y. (2010). Scheduling an operating theatre under human resource constraints, *Computers & Industrial Engineering* **58** : 212–220.

Roland, B. & Riane, F. (2008). A constrained NSGA-II to consider surgeons' preferences while scheduling the operating theatre, International Conference on Information Systems, Logistics and Supply Chain, ILS'08, Madison (WI), USA.

Roland, B. & Riane, F. (2009). Symétrie et préférences des ressources dans la planification des blocs opératoires, Congrès de la société française de Recherche Opérationnelle et d'Aide à la Décision, ROADEF'09, Nancy, France.

Roland, B. & Riane, F. (2010a). Enchères et coopération dans la programmation opératoire, Conférence Internationale de Modélisation et Simulation, MOSIM'10, Hammamet, Tunisia.

Roland, B. & Riane, F. (2010b). Integrating surgeons' preferences in the operating theatre planning, *European Journal of Industrial Engineering* . Accepted, to appear.

Roland, B. & Riane, F. (2010c). When OR techniques help ORs management, International Conference on Information Systems, Logistics and Supply Chain, ILS'10, Casablanca, Morocco. *Professor Salah E. Elmaghraby's Honorable Mention award*.

Ruohonen, T., Neittaanmäki, P. & Teittinen, J. (2006). Simulation model for improving the operation of the emergency department of special health care, *Proceedings of the 38th conference on Winter simulation*, California, pp. 453–458.

Santibanez, P., Begen, M. & Atkins, D. (2007). Surgical block scheduling in a system of hospitals : An application to resource and wait list management in a British Columbia health authority, *Health Care Management Science* **10** : 269–282.

Saporta, G. (2006). *Probabilités, analyse des données et statistique*, second edn, Technip, Paris.

Sciomachen, A., Tanfani, E. & Testi, A. (2005). Simulation models for optimal schedules of operating theatres, *International Journal of Simulation* **6**(12-13) : 36–34.

Shao, J. (2003). *Mathematical Statistics*, second edn, Springer Science, New York.

Stewart, W. (1994). *Introduction to the Numerical Solution of Markov Chains*, Princeton University Press, Princeton, New Jersey.

Stewart, W. (2000). Numerical analysis methods, *in* G. Haring, C. Lindemann & M. Reise (eds), *Performance Evaluation : Origins and Directions*, Vol. 1769 of *Lecture Notes In Computer Science*, Springer-Verlag, pp. 355–376.

Tan, Y., ElMekkawy, T., Peng, Q. & Oppenheimer, L. (2007). Mathematical programming for the scheduling of elective patients in the operating room department, 2007 CDEN and CCEE Conference.

Tancrez, J.-S., Roland, B., Cordier, J.-P. & Riane, F. (2008). Etude de la perturbation par les urgences du planning opératoire, Conférence de Gestion et Ingénierie des SystèmEs Hospitaliers, GISEH'08, Lausanne, Suisse. *Meilleure communication d'étudiant-Docteur*. Egalement publié dans *Logistique & Management* (à paraître).

Tancrez, J.-S., Roland, B., Cordier, J.-P. & Riane, F. (2009). How stochasticity and emergencies disrupt the surgical schedule, *in* S. McClean, P. Millard, E. El-Darzi & C. Nugent (eds), *Intelligent Patient Management*, Vol. 189 of *Studies in Computational Intelligence*, Springer, pp. 221–239.

Testi, A., Tanfani, E. & Torre, G. (2007). A three-phase approach for operating theatre schedules, *Health Care Management Science* **10** : 163–172.

Trilling, L. (2006). *Aide à la décision pour le dimensionnement et le pilotage de ressources humaines mutualisées en milieu hospitalier*, PhD thesis, Institut National des Sciences Appliquées de Lyon.

van Oostrum, J., Van Houdenhoven, M., Hurink, J., Hans, E., Wullink, G. & Kazemier, G. (2008). A master surgical scheduling approach for cyclic scheduling in operating room departments, *OR Spectrum* **30** : 355–374.

VanBerkel, P. & Blake, J. (2007). A comprehensive simulation for wait time reduction and capacity planning applied in general surgery, *Health Care Management Science* **10** : 373–385.

Velasquez, R. & Melo, M. (2006). A set packing approach for scheduling elective surgical procedures, *Operations Research Proceedings* .

Wang, T., Guinet, A. & Besombes, B. (2008). Simulation modeling of emergency service with the impact of inpatient bed resource, International Conference on Information Systems, Logistics and Supply Chain (ILS'08), Madison, WI, USA.

Wolsey, L. (1998). *Integer Programming*, Wiley-Interscience.

Wooldridge, M. (2002). *An Introduction to MultiAgent Systems*, John Wiley & Sons.

Wullink, G., Van Houdenhoven, M., Hans, E., van Oostrum, J., van der Lans, M. & Kazemier, G. (2007). Closing emergency operating rooms improves efficiency, *Journal of Medical Systems* **31** : 543–546.