Planification des blocs opératoires avec prise en compte des aléas

Mehdi Lamiri

Ecole Nationale Supérieure des Mines
SAINT-ETIENNE

N° d'ordre : 446 GI

THÈSE

présentée par

Mehdi Lamiri

Pour obtenir le grade de Docteur
de l'École Nationale Supérieure des Mines de Saint-Étienne

Spécialité : Génie Industriel

Planification des blocs opératoires avec prise en compte des aléas

Soutenue à Saint-Étienne le 4 octobre 2007

Membres du jury

Yves Dallery	Professeur, École Centrale Paris	Président
Alain Guinet	Professeur, INSA de Lyon	Rapporteur
Michel Gourgand	Professeur, Université Blaise Pascal	Rapporteur
Nadine Meskens	Professeur, Facultés Universitaires Catholiques de Mons	Examinateur
Xiaolan Xie	Professeur, École des Mines de Saint-Étienne	Co-directeur
Alexandre Dolgui	Professeur, École des Mines de Saint-Étienne	Co-directeur
Frédéric Grimaud	Maître-assistant, École des Mines de Saint-Étienne	Co-encadrant

● Spécialités doctorales : **Responsables :**

SCIENCES ET GENIE DES MATERIAUX	**J. DRIVER** Directeur de recherche – Centre SMS
MECANIQUE ET INGENIERIE	**A. VAUTRIN** Professeur – Centre SMS
GENIE DES PROCEDES	**G. THOMAS** Professeur – Centre SPIN
SCIENCES DE LA TERRE	**B. GUY** Maitre de recherche – Centre SPIN
SCIENCES ET GENIE DE L'ENVIRONNEMENT	**J. BOURGOIS** Professeur – Centre SITE
MATHEMATIQUES APPLIQUEES	**E. TOUBOUL** Ingénieur – Centre G2I
INFORMATIQUE	**O. BOISSIER** Professeur – Centre G2I
IMAGE, VISION, SIGNAL	**JC. PINOLI** Professeur – Centre CIS
GENIE INDUSTRIEL	**P. BURLAT** Professeur – Centre G2I
MICROELECTRONIQUE	**Ph. COLLOT** Professeur – Centre CMP

● Enseignants-chercheurs et chercheurs autorisés à diriger des thèses de doctorat (titulaires d'un doctorat d'Etat ou d'une HDR)

BATTON-HUBERT	Mireille	MA	Sciences & Génie de l'Environnement	SITE
BENABEN	Patrick	PR 2	Sciences & Génie des Matériaux	SMS
BERNACHE-ASSOLANT	Didier	PR 1	Génie des Procédés	CIS
BIGOT	Jean-Pierre	MR	Génie des Procédés	SPIN
BILAL	Essaïd	DR	Sciences de la Terre	SPIN
BOISSIER	Olivier	PR 2	Informatique	G2I
BOUDAREL	Marie-Reine	MA	Sciences de l'inform. & com.	DF
BOURGOIS	Jacques	PR 1	Sciences & Génie de l'Environnement	SITE
BRODHAG	Christian	MR	Sciences & Génie de l'Environnement	SITE
BURLAT	Patrick	PR 2	Génie industriel	G2I
CARRARO	Laurent	PR 1	Mathématiques Appliquées	G2I
COLLOT	Philippe	PR 1	Microélectronique	CMP
COURNIL	Michel	PR 1	Génie des Procédés	SPIN
DAUZERE-PERES	Stéphane	PR 1	Génie industriel	CMP
DARRIEULAT	Michel	ICM	Sciences & Génie des Matériaux	SMS
DECHOMETS	Roland	PR 1	Sciences & Génie de l'Environnement	SITE
DESRAYAUD	Christophe	MA	Mécanique & Ingénierie	SMS
DELAFOSSE	David	PR 2	Sciences & Génie des Matériaux	SMS
DOLGUI	Alexandre	PR 1	Génie Industriel	G2I
DRAPIER	Sylvain	PR 2	Mécanique & Ingénierie	CIS
DRIVER	Julian	DR	Sciences & Génie des Matériaux	SMS
FOREST	Bernard	PR 1	Sciences & Génie des Matériaux	CIS
FORMISYN	Pascal	PR 1	Sciences & Génie de l'Environnement	SITE
FORTUNIER	Roland	PR 1	Sciences & Génie des Matériaux	CMP
FRACZKIEWICZ	Anna	MR	Sciences & Génie des Matériaux	SMS
GARCIA	Daniel	CR	Génie des Procédés	SPIN
GIRARDOT	Jean-Jacques	MR	Informatique	G2I
GOEURIOT	Dominique	MR	Sciences & Génie des Matériaux	SMS
GOEURIOT	Patrice	MR	Sciences & Génie des Matériaux	SMS
GRAILLOT	Didier	DR	Sciences & Génie de l'Environnement	SITE
GROSSEAU	Philippe	MR	Génie des Procédés	SPIN
GRUY	Frédéric	MR	Génie des Procédés	SPIN
GUILHOT	Bernard	DR	Génie des Procédés	CIS
GUY	Bernard	MR	Sciences de la Terre	SPIN
GUYONNET	René	DR	Génie des Procédés	SPIN
HERRI	Jean-Michel	PR 2	Génie des Procédés	SPIN
KLÖCKER	Helmut	MR	Sciences & Génie des Matériaux	SMS
LAFOREST	Valérie	CR	Sciences & Génie de l'Environnement	SITE
LE COZE	Jean	PR 1	Sciences & Génie des Matériaux	SMS
LI	Jean-Michel	EC (CCI MP)	Microélectronique	CMP
LONDICHE	Henry	MR	Sciences & Génie de l'Environnement	SITE
MOLIMARD	Jérôme	MA	Sciences & Génie des Matériaux	SMS
MONTHEILLET	Frank	DR 1 CNRS	Sciences & Génie des Matériaux	SMS
PERIER-CAMBY	Laurent	PR1	Génie des Procédés	SPIN
PIJOLAT	Christophe	PR 1	Génie des Procédés	SPIN
PIJOLAT	Michèle	PR 1	Génie des Procédés	SPIN
PINOLI	Jean-Charles	PR 1	Image, Vision, Signal	CIS
STOLARZ	Jacques	CR	Sciences & Génie des Matériaux	SMS
SZAFNICKI	Konrad	CR	Sciences de la Terre	SITE
THOMAS	Gérard	PR 1	Génie des Procédés	SPIN
VALDIVIESO	Françoise	CR	Génie des Procédés	SPIN
VALDIVIESO	François	MA	Sciences & Génie des Matériaux	SMS
VAUTRIN	Alain	PR 1	Mécanique & Ingénierie	SMS
VIRICELLE	Jean-Paul	MR	Génie des procédés	SPIN
WOLSKI	Krzysztof	CR	Sciences & Génie des Matériaux	SMS
XIE	Xiaolan	PR 1	Génie industriel	CIS

Glossaire :

PR 1	Professeur 1ère catégorie
PR 2	Professeur 2ème catégorie
MA(MDC)	Maître assistant
DR (DR1)	Directeur de recherche
Ing.	Ingénieur
MR(DR2)	Maître de recherche
CR	Chargé de recherche
EC	Enseignant-chercheur
ICM	Ingénieur en chef des mines

Centres :

SMS Sciences des Matériaux et des Structures
SPIN Sciences des Processus Industriels et Naturels
SITE Sciences Information et Technologies pour l'Environnement
G2I Génie Industriel et Informatique
CMP Centre de Microélectronique de Provence
CIS Centre Ingénierie et Santé

Remerciements

Je tiens, tout d’abord, à exprimer ma profonde gratitude à mes directeurs de thèse : Xiaolan XIE, Alexandre DOLGUI et Frédéric GRIMAUD. Leurs conseils, leur confiance et leurs qualités humaines ont largement contribué à l’aboutissement de ce travail. Je remercie tout particulièrement Xiaolan XIE pour sa disponibilité, son encouragement et l’intérêt qu’il a manifesté pour cette thèse.

Je remercie vivement Messieurs Alain GUINET, Professeur à l’INSA de Lyon, et Michel GOURGAND, Professeur à l’Université Blaise Pascal, pour avoir accepté d’étudier mes travaux et en être les rapporteurs.

Je suis extrêmement reconnaissant à Madame Nadine MESKENS, Professeur à la Faculté Universitaire Catholique de Mons, et Monsieur Yves DALLERY, Professeur à l’École Centrale Paris, d’avoir examiner mes travaux et de participer au jury de ma thèse.

Je remercie également les membres du Centre Génie Industriel et Informatique et du Centre Ingénierie et Santé que j’ai côtoyés durant ces trois dernières années. Je les remercie pour leur accueil, leur soutient, et pour avoir contribué à créer un cadre de travail aussi agréable. Un merci particulier pour Marie Line, Liliane et Gabrielle pour leur sympathie et leur gentillesse.

Enfin, ma gratitude et mes remerciements s’adressent à ma famille qui m’a toujours encouragé et soutenu dans les moments difficiles.

Table des matières

Liste des tableaux

Liste des figures

Introduction

Dans un contexte marqué par la rationalisation des coûts et la maîtrise des dépenses, les établissements hospitaliers doivent aujourd'hui relever un défi majeur lié à leur indispensable évolution. La production des soins longtemps centrée sur l'excellence médicale doit dorénavant concilier l'excellence médicale et l'excellence managériale. Face à ce défi, les gestionnaires des hôpitaux doivent assurer une utilisation efficace et plus rationnelle des ressources (humaines et matérielles) disponibles tout en assurant une bonne qualité de service vis-à-vis des patients.

Le bloc opératoire constitue l'un des secteurs les plus coûteux et les plus importants dans un établissement hospitalier. L'optimisation de son fonctionnement est alors l'une des premières préoccupations aussi bien des gestionnaires que des acteurs y exerçant leurs activités. En plus, étant donné les nombreux aléas, le nombre élevé d'acteurs humains, la difficulté de standardisation et de coordination des interventions chirurgicales, ce secteur est également le plus complexe à gérer et à planifier.

Le problème de planification du bloc opératoire a fait l'objet de nombreuses recherches ces dernières années donnant ainsi naissance à une littérature abondante sur le sujet. Cette littérature montre une utilisation prépondérante des approches déterministes qui font abstraction de toutes sortes d'aléas. Or, le bloc opératoire est sujet à nombreuses formes d'aléas qui concernent essentiellement la chirurgie d'urgence et les durées d'interventions, et qui peuvent perturber son fonctionnement. Ainsi, les objectifs de la planification ne sont pas atteints dès lors que ces aléas ne sont pas pris en compte.

À ce titre, l'objectif de notre travail de recherche est de développer des modèles et méthodes pour la planification des activités chirurgicales tout en tenant en compte des aléas importants.

Au terme de nos travaux de thèse, nous présentons dans ce mémoire les études réalisées et les principaux résultats obtenus. Le mémoire est structuré comme suit :

Dans le premier chapitre, nous décrivons le fonctionnement du bloc opératoire, les différentes ressources qu'il mobilise et les différents modes de gestion utilisés. Nous mettons aussi l'accent sur les différentes formes d'aléas qui caractérisent l'environnement du bloc

opératoire, et leur impact sur le bon fonctionnement de ce dernier. En particulier, nous faisons la distinction entre deux types d'activités : une activité programmée qu'on peut planifier à l'avance (*les interventions électives*) et une activité d'urgence non planifiable (*les interventions urgentes*).

Le deuxième chapitre dresse un état de l'art des méthodes et approches développées pour appréhender les principales problématiques soulevées par la gestion des blocs opératoires : dimensionnement des ressources, planification des activités chirurgicales, ordonnancement des activités chirurgicales, gestion des ressources humaines, et évaluation de la performance.

Le troisième chapitre présente un modèle stochastique pour la planification du bloc opératoire ; un modèle qui intègre l'aspect incertain de la chirurgie d'urgence. Par contre, les durées des interventions programmées sont déterministes. Dans ce modèle de planification, nous nous intéressons à la détermination des dates d'intervention des patients électifs. L'affectation des patients à des salles opératoires spécifiques n'est pas prise en compte. Les salles opératoires sont supposées polyvalentes et seule la capacité globale de l'ensemble des salles opératoires est prise en compte. Nous formulons le problème sous la forme d'un programme mathématique stochastique et nous proposons deux méthodes de résolution : une méthode « exacte » qui combine la simulation Monte Carlo et la programmation en nombres mixtes, et une méthode heuristique basée sur la relaxation lagrangienne. Les performances de celles-ci sont évaluées et comparées.

Dans le quatrième chapitre, nous proposons des extensions au modèle de planification pour permettre une affectation des patients aux salles opératoires, et la prise en compte d'une structure coût plus élaborée pour ces dernières. Nous proposons aussi une approche de résolution basée sur la génération de colonnes. Nous testons l'efficacité et les performances de différentes stratégies de génération de colonnes.

Le cinquième chapitre propose une extension supplémentaire au modèle de planification pour tenir compte non seulement des incertitudes dues à la chirurgie d'urgence mais aussi aux durées aléatoires des interventions électives. A cause de cette nouvelle extension, l'approche de génération de colonnes présentée au chapitre 4 n'est pas applicable. Une nouvelle approche combinant la simulation Monte Carlo et la génération de colonnes est présentée. Afin d'améliorer les performances de cette approche, nous présentons des stratégies de génération de colonnes plus élaborées qui exploitent la structure du problème.

Enfin, nous présentons un bilan final de notre travail et quelques perspectives de recherche pour conclure ce mémoire.

Chapitre 1

Gestion des blocs opératoires : Contexte et problématique

Dans ce chapitre, nous présentons le contexte général de nos travaux. Nous décrivons le fonctionnement du bloc opératoire, les enjeux liés à sa planification, et les différentes formes d'aléas auxquelles il doit faire face. En particulier, nous montrons l'importance de la prise en compte des aléas lors de la planification du bloc opératoire.

---

1. 1 Introduction

Le bloc opératoire constitue un élément essentiel du plateau technique d'un hôpital, en raison de sa haute technicité, de l'investissement financier qu'il représente, de l'importance de la ressource humaine qu'il mobilise, des enjeux en termes de sécurité des patients et d'attractivité des établissements. Dans le cadre de la tarification à l'activité, c'est aussi une source de recettes pour les établissements.

Or, la gestion de ce secteur doit prendre en compte les besoins et les contraintes des chirurgiens, des anesthésistes, des infirmières, l'articulation avec les activités de stérilisation, de brancardage et de logistique (approvisionnement en matériels divers), concilier activité programmée et activité en urgence et enfin faire face à différentes sortes d'aléas qui peuvent survenir.

Dans ce chapitre, nous présentons le fonctionnement du bloc opératoire, les différentes ressources humaines et matérielles qu'il mobilise et les différents modes de gestion utilisés. Nous mettons aussi l'accent sur les différentes formes d'aléas qui caractérisent l'environnement du bloc opératoire, et leur impact sur le bon fonctionnement de ce dernier.

1. 2 Description des blocs opératoires

Le bloc opératoire est le secteur de l'hôpital où on fournit des soins chirurgicaux aux patients. La nature de ces soins nécessite une multitude d'équipements médicaux et de compétences humaines.

Les ressources matérielles constituant le bloc opératoire peuvent être représentées, de manière agrégée, sous la forme d'un système à deux étages : le premier correspond à l'ensemble des *salles opératoires*, le second à la *salle de réveil* (appelée aussi salle de surveillance post-interventionnelle, SSPI). Les salles opératoires sont les salles où les interventions chirurgicales, proprement dites, sont réalisées. Elles doivent être équipées du matériel chirurgical et anesthésique nécessaire pour le bon déroulement de l'intervention chirurgicale. La salle de réveil contient des lits pour accueillir les patients opérés en phase de réveil, et des dispositifs médicaux permettant de contrôler et de suivre l'état du patient.

Le coût d'acquisition de ces ressources matérielles ainsi que le coût de leur maintenance représente une part non négligeable du budget de l'hôpital.

Le bloc opératoire est aussi un lieu de forte concentration de compétences humaines. En effet, une intervention chirurgicale nécessite l'implication d'un nombre important d'acteurs de différentes spécialités (chirurgiens, anesthésistes, infirmières, brancardiers…) qui interviennent dans le même lieu, de manière séquentielle ou parallèle, pour réaliser un ensemble d'activités à l'aide des équipements médicaux adaptés.

Cette multitude d'acteurs impose une coordination et une gestion rigoureuse afin d'assurer non seulement une bonne qualité de soin pour les patients, mais aussi une utilisation efficace des ressources mises à disposition. D'autant plus, cette gestion doit tenir compte des priorités et des contraintes des différents acteurs impliqués.

On peut distinguer deux types de blocs opératoires : mono et pluridisciplinaires. Un bloc opératoire mono disciplinaire est dédié aux activités chirurgicales d'un service donné (Ophtalmologie, Orthopédie, ORL, etc.). Ce type de blocs est généralement caractérisé par un niveau de standardisation faible. Cependant, de nos jours les établissements hospitaliers tendent de plus en plus vers la mutualisation des ressources et optent pour des blocs pluridisciplinaires capables de traiter des patients émanant des différents services.

Un nombre croissant d'établissements envisagent la mise en place des blocs opératoires multidisciplinaires de grande taille où les services n'auront plus leurs propres salles opératoires et leurs propres personnels, mais devront travailler sur un site commun avec des règles de fonctionnement communes et une gestion centralisée (Kharraja, 2003).

Cette tendance de mutualisation des ressources met encore l'accent sur la nécessité d'une organisation et une gestion efficiente du bloc opératoire afin d'assurer une exploitation optimale des moyens matériels et humains disponibles.

1. 3 Processus opératoire

Afin de montrer l'interaction du bloc opératoire avec le reste des secteurs de l'hôpital, nous présentons les différentes phases du *processus opératoire*. Ce dernier peut être décomposé en trois grandes phases :

- La phase pré-opératoire s'étend de la prise en charge du patient jusqu'à son transfert au bloc opératoire pour l'intervention. Durant cette phase, le patient subit des consultations chirurgicales et anesthésiques. Pendant cette phase, une date « provisoire » d'intervention est proposée au patient. Cette date peut être modifiable ou non selon la politique du bloc opératoire.

- La phase per-opératoire correspond au séjour du patient dans le bloc opératoire. Elle couvre la période où le patient entre au bloc jusqu'à ce qu'il quitte la salle de réveil. Le jour de l'intervention, le patient est d'abord transporté par des brancardiers depuis sa chambre à l'hôpital jusqu'à la salle opératoire, il sera ensuite anesthésié et finalement opéré par une équipe chirurgicale. Notons que diverses activités de stérilisation et de préparation des consommables sont réalisées juste avant et après l'intervention. Une fois l'intervention chirurgicale terminée, le patient est transféré à la salle de réveil. Il y séjourne jusqu'au moment où l'anesthésiste l'autorise à retourner dans sa chambre ou il est transféré dans l'unité de soins intensifs et de réanimation.

- La phase post-opératoire : à partir de la salle de réveil, le patient est transféré soit vers sa chambre soit vers l'unité des soins intensifs et réanimation si son état présente des complications. Cette phase recouvre l'ensemble des soins nécessaires suite à l'intervention.

Pour une description plus détaillée des différentes phases du processus opératoire ainsi que des différentes ressources (humaines et matérielles) impliquées dans la réalisation de chacune de ces phases, nous faisons référence aux travaux de Chaabane (Chaabane, 2004).

D'après la description du processus opératoire, on voit bien que le bloc opératoire représente aussi une interface entre plusieurs autres secteurs de l'hôpital, tel que les services

d'hospitalisation, l'unité des soins intensifs, les activités de stérilisation, de brancardage et de logistique (approvisionnement en consommables et en matériels divers).

1. 4 Planification et ordonnancement du bloc opératoire

À fin de gérer et coordonner les différentes activités du bloc opératoire, les établissements hospitaliers utilisent un outil de gestion appelée *programmation opératoire* (Kharraja, 2003 ; Jebali, 2004 ; Chaabane, 2004).

Le terme « programmation opératoire » a des significations différentes d'un établissement à un autre. Mais, la finalité est la même : établir un planning prévisionnel des interventions à réaliser sur un horizon donné, allant d'une journée à une ou plusieurs semaines, en leur attribuant des ressources (humaines et matérielles) et en fixant l'ordre de leur exécution. Ce planning est généralement appelé *programme opératoire*.

Plus précisément, le programme opératoire est une sorte d'agenda précisant pour chaque salle opératoire les patients qui y seront opérés, leur ordre et leurs heures de passage. L'horizon temporel peut aller d'une journée jusqu'à plusieurs semaines. Les interventions programmées peuvent émaner d'un ou plusieurs services chirurgicaux selon le type de bloc mono ou pluridisciplinaire. Les demandes d'intervention peuvent être connues un ou plusieurs mois à l'avance comme dans l'heure qui précède l'établissement du programme opératoire selon la politique adoptée au bloc.

Le programme opératoire représente non seulement un outil de gestion interne au bloc opératoire, mais aussi une source d'informations pour d'autres activités de l'hôpital, tel que les services d'hospitalisation, les activités de stérilisation, de brancardage, etc.

La programmation opératoire se décompose en deux sous-problèmes (Magerlein et Martin, 1978 ; Fei, 2006):

- Planification à l'avance (*advance scheduling*) qui consiste à fixer dans le futur une date d'intervention pour chaque patient. Selon (Margerlein et Martin, 1978), le problème de planification à l'avance peut se séparer en deux catégories de problèmes selon le nombre et le type de contraintes de ressources prises en compte : dans une catégorie, on ne prend en compte que la durée d'ouverture des salles opératoires; dans l'autre catégorie, on prend en compte la durée d'ouverture des salles opératoires ainsi que la disponibilité des lits dans les services d'hospitalisation.

- Ordonnancement (*allocation scheduling*) qui consiste à déterminer un ordonnancement des interventions dans les salles opératoires pour une journée.

Le programme opératoire n'est autre que l'aboutissement d'une planification et d'un ordonnancement des interventions chirurgicales. Ces deux tâches peuvent être réalisées de manière séquentielle (hiérarchique) ou de manière simultanée, selon la politique utilisée par l'hôpital.

L'élaboration du programme opératoire est une tâche très complexe et dont le processus varie considérablement d'un hôpital à un autre. Cependant, la littérature fait état de trois approches de construction de programme opératoire (Patterson, 1996) (Kharraja, 2003) :

- **Programmation par pré-allocation de plages horaires** (*Block Scheduling*) : elle consiste à allouer, au préalable, des plages horaires à chaque chirurgien. Une plage horaire désigne une salle opératoire donnée qui est réservée à l'utilisation exclusive d'un chirurgien donné. Chaque chirurgien est responsable des plages horaires qui lui sont allouées ; il y place ses interventions comme bon lui semble. Dans certains établissements, on ne parle pas de chirurgien mais plutôt d'une unité chirurgicale (groupe de chirurgiens), et dans ce cas les plages horaires sont allouées aux unités chirurgicales. Les plages horaires forme le *squelette* du programme opératoire, appelée plan directeur d'allocation des plages horaires PDA (de l'anglais *Master Surgical Schedule*). Une fois le PDA déterminé, il est figé pour une période de temps donnée (allant d'un mois jusqu'à une année) jusqu'à une nouvelle mise à jour.

Avec cette approche de programmation, on peut distinguer deux classes de problèmes :

- La conception d'un PDA qui respecte les volumes d'activités, les disponibilités et les préférences des chirurgiens. Nous présentons dans le Chapitre 2 quelques travaux qui ont traité ce problème.
- La Planification et l'ordonnancement des interventions d'un chirurgien donné dans les plages horaires réservées au dit chirurgien.

Les inconvénients majeurs de cette approche sont (Kharraja, 2003) :

- La difficulté de construction du PDA. Le placement et la taille des plages horaires sont en effet les facteurs déterminants de la qualité du programme opératoire. Des plages surdimensionnées apportent du confort pour les chirurgiens mais conduisent à une performance médiocre (sous-utilisation des salles opératoires) et des plages justes ou sous dimensionnées risquent, en cas d'aléas, de provoquer des dépassements et une désorganisation du bloc, et par la suite des tensions entre chirurgiens;
- La perte de flexibilité et la dégradation des performances productives du bloc. Comme les plages horaires sont figées, il peut arriver que certains chirurgiens doivent refuser des interventions alors que d'autres ne remplissent pas leurs plages horaires.

- **Programmation ouverte** (*Open Scheduling*) : elle consiste à proposer un programme opératoire vierge de toute contraintes de placement ; aucun chirurgien ne peut à priori réserver des créneaux ou plages horaires pour ses propres interventions. On distingue deux approches pour gérer le programme opératoire :

- Le remplissage se fait de manière chronologique selon la règle « premier arrivé, premier servi ». Cette règle s'applique durant toute la période de construction du planning. Il s'agit d'un « *agenda collectif* » relié à un système d'information qui permet d'estimer les durées d'interventions et d'insérer des nouvelles interventions dans le programme opératoire.

Cet agenda peut être géré par les secrétariats des différents services chirurgicaux. Dans le cas du placement d'une intervention occasionnant un dépassement d'horaire, le responsable du bloc est alerté, ce qui déclenche un processus de régulation dont l'objectif est l'obtention d'une solution de résolution négociée.

Cette technique présente l'avantage d'être extrêmement simple à mettre en œuvre. Cependant, elle présente plusieurs inconvénients. Elle favorise les services chirurgicaux qui ont une activité planifiable sur le moyen et long terme, par exemple l'ophtalmologie, la chirurgie plastique, et défavorise les autres chirurgiens qui ne peuvent pas prédire leur agenda à moyen terme. De plus, comme le souligne (Kontak-Forsyth et Grant, 1995), cette pratique engendre un fort taux de déprogrammation, une sous-utilisation des ressources, des dépassements horaires importants et engendre de fortes tensions entre les chirurgiens ou les services de chirurgie.

- Le programme opératoire est élaboré à travers un processus de négociation entre les différents acteurs du bloc opératoire. Chaque semaine n, le *conseil de bloc* se réunit pour établir le programme opératoire de la semaine $n+1$. Le programme opératoire est généralement obtenu en harmonisant plusieurs pré-plannings. La difficulté majeure, à ce stade, consiste à trouver un programme opératoire qui permet une exploitation optimale du bloc tout en satisfaisant les contraintes et les souhaits des différents acteurs. L'élaboration d'un tel programme repose très souvent sur la compétence et l'expertise du responsable du bloc.

- **Programmation par pré-allocation et ajustement de plages horaires** (*Modified Block Scheduling*) : elle combine les deux approches précédentes. Partant d'un PDA, deux pratiques sont généralement utilisées :

- Une partie des plages horaires est allouée aux chirurgiens, l'autre partie reste commune pour tous les chirurgiens (*Unassigned bloc*). Les plages non allouées vont servir à palier les surcharges de travail ponctuelles de certains chirurgiens. Elles sont gérées par une

programmation ouverte, soit selon la règle premier arrivé, premier servi, soit à travers un processus de négociation.

- Si à une certaine date (Bloc release time) il y a des plages horaires qui ne sont pas bien exploitées par leurs chirurgiens respectifs, alors ces plages seront ajustées ou banalisées par le responsable de bloc afin d'en faire profiter d'autres chirurgiens, et par conséquent maximiser l'utilisation des blocs opératoires.

Cette approche de programmation présente l'avantage de combiner les deux approches précédentes. Cependant, son inconvénient réside dans les plages horaires inexploitées et le grand effort de synchronisation et de coordination nécessaire pour sa mise en œuvre.

1. 5 Bloc opératoire en environnement incertain

Le bloc opératoire fonctionne dans un environnement incertain. En effet, il est sujet à différentes formes d'aléas qui concernent essentiellement : le processus de demande, les durées des interventions chirurgicales et la disponibilité des ressources (humaines et matérielles).

Processus de demande

Les demandes pour interventions chirurgicales peuvent survenir de manière imprévue. On ne peut, en aucun cas prévoir quand la demande de soin se présente.

Certaines demandes peuvent être mises en attente et planifiées pour des dates futures. Ces demandes sont généralement des interventions qui ne représentent pas de caractère urgent ; elles peuvent être différées sans danger pour le patient. Dans le langage médical, ce type de demande est généralement désigné par la *chirurgie réglée*, ou *programmée*.

Cependant, une partie non négligeable des demandes relève du domaine de l'urgence et nécessite une prise en charge immédiate. Ce type de demande est, par nature, difficilement prévisible, et par conséquent non planifiable ; il est généralement désigné par la *chirurgie d'urgence*. Selon la structure juridique de l'établissement hospitalier, ce dernier peut se trouver dans l'obligation d'accepter tous les patients urgents, tel est le cas des établissements publics français. Dans une telle situation, on est obligé d'insérer les patients urgents dans le planning déjà établi, ce qui occasionne parfois des dysfonctionnements et généralement des coûts d'exploitation supplémentaires.

Selon une étude réalisée dans un centre hospitalier québécois (Lafon et Landry, 2001), 69% des perturbations engendrant des modifications dans le programme opératoire sont dues à la chirurgie d'urgence. En se référant à un hôpital canadien, Carter (2006) signale que 50% des

chirurgies cardiaques sont réalisées dans le cadre de chirurgie d'urgence. Une autre étude effectuée dans quatre hôpitaux belges (Rossi-Turk, 2002) révèle qu'en moyenne 20% de l'activité du bloc opératoire est une activité d'urgence.

Donc, une bonne programmation des interventions dépend directement de la capacité à anticiper les besoins de la chirurgie d'urgence. D'où la nécessité d'en tenir compte lors de la planification des activités du bloc opératoire.

Un autre type d'aléas caractérisant le processus de demande est le report ou l'annulation d'une intervention déjà programmée (planifiée). En effet, dans certaines situations, une évolution de l'état de patient peut imposer des analyses ou des tests supplémentaires qui conduisent à un report de l'intervention.

Durée d'intervention

La durée d'une intervention chirurgicale est sujette à des variations non négligeables selon le type d'intervention, le niveau d'expertise du chirurgien, l'état du patient, la technique d'anesthésie, etc.

La durée de l'intervention correspond à la durée de séjour en salle opératoire. Cette durée est composée principalement de : une durée de préparation à l'intervention (préparation de la salle, anesthésie, etc.), une durée de l'acte chirurgical (qui varie en fonction de sa nature, du chirurgien, et du patient), une durée de nettoyage et de re-conditionnement de la salle. Cette durée de séjour, fonction de multiples paramètres, est désignée de manière agrégée par la *durée opératoire*.

La variabilité des durées opératoires engendre très souvent des modifications dans le planning des activités du bloc. Ce qui occasionne une dégradation de la qualité de service vis-à-vis des patients (des délais d'attente, report d'intervention…) et des coûts d'exploitation supplémentaires (heures supplémentaires…). Donc, une exploitation efficace du bloc opératoire est fortement liée à l'intégration de l'aspect stochastique des durées opératoires dans la planification des activités, afin de proposer des plannings robustes.

Disponibilité des ressources matérielles et humaines

L'indisponibilité des ressources matérielles est généralement due à un sous dimensionnement des ressources. On cite, comme exemple, le cas d'un sous dimensionnement de la salle de réveil qui va bloquer le flux des patients et entraînant, par la suite, l'indisponibilité d'une ou plusieurs salles opératoires. Concernant les ressources humaines, leurs indisponibilités sont dues soit à un sous dimensionnement soit à des pratiques organisationnelles. Les chirurgiens,

par exemple, sont généralement affilés à différents établissements et il n'est pas rare qu'ils arrivent en retard à leurs rendez-vous.

Les différents aléas entourant le fonctionnement du bloc opératoire peuvent être classés en deux catégories. Dans une catégorie, on trouve les aléas dus à un disfonctionnement ou à des pratiques organisationnelles défaillantes ; ce qui entraîne des indisponibilités des ressources humaines ou matérielles. Dans une deuxième catégorie, il y a les aléas dus à la nature même du monde médical ; elle regroupe les aléas concernant le processus de demande et la durée opératoire.

Comme nous allons le voir en chapitre 2, rares sont les travaux qui prennent en compte les aléas dans la planification du bloc opératoire. Dans ce travail, nous intégrons les aléas concernant les durées opératoires et le processus de demande lors de la planification du bloc opératoire.

1. 6 Problématique

Dans ce travail, nous nous focalisons sur le problème de planification du bloc opératoire avec la prise en compte des incertitudes liées au processus de demande (chirurgie d'urgence) et au processus chirurgical (durée opératoire incertaine).

Rappelons que le problème de planification consiste à déterminer pour chaque patient la date d'intervention et la salle opératoire correspondante. Le problème d'ordonnancement des interventions dans une salle donnée n'est pas considéré dans ce travail.

Nous supposons que la salle de réveil ainsi que les ressources en aval et en amont du bloc opératoire (lits d'hospitalisation, salle de soins intensifs…) sont convenablement dimensionnées pour permettre une utilisation efficace du bloc. Dans ce contexte, les performances du bloc sont essentiellement liées à une planification efficace des interventions. Une planification qui doit, surtout, faire face aux différentes sortes d'aléas qui entourent le fonctionnement du bloc.

Dans ce travail, nous considérons un bloc opératoire (mono ou pluridisciplinaire) traitant des patients dans le cadre d'une chirurgie réglée (programmée) ou d'urgence. Pour ce faire, nous distinguons deux catégories de patients : les patients *électifs* et les patients *urgents*.

Patient électif : Le terme « *patient électif* » est le calque de l'anglais « *elective patient* » ; nous l'utilisons pour désigner tout patient ne présentant pas de caractère urgent ; et qui, par conséquent, peut être mis en attente sans danger. Les interventions chirurgicales des patients électifs peuvent être planifiées à l'avance ; elles forment ce qu'on appelle la chirurgie réglée,

programmée, ou encore *élective*. Notons que la chirurgie élective peut être réalisée aussi bien dans le cadre d'une hospitalisation complète qu'en ambulatoire.

Patient urgent : Dans le langage médical le terme « urgent » a plusieurs significations et on trouve plusieurs dégrées d'urgence : urgence vitale, urgence non vitale, etc. Dans ce travail, nous désignons par patient urgent tout patient nécessitant une intervention chirurgicale le jour même de son arrivée. Cette catégorie de patients regroupe aussi les « *retours en bloc* », c'est-à-dire les patients ayant déjà subi une intervention mais qui présentent des complications nécessitant un retour au bloc opératoire. Contrairement aux patients électifs les patients urgents ne peuvent pas être planifiés à l'avance de fait de leur caractère difficilement prévisible.

Le report ou l'annulation d'une intervention déjà planifiée représente une autre forme d'aléas qui entachent le processus de demande. Cependant, ce type d'aléas n'est pas pris en compte dans ce travail.

Donc, le problème qui se pose au stade de planification est comment planifier les patients électifs, tout en tenant compte de caractère incertain de la chirurgie d'urgence. Dans un premier temps nous supposons que les patients électifs ont des durées opératoires déterministes et nous nous concentrons sur l'aspect aléatoire dû à la chirurgie d'urgence (Chapitres 3 et 4). Ensuite, nous considérons le problème de planification avec la prise en compte de la chirurgie d'urgence et des durées opératoires aléatoires (Chapitre 5).

1. 7 Conclusion

Dans ce chapitre, nous avons décrit l'environnement du bloc opératoire, les différentes ressources qu'il mobilise, et la complexité de son fonctionnement. Nous avons, en particulier, présenté les différentes formes d'aléas qui peuvent survenir et la nécessité de les prendre en compte en phase de planification ; ce qui représente les motivations de ce travail.

Chapitre **2**

Gestion des blocs opératoires : État de l’art

Dans ce chapitre, nous présentons un panorama des approches de modélisation et de résolution employées dans la littérature pour résoudre des problèmes relevant de la gestion des blocs opératoires. Ces problèmes sont classés en cinq catégories : dimensionnement des ressources, planification des activités chirurgicales, ordonnancement des activités chirurgicales, gestion des ressources humaines, et évaluation de performances.

Ce chapitre propose un éclairage sur l’état de l’art des méthodes et approches développées pour appréhender différentes problématiques soulevées par la gestion des blocs opératoires. Dans un premier temps, nous présentons des travaux traitant des problèmes liés au dimensionnement des ressources du bloc opératoire. Dans la seconde section, nous dressons un large panorama des travaux considérant le problème de planification des activités chirurgicales. Nous distinguons deux catégories de travaux : la première catégorie concerne la répartition de la capacité du bloc opératoire entre les différents chirurgiens (ou unité chirurgicale), et la deuxième catégorie s’intéresse à déterminer la salle et le jour d’intervention pour chaque patient. La troisième section expose des travaux abordant l’ordonnancement des activités chirurgicales. On y trouve des travaux centrés sur les salles opératoires et d’autres incluant des ressources en aval et/ou en amont des salles. Dans la quatrième section, nous abordons les travaux ayant trait à la gestion des ressources humaines, que ce soit de manière quantitative ou qualitative. Dans la cinquième section, nous présentons des travaux qui s’intéressent soit à établir des indicateurs de performance soit à les évaluer afin de mesurer ou de prévoir l’efficacité du bloc opératoire.

2. 1 Dimensionnement des ressources

Le dimensionnement des ressources représente un enjeu de première importance pour les établissements hospitaliers, car il détermine la qualité et la quantité des moyens mis en place pour assurer la production des soins chirurgicaux. Ces ressources, généralement très lourdes en coût d'investissement et coût d'utilisation, doivent être en adéquation avec les objectifs en terme de volume et de type d'activités à réaliser. Un sous-dimensionnement des ressources constitue un handicap pour atteindre les objectifs ; alors qu'un sur-dimensionnement conduit à une sous utilisation des ressources disponibles et par conséquent un coût additionnel pour l'hôpital.

Le problème de dimensionnement du bloc opératoire consiste à déterminer le nombre et la nature des ressources humaines et matérielles à mobiliser. Ces décisions ont généralement un horizon qui dépasse l'année. Les travaux rencontrés dans la littérature considèrent essentiellement le dimensionnement des salles opératoires et de la salle de réveil, c-à-d la détermination du nombre de salles opératoires et du nombre de lits de réveil.

La simulation à événements discrets est l'outil le plus utilisé pour ce problème ; cependant, certains travaux utilisent d'autres outils tel que la programmation mathématique ou des modèle statistiques.

Vissers *et al*., (1998) proposent une approche pour dimensionner les besoins des spécialités chirurgicales en ressources (personnel infirmier, capacité en salles opératoires, lits d'hospitalisation). Le dimensionnement est basé sur l'utilisation actuelle des ressources et de l'évolution de la population, des demandes de soins et de la part du marché. Blake et Carter (2002) proposent une approche multicritères pour déterminer les ratios de différentes interventions et les volumes d'activités à réaliser afin de satisfaire des objectifs fixés par les autorités sanitaires régionales. Dexter *et al*., (1999a) utilisent les séries chronologiques pour estimer les besoins futurs des différentes spécialités chirurgicales afin de déterminer les besoins en capacité en salles opératoires.

En anticipant une augmentation de volume d'activités chirurgicales, (Lovejoy et Li, 2002) propose un approche pour choisir entre : la construction des nouvelles salles opératoires ou l'augmentation des durées d'ouverture des salles existantes. La solution doit satisfaire les besoins de toutes les parties prenantes (managers, chirurgiens et patients).

La simulation représente l'outil le plus utilisé pour les problèmes de dimensionnement des ressources pour satisfaire les besoins futurs, lors de la construction d'un nouvel hôpital ou un nouveau bloc opératoire, ou accompagner les projets de ré-ingénierie et de modernisation (Jun *et al*., 1999).

Hopkins *et al*. (1982) et Goldman et Knappenberger (1968) utilisent la simulation pour déterminer le nombre de salles opératoires nécessaire pour assurer les besoins futurs en terme de volume d'activités. Dans le même contexte, Currie *et al*. (1984) utilisent la simulation pour estimer le nombre de salles opératoires et de lits de réveils nécessaires pour faire face à une augmentation de 20% dans la demande pour les soins chirurgicaux. Kwak *et al*. (1975) déterminent le nombre de lits de réveil nécessaires en réponse à une expansion de la capacité des salles opératoires. Kuzdrall *et al*. (1981) utilisent la simulation pour évaluer les besoins en salles opératoires et lits de réveils en fonction de différents modèles de planification et d'ordonnancement. Olson et Dux (1994) étudient le choix d'augmenter de 7 à 8 le nombre de salles dans un bloc opératoire. Les auteurs concluent que ce choix permet de satisfaire les besoins, mais seulement pour une ou deux années à venir. Cependant, la création d'un centre de chirurgie ambulatoire (pour séparer la chirurgie ambulatoire de la chirurgie en hospitalisation) permettra de mieux satisfaire les besoins futurs de l'hôpital. Marcon *et al*. (2003b) et Smolski *et al*. (2002) traitent le dimensionnement du nombre de lits en salle de réveil et le nombre de brancardiers. A l'aide de la simulation les auteurs déterminent le nombre minimum de lits de réveil et définissent la stratégie de gestion des brancardiers. De manière similaire, Dussauchoy *et al*. (2003) s'intéressent au dimensionnement des salles opératoires et de la salle de réveil. Ils testent plusieurs scénarios combinant différents modèles de charge et différentes stratégies d'ordonnancement des interventions. Gourgand *et al*. (2005) utilisent la simulation pour dimensionner les besoins en brancardiers pour un futur bloc opératoire. Albert *et al*. (2007) proposent un outil de simulation pour l'aide à la décision dans la phase de ré-ingénierie des blocs opératoires.

Amladi (1984) et Ramis *et al*. (2001) utilisent la simulation pour le dimensionnement des ressources nécessaires pour une nouvelle unité de chirurgie ambulatoire, en considérant le nombre de patients admis, le temps d'attente et les ressources nécessaires pour les interventions. Meier *et al*. (1985) étudient plusieurs scénarios (modèles de charge) pour une unité de chirurgie ambulatoire. Les auteurs concluent que la capacité actuelle en salles opératoires est suffisante ; elle permet de satisfaire la demande durant les cinq années à venir. Dans une autre étude, Iskander et Carter (1991) montrent que le nombre des salles opératoires actuelles est suffisant pour couvrir une augmentation dans la demande, mais ils suggèrent d'augmenter le nombre des lits de réveils.

Nous trouvons dans la littératures des travaux qui proposent un dimensionnement en s'appuyant sur l'expérience des spécialistes et les techniques de benchmarking (Broun, 2002).

Nous remarquons que la littérature est relativement riche en ce qui concerne les travaux portant sur le dimensionnement du bloc opératoire. Cependant, rares sont les travaux qui considèrent des ressources autres que les salles opératoires et les lits de réveils, tels que les

ressources humaines (Trilling, 2006), les équipements et les instruments médicaux spécifiques, (Reymondon *et al.*, 2006) etc. Nous présumons que ces ressources sont considérées à l'échelle de tout l'hôpital et pas seulement au niveau du bloc opératoire.

2. 2 Planification des activités chirurgicales

La planification des activités chirurgicales consiste à établir un planning qui spécifie, sur un horizon donné, les interventions qui seront réalisées chaque jour et dans chaque salle opératoire. Selon l'approche de programmation utilisée par l'établissement hospitalier (*open scheduling, block scheduling* (Section 1.4)), la planification est réalisée différemment.

Dans le cas où une approche *open scheduling* est utilisée, la planification revient à affecter les interventions aux salles opératoires sur un horizon donné. Avec une approche *block scheduling*, des plages horaires sont affectées aux chirurgiens. Une plage horaire désigne une salle opératoire qui est réservée pour une durée de temps à l'utilisation d'un chirurgien donné. Chaque chirurgien place à sa convenance ses interventions dans les plages qui lui sont allouées. Avec ce mode de fonctionnement, le problème de planification est plutôt un problème d'*allocation des plages horaires* aux différents chirurgiens.

2. 2. 1 Planification des interventions chirurgicales

Le problème de planification consiste à affecter les interventions chirurgicales aux salles opératoires sur un horizon de planification (généralement une semaine). L'objectif est généralement de minimiser un ensemble des coûts (sous et/ou sur utilisation des salles opératoires, d'hospitalisation ou d'attente, des pénalités de non satisfaction des préférences, etc.) sous des contraintes de ressources (tel que les capacités des salles opératoires, la disponibilité de certains équipements médicaux spécifiques, la disponibilité des chirurgiens et leurs préférences,…) et/ou des contraintes relatives aux patients tel que les dates limites à ne pas dépasser.

Les travaux traitant ce problème font appel le plus souvent à des outils de programmation mathématique (programmation linéaire en nombre entiers ou mixte, programmation multi objectifs, des heuristiques). Cependant, on trouve aussi quelques travaux utilisant des outils tel que la simulation à événements discrets (Jones *et al.*, 1983; Sahney *et al.*, 1976) et les systèmes à base de connaissances (Bharadwaj *et al.*, 1999). Dans ce qui suit nous présentons des travaux utilisant la programmation mathématique.

Dexter *et al.*, (1999b) proposent d'utiliser des heuristiques de bin packing « off-line » pour la planification des interventions ainsi que l'ajout des interventions supplémentaires dans un planning déjà établi. La technique consiste à affecter les interventions aux salles opératoires

(ou plage horaires) en maximisant le remplissage de chaque salle et minimisant le nombre de salles utilisées. Les contraintes prises en compte portent sur les capacités des salles opératoires. Cependant, ces contraintes peuvent être légèrement violées (autoriser des dépassements horaires de l'ordre de 15 minutes). Les auteurs comparent différentes heuristiques classiques de *bin-packing* et concluent que l'heuristique « best-fit décroissant», est la meilleure ; elle assure le meilleur taux d'occupation des salles.

Guinet et Chaabane (2003) proposent un programme linéaire en nombres entiers permettant d'affecter sur un horizon d'une semaine des interventions aux salles opératoires. L'objectif est de trouver un planning qui minimise les coûts des heures supplémentaires et les journées d'hospitalisation des patients. Les contraintes prisent en compte portent sur la capacité des salles opératoires en terme d'heures normales et supplémentaires, les dates d'hospitalisation et les dates limites des patients, le nombre maximal d'interventions par chirurgien par jour, et l'adéquation des salles opératoires. Les auteurs proposent une résolution heuristique pour le problème. Les auteurs supposent que les interventions ont des durées déterministes, égales à l'une des quatre valeurs suivantes : 1, 2, 3 ou 4 heures (intervention mineure, intermédiaire ou majeure).

Ogulata et Erol (2003) considèrent comme objectifs de minimiser les jours d'attentes des patients, et d'équilibrer la charge entre les différents groupes chirurgicaux, et entre les différentes catégories de chirurgie. Le problème est décomposé en trois phases : sélectionner un sous-ensemble de patients (parmi les patients en attente) qui seront opérés pour la semaine à venir, affecter les patients aux groupes chirurgicaux et enfin chaque groupe chirurgical affecte ces patients aux salles jours.

Persson et Persson (2005) prennent comme objectif de minimiser les coûts de non planification des patients. Les contraintes considérées portent sur : la capacité en heures normales des salles opératoires, la disponibilité et la qualification des chirurgiens, les patients associés à un chirurgien donné doivent être affectés à la même salle, les patients affectés à la même salle doivent être associés au même groupe chirurgical. Les auteurs proposent un programme en nombres entiers et une résolution par CPLEX.

Fei *et al*., (2006) traitent le problème de planification d'un ensemble d'interventions programmées pour une semaine dans un bloc opératoire, sur un ensemble de salles opératoires identiques, avec un objectif de minimiser les coûts des heures supplémentaires et de sous-utilisation des salles opératoires. Les contraintes considérées portent sur les dates limites d'interventions et la capacité maximale des salles opératoires. Les auteurs utilisent une approche de génération de colonnes pour la résolution du problème ; ils proposent une méthode exacte (*branch-and-price*) ainsi qu'une heuristique.

Marcon *et al.* (2003a) proposent un outil d'aide à la négociation pour l'élaboration du planning opératoire. Une première phase statique, consiste à établir un planning qui maximise l'occupation des salles opératoires (sans dépasser leurs capacités en heures normales) tout en lissant leurs charges. Ce lissage a pour but de minimiser le risque de dépassement horaire (appelé risque de non réalisation) à cause des aléas des durées interventions. Une deuxième phase dynamique, consiste à évaluer ce risque chaque fois qu'une intervention s'achève, et décider de maintenir le planning ou de le modifier. La modification du planning se fait par négociation entre les différents acteurs du bloc opératoire, et a pour objectif de maintenir le risque de dépassement à un niveau acceptable.

Contrairement aux travaux présentés ci-dessus, d'autres travaux ont considéré des contraintes relatives aux ressources externes au bloc opératoire tel que la salle des soins intensifs et les lits d'hospitalisation. Cependant, ces travaux considèrent un horizon de planification d'un jour.

Ozkarahan (2000) propose un modèle linéaire multi objectifs en variables mixtes pour l'affectation des patients aux salles opératoires sur un horizon d'un jour. Les objectifs considérés sont : minimiser le dépassement horaire et la sous-utilisation des salles opératoires, maximiser les préférences des chirurgiens concernant les salles opératoires ainsi que les interventions à réaliser, et minimiser la rupture en lits dans la salles des soins intensifs.

Jebali *et al*. (2006) considèrent la planification sur un horizon d'un jour. Il s'agit de déterminer les interventions chirurgicales qui seront réalisées en un jour donné, ainsi que leurs affectations aux différentes salles opératoires. Le planning journalier doit minimiser les coûts des heures supplémentaires et de sous-utilisation des salles opératoires, ainsi que les durées d'hospitalisation des patients (période de temps entre la date d'hospitalisation et la date d'intervention). Les contraintes considérées portent sur la capacité des salles opératoires en heures normales et supplémentaires, la capacité maximale de travail par chirurgien, l'adéquation des salles opératoires et la capacité de la salle des soins intensifs en terme de nombre de patient par jour. Les auteurs modélisent le problème sous la forme d'un programme en nombres entiers et utilisent CPLEX pour sa résolution.

Dans un autre travail (Jebali, 2004), l'auteur étend cette approche pour considérer un horizon d'une semaine et des contraintes relatives au nombre de lits d'hospitalisation. L'auteur propose un modèle en nombres entiers et une heuristique pour la résolution.

Il est à noter que les approches considérant des ressources auxiliaires supposent une connaissance fine du processus chirurgical telle que le passage ou non d'un patient en salle de soins intensifs, la durée de son séjour dans cette salle, le nombre de jour d'hospitalisation nécessaire, etc.

Certains travaux traitent le problème de planification et d'ordonnancement de manière conjointe ; il s'agit de déterminer simultanément la date, l'heure et la salle d'intervention pour chaque patient. Velasquez et Melo (2005) proposent un modèle mathématique dont l'objectif est de maximiser une fonction qui représente des préférences relatives aux salles, dates et heures d'interventions, sous des contraintes de capacité telles que le nombre de salles opératoires et leurs capacités en heures normales, la disponibilité des chirurgiens et des équipements médicaux nécessaires. Les auteurs utilisent une approche de génération de colonnes pour la résolution du problème. Roland *et al*. (2007) considèrent le même problème avec un objectif de minimiser le nombre de salles opératoires ouvertes ainsi que les heures supplémentaires, sous des contraintes relatives à la disponibilité des chirurgiens et d'autres ressources humaines. Les auteurs modélisent le problème sous la forme d'un programme en nombres entiers et proposent un algorithme génétique pour sa résolution. Notons que dans ces deux travaux traitant de manière intégrée la planification et l'ordonnancement, les auteurs se restreignent à résoudre des problèmes de petite taille avec un horizon de planification d'une journée ; ceci est probablement dû à la forte complexité des problèmes.

Nous remarquons que la grande majorité des travaux suppose que les salles opératoires soient dédiées seulement à la chirurgie élective, et que les interventions ont des durées déterministes. Dans la littérature peu de travaux ont considéré le problème de planification avec prise en compte des incertitudes (Hans *et al*., 2006 ; Gerchak *et al*., 1996).

Hans *et al*. (2006) traitent la planification avec un objectif de maximiser l'utilisation des salles opératoires et minimiser le risque de dépassement horaire dû aux incertitudes concernant les durées d'interventions. Les auteurs proposent d'affecter un *temps de sécurité* à chaque salle opératoire pour absorber la variabilité des durées opératoires. Le « temps de sécurité » en une salle donnée est une fonction de la variabilité des durées d'interventions affectées à la dite salle. Les contraintes prises en compte lors de l'affectation des patients concernent l'adéquation des salles opératoires et la disponibilité des équipes médicales. Les auteurs proposent plusieurs heuristiques pour la résolution du problème. Cependant, les auteurs font l'hypothèse que les durées d'interventions affectées à une salle donnée ont la même variance et que la somme de ces durées suit une loi normale ; concernant la chirurgie d'urgence elle n'est pas considérée.

Gerchak *et al*. (1996) considèrent le problème d'admission de nouveaux patients électifs en début de journée pour un bloc opératoire partagé avec la chirurgie d'urgence. Les auteurs ont mis en évidence les caractéristiques de la politique optimale. Toutefois, il s'agit d'un modèle mono-période et qui ne spécifie la salle d'intervention.

À notre connaissance, il n'y a pas dans la littérature des travaux qui ont traité le problème de planification avec prise en compte de la chirurgie d'urgence et encore moins en considérant des durées d'interventions aléatoires.

2. 2. 2 Allocation des plages horaires

Des nombreux travaux ont considéré le problème d'affectation des plages horaires aux chirurgiens (spécialités ou groupes chirurgicaux). Le problème consiste à établir un « plan d'allocation des plages horaires » (PDA) qui spécifie le nombre et le type de salles opératoires, leurs durées d'ouverture, et le chirurgien prioritaire.

L'objectif est généralement de maximiser l'utilisation des salles opératoires tout en assurant une répartition équitable entre les chirurgiens et qui respecte leurs disponibilités et leurs préférences. Dans la littérature, plusieurs approches sont proposées ; elles diffèrent essentiellement : par la manière d'estimer les besoins de chaque chirurgien (prévision des besoins futurs, ou la charge actuelle), les ressources considérées (salles opératoires, ressources auxiliaires,…) et la prise en compte ou non du types d'interventions qui seront réalisées dans chaque plage horaire.

Macario *et al.* (2001) étudient l'impact de la répartition des plages horaires en fonction de l'apport (profit) de chaque cas chirurgical avec l'objectif d'augmenter le profit total. Les chirurgiens qui réalisent les interventions les plus rentables ont des plages horaires plus importantes.

Blake et Donald (2002) proposent un modèle de programmation en nombres entiers pour l'affectation des salles opératoires aux différentes spécialités afin de trouver une répartition équitable entre les spécialités. Les besoins des différentes spécialités sont estimés par les charges actuelles. Kharraja et Marcon (2003) étudient le même problème mais en prenant en considération les opérations à réaliser par chaque chirurgien.

D'autres travaux font des prévisions sur un horizon d'une année pour estimer les besoins des chirurgiens. Strum et *al*. (1999) proposent un modèle mathématique pour l'affectation des salles opératoires aux différents groupes chirurgicaux pour minimiser les coûts du personnel en se basant sur la prévision de la moyenne et de l'écart type de la demande des douze mois à venir. Vissers (1998) propose une approche d'allocation des salles opératoire aux différentes spécialités en se basant sur l'utilisation actuelle, l'analyse du besoin de la population, la demande et le développement de la part du marché.

D'autres travaux considèrent non seulement les salles opératoires, lors de l'affectation des plages horaires, mais aussi des ressources auxiliaires tel que le nombre de lits d'hospitalisation, ou de la salle des soins intensifs. Ces travaux généralement classent les

interventions par catégorie selon leurs besoins en ressources médicales et leurs caractéristiques (durée d'intervention, durée de séjour dans la salle des soins intensifs…). Chaque catégorie d'interventions est associée à un ou plusieurs chirurgiens (ou spécialités chirurgicales). Le problème consiste à affecter les plages horaires aux chirurgiens et à déterminer le nombre et les catégories d'interventions qui seront réalisées dans chacune d'elles. Cependant, il est à noter que ces travaux supposent que les besoins en capacité (en heures de bloc) pour chaque chirurgien sont préalablement déterminés.

Visser *et al.* (2005) proposent une approche par programmation en nombres mixtes pour la construction de PDA pour le service de chirurgie « cardiothoracique », avec comme objectif l'optimisation de l'utilisation des ressources auxiliaires tels que : les lits dans la salle des soins intensifs, lits d'hospitalisation et les personnels infirmiers. Santibanez *et al*. (2004) utilisent aussi la programmation en nombres entiers pour maximiser le nombre de patients opérés tout en lissant la charge des ressources auxiliaires : salle de réveil et salle des soins intensifs. Les contraintes considérées sont: équités entre les différentes catégories d'interventions, l'adéquation des salles opératoires, et la capacité des ressources auxiliaires. Van Oostrum *et al*. (2005) utilisent une approche de génération de colonnes pour maximiser l'utilisation des salles opératoires tout en lissant la charge engendrée dans les lits d'hospitalisation et de la salle des soins intensifs.

Beliën et Demeulemeester (2006) traitent le problème d'affectation des plages horaires avec la prise en compte de l'aspect aléatoire des durées de séjours dans le service d'hospitalisation. Il s'agit d'affecter les plages aux chirurgiens de sorte à lisser le nombre moyen des patients séjournant dans le service d'hospitalisation après leurs interventions. Les auteurs modélisent le problème sous la forme d'un programme en nombres mixtes et proposent plusieurs heuristiques pour la résolution.

Afin de remedier aux aléas tel que la variation de la demande hebdomadaire de chaque chirurgien ou à l'insertion des interventions supplémentaires au cours de la semaine, Hammami (2006) propose une approche permettant la construction et l'allocation des plages horaires d'une façon flexible. L'auteur propose de classer les chirurgiens en groupes (pool), tout en minimisant le nombre de groupes. Ensuite, il convient de construire des plages individuelles pour chaque chirurgien et des plages communes pour chaque groupe de chirurgiens. Les plages communes à un groupe permettent d'absorber les variations de l'activité du dit groupe. L'auteur modélise ce problème sous la forme d'un programme linéaire en nombres binaires. Toutefois, il faut noter que cette approche ne permet pas d'équilibrer la charge entre les différentes salles opératoires.

2. 3 Ordonnancement des interventions chirurgicales

L'ordonnancement consiste à donner un ordre de passage aux interventions planifiées sur les différentes ressources du bloc opératoire. La littérature fait état des deux catégories de travaux selon les ressources considérées :

- Ordonnancement centré sur les salles opératoires : les ressources sont essentiellement les salles opératoires.
- Ordonnancement du bloc opératoire : les ressources considérées sont les salles opératoires, les lits de la salle de réveil, et éventuellement d'autres ressources tel que les brancardiers, les équipes de nettoyages, les chirurgiens, les infirmiers, et la salle des soins intensifs.

2. 3. 1 Ordonnancement centré sur les salles opératoires

Étant donné des patients planifiés pour chaque salle opératoire, le problème d'ordonnancement consiste à déterminer la séquence et les heures de passage (starting time) des patients. On distingue deux types de travaux : ordonnancement conjoint (simultané) de plusieurs salles et ordonnancement d'une salle.

Ordonnancement d'une seule salle opératoire

L'ordonnancement est réalisé pour une salle donnée, sans se soucier des activités réalisées dans les autres salles. Les interventions ont des durées aléatoires et sont réalisées par différents chirurgiens (ou équipes chirurgicales). L'objectif de l'ordonnancement est généralement de minimiser le coût de sous-utilisation de la salle et les coûts d'attentes des chirurgiens. Les heures de passage des patients en salle permettent aux chirurgiens d'organiser leurs activités. Cependant, ces heures ne sont pas toujours respectées dû à la variabilité des interventions chirurgicales. En effet, si les interventions précédentes sont plus longues que prévu, l'heure de passage du patient suivant est retardée, entraînant ainsi un temps d'attente pour le chirurgien et le patient. D'un autre côté, si les interventions précédentes sont plus courtes que prévu, l'intervention suivante ne peut pas commencer vu l'indisponibilité du chirurgien. Ce qui engendre, par conséquent, une sous-utilisation de la salle opératoire. Donc, l'estimation des heures de passage doit assurer un compromis entre les coûts d'attente des chirurgiens et les coûts de sous-utilisation de la salle.

De manière plus générique, le problème peut être formulé comme suit : étant donné un serveur (une ressource) et un ensemble de clients (tâches), le problème consiste à trouver un ordonnancement qui minimise le coût de sous (et sur) utilisation des serveurs et le coût

d'attente des clients. La salle opératoire et les chirurgiens représentent respectivement le serveur et les clients.

Lebowitz (2003) propose d'approximer les durées d'intervention par leurs moyennes en se basant sur les informations passées. Il compare par simulation numérique quelques règles de séquencement et conclut que le fait de traiter les patients par ordre croissant des durées moyennes permet de réduire le temps d'attente de chirurgiens, de sous et sur utilisation de la salle.

Weiss (1990) considère le problème d'ordonnancement avec des durées d'interventions aléatoires afin de minimiser le coût de sous-utilisation de la salle et les coûts d'attentes des chirurgiens. Dans un premier temps, l'auteur s'est intéressé à estimer les heures de passage pour une séquence d'interventions donnée. L'auteur a commencé par traiter une séquence de deux interventions, et il a montré que le problème d'ordonnancement est similaire au problème du *vendeur des journaux* (*Newsboy*) en gestion de stock. En effet, les coûts d'attente et de sous-utilisation sont équivalents respectivement aux coûts de rupture et de surplus. Quant à l'heure de passage du deuxième patient, elle est équivalente à la quantité à commander dans le problème du vendeur des journaux. Sur la base de ce résultat, l'auteur a proposé une heuristique qui permet d'estimer l'heure de passage de chaque patient de manière séquentielle. Dans un second temps, l'auteur a proposé une heuristique pour déterminer la séquence des patients : traiter les patients par ordre croissant de la variance des temps opératoires. Selon l'auteur cette heuristique permet de trouver des solutions de bonne qualité.

Denton et Gupta (2003) et Denton *et al*., (2007) ont considéré le même problème d'ordonnancement mais avec comme objectif la minimisation des coûts d'attente des chirurgiens, de sous-utilisation et sur-utilisation de la salle opératoire. Dans un premier travail (Denton et Gupta, 2003), les auteurs ont supposé que la séquence des patients est déterminée au préalable ; ils ont formulé le problème sous la forme d'un programme linéaire stochastique (*two-stage stochastic linear program*) et ont proposé une méthode de résolution exacte (*L-shaped method*). Dans un second article (Denton *et al*., 2007), les auteurs ont comparé numériquement leur approche et plusieurs heuristiques pour le séquencement des patients. Ils concluent que la meilleure consiste à réaliser les interventions par ordre croissant des variances, confirmant ainsi le résultat de Weiss (1990). Dans un autre article (Denton *et al*., 2006), les auteurs utilisent la simulation Monte Carlo et le recuit simulé pour la résolution du problème d'ordonnancement.

On constate que le problème d'ordonnancement est scindé en deux phases : le séquencement des interventions ensuite l'estimation de leurs heures de passage. Ceci est probablement dû à la difficulté du problème conjoint. On remarque aussi que seulement les patients planifiés sont considérés, il n'y a pas des activités non programmées.

Ordonnancement conjoint de plusieurs salles opératoires

D'autres travaux ont considéré le problème d'ordonnancement d'un ensemble de salles de manière conjointe. Étant donné un ensemble de salles, et pour chacune d'elles il y a un ensemble de patients planifiés. Il s'agit de séquencer les interventions dans une salle donnée en tenant en compte des séquencements dans les autres salles.

Certaines interventions nécessitent un équipement médical spécifique (microscope…) ou un type de personnel médical très spécialisé qui est disponible en quantité limitée. Donc, ces interventions ne peuvent pas être réalisées en parallèle (bien évidemment dans des salles différentes). Dexter et Traub (2000) proposent une méthode statistique pour évaluer la probabilité qu'une intervention va durer plus qu'une autre, ce qui permet de séquencer les interventions de manière à éviter un conflit de ressources. Sier *et al.*, (1997) ont considéré le problème d'ordonnancement des interventions avec durées déterministes dans plusieurs salles tout en prenant en compte des contraintes liées à la disponibilité des équipements, l'age du patient, la durée d'interventions,etc. les auteurs ont modélisé le problème sous la forme d'un programme non linéaire en nombres entiers et ont utilisé le recuit simulé pour une résolution approchée.

Lans *et al*. (2006) ont considéré le problème de séquencement des interventions avec durées déterministes dans plusieurs salles. Les salles opératoires sont utilisées pour réaliser les interventions planifiées ainsi que des interventions non programmées (urgentes). Quand un patient urgent arrive, il est opéré dans la première salle disponible. Autrement, dès qu'une intervention encours s'achève, l'intervention urgente peut commencer et l'intervention qui a été planifiée sera décalée. Les auteurs proposent d'ordonnancer les interventions planifiées de manière à diminuer le temps d'attente des interventions urgentes.

Un problème très similaire à celui présenté dans cette section est celui de gestion des rendez-vous (*outpatient appointment problem*) (Cayiril et Veral, 2003) (Klassen et Rohleder, 1996). Il s'agit de déterminer, pour une journée donnée, les heures de rendez-vous des patients pour des soins (chirurgicaux ou non) en ambulatoire. L'objectif est en général de minimiser l'attente du patient, et l'inoccupation des ressources. Les ressources représentent les serveurs et les patients représentent les clients. La littérature fait état de deux types d'ordonnancements :

- ordonnancement *par block* (*block appointment*) : il s'agit de diviser l'horizon de l'ordonnancement en plusieurs périodes (blocks) de tailles égales, et de déterminer l'ensemble de patients qui doivent arriver au début de chaque période,
- et ordonnancement *individuel* (individual appointment scheduling) : on détermine pour chaque patient son rendez-vous sans discrétiser l'horizon de l'ordonnancement.

Ce problème est largement traité dans la littérature ; pour un état de l'art sur le sujet nous faisons référence à (Cayiril et Veral, 2003)). Toutefois, nous présentons dans ce qui suit quelques exemples de ces travaux.

Ho et Lau (1992) ont considéré le problème d'ordonnancement afin de trouver un compromis entre le coût d'attente des patients et celui du personnel médical. Les auteurs ont montré que les temps d'attente (des patients et du personnel) - engendrés par toute règle d'ordonnancement- sont influencés par trois *facteurs environnementaux* : la probabilité de non présence « no-show », le coefficient de variation des durées d'interventions et le nombre de patients à ordonnancer. Ils ont comparé, par simulation, les performances de neuf règles d'ordonnancement (individuel et par block) pour différentes combinaisons des facteurs environnementaux, et ils ont proposé une procédure qui permet d'identifier la règle la plus performante pour chaque combinaison.

Liu et Liu (1998a, 1998b) ont traité le problème d'ordonnancement « par block » avec plusieurs serveurs ; l'objectif est de minimiser les coûts de sous et sur utilisation des serveurs, et les coûts d'attente des clients. Dans un premier travail (Liu et Liu, 1998a), les auteurs ont supposé que les temps de service (ou durées opératoires) sont identiquement distribuées suivant une loi exponentielle, ils ont modélisé le problème sous la forme d'un programme dynamique et ils ont montré que la politique optimale pour ce problème est une politique dynamique à « *charge nominale* » (politique équivalente au base-stock). C'est à dire, pour chaque période, il y a un nombre optimal de patients qui doivent être présents dans le système au début de chaque période. Sur la base de ce résultat, les auteurs ont proposé une méthode approchée qui fournit des solutions proches de l'optimum (obtenu avec une simulation exhaustive). Dans un second article (Liu et Liu, 1998b), les auteurs ont proposé une approche par simulation pour estimer la longueur des périodes et pour ordonnancer les patients.

2. 3. 2 Ordonnancement du bloc opératoire

Étant donné un ensemble d'interventions planifiées pour un jour donné, l'ordonnancement consiste à fixer l'ordre et l'heure de passage des patients programmés ainsi que les différentes ressources mobilisées. Les interventions planifiées peuvent être préalablement affectées à des salles opératoires spécifiques ; cette affectation peut être remise en cause ou non, selon l'approche d'ordonnancement utilisée.

Les ressources considérées sont essentiellement les salles opératoires et les lits de la salle de réveil. Cependant, certains travaux prennent en compte d'autres ressources telles que : les brancardiers, les équipes de nettoyage, les chirurgiens, les infirmiers, et la salle de soins intensifs.

On distingue deux grandes catégories de travaux : la première catégorie suppose que le patient peut commencer son réveil dans la salle opératoire si jamais la salle de réveil est occupée, ce qui peut bloquer la salle opératoire (contrainte de blocage) ; la deuxième catégorie impose que le réveil commence dans la salle de réveil (contrainte de non-attente, ou *no-wait*).

L'objectif de l'ordonnancement est généralement de minimiser le *makespan* (Cmax, heure d'achèvement de la dernière opération) de la salle de réveil ou/et des salles opératoires. Cet objectif est équivalent à minimiser les dépassements horaires dans les salles opératoires et/ou dans la salle de réveil.

Kwak *et al*., (1976) testent par simulation 5 règles d'ordonnancement basées sur les durées moyennes d'intervention et de réveil. Les auteurs ne déterminent pas la meilleure règle d'ordonnancement mais montrent que le choix d'une règle a un impact sur le taux d'utilisation des salles opératoires et de la salle de réveil.

Hsu *et al*., (2003) considèrent le problème d'ordonnancement dans un contexte de chirurgie ambulatoire ; ils modélisent le problème comme un flow-shop à deux étages avec « no-wait ». Le premier étage représente les salles opératoires avec les chirurgiens comme ressources principales ; le deuxième étage représente la salle de réveil avec les infirmiers comme principale ressource. L'objectif est de minimiser le makespan de la salle de réveil ainsi que le nombre des infirmiers dans cette dernière. Les auteurs proposent une heuristique de recherche tabou pour la résolution.

Guinet et Chaabane (2003) étudient le problème d'ordonnancement avec la possibilité de changer l'affectation des patients aux salles et avec l'hypothèse de salles opératoires identiques. Les auteurs modélisent le problème sous la forme d'un flow-shop hybride à deux étages avec des contraintes de « sans attente » entre les étages. L'objectif est de minimiser le makespan (Cmax) de la salle de réveil (i.e. heure de fermeture de la salle de réveil). Les auteurs utilisent une heuristique basée sur l'algorithme de Gilmore et Gomory pour la résolution du problème. Dans (Chaabane *et al.*, 2004), les mêmes auteurs étendent leur approche en considérant les brancardiers comme troisième étage (ressource) d'un flow-shop hybride avec contraintes de précédences. Les auteurs effectuent une étude comparative des plusieurs règles d'ordonnancement et identifient celle qui donne les meilleurs résultats.

Kharraja *et al.* (2002) et Jebali *et al.* (2006) considèrent le problème comme un flow-shop hybride avec contrainte de blocage ; le patient reste en salle opératoire tant qu'un lit de réveil n'est pas disponible. Kharraja *et al.* (2002) proposent une résolution heuristique du problème avec un objectif de minimiser le makespan de la salle de réveil. Jebali *et al.* (2006) considèrent le problème avec un objectif de minimiser les dépassements horaires (makespan) dans les salles opératoires, et proposent une résolution par programmation en nombres mixtes.

Perdomo *et al.* (2006) utilisent la relaxation lagrangienne pour l'ordonnancement des salles opératoires, la salle de réveil ainsi que les activités de nettoyage des salles ; avec une contrainte de « no-wait ». L'objectif à minimiser est une somme des critères : un critère par patient, et qui est fonction de temps d'achèvement de toutes les activités relatives au dit patient. Dans (Augusto et *al*., 2007) le même problème est étendu pour considérer un autre type de ressources les transporteurs (brancardiers) et en autorisant le réveil dans la salle opératoire.

Pham et Klinkert (2006) considèrent le problème d'ordonnancement en tenant compte des salles opératoires (non-identiques), de la salle de réveil, et d'autres ressources telles que l'unité des soins intensifs, les chirurgiens, les infirmiers, etc. Les auteurs modélisent le problème sous la forme d'un job-shop généralisé (multi-mode blocking job shop) avec comme objectif la minimisation du makespan Cmax (afin de maximiser l'utilisation des ressources) ; ils utilisent la programmation en nombres mixtes pour la résolution.

Dans (Fei *et al*., 2006) et (Fei *et al*., 2007) les auteurs proposent un algorithme génétique pour le problème d'ordonnancement pour minimiser le makespan aussi bien de la salle de réveil que celui des salles opératoires, avec une contrainte de blocage (le réveil peut commencer dans la salle opératoire). Dans un premier article (Fei *et al*., 2006), l'ordonnancement ne remet pas en cause l'affectation des patients aux salles opératoires, en supposant que cette affectation respecte déjà la disponibilité des chirurgiens, i.e. les interventions d'un chirurgien donné sont affectés à la même salle. Dans un deuxième article (Fei *et al*., 2007) l'affectation des patients aux salles peut être remise en cause et dans ce cas des contraintes additionnelles concernant les chirurgiens sont considérées, pour éviter que des interventions par le même chirurgien ne soient planifiées pour la même heure.

Cardoen *et al.* (2006) proposent un modèle multi-objectifs pour l'ordonnancement journalier d'un service de chirurgie ambulatoire. La fonction objectif à optimiser est une combinaison de plusieurs critères : le nombre de patients présent en même temps dans la salle de réveil, les préférences relatives aux patients de jeune age, prioritaires, etc. Nombreuses contraintes sont considérées dans le modèle : adéquation des salles opératoires, disponibilité des chirurgiens, des équipements médicaux, le nettoyage après certaines interventions infectieuses, et la nécessité de certains tests médicaux complémentaires. Les auteurs proposent une méthode exacte et une heuristique basée sur la programmation en nombres entiers pour résoudre le problème. Dans (Cardoen *et al.*, 2007), les mêmes auteurs développent une approche de génération de colonnes pour améliorer les performances de la méthode de résolution.

2. 4 Gestion des ressources humaines

Dans le bloc opératoire, travaillent des personnels médicaux et soignant hautement qualifiés. La gestion efficiente de ces personnels représente un défi majeur pour l'hôpital et les coûts engendrés par les personnels représentent une part importante pouvant aller jusqu'à 70% du budget global (Trilling, 2006).

2. 4. 1 Gestion quantitative

Les problèmes liés à la gestion quantitative des personnels peuvent être classés en trois catégories (Warner, 1976 ; Trilling, 2006) :

- Problème de "*Staffing*": c'est un problème de dimensionnement qui consiste à déterminer le nombre de personnes (en équivalent temps plein) nécessaire pour couvrir une charge de travail prévisionnelle.

- Problème de "*Shift scheduling*" (ou *rostering problem*) : c'est un problème de planification dont l'objectif est de former des vacations et de les affecter aux différents personnels.

- Problème de re-affectation : il consiste à réaffecter les personnels aux vacations suite à des perturbations (absence non prévue de personnel, retard pris sur activité, ou l'arrivée d'une activité non prévue etc.).

Le problème de "*Staffing*" :

Des nombreux travaux utilisent la simulation pour dimensionner les ressources humaines essentiellement dans les services d'urgence (Badri et Hollongsworth, 1993; Klafehn et Owens, 1987 ; Klafehn *et al*., 1989 ; Liyanage et Gale, 1995), les unités de soins ambulatoires (Wilt et Goddin, 1989; McHugh, 1989; Swisher *et al*., 1997).

Quelques travaux utilisent des approches de la programmation mathématique pour le problème de *staffing*. Kao et Tung (1981) proposent un modèle linéaire pour ce problème avec comme objectif la minimisation du coût des personnels tout en satisfaisant les prévisions de la demande. Trivedi (1981) utilise la programmation multi objectifs : minimiser les coûts, maximiser la satisfaction des patients, maintenir un certain pourcentage du personnel en temps-plein, et respecter les préférences du personnel. Venkataraman et Brusco (1996) proposent une approche intégrée pour le dimensionnement et la planification des personnels infirmiers.

Le problème de "*Shift scheduling*" :

Le problème de planification consiste à construire des *plannings individuels* des infirmiers pour un horizon donné afin de satisfaire une charge de travail prévisionnelle et des contraintes

sociales et réglementaires. Un planning individuel est une succession de jours travaillés (*days on*) et des jours de repos (*days off*) ; chaque jour travaillé contient une *vacation* tel que « matin », « après-midi », et nuit. Chaque vacation est associée à une heure de début et une heure de fin. En plus, chaque jour est discrétisé en *périodes*. Une période est un intervalle de temps pour lequel une charge de travail est définie.

La planification des personnels doit satisfaire plusieurs contraintes : de charge de travail (définie par un besoin fixé ou une fourchette de besoins), de succession d'activité (une vacation de nuit ne peut pas être suivi par une vacation de matin), de temps de travail, de repos hebdomadaire, de préférences personnelles. Selon l'approche utilisée, l'objectif peut être soit minimiser un coût (salaires, heures supplémentaires,…), maximiser la qualité de service (satisfaction de la demande), ou maximiser la satisfaction des personnels (minimiser le nombre de contraintes flexibles violées, maximiser l'équité entre les différents personnels).

Le problème de planification des personnels infirmiers (*Nurse Scheduling Problem*) est largement traité dans la littérature. Pour un état de l'art sur le sujet, nous faisons référence à (Cheang *et al*., 2003) et (Burke *et al*., 2004).

Ce problème est généralement de grande taille et par conséquent difficile à résoudre de manière exacte. Pour cette raison, plusieurs approches utilisant des heuristiques ou des technique d'intelligence artificielle ont été proposées dans la littérature. Toutefois, certains travaux exploitent les propriétés et la structure du problème pour développer des méthodes exactes efficaces de type *branch-and-bound* (Trivedi et Warner, 1976 ; Bosi et Milano, 2001 ; Ikegami et Niwa, 2003) ou *branch-and-price* (Mason et Smith, 1998 ; Jaumard et al., 1998 ; Bard et Purnomo, 2005a,b).

Les approches basées sur les techniques d'intelligence artificielle essayent de trouver une solution réalisable plutôt que l'optimisation d'un certain critère. Nous rencontrons des travaux utilisant la programmation par contrainte (Okada, 1992 ; Weil *et al*., 1995, Darmoni *et al*., 1995 ; Cheng *et al*., 1997) ou les systèmes experts (Ozkarahan, 1989 ; Chen etYeung, 1993 ; Scott et Simpson, 1998 ; Petrovic *et al*., 2003).

De nombreuses approches heuristiques et méta-heuristiques sont proposées pour une résolution approchée du problème de planification. Nous citons par exemple des travaux utilisant le recuit simulé (Isken et Hancock, 1991 ; Brusco et Jacobs, 1993), la recherche tabou (Dowsland, 1998 ; Burke *et al*., 2003 ; Bellanti *et al*., 2004), et les algorithme génétiques (Tanomaru, 1995 ; Aickelin et Dowsland, 2000 ; Jan *et al*., 2002 ; Aickelin et White, 2004).

Nous remarquons que la littérature est assez abondante en ce qui concerne les problèmes de planification des infirmiers (*Nurse Scheduling Problem*) des unités de soins (autre que le bloc opératoire). Cependant, peu sont les travaux qui traitent de la planification des personnels dans le bloc opératoire (Trilling, 2006 ; Beliën et Demeulemeester, 2007).

Les problèmes rencontrés dans les unités de soins et ceux rencontrés dans les blocs opératoires peuvent être complètement différents (Trilling, 2006). Dans les unités de soins, les activités sont nombreuses et morcelées dans le temps, les besoins en infirmiers sont évalués en fonction de la taille du service, du nombre de patients à soigner, et de leurs gravités. Ces besoins sont généralement approximatifs et exprimés par un besoin minimum et un besoin maximum. Si le niveau minimum n'est pas assuré (ce qui n'est pas souhaitable mais peut arriver), alors la charge supplémentaire est absorbée par l'effectif en place ce qui limite les conséquences sur le travail réalisé. Par contre dans le bloc opératoire, les besoins en infirmiers sont plus facilement quantifiables : une salle ne peut être ouverte que si un nombre donné d'infirmiers sont disponibles. Un sous-effectif n'est donc pas acceptable et conduirait à la fermeture de la salle et au report d'interventions. Dans ce sens, Trilling (2006, Chapitre 6) traite le problème de planification des infirmiers anesthésistes dans un bloc opératoire. L'objectif à maximiser est l'équité de la répartition de la charge entre les différents infirmiers, tout en respectant les réglementations concernant les temps de travail et les préférences personnelles (jours de repos, etc.).

Une autre caractéristique du bloc opératoire est que la charge des personnels est intimement liée au planning des interventions chirurgicales. Mais nous remarquons qu'il y a trop peu de travaux qui abordent ce problème. Beliën et Demeulmeester (2007) proposent une approche intégrée pour la construction des plannings d'infirmiers et l'affection des plages horaires (des salles opératoires) aux chirurgiens. Trilling (2006, Chapitre 7) propose une approche pour l'ajustement des plannings des médecins anesthésistes en fonction des interventions chirurgicales planifiées.

2. 4. 2 Gestion qualitative

En plus des approches quantitatives pour la gestion des ressources humaines, nous trouvons dans la littérature quelques travaux qualitatifs qui visent à améliorer la coordination des différents acteurs du bloc opératoire (chirurgiens, anesthésistes, infirmiers, etc.). L'objectif est d'éliminer les pertes du temps et favoriser les conditions d'une performance optimale du système. Par exemple, Overdyk *et al*. (1998) proposent une méthodologie de travail et une réflexion multidisciplinaire, afin de préciser le rôle et la responsabilité de chaque acteur tout au long du processus opératoire, dans le but d'améliorer l'efficacité du bloc opératoire. Dans le même contexte, Delesie (1998) propose trois étapes pour établir un rapprochement entre les cliniciens et les gestionnaires hospitaliers. Marty (2001) donne des propositions pour

coordonner les activités des acteurs pour une meilleure prestation avec un maximum de sécurité et un minimum de coût des blocs opératoires. De plus, Bleakley *et al*. (2004) proposent une démarche de formation pour améliorer la communication et favoriser la cohésion du groupe (les différents acteurs du bloc) dans un but d'offrir une meilleure qualité de service vis à vis de patient et éviter les erreurs médicales.

2. 5 Évaluation de performances

Afin d'assurer une gestion efficace du bloc opératoire, une étude des performances est nécessaire. L'objectif d'une telle étude peut être : l'évaluation des performance actuelles, l'évaluation d'une action correctrice implantée (Gallivan, 1998 ; Blake *et al*., 1995 ; Lapierre *et al*., 1999), l'identification des goulots d'étranglement (Everett, 2002), ou l'identification des actions correctrices à entreprendre (Dexter et Macario, 2002 ; Mcleod *et al*., 2003).

Plusieurs indicateurs sont utilisés pour évaluer la performance du bloc opératoire (Fei, 2006). Nous citons par exemple les coûts variables per-opératoires (Dexter *et al*., 2002), le cash-flow annuel (Féniès et Rodier, 2006), la durée d'attente moyenne, la durée d'attente maximale, le nombre d'interventions effectuées (Gallivan, 1998 ; Féniès et Rodier, 2006), le taux d'utilisation des salles opératoires (Strum *et al*., 1997), les coûts des programmes opératoires, l'allocation des budgets et le nombre d'équipes chirurgicales (Strum *et al*., 1997 ; Murphy et Sigal, 1985 ; Dexter et Macario, 2002).

Féniès *et al*. (2004) proposent de modéliser le processus opératoire en utilisant une méthodologie de modélisation de l'entreprise intégrée (telle que ARIS, ARchitecture of Integrated information Systems). Selon les auteurs, une telle modélisation permet de (i) déterminer les processus principaux et les facteurs clés de succès, (ii) déterminer les indicateurs de performance, et (iii) concevoir un tableau de bord. Besombes *et al*. (2006) proposent cinq types d'indicateurs pour évaluer l'activité du bloc opératoire ; à savoir les indicateurs : d'activités (nombre et durées d'interventions,…), économiques (ex : coût par intervention), de processus et de maîtrise des risques, de satisfaction des patients et des médecins, et d'apprentissage organisationnel (taux de retard des interventions, qualité des estimations des durées opératoires, etc.). Bounekkar *et al*. (2006) utilisent une approche AMDEC (Analyse des modes de défaillances, de leurs effets et de leurs criticité) et le diagramme cause-effet Ischikawa pour identifier les variables d'actions associées aux différents indicateurs de performance. Le diagramme cause-effet est utilisé pour recenser et classer les facteurs responsables d'un événement donné. L'AMDEC est par la suite utilisée pour caractériser et hiérarchiser ces différents facteurs.

L'évaluation des performances est généralement réalisée moyennant des modèles de simulation ou d'analyse statistique. Par exemple, Strum *et al*. (1997) proposent un modèle

pour évaluer le taux d'utilisation (sur et sous utilisation) des salles opératoires, analyser la qualité des programmes opératoires. En estimant la distribution de probabilité de la durée totale des interventions réalisées dans une salle, les auteurs proposent de déterminer la durée d'ouverture de la dite salle de manière à minimiser les coûts de sous et sur utilisations.

O'Neill et Dexter (2004) présentent une méthode basée sur la DEA « *data envelopment analysis* » pour identifier les meilleures pratiques utilisées par les services chirurgicaux dans les hôpitaux. Cette méthode identifie les interventions les plus (ou les moins) réalisées par les spécialités chirurgicales, et suggère des stratégies pour augmenter le volume des interventions réalisées pour les hôpitaux qui présentent des inefficacités. Ces résultats peuvent être utilisés par les managers pour l'aide à la décision concernant l'allocation des ressources, tel que le recrutement des chirurgiens pour certaines spécialités.

Nous trouvons aussi dans la littérature des travaux utilisant la simulation pour évaluer l'impact prévisionnel d'un changement d'organisation ou de pratiques. Kuzdrall *et al*. (1974) et Schmitz et Kwak (1972) utilisent la simulation Monte Carlo pour évaluer le taux d'utilisation des salles opératoires et de la salle de réveil en fonction d'une augmentation du nombre de lit d'hospitalisation disponibles. Garnett (1998) utilise aussi la simulation pour étudier le fonctionnement du bloc opératoire ; il découvre que, dans le cas étudié, la salle de réveil représente une ressource goulot qui pénalise le fonctionnement de tout le bloc. Murphy et Sigal (1985) évaluent le taux d'utilisation des salles opératoires et montre que ce dernier peut être amélioré en légèrement modifiant les heures d'ouverture des salles. Everett (2002) et Kusters et Groot (1996) proposent un système d'aide à la décision qui permet d'évaluer les performances du bloc et tester l'efficacité de différentes pratiques organisationnelles.

2. 6 Conclusion

Dans ce chapitre, nous avons présenté un panorama des approches de modélisation et de résolution employées dans la littérature pour résoudre différents problèmes soulevés par la gestion du bloc opératoire.

À partir de cette revue de l'état de l'art, nous constatons que le problème de planification des interventions chirurgicales a fait l'objet de nombreux travaux de recherches ; des approches complètes et complémentaires ont été proposées pour apporter une aide à la planification des interventions chirurgicales. Toutefois, nous remarquons que les approches développées ne permettent pas de prendre en compte les aléas qui peuvent survenir durant le fonctionnement du bloc opératoire.

Chapitre 3

Planification avec capacité agrégée

Dans ce chapitre, nous proposons un modèle stochastique pour la planification du bloc opératoire avec prise en compte des aléas dus à la chirurgie d'urgence. Deux catégories de patients sont opérés dans le bloc : les patients électifs et les patients urgents. Les patients électifs sont ceux qui peuvent être planifiés à l'avance. Chaque patient électif engendre un coût qui dépend de la date de son intervention. Les patients urgents arrivent de manière aléatoire durant la journée et doivent être opérés le jour même. Le problème de planification consiste à déterminer, pour un horizon donné, l'ensemble des patients électifs qui seront opérés chaque jour de manière à minimiser les coûts relatifs aux patients ainsi que les coûts de sur-utilisation des salles opératoires. Nous formulons ce problème sous la forme d'un programme mathématique stochastique et nous proposons deux méthodes de résolutions. La première méthode combine la simulation Monte Carlo et la programmation en nombres entiers. Nous étudions les propriétés de convergence de cette méthode et nous montrons que les solutions trouvées convergent de manière exponentielle vers de vraies solutions optimales. La deuxième méthode utilise la technique de relaxation Lagrangienne afin de décomposer le problème en plusieurs sous-problèmes ; un sous-problème par jour. Moyennant des tests numériques, nous évaluons et comparons les performances des deux méthodes d'optimisation et nous montrons que des gains significatifs peuvent être réalisés en utilisant un modèle de planification stochastique.

(Lamiri et Xie, 2006 ; Lamiri *et al*., 2006b ; Lamiri *et al*., 2007b)

3. 1 Introduction

Dans nombreux établissements hospitaliers, le bloc opératoire est utilisé pour assurer deux types d'activités : une activité programmée et une activité d'urgence. L'activité programmée représente la chirurgie réglée (les patients électifs) qu'on peut planifier à l'avance. L'activité d'urgence représente les patients qui arrivent de manière aléatoire durant la journée et qui nécessitent une intervention chirurgicale le jour même. Ces deux types d'activités ont des caractéristiques différentes, une est planifiable tandis que l'autre est par nature imprévisible et non planifiable, mais les deux mobilisent le même ensemble de ressources.

Une gestion efficiente du bloc opératoire est fortement liée à l'intégration du caractère incertain de la chirurgie d'urgence lors de la planification des interventions chirurgicales. Cependant, comme nous l'avons vu dans le chapitre précédent, les approches et modèles de planification existants ne tiennent pas compte de ce type d'incertitudes.

Dans ce chapitre, nous présentons un modèle stochastique pour la planification du bloc opératoire ; un modèle qui intègre l'aspect incertain de la chirurgie d'urgence. Au niveau de ce modèle de planification, nous nous intéressons à la détermination des dates d'intervention des patients. L'affectation des patients à des salles opératoires spécifiques n'est pas prise en compte. Par ailleurs, les salles opératoires sont supposées polyvalentes, une intervention peut être réalisée dans n'importe quelle salle opératoire, et seule la capacité agrégée de l'ensemble des salles opératoires est prise en compte.

Le problème de planification est formulé sous la forme d'un programme mathématique stochastique. Nous proposons deux méthodes de résolution :

- une méthode « presque » exacte qui combine la simulation Monte Carlo et la programmation en nombres mixtes,
- et une méthode heuristique basée sur la relaxation lagrangienne.

3. 2 Modèle et formulation mathématique

3. 2. 1 Modèle proposé

Nous considérons la planification des activités du bloc opératoire sur un horizon de planification discrétisé en H périodes, une période représente un jour. Deux types de patients sont considérés : les patients électifs qui sont connus et à programmer à l'avance sur les H jours, et les patients urgents, non programmables, qui arrivent de manière aléatoire et qui nécessitent un intervention chirurgicale le jour de leur arrivée.

Supposons qu'il y ait N patients électifs. Pour chaque patient électif $i \in \{1,\ldots, N\}$, nous connaissons :

- d_i, la durée opératoire : c'est la durée d'intervention du patient i ; elle inclut le temps de préparation, l'intervention chirurgicale, et la durée de nettoyage et de reconditionnement de la salle;
- e_i, date au plus tôt de l'intervention i ; le patient ne peut pas être opéré avant cette date.

Pour chaque patient électif i, nous disposons d'un ensemble de coûts a_{it} ($t \in \{1,\ldots, H, H+1\}$. a_{it} représente le coût de réalisation de l'intervention du patient i en jour t. Une période fictive $H+1$ est ajoutée à l'horizon de planification afin de regrouper les patients électifs qui ne seront pas programmés dans l'horizon actuel. Le coût $a_{i(H+1)}$ représente alors le coût de la non-programmation de l'intervention i. Cet ensemble des coûts peut représenter des « vrais » coûts financiers, ainsi que des préférences.

Au niveau de ce chapitre, le problème de planification consiste à déterminer les dates d'intervention des patients. L'affectation des patients à des salles opératoires spécifiques n'est pas prise en compte. Par ailleurs, nous supposons que les salles opératoires sont identiques, une intervention peut être réalisée dans toute salle opératoire, et seule la capacité globale de l'ensemble des salles opératoires est prise en compte.

Soit T_t la capacité régulière en nombre d'heures des salles opératoires pour le jour t. Si les interventions programmées et les interventions urgentes (non programmées) du jour t dépassent cette capacité, des coûts d'heures supplémentaires sont engendrés. Soit c_t le coût pour une heure supplémentaire au jour t.

Les patients urgents arrivent de manière aléatoire et doivent être pris en charge le jour de leur arrivée. Plus précisément, les interventions urgentes d'un jour donné doivent être réalisées en plus des interventions planifiées pour cette journée, quelle que soit la capacité régulière disponible. Soit W_t la durée totale de toutes les interventions urgentes réalisées le jour t. Cette durée est considérée comme une variable aléatoire. Nous supposons que sa distribution $f_{W_t}(.)$ peut être estimée à partir des données historiques.

Le problème de planification consiste à déterminer les dates d'intervention des patients de manière à minimiser les coûts liés aux dates d'intervention des patients électifs ainsi les coûts moyens des heures supplémentaires associées à l'utilisation des salles opératoires.

3. 2. 2 Formulation mathématique

Les variables de décision sont $x_{it} \in \{0, 1\}$ avec $x_{it} = 1$ si le patient i est affecté à la période (jour) t, et 0 sinon. Par convention $x_{i(H+1)} = 1$ signifie que le patient i est affecté à la période fictive H+1, c'est-à-dire il ne sera pas opéré durant l'horizon actuel.

Le problème de planification peut être formulé de la manière suivante :

(P) $$J^* = \text{Minimiser } J(X) = \sum_{i=1}^{N} \sum_{t=e_i}^{H+1} a_{it}\, x_{it} + \sum_{t=1}^{H} c_t\, O_t \qquad (3.1)$$

sous contraintes:

$$O_t = E_{W_t}\left[\left(W_t + \sum_{i=1}^{N} d_i x_{it} - T_t\right)^+\right], \ \forall\, t = 1,\ldots, H \qquad (3.2)$$

$$\sum_{t=e_i}^{H+1} x_{it} = 1, \ \forall\, i = 1,\ldots, N \qquad (3.3)$$

$$x_{it} \in \{0, 1\}, \ \forall\, i = 1,\ldots, N, \ \forall\, t = 1,\ldots, H, H+1 \qquad (3.4)$$

où $E_{W_t}[.]$ est l'espérance mathématique, elle est déterminée par rapport à la distribution de W_t, et $(y)^+ = \max\{0, y\}$.

Dans cette formulation, le critère $J(X)$ à minimiser représente le coût total du planning X, c'est la somme des coûts relatifs aux dates d'intervention des patients et les coûts des dépassements horaires moyens. Les contraintes (3.2) définissent formellement le dépassement horaire moyen O_t en chaque jour t. Les contraintes (3.3) garantissent que chaque patient électif est soit planifié pour un jour t soit non planifié (affecté à la période fictive H+1).

Le problème de planification (P) est un problème stochastique (non linéaire) combinatoire.

3. 2. 3 Étude de complexité

Dans cette section, nous étudions la complexité du problème de planification.

Théorème 3.1 : Le problème de planification (3.1)-(3.4) est un problème NP-difficile au sens fort.

Preuve : La preuve est basée sur une transformation polynomiale du problème de 3-*Partition*. Le problème de 3-*Partition*, connu comme étant NP-difficile au sens fort, est défini comme suit (Gray et Janson, 1979) : Étant donné un ensemble $A = \{a_i\}_{1 \le i \le 3z}$ de 3z entiers tel que $\sum_{i=1}^{3z} a_i = zB$ et $B/4 < a_i < B/2$. Existe-t-il une partition de A en z triplets disjoint $A_1,\ldots, A_z$,

tel que $\sum_{i\in A_t} a_i = B$ pour $1\leq t \leq z$? Une réponse affirmative à cette question signifie que le problème de 3-Partition admet une solution.

Afin de prouver la NP complexité du problème de planification (P), nous définissons une transformation polynomiale du problème de 3-Partition en un problème de décision : existe-t-il un planning (solution) faisable X^* pour le problème de planification tel que $J(X^*)= 0$?

À chaque instance du problème de 3-Partition, nous associons une instance (P1) du problème de planification ; cette instance est définie comme suit. L'horizon H égal à z, et le nombre des patients électifs N est égal à $|A|$. Pour chaque période t, la capacité régulière, la durée totale des interventions urgentes, les coûts des heures supplémentaires sont comme suit :

$$T_t = B,\ W_t = 0,\ c_t = 1,\ \forall\ t\in\{1,\ldots, H\}.$$

Le coût relatif à chaque patient électif i ($i\in\{1,\ldots, |A|\}$) est comme suit : $a_{it} = 0$, $\forall\ t\in\{1,\ldots, H\}$ et $a_{i(H+1)}=1$. Pour chaque patient i, la date au plus tôt est $e_i = 1$ et la durée d'intervention est $d_i = a_i$.

Maintenant, nous montrons que le problème de planification (P1) admet un planning faisable X^* avec $J(X^*)=0$, si et seulement si le problème de 3-Partition associé admet une solution.

Condition suffisante : Supposons que le problème de 3-Partition a une solution $\bigcup_{1\leq t\leq z} A_t = A$, avec $\sum_{i\in A_t} a_i = B$ pour $1\leq t \leq z$. Considérons le planning X^* construit de la manière suivante : les patients regroupés en un sous-ensemble A_t sont planifiés pour la période t. Comme $\sum_{i\in A_t} a_i = B$ (pour $1\leq t \leq z$), il est évident que $J(X^*)= 0$.

Condition nécessaire : Supposons qu'il existe un planning X^* avec $J(X^*)= 0$. Désignons par A_t l'ensemble des patients planifiés en période t ($\forall\ t\in\{1,\ldots, H\}$). Comme $B/4 < a_i$ pour tout i, donc chaque sous-ensemble A_t contient exactement 3 éléments. Puisque $J(X^*)= 0$, alors on a $\bigcup_{1\leq t\leq z} A_t = A$ et $\sum_{i\in A_t} a_i = B$ pour $1\leq t \leq z$. Par conséquent, le problème de 3-Partition a une solution. □

Le problème de planification (3.1)-(3.4) demeure NP-difficile même pour un nombre fixe de périodes. Avec un nombre de périodes égal à deux, le problème reste combinatoire par rapport à N.

Théorème 3.2 : Le problème de planification (3.1)-(3.4) avec un horizon de deux périodes (H=2) est un problème NP-difficile.

Preuve : La preuve est basée sur une transformation polynomiale du problème de 2-*Partition*, problème connu comme étant NP-difficile, est qui se défini comme suit (Gray et

Janson, 1979) : Étant donné un ensemble d'entiers $A = \{a_i\}$ tel que $\sum_{i \in A} a_i = 2B$. Existe-t-il une partition de A en deux sous-ensembles disjoints A_1 et A_2 tel que $\sum_{i \in A_1} a_i = \sum_{i \in A_2} a_i = B$?

À chaque instance du problème de 2-Partition, nous associons une instance (P2) du problème de planification ; cette instance est définie de la même manière que (P1), définie dans la preuve du théorème 3.1. La seule différence est que l'horizon H égal à 2.

Instance (P2) : $H = 2$, $T_t = B$, $W_t = 0$, $c_t = 1$, $\forall\ t \in \{1,\ldots, H\}$. Le nombre des patients électifs N est égal à $|A|$. Pour chaque patient électif i ($i \in \{1,\ldots, |A|\}$), nous avons (la date au plus tôt) $e_i = 1$, (la durée d'intervention) $d_i = a_i$., et les coûts $a_{it} = 0$, $\forall\ t \in \{1,\ldots, H\}$ et $a_{i(H+1)} = 1$.

Problème de décision : Existe-t-il un planning X^* pour le problème (P2) tel que $J(X^*) = 0$?

Nous montrons que (P2) a une solution si et seulement si le problème de 2-Partition associé a une solution. Il est évident qu'un planning X^* avec $J(X^*) = 0$ existe si et seulement si le problème de 2-Partition a une solution. Ceci peut être facilement prouvé en utilisant les mêmes arguments que ceux de la preuve du théorème 3.2. □

Pour récapituler, le problème de planification est un problème NP-difficile au sens fort à cause de sa combinatoire. À cette difficulté s'ajoute la non linéarité due aux espérances mathématiques.

3. 2. 4 Évaluation de la fonction économique

Dans cette section nous présentons deux méthodes pour l'évaluation de la fonction économique (critère) $J(X)$ pour un planning X donné.

Étant donné un planning X, désignons par D_t, $t \in \{1,\ldots,H\}$, la durée totale des interventions électives planifiées pour la période t, concrètement $D_t = \sum_{i=1}^{N} d_i x_{it}$. L'évaluation du critère $J(X)$ peut être réalisée moyennant l'une des deux méthodes suivantes : méthode analytique ou méthode par simulation Monte Carlo.

La première méthode est une méthode analytique qui fournit une évaluation exacte du critère. Pour chaque période t, le dépassement horaire moyen est déterminé de manière exacte -par rapport à la distribution de W_t - par évaluation de l'espérance mathématique :

$$O_t = E_{W_t}\left[\left(W_t + D_t - T_t\right)^+\right] = \int_{T_t - D_t}^{+\infty} \left(w + D_t - T_t\right) f_{W_t}(w)\, dw$$

Le critère est alors de la forme suivante :

$$J(X) = \sum_{i=1}^{N} \sum_{t=e_i}^{H+1} a_{it} x_{it} + \sum_{t=1}^{H} c_t \int_{T_t - D_t}^{+\infty} \left(w + D_t - T_t \right) f_{W_t}(w)\, dw$$

Notons que le dépassement horaire moyen est une espérance dont l'évaluation nécessite une intégration numérique, sauf pour quelques lois de distribution où l'on dispose d'une expression explicite.

La deuxième méthode est basée sur la simulation Monte Carlo ; elle fournit une estimation du critère. Dans un premier temps, L échantillons indépendants identiquement distribués W_t^1, …, W_t^L sont aléatoirement générés pour chaque W_t ($t \in \{1,\ldots,H\}$). Ensuite, pour chaque période t, le dépassement horaire moyen est estimé par une moyenne empirique basée sur les échantillons précédemment générés :

$$O_t \approx O_{tL} = (1/L) \sum_{l=1}^{L} \left(W_t^l + D_t - T_t \right)^+$$

Le critère est alors estimé comme suit :

$$J(X) \approx J_L(X) = \sum_{i=1}^{N} \sum_{t=e_i}^{H+1} a_{it} x_{it} + \sum_{t=1}^{H} c_t O_{tL}$$

Remarquons que par la loi des grands nombres O_{tL} converge vers O_t lorsque le nombre d'échantillons L croit. Autrement, en utilisant un nombre de scénarios élevé on peut obtenir une estimation assez précise du critère.

3. 3 Optimisation Monte Carlo

Le principe de simulation *Monte Carlo* consiste à utiliser un échantillon discret obtenu par simulation pour approximer une distribution de probabilité que l'on ne sait pas calculer analytiquement (Hammersley et Handscomb, 1964; Calos et Whitlock, 1986; Decker, 1991). Typiquement, cet échantillon pourra être utilisé pour approximer des espérances mathématiques par des moyennes empiriques. C'est ce principe que nous mettons en œuvre pour développer une méthode de résolution pour le problème de planification.

3. 3. 1 La méthode d'optimisation

L'idée de base de cette méthode est d'utiliser la simulation Monte Carlo pour approximer le problème de planification stochastique (P) par un problème d'optimisation déterministe. Cette approche est connue dans la littérature sous différents noms, tels que « *sample average approximation method* » (Kleywegt *et al*., 2001), « *stochastic counterpart optimization* », et « *sample-path optimization* » (Gürkan *et al*., 1999, Plambeck *et al*., 1996).

Concrètement, on génère, pour chaque période $t \in \{1,\ldots,H\}$, K échantillons aléatoires indépendants et identiquement distribués (i.i.d) $W_t^1, \ldots, W_t^K$ pour chaque variable aléatoire W_t.

Chaque vecteur $[W_1^k, \ldots, W_H^k]$ représente un scénario, $k \in \{1,\ldots, K\}$. C'est aussi un échantillon aléatoire (i.i.d) du vecteur aléatoire $W = [W_1, \ldots, W_H]$.

En utilisant les scénarios générés, les espérances mathématiques (3.2) évaluant les dépassements horaires peuvent être approximées par des moyennes empiriques. Avec cette estimation le problème de planification (P) peut être approximé par le problème déterministe suivant :

$$(\mathrm{P_K}) \qquad J_K^* = \text{Minimiser } J_K(X) = \sum_{i=1}^{N}\sum_{t=e_i}^{H+1} a_{it}\, x_{it} + \sum_{t=1}^{H} c_t\, O_{tK} \tag{3.5}$$

sous contraintes:

$$O_{tK} = \frac{1}{K}\sum_{k=1}^{K}\left(W_t^k + \sum_{i=1}^{N} d_i x_{it} - T_t \right)^+, \ \forall\, t = 1,\ldots, H \tag{3.6}$$

$$\sum_{t=e_i}^{H+1} x_{it} = 1, \ \forall\, i = 1,\ldots, N \tag{3.7}$$

$$x_{it} \in \{0, 1\}, \forall\, i = 1,\ldots, N, \ \forall\, t = 1,\ldots, H, H+1 \tag{3.8}$$

Le critère $J_K(X)$ à minimiser est maintenant un *estimateur* du critère exact $J(X)$. Les contraintes (3.6) fournissent une estimation des heures supplémentaires en chaque période t, en se basant sur les scénarios générés.

Le problème approximé $(\mathrm{P_K})$ peut être reformulé sous la forme du problème linéaire suivant :

$$(\mathrm{P_K}) \qquad J^* = \text{Minimiser } J(X) = \sum_{i=1}^{N}\sum_{t=e_i}^{H+1} a_{it}\, x_{it} + \sum_{t=1}^{H} c_t\, O_{tK} \tag{3.9}$$

sous contraintes:

$$O_{tK} = \frac{1}{K}\sum_{k=1}^{K} O_t^k, \ \forall\, t = 1,\ldots, H \tag{3.10}$$

$$O_t^k \geq W_t^k + \sum_{i=1}^{N} d_i x_{it} - T_t, \forall\, t = 1,\ldots, H, \ \forall\, k = 1,\ldots, K \tag{3.11}$$

$$\sum_{t=e_i}^{H+1} x_{it} = 1, \ \forall\, i = 1,\ldots, N \tag{3.12}$$

$$x_{it} \in \{0, 1\}, \ \forall\, i = 1,\ldots, N, \ \forall\, t = 1,\ldots, H, H+1 \tag{3.13}$$

$$O_t^k \geq 0, \forall\, t = 1,\ldots, H, \ \forall\, k = 1,\ldots, K \tag{3.14}$$

(P_K) est un problème de programmation linéaire en variables mixtes, qui peut être résolu en utilisant une des méthodes d'optimisation déterministe classiques, tel que *branch and bound* ou *branch and cut*.

Le problème de planification (P) peut alors être résolu par l'algorithme suivant :

- Étape 1. Générer, pour chaque période *t*, *K* échantillons aléatoires indépendants W_t^1, …, W_t^K de W_t la capacité aléatoire utilisée par les urgences.
- Étape 2. Résoudre le problème approximé (P_K). Soit X_K^* la solution optimale obtenue.
- Étape 3. Évaluer la valeur du critère exact $J\left(X_K^*\right)$ (le coût exact) en utilisant l'une des méthodes présentées dans la section précédentes (c'est-à-dire analytiquement ou par simulation).

Notons que si le critère exact est évalué moyennant la simulation Monte Carlo, c-à-d $J\left(X_K^*\right) \approx J_L\left(X_K^*\right)$, alors les L scénarios utilisés pour l'évaluation sont générés indépendamment de ceux utilisés pour l'optimisation et leur nombre est généralement beaucoup plus élevé que *K*.

La solution optimale X_K^* du problème approximé représente une estimation de la solution optimale du problème exact (P). Elle peut être aussi considérée comme une variable aléatoire ; dans le sens où si on génère un autre ensemble de scénarios, on peut trouver une autre solution optimale.

3. 3. 2 Propriétés de convergence

La valeur optimale $J_K^* := J_K\left(X_K^*\right)$ et la solution optimale X_K^* du problème approximé (P_K) représentent des estimateurs de leurs contreparties J^* et X^* du problème exact (P). Dans cette section, nous étudions les propriétés de convergence de ces deux estimateurs.

Le critère exact $J(X)$ peut être re-exprimé de la manière suivante :

$$J\left(X\right) = E\left[G\left(X, W\right)\right]. \tag{3.15}$$

où $G(X,W) = \sum_{i,t} a_{it} x_{it} + \sum_t c_t \left(W_t + \sum_i x_{it} d_i - T_t\right)^+$, $W = [W_1, \ldots, W_H]$ est un vecteur de variables aléatoires, et par conséquent $G(X, W)$ est une variable aléatoire.

Les scénarios $W^k = [W_1^k, \ldots, W_H^k]$, $k \in \{1,\ldots,K\}$, représentent des échantillons aléatoires i.i.d du vecteur aléatoire *W*, et $G(X, W^k)$ représentent des échantillons aléatoires i.i.d de $G(X, W)$. Le critère estimé est alors

$$J_K(X) = \frac{1}{K}\sum_{k=1}^{K} G\left(X, W^k\right) \tag{3.16}$$

$$= \frac{1}{K}\sum_{k=1}^{K}\left[\sum_{i,t} a_{it}x_{it} + \sum_t c_t\left(W_t^k + \sum_i x_{it}d_i - T_t\right)^+\right]$$

Désignons par S l'ensemble des solutions faisables du problème exact (P) et par S^* l'ensemble des solution optimales. Le problème approximatif (P_K) consiste à identifier une solution X_K^* qui minimise le critère estimé $J_K(X)$, c'est-à-dire :

$$X_K^* = \underset{X\in\{X^1, X^2, \ldots X^m\}}{\operatorname{argmin}} J_K(X)$$

Notons que le critère estimé utilise toujours le même ensemble d'échantillons W_t^k ($t\in\{1,\ldots, H\}$, $k\in\{1,\ldots, K\}$) , c'est-à-dire le même flux de nombres aléatoires. Cette stratégie est généralement connue sous le nom de *nombres aléatoires communs* (Commun Random Numbers scheme, *CRN*) ; c'est l'une des techniques les plus utilisées pour la réduction de la variance en simulation (Law et Kelton, 1991).

Théorème 3.3 : Lorsque K tend vers l'infini,

- (i) la valeur optimale estimée J_K^* converge avec une probabilité de 1 vers la valeur optimale exacte J^*.
- (ii) la solution optimale estimée X_K^* converge avec une probabilité de 1 vers une solution optimale exacte X^*.

Preuve : Dans un premier temps nous allons montrer que le critère estimé $J_K(X)$ converge vers le critère exact $J(X)$, avec probabilité 1, pour toute solution faisable $X\in S$.

Soit $X\in S$ une solution faisable. Comme $G(X, W^k)$, $k\in\{1,\ldots,K\}$, sont des échantillons aléatoires i.i.d de $G(X, W)$, donc selon la loi des grand nombres, $(1/K)\sum_{k=1}^{K} G\left(X, W^k\right)$ converge avec probabilité 1 vers $E_W[G(X, W)]$ lorsque K tends vers l'infini. Par conséquent, pour tout $\varepsilon>0$, il existe un entier $\bar{K}_X > 0$ tel que

$$\left|J_K(X) - J(X)\right| = \left|\frac{1}{K}\sum_{k=1}^{K} G\left(X, W^k\right) - E_W\left[G\left(X, W\right)\right]\right| < \varepsilon,\ \forall K \geq \bar{K}_X .$$

Soit $\bar{K} = \max_{X\in S} \bar{K}_X$. Donc pour tout $X\in S$, on a :

$$\left|J_K(X) - J(X)\right| < \varepsilon,\ \forall K \geq \bar{K} . \tag{3.17}$$

C'est-à-dire, pour toute solution faisable X, le critère estimé $J_K(X)$ converge avec probabilité 1 vers le critère exact $J(X)$, lorsque K croit. Donc,

$$\left| J_K\left(X_K^*\right) - J\left(X_K^*\right) \right| \le \varepsilon\,,\ \forall K \ge \bar{K}\,.$$

En plus, pour tout $K > \bar{K}$,

$$\begin{aligned}
\left| J_K\left(X_K^*\right) - J_K\left(X^*\right) \right| &= J_K\left(X^*\right) - J_K\left(X_K^*\right) \quad (\text{Car } X_K^* \text{ est la solution optimale du } (P_K)) \\
&\le J_K\left(X^*\right) - J_K\left(X_K^*\right) + J\left(X_K^*\right) - J\left(X^*\right) \quad (\text{Car } X^* \text{ est la solution optimale du (P)}) \\
&\le \left| -J_K\left(X_K^*\right) + J\left(X_K^*\right) \right| + \left| J_K\left(X^*\right) - J\left(X^*\right) \right| \\
&\le 2\,\varepsilon. \qquad (3.18)
\end{aligned}$$

Il en découle alors :

$$\begin{aligned}
\left| J_K\left(X_K^*\right) - J\left(X^*\right) \right| &= \left| J_K\left(X_K^*\right) - J_K\left(X^*\right) + J_K\left(X^*\right) - J\left(X^*\right) \right| \\
&\le \left| J_K\left(X_K^*\right) - J_K\left(X^*\right) \right| + \left| J_K\left(X^*\right) - J\left(X^*\right) \right| \\
&\le 3\,\varepsilon. \qquad (\text{À partir de (3.17) et (3.18)})
\end{aligned}$$

En plus nous avons

$$\begin{aligned}
\left| J\left(X_K^*\right) - J\left(X^*\right) \right| &= J\left(X_K^*\right) - J\left(X^*\right) \quad (\text{Car } X^* \text{ est la solution optimale du (P)}) \\
&\le J\left(X_K^*\right) - J\left(X^*\right) - J_K\left(X_K^*\right) + J_K\left(X^*\right) \quad (\text{Car } X_K^* \text{ est la solution optimale du } (P_K)) \\
&\le \left| J\left(X_K^*\right) - J_K\left(X_K^*\right) \right| + \left| -J\left(X^*\right) + J_K\left(X^*\right) \right| \\
&\le 2\,\varepsilon. \quad (\text{À partir de (3.17)})
\end{aligned}$$

Pour récapituler, pour tout $\varepsilon > 0$, il existe un entier $\bar{K}$ fini tel que pour tout $K > \bar{K}$,

$$\left| J_K\left(X_K^*\right) - J\left(X^*\right) \right| \le 3\,\varepsilon\,, \qquad (3.19)$$

$$\left| J\left(X_K^*\right) - J\left(X^*\right) \right| \le 2\,\varepsilon\,. \qquad (3.20)$$

L'inégalité (3.19) signifie que, lorsque K croit, le coût estimé $J_K(X_K^*)$ de la solution optimale estimée converge avec probabilité 1 vers le coût optimal exact $J(X^*)$. L'inégalité (3.20) signifie que le coût exact de la solution optimale estimée converge avec probabilité 1 vers le coût optimal exact, lorsque K croit. Donc, X_K^* converge vers une solution optimale $X^* \in S^*$. □

Dans le reste de cette section, nous montrons que la solution optimale estimée X_K^* converge avec une vitesse exponentielle vers une vraie solution optimale du problème exact (P), et que la vitesse de convergence est maximale à cause de l'utilisation du « *commun random numbers* ». Pour ce faire, nous commençons par introduire quelques résultats qui nous seront utiles par la suite.

Considérons une séquence $\{y_n, \forall\ n \geq 1\}$ de variables aléatoires indépendantes identiquement distribuées et $Y_n = \sum_{i=1}^{n} y_i$. Introduisons maintenant :

- la fonction génératrice des moments : $M(\lambda) = E[exp(\lambda\, y_1)]$,
- la fonction génératrice des cumulants : $\Lambda(\lambda) = log\ M(\lambda)$,
- et la transformée de Fenchel-Legendre de $\Lambda(\lambda)$: $\Lambda^*(z) = sup_{\lambda \geq 0}\{\lambda z - \Lambda(\lambda)\}$.

Lemme 3.1 (Dembo et Zeitouni, 1993 ; Lemme 2.2.5) : Si $M(\lambda)$ existe dans un voisinage $(-\varepsilon, \varepsilon)$ de $\lambda = 0$ pour un $\varepsilon > 0$, alors

- $\Lambda(\lambda)$ est une fonction convexe
- $\Lambda(\lambda)$ est différentiable sur $(-\varepsilon, \varepsilon)$ et $\Lambda'(0) = E[y_1]$

Lemme 3.2 (Xie, 1997 ; Lemme 4.3) : Si $M(\lambda)$ existe dans un voisinage $(-\varepsilon, \varepsilon)$ de $\lambda = 0$ pour un $\varepsilon > 0$, alors

- $P[Y_n / n \geq z] \leq \exp(-n\, \Lambda^*(z)), \forall\ z \geq E[y_1]$
- $P[Y_n / n \leq z] \leq \exp(-n\, \Lambda^*(z)), \forall\ z \leq E[y_1]$

Théorème 3. 4 (Dai et Chen, 1997, Théorème 2.1) : Soit Z_i, i=1, 2,…, n, des variables aléatoires et $g(Z_1, Z_2,\ldots, Z_n)$ une fonction continue à droite et superadditive. L'espérance $E[g(Z_1, Z_2,\ldots, Z_n)]$ est maximisée en échantillonnant tout les Z_i selon le « commun random numbers ».

Supposons que le problème (P) admet m solutions faisables $S = \{X^1, X^2, \ldots, X^m\}$, et parmi elles il y a n solutions optimales. Sans perte de généralités, nous supposons que

$$J\left(X^1\right) = J\left(X^2\right) = ... = J\left(X^n\right) < J\left(X^{n+1}\right) \leq ... \leq J\left(X^m\right).$$

Rappelons que pour une solution faisable X donnée, $\{G(X, W^k), k \geq 1\}$ est une séquence de variables aléatoires indépendantes identiquement distribuées (i.i.d), et $E[G(X, W^k)] = J(X)$.

Pour tout $(u, v) \in (\{1,\ldots,n\},\{n+1,\ldots,m\})$, $\{G(X^u,W^k) - G(X^v,W^k), k \geq 1\}$ est une séquence de variables aléatoires i.i.d.

Définissons maintenant, pour tout $(u, v) \in (\{1,\ldots,n\},\{n+1,\ldots,m\})$, une fonction génératrice des moments $M_{uv}(\lambda) = E[exp(\lambda\,(G(X^u,W^1) - G(X^v,W^1)))]$, une fonction génératrice des cumulants $\Lambda_{uv}(\lambda) = \log M_{uv}(\lambda)$ et sa transformée de Fenchel-Legendre $\Lambda^*_{uv}(z) = \sup_{\lambda \geq 0}\{\lambda z - \Lambda_{uv}(\lambda)\}$.

Hypothèse (A) : Pour tout$(u, v) \in (\{1,\ldots,n\},\{n+1,\ldots,m\})$, $M_{uv}(\lambda)$ existe dans un voisinage $(-\varepsilon, \varepsilon)$ de $\lambda = 0$ pour un $\varepsilon > 0$.

Lemme 3.3 : Sous l'hypothèse (A), si X^u et X^v sont deux solutions faisables avec $J\left(X^u\right) < J\left(X^v\right)$, alors

$$P\left[J_K\left(X^u\right) > J_K\left(X^v\right)\right] \leq \exp\left(-K\ \Lambda_{uv}^*\right)$$

où $\Lambda_{uv}^* = \Lambda_{uv}^*(0) = \sup_{\lambda \geq 0}\ \left\{-\log M_{uv}(\lambda)\right\}$.

En d'autres termes; la probabilité $P\left[J_K\left(X^u\right) > J_K\left(X^v\right)\right]$ converge vers zéro avec une vitesse exponentielle.

Preuve : Grâce à l'équation (3.16), nous avons

$$P\left[J_K\left(X^u\right) > J_K\left(X^v\right)\right] = P\left[\frac{1}{K}\sum_{k=1}^{K}\left(G(X^u,W^k) - G(X^v,W^k)\right) > 0\right].$$

Comme $\{G(X^u,W^k) - G(X^v,W^k),\ k \geq 1\}$ est une séquence de variables aléatoires i.i.d et $E[G(X^u,W^1) - G(X^v,W^1)] < 0$, donc par application du Lemme 3.2, nous obtenons

$$P\left[\frac{1}{K}\sum_{k=1}^{K}\left(G(X^u,W^k) - G(X^v,W^k)\right) > 0\right] \leq \exp\left(-K\ \Lambda_{uv}^*(0)\right),$$

ce qui achève la preuve. □

Théorème 3.5 : Sous l'hypothèse (A), il existe des constantes positive $\gamma > 0$, $\eta > 0$ tel que

$$P\left(X_K^* \in S^*\right) \geq 1 - \gamma e^{-\eta K}$$

Preuve :

$$\begin{aligned} P\left(X_K^* \notin S^*\right) &= P\left[\min_{u=1..n} J_K\left(X^u\right) > \min_{v=n+1...m} J_K\left(X^v\right)\right] \\ &\leq \sum_{v=n+1}^{m} P\left[\min_{u=1..n} J_K\left(X^u\right) > J_K\left(X^v\right)\right] \\ &\leq \sum_{v=n+1}^{m}\prod_{u=1}^{n} P\left[J_K\left(X^u\right) > J_K\left(X^v\right)\right] \end{aligned}$$

En appliquant le lemme 3.3, nous obtenons :

$$\begin{aligned} P\left(X_K^* \notin S^*\right) &\leq \sum_{v=n+1}^{m}\prod_{u=1}^{n} \exp\left(-K\ \Lambda_{uv}^*\right) \\ &= \sum_{v=n+1}^{m} \exp\left(-K\ \sum_{u=1}^{n}\Lambda_{uv}^*\right) \\ &\leq (m-n)\ \exp\left(-K\ \min_{v=n+1...m}\left(\sum_{u=1}^{n}\Lambda_{uv}^*\right)\right) \end{aligned}$$

Introduisons maintenant

$$\eta = \min_{v=n+1\ldots m}\left(\sum_{u=1}^{n}\Lambda_{uv}^{*}\right) \qquad (3.21)$$

Il en découle que

$$P\left(X_K^{*}\in S^{*}\right)=1-P\left(X_K^{*}\notin S^{*}\right)\ \geq 1-\gamma\ \exp\left(-K\ \eta\right)$$

avec $\gamma = m-n>0$. Il nous faut maintenant prouver que $\eta>0$.

Soit $(u, v)\in(\{1,\ldots,n\},\{n+1,\ldots,m\})$, $\Lambda_{uv}(\lambda)$ est une fonction convexe et différentiable sur $(-\varepsilon, \varepsilon)$ avec $\Lambda'_{uv}(0) = E[G(X^u,W^1) - G(X^v,W^1)]$, d'après le lemme 3.1.

Comme $\Lambda_{uv}(0) = 0$ et $\Lambda'_{uv}(0)< 0$, (car $E[G(X^u,W^1) - G(X^v,W^1)] < 0$), donc pour tout $(u, v)\in(\{1,\ldots,n\},\{n+1,\ldots,m\})$

$$\Lambda_{uv}^{*}=\sup_{\lambda\geq 0}\left\{-\Lambda_{uv}\left(\lambda\right)\right\}>0$$

Il en résulte alors que $\eta = \min_{v=n+1\ldots m}\left(\sum_{u=1}^{n}\Lambda_{uv}^{*}\right)>0$ □

Théorème 3.6 : Sous l'hypothèse (A), le taux de convergence η de $P\left(X_K^{*}\in S^{*}\right)$ est maximisé grâce à l'utilisation de « *commun random numbers* ».

Preuve : Rappelons tout d'abord que :

$$\eta = \min_{v=n+1\ldots m}\left(\sum_{u=1}^{n}\Lambda_{uv}^{*}\right)$$

Nous allons commencer par prouver que, pour tout $(u, v)\in(\{1,\ldots,n\},\{n+1,\ldots,m\})$, Λ_{uv}^{*} est maximisé en utilisant le « commun random numbers ». D'après la définition de Λ_{uv}^{*}, nous avons

$$\Lambda_{uv}^{*}=\sup_{\lambda\geq 0}\left\{-\Lambda_{uv}\left(\lambda\right)\right\}$$

avec $\Lambda_{uv}\left(\lambda\right)=\log M_{uv}\left(\lambda\right)=\log\ E\left[\exp\left(\lambda\left(G(X^{u},W^{1})-G(X^{v},W^{1})\right)\right)\right]$

Donc, maximiser Λ_{uv}^{*} revient à minimiser $\Lambda_{uv}\left(\lambda\right)$, et par conséquent minimiser $E\left[\exp\left(\lambda\left(G(X^{u},W^{1})-G(X^{v},W^{1})\right)\right)\right]$.

Définissons une fonction $g(,)$ comme suit :

$$g(a, b) = -\exp(\lambda(a - b)).$$

On a alors

$$E\left[\exp\left(\lambda\left(G(X^u,W^1)-G(X^v,W^1)\right)\right)\right]=-E\left[g\left(G(X^u,W^1),G(X^v,W^1)\right)\right].$$

On peut facilement montrer que que $g(.,.)$ est une fonction superadditive. Donc, d'après le Théorème 3.4, échantillonner $G(X^u,W^1)$ et $G(X^v,W^1)$ en utilisant le « commun random numbers » permet de maximiser

$$E\left[g\left(G(X^u,W^1),G(X^v,W^1)\right)\right],$$

et par conséquent minimiser

$$E\left[\exp\left(\lambda\left(G(X^u,W^1)-G(X^v,W^1)\right)\right)\right].$$

Il en résulte alors que le « *commun random numbers* » maximise Λ_{uv}^* pour tout $(u, v)\in(\{1,\ldots,n\},\{n+1,\ldots,m\})$; et par la suite il maximise le taux de convergence η ; grâce à (3.17). □

En conclusion, la solution X_K^* obtenue en résolvant le problème approximé (P_K) converge en exponentielle vers une solution optimale du problème exact (P). Ceci signifie qu'une solution « presque » optimale peut être obtenue en utilisant un nombre de scénarios de taille modeste.

Cependant, comme nous allons le constater en section 3.5, l'inconvénient de la méthode d'optimisation Monte Carlo est qu'elle nécessite la résolution d'un programme en variables mixtes (P_K) ; un programme difficile à résoudre pour des problèmes de grande taille. Dans la section suivante, nous présentons une heuristique basée sur la relaxation lagrangienne pour la résolution du problème de planification (P).

3. 4 Relaxation Lagrangienne

L'idée générale derrière une approche de relaxation lagrangienne est de décomposer un problème de grande taille en plusieurs sous-problèmes de plus petite taille. Ceci est généralement fait en relaxant des contraintes dites *couplantes* (Minoux, 1983 ; Wolsey, 1998).

Les contraintes d'affectation (3.3) dans le problème (P) sont des contraintes couplantes, puisqu'elles relient toutes les périodes pour assurer que chaque patient soit planifié une et une seule fois. L'idée de base de cette méthode consiste à relâcher ces contraintes (3.3) et pénaliser leurs éventuelles violations par des « *multiplicateurs Lagrangiens* » π_i, $i\in\{1,\ldots, N\}$.

Pour un vecteur de multiplicateur π donné, nous obtenons alors le programme relaxé $PR(\pi)$, problème du relaxation lagrangienne, suivant:

$$L(\pi) = \text{Minimiser } L(X, \pi) = \sum_{i=1}^{N}\sum_{t=e_i}^{H+1} a_{it}\ x_{it} + \sum_{t=1}^{H} c_t\ E\left[\left(W_t + \sum_{i=1}^{N} d_i x_{it} - T_t\right)^+\right]$$

$$+\sum_{i=1}^{N}\pi_i\left(1-\sum_{t=e_i}^{H+1} x_{it}\right)$$

Sous les contraintes : $x_{it}\in\{0, 1\},\ \forall\ i\in\{1,\ldots, N\}, t\in\{1,\ldots,H+1\}$

$L(X, \pi)$ et $L(\pi)$ sont généralement appelées fonction de *Lagrange* et fonction *duale*, respectivement (Minoux, 1983).

Pour chaque période t, définissons $I_t = \{i / e_i \le t\}$ l'ensemble des patients qui peuvent être planifié dans cette période. Cet ensemble regroupe les patients candidats pour cette période, c-à-d ceux avec une date au plus tôt inférieure ou égale à t. Par convention, on suppose que $c_{H+1} = 0$ et $T_{H+1} = \infty$. Par simple arrangement des termes, le programme relaxé $PR(\pi)$ est re-exprimé comme suit :

$$L(\lambda) = \sum_{i=1}^{N}\pi_i +$$

$$\sum_{t=1}^{H+1}\min_{x_{it}\in\{0,1\}}\left\{\sum_{i\in I_t}\left(a_{it}-\pi_i\right)x_{it} + c_t\ E\left[\left(W_t + \sum_{i\in I_t} d_i x_{it} - T_t\right)^+\right]\right\}$$

La résolution du problème $PR(\pi)$ se ramène ainsi à la résolution de H+1 sous-problèmes indépendants $PR_t(\pi)$ de type :

$$L_t(\pi) = \min_{x_{it}\in\{0,1\}}\sum_{i\in I_t}\tilde{a}_{it}x_{it} + c_t\ E\left[\left(W_t + \sum_{i\in I_t} d_i x_{it} - T_t\right)^+\right]$$

où $\tilde{a}_{it} = a_{it} - \pi_i$.

Chaque sous-problème est maintenant relatif à une période. Nous présentons dans la section 3.4.1 un algorithme de programmation dynamique pour sa résolution.

Quel que soit le vecteur de multiplicateurs π, si $X(\pi)$ désigne une solution optimale du problème relaxé $PR(\pi)$ alors $L(\pi) = L(X(\pi),\pi)$ est une borne inférieure de la valeur de la fonction économique de (P) à l'optimum, J^*. La recherche du vecteur de multiplicateurs λ qui donnera la plus grande borne inférieure $L(\pi)$ revient à résoudre, le problème suivant, noté (PD), et appelé *problème dual lagrangien*

$$L^* = \max_{\lambda_i\in\mathbb{R}} L(\pi) = \sum_{i=1}^{N}\pi_i + \sum_{t=1}^{H+1} L_t(\pi)$$

La fonction duale $L(\pi)$ est une fonction concave non partout différentiable. Dans la section 3.4.3, nous présentons une méthode de sous-gradient pour la résolution du problème dual.

La solution optimale $X(\pi)$ du problème relaxé $PR(\pi)$ est généralement non réalisable pour le problème initiale (P). Comme les contraintes d'affectation ont été relâchées, certains patients peuvent être affectés à plusieurs périodes et d'autres affectés à aucune période.

Remarque : Dans le cas général, si la solution optimale d'une relaxation lagrangienne est réalisable pour le problème initial, alors nous ne pouvons pas conclure qu'elle est optimale pour ce problème initial. En effet, si certaines contraintes relâchées sont des inéquations et si cette solution ne vérifie pas l'une d'elle à l'égalité, alors la valeur de la fonction économique de la relaxation lagrangienne diffère de celle du programme initial. Notre relaxation est particulière dans le sens que nous n'avons relâché que des équations. Donc, contrairement au cas général, s'il existe un vecteur lagrangien π tel que la solution $X(\pi)$ est réalisable pour (P) alors elle est optimale pour (P).

Le schéma de l'algorithme de relaxation lagrangienne est comme suit :

Initialiser le vecteur des multiplicateurs π^0.
$n \leftarrow 0$
Répéter
  Résoudre PR(π^n)
  Si la solution X^{π^n} est réalisable pour (P) **alors**
    $X(\pi^n)$ est optimale pour (P), STOP
  Fin si
  Exploiter $X(\pi^n)$
  Mise à jour des multiplicateurs : calculer π^{n+1}
  $n \leftarrow n + 1$
Jusqu'à ce que le test d'arrêt soit vérifié

Rappelons que la résolution PR(π^n) se décompose en la résolution de H+1 sous-problèmes indépendants $PR_t(\pi)$.

Exploiter la solution $X(\pi^n)$ consiste à

- mettre à jour *LB*, la meilleure borne inférieure trouvée jusqu'à présent, si $L(\pi^n) > LB$
- construire une solution réalisable de (P) à partir de $X(\pi^n)$ (voir section 3.4.2). Si nécessaire, mettre à jour la meilleure borne supérieure *UB*, valeur de la meilleure solution réalisable de (P) trouvée jusqu'à présent.

Le test d'arrêt est intimement lié à la méthode de mise à jour des multiplicateurs. Il sera exposé dans la section 3.4.3.

3. 4. 1 Résolution des sous-problèmes

Dans cette section, nous présentons un algorithme de programmation dynamique que nous avons développé pour la résolution du sous-problème $PR_t(\pi)$.

Rappelons, tout d'abord, la formulation du sous-problème :

$$L_t(\pi) = \min_{x_{it}\in\{0,1\}} \sum_{i\in I_t} \tilde{a}_{it} x_{it} + c_t \; E\left[\left(\sum_{i\in I_t} d_i x_{it} - T_t + W_t\right)^+\right]$$

avec $\tilde{a}_{it} = a_{it} - \pi_i$ (coût *modifié*).

Pour un vecteur de multiplicateurs π donné, le sous-problème $PR_t(\pi)$ consiste à identifier l'ensemble des patients qui seront opérés en jour t.

Le sous-problème $PR_t(\pi)$ peut être considéré comme un problème de sac à dos stochastique, où la capacité du sac est une variable aléatoire T_t - W_t. La capacité du sac peut être dépassée, mais le dépassement est pénalisé via le coût c_t. Donc, le sous-problème consiste à sélectionner l'ensemble des patients qui minimise le coût de dépassement moyen plus les coûts (modifiés) relatifs aux patients sélectionnés. Dans ce qui suit nous présentons une méthode de résolution pour cette nouvelle variante du problème de sac à dos, le sous-problème $PR_t(\pi)$.

Le sous-problème peut être reformulé comme suit :

$$L_t(\lambda) = \text{Min} \sum_{i\in I_t} \tilde{a}_{it} x_{it} + c_t \; E\left[\left(W_t + D - T_t\right)^+\right]$$

$$\text{Sous contraintes :} \quad \sum_{i\in I_t} d_i x_{it} = D \qquad (3.22)$$

$$0 \le D \le D_{max}$$

$$x_{it} \in \{0,1\}, \qquad \forall i \in I_t$$

avec $D_{max} = \sum_{i\in I_t} d_i$, c'est-à-dire la durée totale de toutes les chirurgies électives candidates pour le jour t.

Définissons maintenant, pour tout $0 \le D \le D_{max}$, le problème du sac à dos binaire suivant :

$$h(D) = \min\left\{\sum_{i\in I_t} \tilde{a}_{it} x_{it} : \sum_{i\in I_t} d_i x_{it} \le D, x_{it} \in \{0,1\} \forall i \in I_t\right\} \qquad (3.23)$$

Nous observons que le sous-problème $PR_t(\pi)$ est équivalent à

$$L_t(\lambda) = \min_{0\le D\le D_{\max}} h(D) + c_t \; E\left[\left(W_t + D - T_t\right)^+\right]$$

Notons que la contrainte d'égalité (3.22) peut être remplacée par une contrainte d'inégalité (3.23) parce que le coût du dépassement horaire $c_t\ E\left[\left(W_t + D - T_t\right)^+\right]$ est croissant en fonction de *D*.

La quantité $c_t\ E\left[\left(W_t + D - T_t\right)^+\right]$ peut être évaluée soit analytiquement soit par simulation (section 3.2.4) pour tout *D*. Les différentes valeurs de *h*(*D*) peuvent être obtenues en résolvant le problème du sac à dos (3.23) avec $D = D_{max}$.

En effet, quand on utilise la programmation dynamique pour la résolution du problème de sac à dos avec $D = D_{max}$, on résout implicitement tous les problèmes de sac à dos réduits avec *D* variant de 0 à D_{max}.

Étant donné une paire d'entiers *m* ($1 \le m \le |\ I_t\ |$) et *D* ($0 \le D \le D_{max}$), définissons le problème de sac à dos réduit suivant :

$$h_m(D) = \min\left\{\sum_{i=1}^{m} \tilde{a}_{it} x_{it} : \sum_{i=1}^{m} d_i x_{it} \le D,\ x_{it} \in \{0,1\}\right\}$$

L'algorithme de programmation dynamique procède en $|\ I_t\ |$ étapes, pour *m* croissant de 1 à $|\ I_t\ |$. À chaque étape les valeurs $h_m(D)$ (pour *D* croissant de 0 à D_{max}) sont déterminées en utilisant l'équation de récurrence suivante :

$$h_m\left(D\right) = \begin{cases} h_{m-1}\left(D\right) & \text{pour } 0 \le D \le d_m - 1 \\ min\left(h_{m-1}\left(D\right),\ h_{m-1}\left(D - d_m\right) + \tilde{a}_{mt}\right) & \text{pour } d_m \le D \le D_{max} \end{cases}$$

avec les conditions initiales suivantes :

$$h_1\left(D\right) = \begin{cases} 0 & \text{for } 0 \le D \le d_1 - 1 \\ \min\left(0, \tilde{a}_{1t}\right) & \text{for } d_1 \le D \le D_{\max} \end{cases}$$

Donc, on voit bien qu'à la dernière étape on obtient toutes les valeurs $h_{I_t}(D)$, c'est-à-dire *h*(*D*). Par conséquent, la solution optimale du sous-problème $PR_t(\lambda)$ est la solution correspondant à l'état $h_{|I_t|}(D^*)$, avec

$$D^* = \underset{0 \le D \le D_{max}}{arg\ min}\ h_{|I_t|}\left(D\right) + c_t E\left[\left(W_t + D - T_t\right)^+\right]$$

L'algorithme développé pour la résolution du sous-problème est similaire à celui de la programmation dynamique pour le problème de sac à dos classique (Martello et Toth, 1990), mais il tient compte du coût non-linéaire à la dernière étape de l'algorithme. Dans cette section, la valeur de D_{max} a été fixé à $\sum_{i \in I_t} d_i$. Cependant, les expérimentations numériques montrent qu'il suffit de la fixer à des valeurs beaucoup plus faibles, de l'ordre de la capacité régulière T_t.

3. 4. 2 Construction des plannings réalisables

Comme les contraintes d'affectation ont été relâchées, la solution $X(\pi)$ du problème de relaxation lagrangienne PR(π), appelée aussi *solution duale*, fournit généralement un planning non réalisable. En effet, certains patients peuvent être affectés à plusieurs périodes et d'autres affectés à aucune période. Nous présentons dans cette section deux heuristiques que nous employons tout au long de l'algorithme de relaxation lagrangienne pour construire un planning réalisable de « bonne » qualité.

Avec les approches de relaxation lagrangienne, la méthode qui est la plus souvent proposée consiste à exploiter la meilleure solution duale, celle associée à la meilleure borne inférieure. Une telle solution duale est généralement caractérisée par un faible nombre de contraintes violées ; et par conséquent elle peut être transformée en une solution réalisable moyennant quelques modifications mineures. Mais, malheureusement, cette solution réalisable n'est pas nécessairement de bonne qualité. Pour cette raison, nous proposons de considérer toutes les solutions duales générées. À partir de chaque solution duale, un planning réalisable est construit. Le planning obtenu est ensuite amélioré par une heuristique de *recherche locale*. À la fin de l'algorithme le meilleur planning est sélectionné comme solution finale.

Dans le reste de cette section, nous commençons par présenter une heuristique gloutonne, dite *optimisation séquentielle*, que nous avons mis au point afin d'obtenir des solutions réalisables du problème, ensuite, nous présentons une procédure d'amélioration, dite *heuristique de recherche locale*.

Optimisation séquentielle

Étant donné une solution duale, l'heuristique construit un planning réalisable en réaffectant les patients qui violent les contraintes d'affectation. Les étapes principales de cette méthode sont les suivantes :

- les patients qui respectent les contraintes d'affectation, c'est-à-dire affectés exactement une fois, sont planifiés comme suggéré par la solution duale, et forment ainsi un planning « *partiel* »,
- les patients violant les contraintes d'affectation sont placés dans une liste et seront ensuite insérés un par un dans le planning partiel.

À chaque itération, un patient de la liste est considéré : on détermine la *meilleure* période pour le planifier, sans remettre en cause l'affectation des patients qui sont déjà affectés, ensuite, on l'affecte à cette période. La meilleure période correspond à la période qui engendre une augmentation du coût minimale. À l'itération suivante un nouveau patient est considéré et

ainsi de suite. L'ordre de traitement des patients dans la liste est arbitraire. L'heuristique s'arrête lorsque tous les patients sont insérés dans le planning.

Heuristique de recherche locale

Très souvent, les solutions réalisables fournies par l'heuristique d'optimisation séquentielle sont de qualité moyenne. Nous proposons ici une méthode d'amélioration itérative basée sur le principe de recherche locale.

Nous définissons la notion de *voisinage* d'une solution comme étant l'ensemble de toutes les solutions réalisables obtenues en modifiant l'affectation d'un patient. Partant d'une solution réalisable, l'heuristique progresse d'une solution vers sa meilleure voisine, jusqu'à ce qu'il n'y a plus d'amélioration.

Plus précisément, à chaque itération, on détermine pour chaque patient *i* le gain qui peut être réalisé en le réaffectant à une autre période, sans changer l'affectation des autres patients. Ensuite, on réaffecte le patient qui fournit le gain maximal. Ce processus itératif s'arrête lorsque il n'y a aucune réaffectation améliorante.

3. 4. 3 Résolution du problème dual : Algorithme du sous-gradient

La méthode du sous-gradient est une méthode itérative pour maximiser $L(\pi)$, fonction économique du problème relaxé. À chaque itération *n*, les multiplicateurs lagrangiens sont mis à jour de la manière suivante :

$$\pi^{n+1} = \pi^n + \alpha^n \ \nabla(\pi^n)$$

π^n et α^n sont respectivement la valeur du vecteur π et le pas de déplacement à l'itération *n*. $\nabla(\pi^n)$ est le gradient de la fonction duale $L(\pi)$ au point π^n. La $i^{\text{ème}}$ composante du sous-gradient $\nabla(\pi^n)$ est la suivante

$$\left[\nabla(\pi)\right]_i = 1 - \sum_{t=e_i}^{H+1} x_{it}(\pi), \quad i = 1,\ldots, N$$

Remarquons que la composante $[\nabla(\pi^n)]_i$ est relative au patient électif *i*. Elle peut être égale soit à 0, si la contrainte d'affectation est respectée, soit à 1 si le patient *i* n'est affecté à aucune période, soit à un nombre strictement négatif indiquant que le patient *i* est affectée à plusieurs périodes. Par conséquent, si le sous-gradient est identiquement nul, la solution est une solution réalisable pour (P). Vu que nous n'avons relâché que des équations, elle est optimale pour (P), et l'algorithme s'arrête.

La convergence et les performances de la méthode du sous-gradient ont fait l'objet de plusieurs études théoriques. Nous citons par exemple (Polyack, 1967, Polyack 1969, Held *et*

al., 1974, Goffin, 1977). Le résultat fondamental de ces études est que la suite $L(\pi^n)$ converge vers l'optimum L^* si le pas de déplacement vérifie les deux conditions suivantes : $\alpha^n \to 0$ et $\sum_n^{+\infty} \alpha^n \to +\infty$ quand $n \to +\infty$.

En pratique, le pas de déplacement α^n est déterminé suivant la formule (Fischer, 1981) :

$$\alpha^n = \beta \frac{\bar{L} - L(\pi^n)}{\| \nabla(\pi^n) \|^2}$$

où $\bar{L}$ est une estimation (par excès ou par défaut) de la valeur optimale L^* et β un coefficient (coefficient de relaxation) vérifiant $0 < \beta \leq 2$. $\|\nabla(\pi^n)\|$ est la norme euclidienne du sous-gradient $\nabla(\pi^n)$.

Le pas α^n reflète ainsi non seulement la qualité de la solution optimale lagrangienne courante, mais aussi la violation des contraintes relâchées.

Comme estimation de l'optimum de L^* nous prenons $\bar{L} = UB$, où UB est la meilleure borne supérieure obtenue dans les étapes précédentes du calcul; valeur de la meilleure solution réalisable de (P) trouvée jusqu'à présent.

Le coefficient β est initialisé à 2. Ensuite, il est régulièrement diminué au cours de l'exécution de l'algorithme, permettant ainsi de réduire la taille de l'intervalle dans lequel évoluent les multiplicateurs lagrangiens.

Outre le cas où le sous-gradient est identiquement nul, il existe deux tests d'arrêt. Le premier est la limitation du nombre d'itérations. Le second est un test d'optimalité : si $L(\pi^n) = UB$ alors la solution réalisable de (P) qui est associée à UB est une solution optimale.

3. 5 Expérimentations numériques

Les tests numériques présentés dans cette section ont été réalisés sur un PC équipé d'un processeur Pentium 4 à 3.0 GHz avec une mémoire de 512 Mo, utilisant un système d'exploitation Windows XP. Les algorithmes ont été programmés en MS Visual C++.

Les performances des deux méthodes d'optimisations (optimisation Monte Carlo et méthode de relaxation lagrangienne) ont été évaluées par des tests sur des instances des problèmes générées de manière aléatoire.

3. 5. 1 Génération des instances

Le nombre de jours H est égal à 5 (un horizon d'une semaine). La capacité régulière agrégée en une période t dépend du nombre des salles disponibles et de leurs capacités régulières en

période t. Dans cette étude numérique, nous considérons des problèmes où la capacité régulière d'une salle opératoire est toujours de 8 heures, et le nombre de salles disponibles est le même pour toutes les périodes de l'horizon de planification. La capacité régulière agrégée est T_t = nombre de salles disponibles $\times$ 8 heures.

Des problèmes avec 2, 4, 8, 12 et 16 salles opératoires sont considérées. Pour des problèmes avec 4 salles par exemple, T_t = 32 heures. Notons que le nombre de salles est utilisé juste comme paramètre pour contrôler la taille du problème, et que seulement la capacité agrégée est considérée.

La capacité aléatoire utilisée par la chirurgie d'urgence W_t est supposée suivre une loi exponentielle avec une moyenne qui dépend du nombre des salles. Cette moyenne est fixée comme suit : $E[W_t]$ = nombre de salles disponibles $\times$ 1,5 heures. Par exemple, pour un problème avec 4 salles, W_t est exponentiellement distribuée avec une moyenne de 6 heures. Le coût des heures supplémentaires c_t, est fixé à 500 €/heure.

Les durées d_i des chirurgies électives sont générées de manière aléatoire et uniforme à partir de l'intervalle [0.5 heure, 3 heures]. Ces durées sont multiples de 5 minutes.

Les dates au plus tôt sont générées aléatoirement à partir de l'intervalle $\{1,\ldots, H\}$. Afin de tenir compte des patients qui ont une date au plus tôt $e_i = 1$ et qui ont été rejetés du planning précèdent, nous introduisons un nouveau paramètre e_i', date au plus tôt *effective*. Les dates au plus tôt sont générées en deux étapes. Dans un premier temps, on génère pour chaque patient i une date au plus tôt effective e_i'. Les e_i' sont générées aléatoirement et uniformément à partir de l'ensemble $\{-2, \ldots, H\}$. Ensuite, les patients avec $e_i' \leq 0$ vont avoir leur $e_i = 1$; pour les autres patients les valeurs de e_i seront les mêmes que e_i' ($e_i = 1$ si $e_i' < 1$; $e_i = e_i'$ si non).

Les coûts d'affectations a_{it} des patients électifs sont considérés comme croissants en fonction du jour t. Ils sont définis comme suit :

$$a_{it} = (t - e_i') \times c \quad \text{pour } t \in \{e_i, \ldots H, H+1\}.$$

Le paramètre c peut être interprété comme un coût d'hospitalisation par jour ou une pénalité par jour d'attente. Il est fixé à 100€.

Les patients électifs sont générés un par un jusqu'à ce que la somme des durées opératoires dépasse la somme des capacités régulières agrégée sur tout l'horizon de planification.

3. 5. 2 Performances de l'optimisation Monte Carlo

Dans cette section nous évaluons les performances de la méthode d'optimisation Monte Carlo (section 3.3). Le problème approximé (P_K) est résolu en utilisant le solveur d'optimisation

CPLEX 9.0. Le coût (critère) exact $J(X_K^*)$ de la solution optimale X_K^* du (P_K) est évalué de manière analytique (section 3.2.4).

Dans un premier temps, nous testons la convergence de la solution fournie par l'optimisation Monte Carlo en fonction du nombre de scénarios *K*. Nous considérons une instance d'un problème avec 2 salles opératoires avec 47 patients électifs, et nous résolvons le problème associé (P_K) pour différentes valeurs de *K*. Pour chaque valeur de *K*, nous réalisons 10 essais ; un essai correspond à la génération aléatoire de *K* scénarios et à la résolution du problème approximé (P_K) correspondant. À chaque essai, on obtient une solution optimale « estimée » X_K^*, son coût estimé $J_K(X_K^*)$ ainsi que son coût exact $J(X_K^*)$.

Dans le tableau 3.1, nous présentons l'évolution du coût exact en fonction de *K*. Pour chaque valeur de *K*, nous présentons le maximum, la moyenne, le minimum et l'écart type des coûts exacts obtenus à partir des 10 essais. Les résultats détaillés concernant la résolution du problème approximé (P_K), obtenus par CPLEX sont présentés dans le tableau 3.2. Ces résultats comprennent les coûts estimés et les temps de calcul.

K	Coût optimal maximal (€)	Coût optimal moyen (€)	Coût optimal minimal (€)	Écart type
2	11841,2	10386,0	9568,0	757,4
5	9670,6	9344,2	8949,8	205,7
10	9469,5	9218,2	8967,0	171,2
20	9222,7	9026,9	8945,0	89,3
50	9000,1	8965,7	8943,6	20,5
100	9021,7	8967,5	8945,0	24,1
200	8977,6	8958,5	8945,0	11,6
500	8984,9	8951,9	8945,0	11,8
700	8949,8	8946,5	8943,6	2,9
1000	8965,5	8948,1	8943,6	6,7

TAB. 3.1 - Évolution du coût exact de la solution optimale

K	Coût optimal estimé (€)				Temps de calcul (secondes)			
	Maximum	Moyenne	Minimum	Écart type	Maximum	Moyenne	Minimum	Écart type
2	11154,7	8465,6	5830,1	1725,3	5591,1	1569,1	17,8	1798,6
5	10392,8	9106,5	7918,4	823,9	1494,3	388,2	0,0	546,5
10	9646,8	8952,4	8133,2	497,3	611,2	73,5	0,0	189,5
20	9428,7	8829,6	7915,1	561,8	233,0	44,5	0,1	74,1
50	9562,5	8878,2	8014,1	475,5	623,2	79,4	0,2	192,5
100	9452,3	9022,9	8804,5	243,6	184,7	62,4	1,3	62,8
200	9206,9	8984,6	8826,6	138,4	439,7	130,5	6,8	141,5
500	9078,1	8958,1	8742,8	101,5	4026,2	1545,2	215,7	1151,3
700	9056,7	8960,0	8800,0	73,9	5873,0	2308,3	718,5	1790,7
1000	9032,2	8961,9	8841,9	56,6	4353,8	2617,4	832,8	1333,3

TAB. 3.2 - Évolution du coût estimé et du temps de calcul

Afin d'évaluer la qualité des solutions optimales fournies par l'optimisation Monte Carlo et les bénéfices d'une modélisation stochastique du problème, nous considérons aussi une version déterministe du problème. La version déterministe consiste à remplacer les variables aléatoires W_t par leurs moyennes $E[W_t]$. Le problème de planification (P) est dans ce cas un problème déterministe (problème linéaire en variables mixtes) ; il correspond à une pratique qui consiste à réserver une quantité fixe de la capacité régulière pour l'utilisation exclusive de la chirurgie d'urgence. Soit X^{det} la solution optimale du problème déterministe. Pour l'instance du problème que nous considérons ici, le coût exact $J(X^{det})$ de la solution déterministe est de 9500€.

À partir des résultats présentés en tableau 3.1, les solutions obtenues par l'optimisation Monte Carlo sont meilleures que la solution déterministe, même pour des faibles valeurs de K (K=20). En plus, les solutions fournies avec K=1000 permettent une réduction du coût de l'ordre de 5,8% par rapport à la solution déterministe.

Afin de mieux illustrer la convergence de l'optimisation Monte Carlo, nous présentons en Figure 3.1 l'évolution du coût exact des solutions optimales en fonction du nombre de scénarios. On voit bien que les solutions obtenues sont meilleures que la solution déterministe lorsque K est plus grand que 20. On peut aussi remarquer que les courbes convergent lorsque le nombre de scénarios est plus grand que 200, ce qui confirme les propriétés de convergence établies en section 3.3.2.

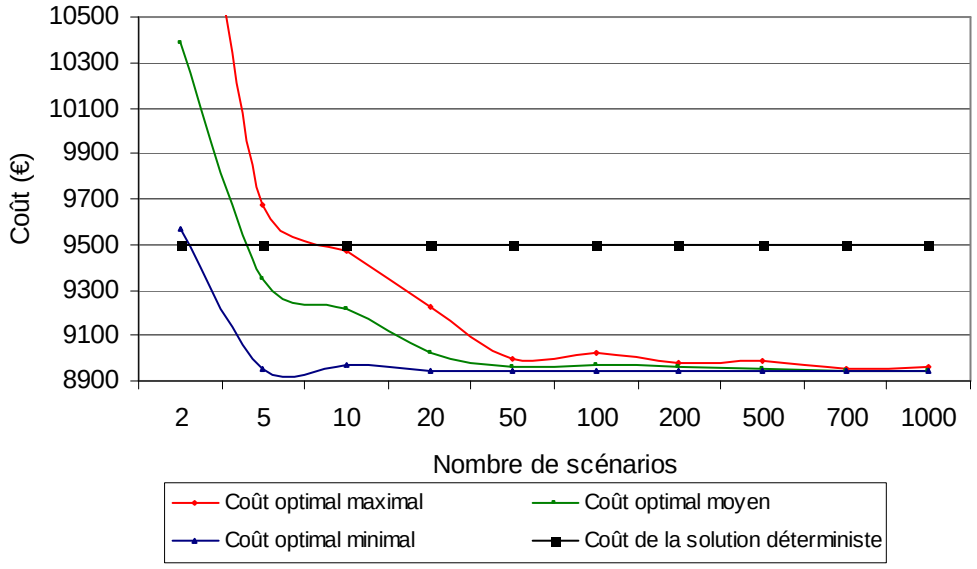

FIG. 3.1 - Évolution du coût exact de la solution optimale

Nous examinons maintenant la vitesse de convergence du coût exact et du coût estimé de la solution optimale. Pour cette fin, nous comparons l'évolution de l'écart type du coût exact et estimé de la solution optimale. Ces résultats, présentés dans les tableaux 3.1 et 3.2, sont repris dans la Figure 3.2. On voit bien que les deux quantités décroissent lorsque K croit ; ce qui confirme la convergence vers une solution optimale du problème de planification (P). On remarque aussi que l'écart type du coût estimé est toujours plus élevé que celui du coût exact. Ceci est dû à un faible nombre de scénarios utilisés pour estimer le coût $J_K(X_K^*)$ de la solution optimale, une convergence lente du critère estimé $J_K(X)$, et une convergence rapide (en exponentielle) de l'optimisation Monte Carlo vers une solution optimale.

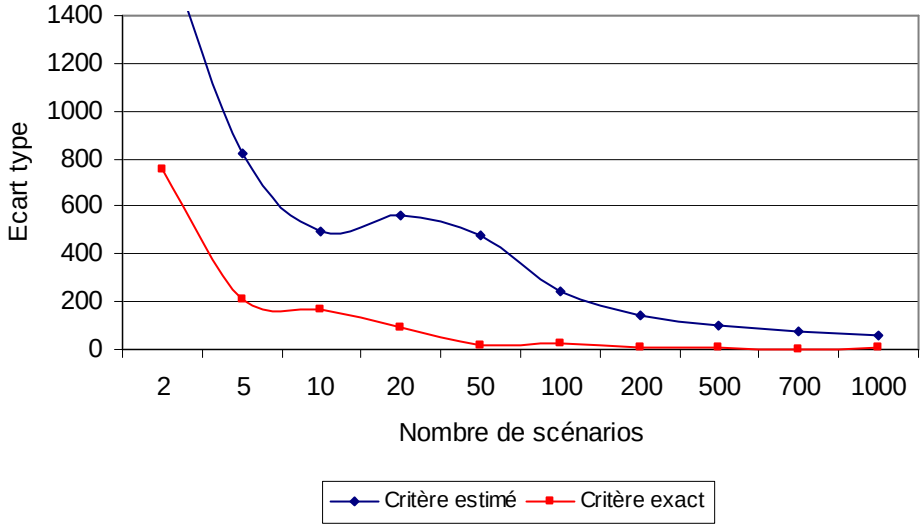

FIG. 3.2 - Évolution des écarts type des coûts exacts et estimés

Examinons maintenant l'influence du nombre de scénarios sur le temps de calcul. La Figure 3.3 présente l'évolution du temps de calcul maximal, moyen et minimal en fonction *K*. Rappelons que ce temps de calcul est le temps nécessaire pour la résolution du problème approximé (P_K) par CPLEX. De manière surprenante, le temps de calcul peut être très important et présente une grande variation pour des faibles valeurs de *K* ; il est faible pour des *K* entre 20 et 200. Pour des *K* supérieurs à 200, le temps de calcul est croissant en fonction de *K*.

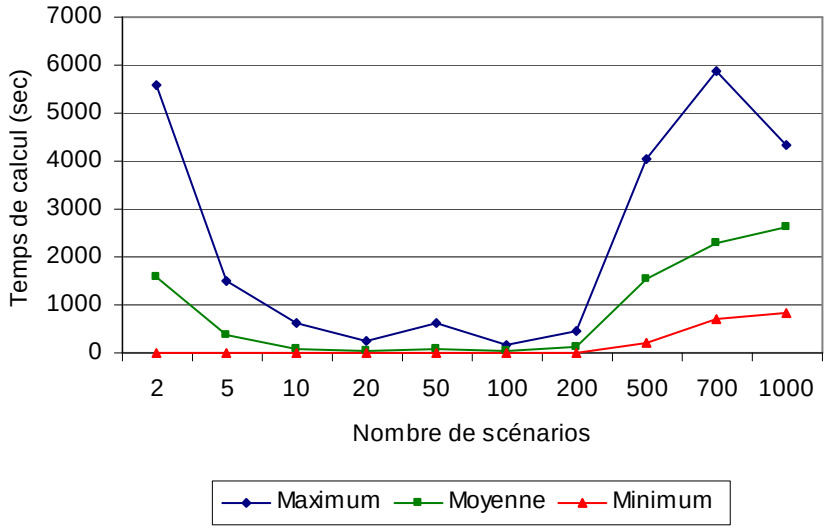

FIG. 3.3 - Évolution du temps de calcul en fonction du nombre de scénarios

Nous allons maintenant examiner l'évolution du temps de calcul pour différentes tailles du problème. Nous considérons des problèmes avec 2, 4 et 8 salles opératoires. Pour chaque taille du problème, 10 instances sont générées aléatoirement. Chaque instance est résolue deux fois avec la méthode d'optimisation Monte Carlo, une fois avec un nombre de scénarios $K = 100$ et une deuxième fois avec $K = 1000$. Les résultats sont présentés dans le tableau 3.3.

On peut remarquer que le temps de calcul augmente considérablement en fonction de la taille du problème. En plus, pour un problème de taille donnée le temps de calcul présente une grande variabilité d'une instance à une autre. En particulier, pour des problèmes avec plus de 2 salles les temps de calcul peuvent dépasser les 3 heures et certaines instances ne peuvent pas être résolues.

Nombre de salles	N	K = 100				K = 1000			
		Max	Moyenne	Min	Écart type	Max	Moyenne	Min	Écart type
2	48,4	474,4	70,9	0,7	142,9	2638,4	373,2	10,7	804,6
4	95,0	454,9	100,8	0,4	177,2	10800,2	1142,1*	7,6	3395,3
8	189,8	7200,5	3809,5	0,4	3645,3	10800,1	2060,3*	16,7	4154,5

* Résultats basés sur 8 instances (2 instances n'ont pas pu être résolues à cause des problèmes de mémoire)

TAB. 3.3 - Évolution du temps de calcul (en secondes) en fonction de la taille du problème

Pour récapituler, l'optimisation Monte Carlo fournie des solutions presque optimales en utilisant un nombre modeste de scénarios. Cependant, l'inconvénient majeur est qu'elle fait appel à un programme linéaire en variables mixte dont la résolution peut être très gourmande en temps de calcul.

3. 5. 3 Performances de la méthode de relaxation lagrangienne

Nous évaluons, dans cette section, les performances de la méthode de relaxation lagrangienne pour des problèmes de différentes tailles. La taille du problème est déterminée par le nombre de salles opératoires. Le gap de dualité « *GAP* » est utilisée comme une mesure de performances afin d'évaluer la qualité des solutions obtenues ; il est défini comme suit :

$$GAP = (UB - LB) / LB.$$

où *UB* est le coût de la meilleure solution réalisable et *LB* est la meilleure borne inférieure fournie par la relaxation lagrangienne.

Pour l'algorithme de sous-gradient le nombre d'itérations est fixé à 100. Les sous-problèmes de relaxation lagrangienne $PR_t(\pi)$ sont résolus comme présenté dans la section 3.4.1 ; avec $D_{\max} = \sum_{i \in I_t} d_i$.

Les résultats numériques sont présentés dans le tableau 3.4 ; ils sont basés sur 10 instances générées aléatoirement (selon le schéma présenté dans la section 3.5.1) pour chaque taille de problème. Ces résultats comprennent, pour chaque taille de problème, le nombre moyen des patients électifs (*N*), le gap de dualité (GAP), l'itération à laquelle la meilleure borne inférieure est obtenue (LB iter), l'itération à laquelle la meilleure solution réalisable est obtenue (UB iter), et le temps de calcul (Temps CPU).

Nous avons aussi testé une version légèrement modifiée de la méthode de relaxation lagrangienne : pour chaque période *t*, le sous-problème $PR_t(\pi)$ est résolu avec

$$D_{\max} = \begin{cases} \sum_{i \in I_t} d_i & \text{si } \sum_{i \in I_t} d_i \leq 1.5 \times T_t \\ 1.5 \times T_t & \text{sinon} \end{cases}. \quad (3.20)$$

Ceci signifie que la durée totale des chirurgies électives planifiées pour le jour *t* ne doit pas dépasser la capacité régulière multipliée par 1,5. C'est une contrainte additionnelle qui ne change rien aux solutions obtenues, mais qui permet une réduction du temps de calcul.

Les résultats concernant les nouveaux temps de calcul (Temps CPU 2) sont présentés dans le même tableau 3.4. Notons que tous les autres résultats sont identiques à ceux obtenus avec la version d'origine de la méthode.

Nombre de Salles	N	GAP (%)	UB iter	LB iter	Temps CPU (sec)	Temps CPU 2 (sec)
2	48,4	0,95	74,1	85,9	1,1	0,4
4	95,0	1,20	79,4	93,3	7,3	2,4
8	189,8	1,10	76,4	90,6	56,0	16,6
12	283,1	1,40	65,5	89,1	199,3	63,7
16	377,0	1,30	68,3	92,7	461,6	148,5

TAB. 3.4 - Résultats de la méthode de relaxation lagrangienne

À partir du tableau 3.4, on peut remarquer que les valeurs de « UB iter » sont toujours différentes de « LB iter » ; ce qui signifie que la meilleure solution réalisable n'est généralement pas obtenue à partir de la meilleure solution duale, celle correspondant à la meilleure borne inférieure. Ce résultat confirme l'intérêt de construire une solution réalisable à partir de chaque solution duale.

On peut aussi remarquer que les GAP varient entre 0,95% et 1,4% ; ce qui signifie que les solutions obtenues sont très proches de l'optimum. Les temps de calcul sont croissants en fonction de la taille de problème mais demeurent acceptables, pour un maximum de l'ordre de 8 minutes pour des problèmes de grande taille (16 salles et à peu près 380 patients).

La version modifiée de la méthode de relaxation lagrangienne permet une amélioration significative des temps de calcul sans affecter la qualité des solutions obtenues. Les temps de calcul ont diminué de plus de 60%. En effet, la résolution d'un sous-problème $PR_t(\pi)$ avec D_{max} fixé comme suggérée par l'équation (3.20) permet de réduire le temps de calcul nécessaire pour la résolution du dit problème, et par conséquent une réduction du temps de calcul total.

3. 6 Conclusion

Dans ce chapitre, nous avons proposé un modèle stochastique pour la planification du bloc opératoire ; un modèle qui intègre de manière explicite les incertitudes liées à la chirurgie d'urgence.

Nous avons développé une méthode de résolution, qu'on peut qualifier d'« exacte », qui combine la simulation Monte Carlo et la programmation en nombre mixtes. Le principe de base de cette méthode consiste à approximer le problème d'optimisation stochastique par un problème d'optimisation déterministe. Ce dernier est un programme linéaire à variables mixtes qui est ensuite résolu en utilisant une méthode d'optimisation déterministe. Nous avons montré que, lorsque le nombre de scénarios utilisés augmente, les solutions optimales du problème approximé convergent avec une vitesse exponentielle vers des vraies solutions optimales du problème stochastique. Les tests numériques réalisés ont confirmé cette convergence, et ont montré qu'une réduction considérable des coûts peut être réalisée moyennant une modélisation stochastique du problème. Ce pendant, l'inconvénient de cette méthode est qu'elle incorpore un programme en nombres mixtes, ce qui limite son utilisation à des problèmes de petite taille.

Pour surmonter ce handicap, nous avons développé une méthode approchée basée sur la technique de relaxation Lagrangienne. Cette méthode fournit des solutions de bonne qualité, avec un faible gap de dualité, et permet de résoudre rapidement des problèmes de grande taille.

Enfin, nous signalons que d'autres travaux (Lamiri *et al*., 2006a), non présentés dans ce mémoire, ont porté sur la résolution approchée de ce même problème de planification stochastique. Dans ces travaux nous avons utilisé la recherche locale, la recherche tabou, et le recuit simulé comme méthodes d'optimisation.

Dans ce chapitre, nous avons supposé que les salles opératoires sont polyvalentes (nous avons agrégé leurs capacités) et que la capacité en heures supplémentaires est illimitée. Ces deux hypothèses seront relâchées dans le chapitre suivant, et nous considérerons aussi les coûts de sous-utilisation des salles lors de la planification.

Chapitre **4**

Planification avec capacité désagrégée

Nous nous intéressons toujours à la planification du bloc opératoire avec prise en compte des aléas dus à la chirurgie d'urgence. Dans ce chapitre, le problème de planification consiste à déterminer l'ensemble des patients électifs qui seront opérés en chaque salle opératoire et en chaque jour sur un horizon de planification donné. L'objectif est de minimiser la somme des coûts relatifs aux patients électifs et des coûts de sur-utilisation et de sous-utilisation des salles opératoires. Nous formulons le problème sous la forme d'un programme stochastique en nombres entiers et nous proposons une approche de résolution basée sur la génération de colonnes. Les principales étapes de cette approche sont les suivantes. Dans un premier temps, nous reformulons le problème de planification sous la forme d'un problème linéaire (en nombres binaires) comportant un nombre très élevé de variables (colonnes), appelé problème maître. Ensuite, nous résolvons la relaxation linéaire du problème maître en utilisant la génération de colonnes. À partir de la solution optimale du problème relâché un planning réalisable est construit et enfin amélioré par optimisation locale. Cette approche fournit des solutions proches de l'optimum ainsi qu'une borne inférieure qui permet d'évaluer la qualité des solutions obtenues. Les expérimentations numériques montrent que cette approche permet de trouver des solutions à moins de 2% de l'optimum pour des problèmes de grande taille (12 salles opératoires et environ 210 patients) en un temps de calcul très court.

(Lamiri et Xie, 2007 ; Lamiri *et al*., 2007c)

4. 1 Introduction

Dans ce chapitre nous étendons le modèle de planification présenté dans le chapitre précédent et nous proposons une méthode de résolution basée sur la génération de colonnes.

Le problème de planification consiste maintenant à déterminer pour chaque patient électif la date d'intervention ainsi que la salle opératoire où l'intervention aura lieu ; bien évidement tout en tenant compte du caractère incertain de la chirurgie d'urgence.

Dans ce nouveau modèle, chaque salle opératoire dispose d'une capacité en heure normale et d'une capacité en heure supplémentaire. Le coût moyen d'utilisation d'une salle opératoire est composé maintenant des coûts de sous-utilisation et de sur-utilisation. Ce dernier, comporte à son tour les coûts des heures supplémentaires et une pénalité de dépassement de la capacité totale (capacité en heures normales et supplémentaires) de la salle.

Dans un premier temps, le problème est formulé sous la forme d'un programme stochastique. Ensuite, nous proposons une formulation *orientée colonnes* qui comporte un nombre très élevé de variables (colonnes). Cette formulation sera résolue par une approche de génération de colonnes.

4. 2 Modèle et formulation mathématique

4. 2. 1 Modèle proposé

Dans ce chapitre nous nous intéressons à la planification des interventions électives dans un bloc opératoire sur un horizon de planification de H périodes (jours). Le bloc opératoire est composé de S salles opératoires qui sont utilisées pour la chirurgie d'urgence et la chirurgie élective.

Nous supposons qu'au début de l'horizon de planification, il y a un ensemble de N patients électifs en attente. Le problème de planification consiste à déterminer l'ensemble de patients qui seront opérés dans chaque salle opératoire et en chaque période de l'horizon. Dans le reste de ce chapitre, nous désignons une salle opératoire $s \in \{1,\ldots, S\}$ en une période (jour) $t \in \{1,\ldots, H\}$ par « salle-jour » (s, t). Les notations suivantes seront utilisées tout au long de ce chapitre :

H : Horizon de planification,

t : (1… H) Indice de période (jour),

S : Nombre de salles opératoires,

s : (1…S) Indice de salle opératoire,

N : Nombre de patients programmables,

i : (1… N) Indice de patient programmable,

d_i : Temps nécessaire pour réaliser l'opération chirurgicale du patient i,

e_i : Date au plus tôt pour opérer le patient i,

a_{its} : Coût d'affectation du patient *i* à la salle-jour (*s*, *t*),

W_{ts} : Variable aléatoire représentant la durée totale des chirurgies d'urgence réalisées en salle-jour (*s*, *t*),

$g_{ts}(.)$: Coût moyen d'utilisation de la salle opératoire *s* en jour *t*.

Les différentes notations et hypothèses relatives au problème de planification sont expliquées en détail dans ce qui suit.

Durée opératoire

À chaque patient électif *i* est associée une durée opératoire (durée d'intervention) d_i. La durée d'intervention est définie par l'intervalle de temps compris entre l'arrivée du patient en salle et la fin de remise en état de la dite salle. Cette durée regroupe la durée de chirurgie, le temps de préparation et de set-up, et le temps de nettoyage de la salle une fois la chirurgie terminée. Les durées opératoires peuvent être estimées en exploitant l'historique des données et/ou l'expertise des chirurgiens et des gestionnaires des salles opératoires (Shukla *et al.*, 1990 ; Wright *et al.*, 1996 ; Levecq *et al*., 2003).

Dans le modèle proposé dans ce chapitre, nous supposons que les durées opératoires d_i sont des données déterministes et discrètes ; multiples d'une certaine unité de temps θ (par exemple $\theta = 5$ minutes). Nous supposons également que la durée opératoire ne dépend pas de la salle opératoire où l'intervention est réalisée.

Remarque : En pratique, le temps de préparation et de set-up peut varier d'une salle à une autre selon la disponibilité de certains équipements médicaux, etc. Par conséquent, la durée opératoire peut dépendre de la salle opératoire. Le modèle proposé dans ce chapitre ainsi que l'approche de génération de colonnes peuvent facilement prendre en compte des durées opératoires dépendantes de salles. Cependant, comme il est difficile d'obtenir des estimations réalistes des durées opératoires pour différentes salles opératoires, et que la différence est relativement mineure, nous nous limitons au cas où les durées sont indépendantes des salles opératoires.

Coûts relatifs aux patients électifs

Chaque patient électif est caractérisé par une date au plus tôt et un ensemble des coûts. La date au plus tôt e_i représente une date avant laquelle le patient ne peut pas être opéré. Elle peut correspondre à une date d'hospitalisation fixée, date de disponibilité des résultats d'analyse etc. À partir de la date au plus tôt, le patient peut être planifié dans n'importe quel jour sur l'horizon de planification et engendre un coût d'« affectation » qui dépend aussi bien du jour d'intervention que de la salle où il sera opéré.

Désignons par a_{its} le coût engendré lorsque le patient électif i est opéré en salle-jour (s, t), pour $t \in \{e_i,\ldots, H\}$ et $s \in \{1,\ldots, S\}$. Un jour « fictif » H+1 est ajouté à l'horizon de planification afin de regrouper les patients rejetés du planning courant ; c'est-à-dire ceux qui ne seront pas planifiés pour l'horizon courant. Donc, $a_{i(H+1)s}$ représente le coût de la non planification (du rejet) du patient i. Bien évidemment, ce coût est indépendant de la salle opératoire ($a_{i(H+1)s} = a_{i(H+1)s'}$, $\forall$ s, $s' \in \{1\ldots S\}$) ; il sera simplement désigné par $a_{i(H+1)}$.

La structure des coûts proposée permet de modéliser plusieurs contraintes. La matrice des coûts $[a_{its}]_{ts}$ peut représenter des coûts d'hospitalisation (Jebali *et al.*, 2006; Guinet et Chaabane, 2003), des pénalités d'attente (Gerchak *et al.*, 1996), les préférences du chirurgien ou du patient concernant la date et/ou la salle d'intervention, des dates limites à ne pas dépasser, la disponibilité ou l'adéquation des salles opératoires, etc.

Nous présentons, maintenant, un exemple d'utilisation de la matrice des coûts pour représenter plusieurs contraintes. Si, par exemple, le patient i est hospitalisé en jour e_i et s'il est en attente de son opération chirurgicale, alors les coûts a_{its} seront croissants en fonction du nombre de jours d'attente. L'accroissement d'un jour à un autre représente le coût d'hospitalisation par jour. Si en plus, le patient ne peut pas être opéré le jour t', à cause de l'indisponibilité du chirurgien par exemple, alors un coût très élevé sera choisi pour $a_{it's}$, $\forall$ $s \in \{1\ldots S\}$. Si le patient peut être opéré dans n'importe quelle salle opératoire, alors les coûts d'affectation seront indépendants des salles opératoires ; $a_{its} = a_{its'}$, $\forall$ s, $s' \in \{1\ldots S\}$. Cependant, si le patient i ne peut pas être opéré dans certaines salles, par manque de certains équipements spécifiques par exemple, alors le coût d'affectation du patient à ces salles sera choisi très élevé. Si en plus le patient doit être opéré avant un jour donné L_i, alors les coûts a_{its} seront très élevés pour tout $t > L_i$. Plusieurs autres situations peuvent être modélisées en ajustant les coûts d'affectation.

Remarque : L'indisponibilité des chirurgiens ou des salles opératoires sont des contraintes fortes qui définissent l'ensemble des salles-jours potentiels pour planifier le patient. Dans ce mémoire, nous avons opté pour les modéliser moyennant des pénalités (coûts d'affectation élevés). Cependant, il est à noter que ce type de contrainte peut être facilement pris en compte en définissant pour chaque patient un ensemble explicite de salles-jours auquel le patient peut être affecté. Le modèle proposé dans ce chapitre ainsi que l'approche de génération de colonnes peuvent facilement intégrer ce genre de restrictions, moyennant quelques modifications mineures. Cependant, par souci de simplicité de notations, nous n'utilisons pas ces restrictions au niveau de ce mémoire.

Remarque : Comme nous l'avons présenté ci-dessus, la matrice des coûts $[a_{its}]_{ts}$ représente plusieurs facteurs qui doivent être pris en compte lors de la planification des salles opératoires : coûts d'hospitalisation, préférences des chirurgiens et des patients, adéquation

des salles opératoires, des contraintes médicales, etc. La détermination de ces coûts est une tâche qui n'est pas facile et qui nécessite la collaboration de plusieurs acteurs. Le développement d'une procédure pour estimer ces coûts est une direction intéressante pour nos futures recherches. Dans ce travail, nous supposons qu'on dispose de cet ensemble de coûts.

Chirurgie d'urgence

Les patients urgents arrivent de manière aléatoire tout au long de la journée et doivent être opérés le plus vite possible le jour même. En d'autres termes, toute la demande pour la chirurgie d'urgence doit être satisfaite le jour d'arrivée quelque soit la capacité disponible.

Les patients urgents sont classés en plusieurs groupes selon leurs besoins médicaux. Une salle opératoire est associée à chaque groupe. Quand un patient urgent arrive, celui ci sera opéré dans la salle correspondante à son groupe dès qu'elle devient disponible. Notons que ces salles opératoires ne sont pas réservées exclusivement aux patients urgents, elles sont utilisées aussi bien pour la chirurgie d'urgence que la chirurgie élective.

Dans ce travail, nous supposons qu'une portion aléatoire de la capacité de chaque salle-jour est utilisée pour réaliser les opérations urgentes. Désignons par W_{ts} la capacité consommée en salle-jour (s, t) par la chirurgie d'urgence. W_{ts} représente la durée totale des opérations urgentes réalisées en salle-jour (s, t) ; elle est considérée comme une variable aléatoire avec une loi de distribution $f_{W_{ts}}(.)$. Cette distribution peut être estimée à partir de l'historique des données.

Coûts d'utilisation des salles opératoires

Chaque salle-jour (s, t) dispose d'une capacité régulière T_{ts} et une capacité en heure supplémentaire V_{ts}. La capacité régulière représente la durée d'ouverture en heures normales de la salle opératoire s en jour t. Si cette capacité n'est pas suffisante on peut avoir recours à l'usage d'heures supplémentaires. Le nombre maximal d'heures supplémentaires autorisées représente la capacité en heures supplémentaires. La somme de ces deux capacités ($T_{ts} + V_{ts}$) représente la capacité totale de la salle-jour (s, t). Si la durée totale des interventions programmées plus les chirurgies d'urgence dépasse la capacité régulière, alors des heures supplémentaires sont nécessaires. Nous désignons par c_{ts} le coût par unité de temps d'heures supplémentaires en salle-jour (s, t). Si cette même durée totale dépasse la capacité totale, alors une pénalité supplémentaire est encourue. Soit $\overline{c}_{ts}$ une pénalité par unité de temps associée au dépassement de la capacité totale en salle-jour (s, t).

Dans ce travail, nous imposons que la durée totale des interventions programmées dans une salle-jour ne dépasse pas sa capacité totale (régulière plus en heure supplémentaire). Le dépassement de la capacité totale ne peut être dû qu'à la chirurgie d'urgence.

Si la durée totale des chirurgies programmées en salle-jour (*s*, *t*) est égale à *D*, $D \leq T_{ts} + V_{ts}$, alors le coût moyen de sur-utilisation de la salle-jour peut être exprimé comme suit :

$$c_{ts}\ E_{W_{ts}}\left[\left(W_{ts}+D-T_{ts}\right)^{+}\right]+\ \bar{c}_{ts}\ E_{W_{ts}}\left[\left(W_{ts}+D-T_{ts}-V_{ts}\right)^{+}\right]. \tag{4.1}$$

Les espérances mathématiques sont déterminées relativement à la distribution de W_{ts}, et $(x)^{+} = \max\{0, x\}$.

Une grande partie des coûts relatifs aux salles opératoires sont des coûts fixes, salaires du personnel médical, coûts fixes des équipements médicaux, etc (Denton and Gupta, 2003). Dans ce travail, nous prenons en compte aussi la sous-utilisation des salles opératoires lors de la planification des chirurgies.

Désignons par u_{ts} le coût par unité de temps de sous-utilisation de la salle-jour (s, t). Le coût moyen de sous-utilisation de la salle-jour peut être exprimé comme suit :

$$u_{ts}\ E_{W_{ts}}\left[\left(W_{ts}+D-T_{ts}\right)^{-}\right], \tag{4.2}$$

avec $(x)^{-} = \max\{0, -x\}$.

Donc, si la durée des chirurgies programmées en salle-jour (*s*, *t*) est égale à *D*, alors le coût moyen d'utilisation $g_{ts}(D)$ de la salle opératoire *s* en jour *t* est la somme du coût de sur-utilisation (4.1) et de sous-utilisation (4.2), c'est-à-dire,

$$\begin{aligned} g_{ts}(D) = {} & c_{ts}E\left[\left(W_{ts}+D-T_{ts}\right)^{+}\right]+\bar{c}_{ts}E\left[\left(W_{ts}+D-T_{ts}-V_{ts}\right)^{+}\right] \\ & +u_{ts}E\left[\left(W_{ts}+D-T_{ts}\right)^{-}\right] \end{aligned}. \tag{4.3}$$

Notons que même lorsqu'il n'y a pas de patients électifs planifiés en salle-jour (*s*, *t*), $D = 0$, le coût d'utilisation est $g_{ts}(0)$, qui n'est pas forcement nul.

Si la demande pour la chirurgie d'urgence est très élevée cela va générer un important usage en heures supplémentaires. Cependant, en pratique cela peut résulter en l'annulation de certaines interventions déjà programmées. Malheureusement, le modèle que nous proposons ci-dessous ne permet pas de prendre en compte l'annulation des chirurgies programmées.

4. 2. 2 Formulation mathématique

Le problème de planification des chirurgies consiste à affecter les patients programmables aux différentes salles-jours de telle sorte que la somme des coûts relatifs aux patients et des coûts d'utilisation des salles opératoires soit minimale.

Définissons les variables de décisions $x_{its} \in \{0,1\}$, avec x_{its} égale 1 si le patient électif i est affecté à la salle-jour (s, t) et 0 sinon. Par convention, $x_{i,H+1,s} = 1$ signifie que le patient i est rejeté du planning courant ; il sera re-considéré dans le prochain planning.

Le problème de planification est maintenant formulé par le programme mathématique suivant, appelé problème général (PG) :

$$J^* = \text{Minimiser } J(X) = \sum_{i=1}^{N}\sum_{t=e_i}^{H+1}\sum_{s=1}^{S} a_{its}x_{its} + \sum_{t=1}^{H}\sum_{s=1}^{S}\left[c_{ts}O_{ts} + \overline{c}_{ts}\overline{O}_{ts} + u_{ts}U_{ts}\right] \tag{4.4}$$

sous contraintes :

$$O_{ts} = E_{W_{ts}}\left[\left(W_{ts} + \sum_{i=1}^{N} x_{its}d_i - T_{ts}\right)^+\right], \forall\, t \in \{1,\ldots, H\}, s \in \{1,..., S\} \tag{4.5}$$

$$\overline{O}_{ts} = E_{W_{ts}}\left[\left(W_{ts} + \sum_{i=1}^{N} x_{its}d_i - T_{ts} - V_{ts}\right)^+\right], \forall\, t \in \{1,\ldots, H\}, s \in \{1,..., S\} \tag{4.6}$$

$$U_{ts} = E_{W_{ts}}\left[\left(W_{ts} + \sum_{i=1}^{N} x_{its}d_i - T_{ts}\right)^-\right], \forall\, t \in \{1,\ldots, H\}, s \in \{1,..., S\} \tag{4.7}$$

$$\sum_{i=1}^{N} x_{its}d_i \le T_{ts} + V_{ts}, \forall\, t \in \{1,\ldots, H\}, s \in \{1,..., S\} \tag{4.8}$$

$$\sum_{t=e_i}^{H+1}\sum_{s=1}^{S} x_{its} = 1, \ \forall\, i \in \{1,\ldots, N\} \tag{4.9}$$

$$x_{its} \in \{0,1\}, \forall\, i \in \{0, 1\}, t \in \{1,\ldots, H{+}1\}, s \in \{1,\ldots, S\} \tag{4.10}$$

où les espérances mathématiques sont relatives aux distributions des W_{ts}, $(y)^+ = \max\{0, y\}$, et $(y)^- = \max\{0, -y\}$.

La fonction objectif (4.4) est la somme des coûts relatifs aux patients électifs et les coûts moyens d'utilisations des salles opératoires. Les contraintes (4.5), (4.6) et (4.7) déterminent respectivement, le dépassement moyen de la capacité régulière, le dépassement moyen de la capacité totale, et la sous-utilisation moyenne. Les contraintes (4.8) imposent que pour chaque salle-jour la durée totale des chirurgies électives programmées ne dépasse pas la capacité totale. Les contraintes (4.9) sont les contraintes d'affectation ; elles assurent que chaque patient programmable est planifié exactement une fois.

Le problème de planification (PG) est un problème combinatoire stochastique. En utilisant une transformation polynomiale du problème de « *3-Partition* », on peut facilement prouver qu'il s'agit d'un problème NP-difficile au sens fort. Dans la section suivante, nous proposons une approche de génération de colonnes qui fournit des solutions proches de l'optimum.

4. 3 Approche de résolution : Génération de colonnes

Dans cette section, nous proposons une approche de résolution basée sur la génération de colonnes pour le problème général (PG). Les principales étapes de cette approche sont les suivantes. Dans un premier temps, nous allons reformuler le problème de planification sous la forme d'un problème mathématique comportant un nombre très élevé de variables (colonnes), appelé problème maître. La relaxation linéaire du problème maître est ensuite résolue en utilisant la génération de colonnes. À partir de la solution optimale du problème relâché un planning réalisable sera construit et sera enfin amélioré par une optimisation locale.

4. 3. 1 Approche de génération de colonnes

Dans cette section, nous proposons une nouvelle formulation pour le problème de planification. Pour ce faire, nous commençons par introduire le concept de *planning élémentaire*. Un planning élémentaire est un planning relatif à une salle opératoire donnée en un jour donné ; il sera appelé aussi *colonne*. Un planning élémentaire p est défini par les variables binaires suivantes :

$$y_{ip} = \begin{cases} 1 & \text{si le patient } i \text{ est affecté au planning élementaire } p, \\ 0 & \text{sinon.} \end{cases}$$

$$z_{tsp} = \begin{cases} 1 & \text{si le planning élémentaire } p \text{ est affecté à la salle - jour } (s,t), \\ 0 & \text{sinon.} \end{cases}$$

Donc le planning élémentaire p peut être représenté par le vecteur binaire $\left[y_p , z_p \right] = \left[\left(y_{1p} ,..., y_{Np} \right), \left(z_{1p} ,..., z_{(H \times S)p} \right) \right]$. Les N premières composantes, formant le vecteur y_p, représentent les patients électifs affectés au planning p. Les $H \times S$ composantes suivantes, c'est-à-dire z_p, indiquent à quelle salle-jour(s) le planning est affecté.

Soit Ω l'ensemble de tous les plannings élémentaires faisables. Un planning élémentaire p est faisable si les trois conditions suivantes sont vérifiées :

- (i) le planning est affecté à une et une seule salle-jour, c.-à-d., $\sum_{t,s} z_{tsp} = 1$,
- (ii) le planning respecte les dates au plus tôt des patients, c.-à-d., $y_{ip} \times z_{tsp} = 0, \forall\, t < e_i$,

- (iii) la durée totale des chirurgies électives affectées à ce planning ne dépasse pas la capacité totale de la salle-jour à laquelle le planning est affecté, autrement $\sum_i y_{ip} d_i \le (T_{ts} + V_{ts}) z_{tsp}, \quad \forall t, s$.

Notons que l'ensemble des plannings élémentaires, défini ci-dessus, ne concerne que les salles-jours comprises dans l'horizon de planification. Aucun planning élémentaire n'est associé à la période « virtuelle » H+1.

Le coût moyen d'un planning élémentaire peut être exprimé comme suit :

$$C_p = \sum_{t,s} \left[z_{tsp} \sum_i y_{ip} a_{its} + g_{ts} \left(\sum_{i=1}^{N} y_{ip} d_i \right) \right], \tag{4.11}$$

où $g_{ts}\left(\sum_{i=1}^{N} y_{ip} d_i\right)$ est le coût moyen d'utilisation de la salle s en jour t, définie par l'équation (4.3), lorsque la durée totale des chirurgies programmées est égale à $\sum_{i=1}^{N} y_{ip} d_i$. On peut facilement remarquer que le coût du planning est constitué de deux parties : les coûts relatifs aux patients programmés et le coût moyen d'utilisation de la salle opératoire (sous-utilisation et sur-utilisation).

En utilisant la notion des plannings élémentaires, le problème de planification peut être maintenant considéré comme un problème de sélection des plannings élémentaires. Introduisons λ_p, pour tout $p \in \Omega$, comme une variable de décision binaire indiquant si le planning élémentaire p est sélectionné ($\lambda_p = 1$) ou non ($\lambda_p = 0$). Le problème de planification est maintenant formulé comme suit :

$$\text{Min} \sum_{p\in\Omega} C_p \lambda_p \; + \; \sum_i a_{iH+1} (1 - \sum_{p\in\Omega} y_{ip} \lambda_p) \; + \; \sum_{t,s} g_{ts}(0) \times (1 - \sum_{p\in\Omega} z_{tsp} \lambda_p) \tag{4.12}$$

sous contraintes :

$$\sum_{p\in\Omega} y_{ip} \lambda_p \le 1, \; \forall i = 1,..., N \tag{4.13}$$

$$\sum_{p\in\Omega} z_{tsp} \lambda_p \le 1, \; \forall t=1,...,H \; , \; \forall s=1,...,S \tag{4.14}$$

$$\lambda_p \in \{0,1\}, \; \forall p \in \Omega \tag{4.15}$$

Les contraintes (4.13) assurent que chaque patient est inclu au plus dans un planning sélectionné, c'est-à-dire qu'il est affecté au plus à une salle-jour sur l'horizon de planification. Si un patient i est inclus dans un planning sélectionné ($\sum_{p\in\Omega} y_{ip} \lambda_p = 1$), il est affecté à une salle-jour dans l'horizon de planification. Et s'il n'est inclus dans aucun planning sélectionné ($\sum_{p\in\Omega} y_{ip} \lambda_p = 0$), il sera affecté à la période virtuelle H+1.

Les contraintes (4.14) garantissent qu'il y a au plus un planning élémentaire sélectionné pour chaque salle-jour. Par conséquent, certaines salles-jours peuvent ne pas recevoir de plannings, c'est-à-dire, il n'y a aucun patient électif planifié dans ces salles jours.

La fonction objectif (4.12) exprime toujours le coût d'un planning global, mais maintenant exprimé en utilisant les nouvelles variables λ_p. Elle est constituée de trois parties :

- les coûts des plannings élémentaires sélectionnés : $\sum_{p\in\Omega} C_p\lambda_p$
- les coûts engendrés par les patients non planifiés: $\sum_i a_{iH+1}(1-\sum_{p\in\Omega} y_{ip}\lambda_p)$
- les coûts d'utilisation des salles opératoires n'ayant pas reçu de plannings : $\sum_{t,s} g_{ts}(0)\times(1-\sum_{p\in\Omega} z_{tsp}\lambda_p)$

Soit $\tilde{C}_p = C_p - \sum_i a_{iH+1}\, y_{ip} - \sum_{t,s} g_{ts}(0)\, z_{tsp}$, le coût *modifié* du planning élémentaire *p*. En regroupant les termes, le problème de planification (4.12)-(4.15) peut être re-exprimé sous la forme du problème mathématique suivant, appelé problème maître (PM) :

$$\text{(PM):}\quad \text{Min} \sum_i a_{iH+1} + \sum_{t,s} g_{ts}(0) + \sum_{p\in\Omega} \tilde{C}_p\lambda_p \tag{4.16}$$

sous contraintes :

$$\sum_{p\in\Omega} y_{ip}\lambda_p \le 1,\ \forall\, i \tag{4.17}$$

$$\sum_{p\in\Omega} z_{tsp}\lambda_p \le 1,\ \forall\, t, s \tag{4.18}$$

$$\lambda_p \in \{0,1\},\ \forall p \in \Omega \tag{4.19}$$

Notons que le problème maître est un problème linéaire (en nombres entiers) contrairement à la formulation initiale (PG, problème général) où la fonction objectif est non linéaire. En effet, les quantités non linéaires sont maintenant incorporées dans les coûts des colonnes $\tilde{C}_p$. On voit bien aussi que les colonnes ne sont autres que les plannings élémentaires [y_p, z_p].

Le problème maître (PM) est un programme mathématique comportant un très grand nombre de variables (colonnes), et ne peut pas être résolu directement. Nous allons commencer par résoudre la relaxation linéaire du PM, appelé problème maître linéaire (PML), en utilisant la méthode de génération de colonnes. Le problème maître linéaire est obtenu à partir de PM en remplaçant les contraintes d'intégralité $\lambda_p \in \{0,1\}$ par $\lambda_p \ge 0$.

Afin de résoudre le problème maître linéaire PML, on commence par résoudre un problème maître restreint (PMR) qui considère seulement un sous-ensemble $\Omega' \subset \Omega$ de colonnes. Le sous-ensemble Ω' peut être initialisé, par exemple, avec des colonnes générées de manière

aléatoire ou en utilisant une heuristique. Ensuite, des colonnes supplémentaires sont intégrées seulement si cela est nécessaire, selon le schéma suivant.

Soient π_i et π_{ts} les solutions duales optimales d'un PMR, associées respectivement aux contraintes (4.17) et (4.18). Le problème de *génération de colonnes* (ou *problème générateur*)

$$\sigma_p = \min_{\{y_{ip}, z_{tsp}\}} \tilde{C}_p - \sum_i \pi_i y_{ip} - \sum_{t,s} \pi_{ts} z_{tsp} \quad \text{tel que} \quad p \in \Omega$$

permet d'identifier une colonne p qui a un coût réduit σ_p minimal. Si le coût réduit σ_p est inférieur à 0, alors la colonne p (c.-à-d., le planning élémentaire [y_p, z_p]) peut être ajoutée au problème restreint PMR. Si σ_p est supérieure ou égale à 0, alors la solution optimale actuelle du PRM est aussi optimale pour le problème maître linéaire PML.

Il est évident que le coût optimal du problème maître linéaire PML représente un borne inférieure du coût optimal du problème maître PM, et par conséquent une borne inférieure du coût optimal du problème de planification initial PG.

4. 3. 2 Résolution du problème de génération de colonnes

Le problème de génération de colonnes consiste à identifier la colonne (le planning élémentaire) qui a le coût réduit minimal parmi les colonnes faisables. À partir de la définition des planning faisables, le problème de génération de colonnes (*problème générateur*) peut être formulé comme suit :

$$\text{Min} \quad \tilde{C}_p - \sum_{i=1}^{N} \pi_i y_{ip} - \sum_{t=1}^{H} \sum_{s=1}^{S} \pi_{ts} z_{tsp} \tag{4.20}$$

sous contraintes :

$$y_{ip} \times z_{tsp} = 0, \ \forall\ t < e_i, \tag{4.21}$$

$$\sum_i y_{ip} d_i \le (T_{ts} + V_{ts}) z_{tsp} \tag{4.22}$$

$$\sum_{t,s} z_{tsp} = 1 \tag{4.23}$$

$$y_{ip} \in \{0, 1\}, \ z_{tsp} \in \{0, 1\}, \ \forall\ i, t, s \tag{4.24}$$

avec $\tilde{C}_p = \sum_{t,s} \left[z_{tsp} \sum_i y_{ip} a_{its} + g_{ts} \left(\sum_{i=1}^{N} y_{ip} d_i \right) \right] - \sum_i a_{iH+1} y_{ip} - \sum_{t,s} g_{ts}(0) z_{tsp}$.

Étant donné les solutions duales optimales, π_i et π_{ts}, du problème maître restreint actuel, la fonction objectif (4.20) détermine le coût réduit d'une colonne. Les contraintes (4.21) garantissent qu'aucun patient n'est planifié avant sa date au plus tôt. Les contraintes (4.22) garantissent que la durée totale des chirurgies électives affectées à la colonne ne dépasse pas

la capacité totale (capacité régulière et capacité en heure supplémentaire) de la salle-jour à laquelle la colonne est associée. La contrainte (4.23) assure que la colonne est associée à une et une seule salle-jour.

Le problème de génération de colonnes (4.20)-(4.24) peut être décomposé en $H \times S$ sous-problèmes, un par salle-jour. Il peut être résolu en résolvant le sous-problème de génération de colonne (GC_{ts}) pour chaque salle-jour :

$$(CG_{t,s}) : \text{ Min} \sum_{i=1}^{N} \tilde{a}_{its} y_{ip} + g_{ts}\left(\sum_{i=1}^{N} y_{ip} d_i\right) - g_{ts}(0) - \pi_{ts} \tag{4.25}$$

sous contraintes :

$$y_{ip} = 0\,, \ \forall i \text{ avec } t < e_i \tag{4.26}$$

$$\sum_i y_{ip} d_i \le (T_{ts} + V_{ts}), \quad \forall t, s \tag{4.27}$$

$$y_{ip} \in \{0, 1\}, \forall\, i = 1,..., N \tag{4.28}$$

où $\tilde{a}_{its} = a_{its} - a_{iH+1} - \pi_i$.

Donc, le problème de génération de colonnes peut être résolu en résolvant $H \times S$ sous-problèmes de génération de colonne (GC_{ts}), et ensuite choisir la solution (colonne) qui a le coût réduit le plus petit.

Le problème de génération de colonnes permet d'identifier la colonne qui a le coût réduit minimal. Donc, s'il existe une colonne qui un coût réduit négatif, elle sera forcément identifiée. C'est cette propriété qui garantit l'optimalité de la solution du problème maître linéaire PML (Barnhart *et al*., 1998). Cependant, on peut également choisir d'ajouter une colonne quelconque qui a un coût réduit négatif.

Dans ce travail, nous utilisons trois stratégies pour ajouter les colonnes au problème maître restreint PMR. La première stratégie n'est autre que la règle classique de Dantzig (Lübbecke and Desrosiers, 2005) ; elle consiste à ajouter la colonne qui a le coût réduit négatif le plus petit (*best-negative strategy*). La seconde stratégie consiste à ajouter une colonne pour chaque salle-jour si son coût réduit est négatif (*all-negative strategy*). Pour une salle-jour donnée la colonne sélectionnée est celle qui a le coût réduit négatif le plus petit. La troisième stratégie consiste à résoudre les sous- problèmes (GC_{ts}) dans un ordre aléatoire, et dès qu'un sous-problème fournit une colonne améliorante on s'arrête (*first-negative strategy*). Bien évidemment, les colonnes qui ont un coût réduit supérieur ou égal à 0 ne sont jamais ajoutées, quelque soit la stratégie utilisée.

Dans le reste de cette section, nous présentons un algorithme de programmation dynamique pour la résolution du sous-problème de génération de colonnes. Pour un ensemble de

multiplicateurs π_i et π_{ts} donnés, le sous-problème (GC_{ts}) consiste à déterminer les patients électifs qui seront opérés dans la salle-jour (s, t).

Dans la fonction objectif (4.25), le terme $-g_{ts}(0)-\pi_{ts}$ est une constante, par conséquent il sera supprimé du critère. Définissons pour chaque jour t, l'ensemble $I_t = \{i / e_i \leq t\}$ de patients électifs qui peuvent être planifiés pour ce jour.

Les patients n'appartenant pas à l'ensemble I_t ne peuvent pas être opérés en jour t ; par conséquent, $y_{ip} = 0,\ \forall i \notin I_t$. Le sous-problème (GC_{ts}) peut être formulé comme suit :

$$\text{Min} \sum_{i \in I_t} \tilde{a}_{its} y_{ip} + g_{ts}(D) \tag{4.29}$$

sous contraintes:

$$\sum_{i \in I_t} y_{ip} d_i = D \tag{4.30}$$

$$0 \leq D \leq D_{max} \tag{4.31}$$

$$y_{ip} \in \{0, 1\},\ \forall i \in I_t, \tag{4.32}$$

avec $D_{max} = T_{ts} + V_{ts}$, la capacité régulière plus la capacité en heure supplémentaire.

Remarque : Des contraintes supplémentaires relatives à la disponibilité des chirurgiens, des patients, ou des salles opératoires peuvent être facilement prises en compte. En effet, pour chaque salle-jour (s, t) on peut définir un ensemble I_{st} des patients candidats qui respecte ces contraintes supplémentaires. Ensuite, il suffit d'utiliser l'ensemble I_{st} à la place de I_t.

Pour tout $0 \leq D \leq D_{max}$, définissons le problème du sac à dos binaire (avec contrainte d'égalité) suivant :

$$h(D) = \min\left\{\sum_{i \in I_t} \tilde{a}_{its} y_{ip} : \sum_{i \in I_t} d_i y_{ip} = D, y_{ip} \in \{0,1\} \forall i \in I_t\right\} \tag{4.33}$$

Il en découle donc que le sous-problème (GC_{ts}) est équivalent à :

$$\min_{0 \leq D \leq D_{\max}} \left\{h(D) + g_{ts}(D)\right\} \tag{4.34}$$

Les différentes valeurs de $g_{ts}(D)$ peuvent être évaluées par intégration numérique par rapport à la distribution de W_{ts}. Les différentes valeurs de $h(D)$ peuvent être déterminées en utilisant la programmation dynamique pour la résolution du problème (4.33) avec $D = D_{max}$.

Étant donné une paire d'entiers m ($1 \leq m \leq |\ I_t\ |$) et D ($0 \leq D \leq D_{max}$), définissons le problème réduit suivant :

$$h_m(D) = \min\left\{\sum_{i=1}^{m} \tilde{a}_{its} y_{ip} : \sum_{i=1}^{m} d_i y_{ip} = D, y_{ip} \in \{0,1\}\right\} \tag{4.35}$$

L'algorithme de programmation dynamique consiste à considérer | I_t | étapes (pour m variant de 1 à | I_t |). A chaque étape, les valeurs de $h_m(D)$ (pour D variant de 0 à D_{max}) sont déterminées en utilisant l'équation de récurrence suivante :

$$h_m(D) = \begin{cases} h_{m-1}(D) & \text{pour } 0 \le D \le d_m - 1 \\ min\left(h_{m-1}(D),\ h_{m-1}(D - d_m) + \tilde{a}_{mts}\right) & \text{pour } d_m \le D \le D_{max} \end{cases}$$

avec les conditions initiales suivantes :

$$h_1(D) = \begin{cases} 0 & \text{pour } D = 0 \\ \tilde{a}_{1ts} & \text{pour } D = d_1 \\ +\infty & \text{pour } D = 1 \ldots D_{max}, D \ne d_1 \end{cases}$$

La solution optimale du sous-problème (GC_{ts}) est la solution correspondant à l'état $h_{|I_t|}(D^*)$, avec

$$D^* = \underset{0 \le D \le D_{max}}{arg\,min}\ h_{|I_t|}(D) + g_{ts}(D) \tag{4.36}$$

La méthode de résolution que nous avons développée est similaire à celle de l'algorithme de programmation dynamique pour le sac à dos classique, mais qui tient compte de la contrainte d'égalité et du coût non linéaire.

4. 3. 3 Construction d'une « bonne » solution réalisable

Étant donné une solution λ_p optimale du problème maître linéaire (PML), la solution du problème général (PG) peut en être déduite en utilisant l'expression suivante :

$$x_{its} = \sum_{p \in \Omega} y_{ip} z_{tsp} \lambda_p$$

Si les variables λ_p sont toutes entières, alors les x_{its} sont à leur tour aussi entières. Ceci est équivalent à dire que toute solution entière du PML a une solution correspondante faisable pour GP, c'est-à-dire un planning réalisable. Cependant, si la solution du PML est fractionnaire, alors sa correspondante l'est aussi et par conséquent elle fournit un planning non réalisable puisque certains patients sont affectés aux salles-jours de manière fractionnaire.

Donc, à moins que la solution optimale du PML soit entière, la solution fournit par la génération de colonnes n'est pas faisable pour le problème maître PM, et le planning correspondant n'est pas réalisable. Nous avons développé des méthodes heuristiques pour la construction d'un planning réalisable de bonne qualité en exploitant la solution optimale du problème maître linéaire PML. Dans un premier temps, une solution réalisable sera construite. Ensuite, elle sera améliorée en utilisant une ou plusieurs méthodes d'optimisation locale.

4. 3. 3. 1 Construction d'un planning réalisable

Une première approche pour la recherche d'une solution entière consiste à lancer une recherche arborescente classique de type *branch and bound* (ou *branch and cut*) sur le problème maître restreint à l'ensemble Ω' des colonnes explicitées lors du processus de génération colonnes. C'est une approche heuristique car, même si l'ensemble de colonnes Ω' permet de trouver la solution optimale continue du PM à la racine de l'arborescence, on n'est pas assuré d'obtenir la solution optimale aux autres nœuds de l'arbre. En effet, des colonnes, hors Ω', sont susceptibles d'améliorer la solution au cours de la recherche arborescente.

Dans ce qui suit, nous présentons deux autres méthodes heuristiques pour la construction d'une solution entière.

Réaffectation complète

Étant donnée une solution $\{\lambda_p\}$ optimale du problème maître linéaire PML, les principales étapes de cette méthode sont les suivantes. Au début, les patients affectés à des plannings élémentaires (colonnes) sélectionnés, avec $\lambda_p = 1$, sont affectés aux salles-jours associées à ces plannings. Soit X la solution partielle formée par ces affectations. Ensuite, tous les autres patients sont affectés un par un en prenant en compte les patients déjà affectés.

À chaque itération, un patient i non encore affecté est sélectionné. Désignons par $X^{(s', t')}$ les solutions obtenues à partir de X en en ajoutant le patient i dans la salle-jour (s', t'), tout en gardant les affectations des patients précédemment affectés. Soient $J(X^{(s', t')})$ les coûts de ces nouvelles solutions. Le patient i est finalement affecté à la salle-jour (s_i, t_i) qui minimise $J(X^{(s', t')})$, c'est-à-dire, $(s_i, t_i) = \text{argmin}_{\{(s', t')\}} J(X^{(s', t')})$. Ensuite, X est remplacée par $X^{(s_i, t_i)}$, et l'algorithme continue jusqu'à ce que tous les patients soient affectés. Les patients sont ajoutés un par un dans un ordre arbitraire.

Réaffectation progressive

Désignons par x_{its} la solution du problème général correspondant à la solution optimale du problème maître linéaire PML, c'est-à-dire, $x_{its} = \sum_{p \in \Omega} y_{ip} z_{tsp} \lambda_p$. Dans un premier temps, on sélectionne un patient électif i dont la matrice d'affectations $[x_{its}]_{ts}$ contient des fractions, c'est-à-dire affecté de manière fractionnaire à plusieurs salles-jours. Ensuite, considérons les solutions $X^{(s', t')}$ obtenues à partir de $X = [x_{its}]_{its}$ en réaffectant le patient i seulement à la salle-jour (s', t'), et en gardant les affectations des autres patients même si elles sont fractionnaires.

Le patient i est finalement affecté à la salle-jour (s_i, t_i) qui minimise $J(X^{(s', t')})$, c'est-à-dire $(s_i, t_i) = \text{argmin}_{\{(s', t')\}} J(X^{(s', t')})$. Bien évidemment, afin que la solution soit réalisable. La solution

X est ensuite remplacée par $X^{(s_i, t_i)}$, et le processus continue pour les autres patients ayant des affectations fractionnaires. L'ordre de traitement de ces patients est arbitraire.

4. 3. 3. 2 Heuristiques d'amélioration

Étant donnée une solution faisable (un planning réalisable), celle ci sera améliorée par les heuristiques suivantes.

Optimisation locale par réaffectation

La solution est itérativement améliorée en réaffectant les patients électifs. À chaque itération, on détermine pour chaque patient le meilleur gain qui peut être réalisé en le réaffectant à une autre salle-jour (ou à la période H+1) et en gardant les affectations des autres patients. Bien évidemment, la nouvelle affectation doit satisfaire la date au plus tôt du patient ainsi que les capacités totales des salles-jours. Ensuite, on choisit le patient dont la réaffectation apporte le gain le plus important, et on le réaffecte. Ce processus est répété jusqu'à ce qu'on ne puisse plus améliorer la solution.

Optimisation locale par permutation

Le principe de cette méthode est d'itérativement améliorer la solution en permutant à chaque fois les affectations d'une paire de patients électifs.

À chaque itération, on considère un patient i, et on détermine un patient j pour permuter leurs affectations. Le patient j est celui qui apporte le gain le plus important lorsqu'il est permuté avec le patient i, sans toucher aux affectations des autres patients et tout en satisfaisant les dates au plus tôt des patients et les capacités totales des salles-jours. Ensuite, les deux patients i et j sont permutés. À l'itération suivante, un autre patient est considéré et ainsi de suite. Ce processus est répété jusqu'à ce qu'il n'y a plus de permutations améliorantes.

Optimisation locale orientée périodes

On considère les patients affectés à une salle-jour donnée et les patients affectés à la période H+1 (rejetés du planning courant) et on re-optimise la répartition de ces patients entre la salle-jour et la période H+1.

Désignons par I_{st} l'ensemble des patients affectés à la salle-jour (s, t), et par I_{H+1} l'ensemble des patients affectés à la période H+1. Introduisons maintenant $Z_i \in \{0, 1\}$, pour tout $i \in I_{st} \cup I_{H+1}$, comme variable de décision indiquant si le patient est affecté à la salle-jour (s, t) ($Z_i = 1$) ou à la période H+1 ($Z_i = 0$) dans la nouvelle solution. Le problème de ré-optimisation relatif à la salle-jour (s, t) est formulé comme suit :

(RP_{st}): $$\text{Min } \sum_{i\in I_{st}\cup I_{H+1}} a_{its}Z_i + \sum_{i\in I_{st}\cup I_{H+1}} a_{iH+1}(1-Z_i) + g_{ts}\left(\sum_{i\in I_{st}\cup I_{H+1}} d_i Z_i\right) \quad (4.37)$$

sous contraintes :

$$\sum_{i\in I_{st}\cup I_{H+1}} d_i Z_i \le T_{ts} + V_{ts} \quad (4.38)$$

$$Z_i \in \{0, 1\},\ \forall\ i \in I_{st} \cup I_{H+1}. \quad (4.39)$$

La fonction objectif (4.37) est composée des coûts relatif aux (i) patients affectés à la salle-jour (*s*, *t*), (ii) patients rejetés, et (iii) coûts d'utilisation de la salle opératoire *s* en jour *t*. (4.38) est la contrainte de capacité.

Par simple arrangement des termes, le problème de ré-optimisation peut être reformulé comme suit :

$$\min \sum_{i\in I_{st}\cup I_{H+1}} a_{iH+1} + \sum_{i\in I_{st}\cup I_{H+1}} (a_{its} - a_{iH+1})\ Z_i + g_{ts}\left(\sum_{i\in I_{st}\cup I_{H+1}} d_i Z_i\right)$$

sous les contraintes : (4.38) et (4.39).

Ce problème est similaire au sous-problème de génération de colonnes. Pour sa résolution, on peut utiliser la méthode de programmation dynamique présentée à la section 4.3.2.

À chaque itération, on sélectionne une salle-jour (*s*, *t*), on résout le problème de ré-optimisation (RP_{st}) correspondant, et on re-affecte les patients comme suggéré par la solution du (RP_{st}). À l'itération suivante une autre salle-jour est considérée et ainsi de suite. Le processus s'arrête quand on ne peut plus améliorer la solution. Dans ce travail, nous considérons les salles-jours dans un ordre chronologique.

4. 3. 4 Combinaisons des différentes heuristiques

La méthodologie de résolution proposée pour la résolution du problème de planification est composée de trois phases. Dans la première phase, le problème maître linéaire PML est résolu en utilisant la génération de colonnes : nous utilisons CPLEX LP pour résoudre le problème maître réduit PMR et la programmation dynamique pour les sous-problèmes de génération de colonnes. En seconde phase, on construit une solution faisable (planning réalisable) à partir de la solution optimale du PML. Cette phase peut être réalisée en utilisant l'une des méthodes suivantes : (i) *branch and bound* (**CPLEX IP**) en se restreignant aux colonnes générées lors du processus de génération de colonnes ; (ii) l'heuristique de réaffectation complète (**RC**), l'heuristique de réaffectation progressive (**RP**). La troisième phase a pour but d'améliorer la qualité de la solution faisable en utilisant une combinaison des heuristiques suivantes : optimisation locale (**OL**), optimisation par permutation (**OP**), et optimisation orientée période (**OOP**).

Dans ce travail, nous avons considéré 7 combinaisons de ces méthodes. Ces combinaisons sont présentées en tableau 4.1. Chaque ligne représente une combinaison désignée par M#.

Méthode	Résolution du PML	Construction d'une solution faisable	Amélioration de la solution faisable
M1	Génération de	CPLEX IP	
M2	colonnes :	RC	OL
M3		RP	OL
M4	CPLEX LP	RP	OP, OOP
M5	+	RP	OP, OL, OOP
M6	Programmation	RP	OL, OOP, OP
M7	dynamique	RP	OL, OP

TAB. 4.1 - Les différentes combinaisons

Afin d'évaluer les performances des différentes méthodes, nous utilisons le gap de dualité (GAP) comme mesure de performances :

$$GAP = (UB - LB) / UB$$

UB est le coût de la solution faisable (le planning final) ; il représente une borne supérieure du coût optimal J^* du problème général. *LB* est le coût optimal du problème maître linéaire (PML) fournit par la génération de colonnes, et qui représente une borne inférieure de J^*.

Notons que lors de la résolution du problème maître linéaire (PML), trois stratégies de générations de colonnes peuvent être utilisées pour identifier la(les) nouvelle(s) colonnes entrantes: best-negative, all-negative, et first-negative (section 4.3.2).

4. 4 Expérimentations numériques

Les tests numériques présentés dans cette section ont été réalisés sur un PC équipé d'un processeur Pentium 4 à 2.8 GHz avec une mémoire de 512 Mo, utilisant un système d'exploitation Windows XP. Les algorithmes ont été programmées en MS Visual C++ et faisant appel à la bibliothèque d'optimisation CPLEX 8.1.

4. 4. 1 Génération des instances

Deux classes de problèmes sont considérées. Pour la première classe, les patients électifs peuvent être affectés à n'importe quelle salle opératoire sans aucun coût additionnel, c'est-à-dire, pour chaque patient *i* les coûts d'affectation a_{its} dépendent seulement de la date de chirurgie *t*. Pour la seconde classe, les salles sont de différents types et sont allouées à différentes spécialités. Nous supposons qu'il y a trois spécialités et que les salles sont allouées de manière équitable entre ces spécialités. Par exemple, si le bloc opératoire est composé de 6

salles, alors 2 salles sont allouées à chaque spécialité. Les salles allouées à la même spécialité sont considérées comme identiques. Cependant, des salles allouées à des spécialités différentes sont considérées de types différents. Ces différences peuvent être dues à l'emplacement des salles, à la disponibilité de certains équipements médicaux, etc. Pour cette classe de problèmes, un patient électif appartenant à une spécialité donnée peut être affecté soit à une salle allouée à la spécialité en question, soit à une salle allouée à une autre spécialité, mais dans ce cas il y a un coût supplémentaire, une pénalité. En effet, chaque spécialité peut nécessiter certains équipements médicaux spéciaux ou certaines préparations préliminaires dans la salle opératoire. Donc, ces coûts supplémentaires sont dus à des temps de set-up ou de préparation additionnels.

Des problèmes avec 3, 6, 9 et 12 salles opératoires sont considérées pour chaque classe de problème. Les instances des problèmes sont générées de manière aléatoire comme expliquée ci-dessous.

Le nombre de jours H est égal à 5 (un horizon d'une semaine). Les capacités en heures régulières T_{ts} et les capacités en heures supplémentaires V_{ts} des salles-jours sont respectivement fixées à 8 et à 3 heures. Les coûts des heures supplémentaires c_{ts}, la pénalité de dépassement de la capacité totale $\overline{c}_{ts}$, et les coûts de sous-utilisation u_{ts} sont respectivement égaux à 500 €/heure, 3000 €/heure, et $c_{ts}/1.75$. La capacité aléatoire utilisée par la chirurgie d'urgence W_{ts} en chaque salle-jour est supposée être exponentiellement distribuée avec une moyenne de 2 heures ($E[W_{ts}] = 2$ heures).

Les durées d_i des chirurgies électives sont générées de manière aléatoire et uniforme à partir de l'intervalle [0.5 heure, 3 heures]. Ces durées sont multiples de 5 minutes.

Les dates au plus tôt sont générées aléatoirement à partir de l'intervalle $\{1,\ldots, H\}$. Afin de tenir compte des patients qui ont une date au plus $e_i = 1$ et qui ont été rejetés du planning précèdent, nous introduisons un nouveau paramètre e_i', date au plus tôt *effective*. Les dates au plus tôt sont générées en deux étapes. Dans un premier temps, on génère pour chaque patient i une date au plus tôt effective e_i'. Les e_i' sont générées aléatoirement et uniformément à partir de l'ensemble $\{-2, \ldots, H\}$. Ensuite, les patients avec e_i' négative ou nulle vont avoir leur $e_i=1$; pour les autres patients les valeurs de e_i seront les mêmes que e_i' ($e_i = 1$ si $e_i' < 1$; $e_i = e_i'$ si non).

Les coûts d'affectations a_{its} sont générés de deux manières différentes selon la classe de problème. Pour des problèmes de la première classe, les coûts d'affectation a_{its} d'un patient i sont indépendants de la salle opératoire « s » ; ils dépendent seulement du jour d'intervention t. Pour cette raison, pour un patient i et un jour t donnés, on a $a_{its} = a_{its'}$, $\forall s, s' \in \{1..S\}$.

Les coûts sont considérés comme croissants en fonction du jour *t*. Ils sont définis comme suit :

$$a_{its} = (t - e_i^{'}) \times c \quad \text{pour } t \in \{e_i, \ldots H\},\ s \in \{1..S\},$$

et

$$a_{i(H+1)} = ((H+1 - e_i^{'}) \times c) + 2 \times c.$$

La constante *c* peut représenter un coût d'hospitalisation par jour, une pénalité par jour d'attente, etc. La quantité $2 \times c$ dans le coût $a_{i(H+1)}$ représente une pénalité de non planification du patient. Pour les tests numériques, nous avons fixé *c* à 350 €.

Pour la deuxième classe de problèmes, les coûts d'affectations sont générés comme suit. Pour chaque patient *i*, nous avons

$$a_{its} = (t - e_i^{'}) \times c \quad \text{pour } t \in \{e_i, \ldots H\}$$

si la salle opératoire s est allouée à la spécialité traitant le patient, et si non

$$a_{its} = (t - e_i^{'}) \times c + R \quad \text{pour } t \in \{e_i, \ldots H\}.$$

Les coûts de non planifications $a_{i(H+1)}$ sont pareils à ceux de la première classe.

La constante *R* représente un coût supplémentaire (pénalité) engendré par l'affectation du patient à une salle opératoire pas tout à fait adéquate. Pour nos tests numériques nous avons utilisé différentes valeurs de *R* (100, 200, 300 et 400). Dans le reste de cette section nous considérons *R* égal 100, sauf si c'est explicitement mentionné.

Remarque : Si $R = 0$ alors les patients peuvent être affectés à n'importe quelle salle sans aucun coût additionnel. Donc, un problème avec salles identiques est équivalent à un problème avec salles non-identiques mais avec $R = 0$.

Les paramètres des coûts c_{ts} (coût des heures supplémentaires), u_{ts} (coût de sous-utilisation) et *c* sont similaires à ceux utilisées par (Jebali *et al.*, 2006) et (Guinet and Chaabane, 2003) ; ils reflètent les coûts dans les établissement hospitaliers français. Pour la pénalité $\overline{c}_{ts}$, une valeur élevée a été choisie, et ceci afin de pénaliser le dépassement de la capacité totale des salles opératoires. La pénalité *R* est introduite essentiellement pour des fins expérimentales. En utilisant ce paramètre, on peut varier les coûts d'affectation des patients et par la suite réaliser quelques études de sensibilité.

Les patients électifs sont générés un par un jusqu'à ce que la somme des durées opératoires dépasse un pourcentage τ % donné de la somme des capacités régulières de toutes les salles opératoires sur tout l'horizon de planification. Autrement, le nombre des patients électifs est déterminé de sorte que la charge engendrée par les chirurgies programmables soit égale à τ % de la capacité régulière disponible sur tout l'horizon de planification. Nous considérons des problèmes avec des pourcentages τ à 75 % et à 100 %.

Avec $E[W_{ts}]$ = 2 heures et T_{ts} = 8 heures, la charge due à la chirurgie d'urgence est 25 % de toute la capacité régulière disponible sur l'horizon de planification. Donc, avec un τ = 75 % la charge totale (due à la chirurgie programmable et d'urgence) des salles opératoires est 100 % de la capacité régulière. Et lorsque τ = 100 % la charge totale est à 125 %. Donc, avec τ = 75 % ou 100 %, on a des salles opératoires avec des charges élevées et par conséquent, des problèmes de planification suffisamment difficiles.

On peut facilement voir que, la taille du problème est déterminée par le nombre de salles opératoires et par le pourcentage τ. Pour les tests numériques, nous considérons des problèmes avec les tailles suivantes S = 3, 6, 9, 12 et $\boldsymbol{\tau}$ = 75%, 100%.

4. 4. 2 Évaluation et comparaison des différentes méthodes

Nous commençons par évaluer les performances des 7 méthodes (combinaisons) présentées dans le tableau 4.1. Ces méthodes sont testées avec des problèmes de différentes tailles (S = 3, 6, 9, 12 et $\boldsymbol{\tau}$ = 75 %) et appartenant aux deux classes de problèmes (salles identiques et salles non-identiques avec R = 100). La stratégie classique « best-negative » est utilisée pour la génération de colonnes. Le problème maître restreint démarre avec un ensemble Ω' de colonnes générées aléatoirement, et de sorte qu'elles forment une solution faisable pour le problème général (PG). Autrement, la solution $X = [x_{its}]$, avec $x_{its} = \sum_{p \in \Omega'} y_{ip} z_{tsp}$, vérifie les contraintes (4.8) et (4.9).

Les résultats numériques sont basés sur 10 instances générées aléatoirement, pour chaque taille et classe de problème en considération. Les résultats sont présentés dans les tableaux 4.2 et 4.3. Ces résultats comprennent, pour chaque taille de problème, le nombre moyen des patients programmables (Nombre des patients) et les performances des 7 méthodes (combinaisons). Pour chaque méthode, nous présentons le gap de dualité (GAP), le temps de calcul nécessaire pour la résolution du problème maître linéaire (CPU Linéaire), le temps de calcul nécessaire pour obtenir un « bon » planning réalisable (CPU Entier), et le temps de calcul total (CPU) qui est la somme des deux précédents.

Les résultats numériques montrent que la méthode M1 est très lente. Pour les problèmes de taille supérieure ou égale à 6 salles, la méthode M1 ne peut pas fournir de planning réalisable en un temps de calcul raisonnable ; « CPU Entier » dépasse les 2 heures. Cependant, les autres méthodes fournissent rapidement des plannings réalisables proches de l'optimum, en moins de 7 secondes. En plus, les solutions fournies par la méthode M1 ont des gaps de dualité très élevés en comparaison avec les autres méthodes. À partir de ces résultats, on déduit que la résolution du problème maître en se restreignant aux colonnes générées est gourmande en terme de temps de calcul, et que les plannings fournis peuvent être d'une très mauvaise qualité.

D'après les tableaux 4.2 et 4.3, on peut aussi observer que la Méthode M3, utilisant la re-affectation progressive (RP) pour la construction des plannings réalisables, est nettement meilleure que la méthode M2 qui utilise une réaffectation complète (RC). Ce constat peut être, en effet, expliqué par le fait que la réaffectation progressive préserve la structure de la solution optimale du problème linéaire maître PLM ; alors que la réaffectation complète prend en compte seulement la partie entière de la solution optimale du PLM.

On peut aussi remarquer que les méthodes de M3 à M7 ont des performances comparables. Toutefois, M6 et M7 sont meilleures que les autres. On note aussi que les problèmes avec des salles non-identiques nécessitent un temps de calcul plus long et présentent des gaps de dualité un peu plus élevés, en comparant aux problèmes avec des salles identiques.

Nombre de salles	Nombre de patients	Méthode #	GAP (%)	CPU Linéaire (sec)	CPU Entier (sec)	CPU (sec)
3	53.9	M1	18.31	8.60	43.00	51.60
		M2	3.78		0.15	8.76
		M3	2.72		0.12	8.72
		M4	2.83		0.05	8.65
		M5	2.13		0.08	8.68
		M6	1.52		0.17	8.77
		M7	1.54		0.14	8.75
6	106.9	M1	24.80*	42.76	7820.29	7863.05
		M2	2.81		0.93	43.69
		M3	2.27		0.72	43.48
		M4	1.74		0.15	42.90
		M5	1.69		0.25	43.00
		M6	1.40		1.08	43.84
		M7	1.40		0.80	43.56
9	160.1	M1	-	138.32	>8000	>8000
		M2	3.25		2.58	140.90
		M3	2.26		2.15	140.47
		M4	1.51		0.31	138.63
		M5	1.44		0.61	138.93
		M6	1.15		2.41	140.73
		M7	1.15		2.34	140.66
12	211.8	M1	-	344.30	>8000	>8000
		M2	2.86		6.29	350.60
		M3	1.99		4.70	349.00
		M4	1.33		0.57	344.87
		M5	1.29		1.13	345.44
		M6	0.88		5.45	349.75
		M7	0.88		5.05	349.35

* Based on two instances

TAB. 4.2 - Résultats pour salles identique avec τ = 75%

Nombre de salles	Nombre de patients	Méthode #	GAP (%)	CPU Linéaire (sec)	CPU Entier (sec)	CPU (sec)
3	53.9	M1	0.95	17.34	1.06	18.40
		M2	3.91		0.06	17.40
		M3	1.56		0.06	17.40
		M4	2.29		0.04	17.38
		M5	1.30		0.08	17.43
		M6	0.87		0.12	17.46
		M7	0.93		0.09	17.43
6	106.9	M1	-	76.99	>8000	>8000
		M2	5.68		0.89	77.88
		M3	2.47		0.54	77.53
		M4	2.05		0.15	77.14
		M5	1.86		0.29	77.28
		M6	1.53		0.73	77.72
		M7	1.53		0.63	77.62
9	160.1	M1	-	190.40	>8000	>8000
		M2	5.83		2.31	192.70
		M3	2.82		2.00	192.40
		M4	2.13		0.32	190.72
		M5	2.06		0.62	191.02
		M6	1.58		2.66	193.06
		M7	1.57		2.18	192.58
12	211.8	M1	-	402.84	>8000	>8000
		M2	5.68		5.91	408.76
		M3	3.16		4.68	407.53
		M4	2.17		0.58	403.42
		M5	2.10		1.38	404.22
		M6	1.70		6.01	408.85
		M7	1.70		5.03	407.87

TAB. 4.3 - Résultats pour salles non-identiques avec τ = 75%

Nous comparons maintenant les performances des méthodes M2 à M7 avec des problèmes de tailles (S = 3, 6, 9, 12 et τ = 100 %) et appartenant aux deux classes de problèmes. Les résultats numériques sont présenté dans le tableau 4.4, et sont basés sur 10 instances générées aléatoirement, pour chaque taille et classe de problème en considération. Les résultats comprennent, pour chaque taille de problème, le nombre moyen des patients programmables (Nombre des patients), et le gap de dualité (GAP) de chacune des 6 méthodes. Les temps de calcul de ces méthodes sont proches de ceux présentés dans les tableaux 4.2 et 4.3.

À partir du tableau 4.4, on peut facilement observer que les méthodes M6 et M7 sont meilleures que les autres. Cependant, on remarque que la méthode M6 est légèrement meilleure que M7. Ceci est probablement dû à l'heuristique d'optimisation orientée période (OOP) incorporée dans la méthode M6 et qui s'avère utile lorsque il y a beaucoup de patients rejetés du planning courant , c'est-à-dire affectés au jour « virtuel » H+1. À partir de ces

résultats, on peut aussi noter que les gaps de dualité sont généralement un peu plus élevés en comparaison avec les problèmes de τ = 75%.

Nombre de salles	Nombre de patients	Méthode #	Salles identiques GAP (%)	Salles non-identiques GAP (%)
3	71.40	M2	3.12	3.22
		M3	2.85	1.54
		M4	2.67	1.28
		M5	2.63	1.07
		M6	2.04	0.76
		M7	2.02	0.78
6	142.30	M2	2.77	5.52
		M3	2.28	2.46
		M4	2.00	2.30
		M5	1.95	2.15
		M6	1.31	1.53
		M7	1.35	1.57
9	211.80	M2	2.63	4.86
		M3	2.16	3.36
		M4	1.71	2.56
		M5	1.71	2.52
		M6	1.18	1.93
		M7	1.20	1.89
12	283.10	M2	2.45	4.37
		M3	2.42	3.38
		M4	1.82	2.39
		M5	1.77	2.35
		M6	1.35	1.82
		M7	1.35	1.84

TAB. 4.4 - Le Gap pour des problèmes avec τ = 100%

4. 4. 3 Comparaison des différentes stratégies de génération de colonnes

Dans tous les tests présentés ci-dessus, plus que 65% du temps de calcul (CPU Linéaire) nécessaire pour la résolution du problème maître linéaire PML est consommé pour la résolution du problème générateur. Dans ce qui suit nous comparons l'impact des trois stratégies de générations de colonnes (*best-negative*, *all-negative*, et *first-negative*) sur la résolution du PML. Les méthodes de construction des plannings réalisables ne sont pas considérées ; on s'intéresse seulement à la résolution du PML.

Pour la comparaison des trois stratégies de générations de colonnes, nous utilisons des instances des problèmes avec τ = 75%. Les résultas sont présentés dans le tableau 4.5 ; ils sont basés sur 10 instances pour chaque taille et classe de problème en considération. Les

instances utilisées sont les mêmes que celles de la section précédente. Le problème maître restreint démarre avec un ensemble Ω' de colonnes générées aléatoirement, et de sorte qu'elles forment une solution faisable pour le problème général PG.

En tableau 4.5, nous présentons pour chaque stratégie le nombre de colonnes générées (Nombre de colonnes), le temps de calcul consommé pour la résolution du problème générateur (CPU Générateur), le temps de calcul pour le problème maître restreint PMR (CPU PMR), et le temps de calcul nécessaire pour la résolution de problème maître Linéaire (CPU Linéaire), c'est-à-dire la somme des deux temps précédents. Les nombres moyens des patients sont présentés dans les tableaux 4.2 et 4.3.

		Salles identiques				Salles non-identiques			
Nombre de salles		3	6	9	12	3	6	9	12
Best-Negative	Nombre de colonnes	284.70	562.78	957.30	1288.80	379.90	752.00	1109.20	1468.80
	CPU Générateur	6.09	31.36	101.18	238.23	8.73	46.71	126.40	274.03
	CPU RMP	2.52	11.39	37.14	106.07	8.61	30.28	64.00	128.81
	CPU Linéaire	8.60	42.76	138.32	344.30	17.34	76.99	190.40	402.84
All-Negative	Nombre de colonnes	1034.10	4039.20	9419.40	16694.40	747.60	2637.00	5863.20	9950.00
	CPU Générateur	1.52	7.58	22.62	45.27	1.18	4.92	13.83	27.62
	CPU RMP	1.40	7.94	26.24	49.71	1.34	6.25	17.58	37.17
	CPU Linéaire	2.92	15.52	48.86	94.98	2.52	11.17	31.42	64.79
First-Negative	Nombre de colonnes	400.20	855.00	1269.20	1655.40	684.60	1240.40	1641.80	2210.40
	CPU Générateur	9.08	51.14	144.53	302.08	15.35	72.77	183.55	396.77
	CPU RMP	2.79	15.77	46.29	102.71	4.09	21.07	57.32	130.44
	CPU Linéaire	11.87	66.92	190.82	404.79	19.43	93.84	240.88	527.21

TAB. 4.5 - Performances des différentes stratégies de génération de colonnes (temps en secondes)

Les résultats numériques montrent que la stratégie « all-negative » est mieux que les deux autres, et apporte des gains considérables en terme de temps de calcul.

En utilisant la stratégie « all-negative » ou « first- negative » le nombre de colonnes générées est plus élevé que dans le cas de la stratégie « best-negative ». En comparaison avec la

stratégie « best-negative », la stratégie « all-negative » permet une réduction de temps de calcul « CPU Linéaire », mais la stratégie « first- negative » engendre une augmentation de ce temps de calcul.

La résolution du problème générateur nécessite la résolution de $H \times S$ sous-problèmes de génération de colonnes (CG_{ts}) qui sont assez gourmands en temps de calcul. Avec la stratégie « best-negative » une seule colonne est ajoutée au problème maître restreint PMR, à chaque itération.

En utilisant la stratégie « all-negative », plusieurs colonnes sont ajoutées au PMR à chaque itération ; toutes les colonnes ayant un coût réduit négatif. Donc, cette stratégie n'affecte pas le temps de calcul nécessaire à la résolution d'un problème générateur, mais augmente le temps de résolution du PMR (par itération) parce que sa taille est plus grande. Cependant, avec cette stratégie le nombre total d'itération diminue, et par conséquent, le temps de calcul total « CPU Linéaire » est plus petit.

L'utilisation de la stratégie « first-negative » permet de diminuer le temps de résolution du problème générateur ; parce qu'on ne résout plus tous les $H \times S$ sous-problèmes de génération de colonnes. Cependant, le nombre d'itérations du PMR augmente, ce qui dégrade le temps de calcul total « CPU Linéaire ».

Remarquons aussi qu'en utilisant la stratégie « first-negative » ou « best-negative », les problèmes avec salles non-identiques présentent un temps de calcul « CPU linéaire » et un nombre de colonnes générées plus élevé que pour les problèmes avec salles identiques. Mais l'opposé de ce constat est vrai pour la stratégie « all-negative ».

Nous allons maintenant évaluer les performances des méthodes M1 et M6 en utilisant la stratégie « all-negative » lors de la résolution du PML. Nous considérons des problèmes avec $\boldsymbol{\tau}$= 75%. Les résultats numériques sont présentés dans le tableau 4.6.

En comparaison avec les résultats précédents (Tableaux 4.2 et 4.3), on peut observer que le temps de calcul de la méthode M6 a considérablement diminué grâce à l'utilisation de la stratégie « all-negative ». On voit aussi que la résolution des problèmes avec salles non-identiques est plus rapide que celle des problèmes avec salles identiques ; un comportement opposé à celui observé dans les tableaux 4.2 et 4.3. Les GAPs sont légèrement différents de ceux précédemment présentés, où le PML a été résolu en utilisant la stratégie « all-negative ». Cette différence s'explique par la non unicité de la solution optimale du PML, ce qui conduit à différentes solutions faisables et par conséquent à différents GAPs.

En ce qui concerne la méthode M1, on remarque que les GAPs sont plus faibles que ceux précédemment obtenus. Toutefois, ils demeurent plus élevés que ceux obtenus avec la

méthode M6. Ce résultat s'explique par le fait qu'un plus grand nombre de colonnes est considéré lors de la résolution du programme en nombres entiers pour la construction de la solution faisable. Cependant, le temps de calcul « CPU Time » de la méthode M1 a considérablement augmenté pour les problèmes avec 3 salles identiques. Pour des problèmes avec plus que 3 salles opératoires, CPLEX IP ne permet pas de trouver des solutions faisables à cause du grand nombre de colonnes générées.

Donc, en conclusion la méthode M6 fournit des solutions proches de l'optimum, à 2% de l'optimum, en un court temps de calcul même pour les problèmes de grandes tailles. Nous n'avons pas présenté les résultats de toutes les méthodes proposées, mais nous signalons toutefois, que les méthodes M4, M5 et M7 ont des comportements similaires à la méthode M6.

Méthode #	Nombre de salles	Salles identiques		Salles non-identiques	
		GAP (%)	CPU Time (sec)	GAP (%)	CPU Time (sec)
M1	3	4.95	1141.46	0.43	2.85
	6, 9 and 12	-	-	-	-
M6	3	1.78	3.05	0.77	2.6
	6	1.01	16.34	1.45	11.87
	9	0.81	51.36	1.53	33.58
	12	1.19	99.83	1.57	70.00

TAB. 4.6 - Performances des méthodes M1 et M6 avec la stratégie « all-negative »

Dans le reste de cette section, nous étudions la sensibilité des mesures de performances (GAP et temps de calcul) de la méthode M6 relativement aux coûts d'affectation des patients. Pour ce faire, nous considérons des problèmes avec salles non-identiques, $\boldsymbol{\tau}$ = 75% et différentes valeurs de R (R=200, 300 et 400). Les résultats numériques sont présentés dans le tableau 4.7 ; ils sont basés sur 10 instances aléatoires pour chaque valeur de R et chaque taille de problème en considération. Pour une taille de problème donnée, la différence entre des instances correspondant à des R différents réside seulement au niveau des coûts d'affectation des patients. Les résultats présentés dans le tableau 4.7 comprennent les GAP et les temps de calcul (Temps CPU).

À partir du tableau 4.7, on peut facilement observer que les performances de la méthode M6 ne sont pas sensibles aux changements affectant les coûts relatifs aux patients. La méthode fournit toujours des solutions de bonne qualité en un temps de calcul assez court.

Nombre de salles	*R* = 200		*R* = 300		*R* = 400	
	GAP (%)	CPU Time (sec)	GAP (%)	CPU Time (sec)	GAP (%)	CPU Time (sec)
3	0.63	2.73	0.26	3.46	0.34	2.36
6	1.70	10.75	1.40	10.41	1.44	10.42
9	1.66	29.76	1.31	28.33	1.19	28.50
12	1.30	62.41	1.60	61.10	1.24	55.36

TAB. 4.7 - Performances de la méthode M6 pour différentes valeurs de *R*

4. 4. 4 Caractéristiques du planning « optimal »

Dans cette section, nous illustrons les caractéristiques du planning (presque optimal) obtenu par l'approche de génération de colonnes. Pour ce faire, nous considérons le planning obtenu par la méthode M6 pour deux instances avec 12 salles opératoires et $\boldsymbol{\tau}$ = 75%, une avec salles identiques et l'autre avec salles non-identiques (R = 400). Pour ces deux exemples, il n'y a pas de patients rejetés ; ils sont tous planifiés sur l'horizon de planification.

Dans les Figures 4.1 et 4.2, nous présentons les informations suivantes :

- La demande agrégée du bloc opératoire par jour D_t ; elle est la somme des durées opératoires de tous les patients ayant une date au plus tôt égale à *t*, divisée par S le nombre de salles, c.-à-d., $D_t = \sum_{i:e_i=t} d_i \big/ S$.

- La charge minimale et maximale planifiée (pour la chirurgie programmable) dans les salles, pour chaque jour.

Rappelons que la demande moyenne pour la chirurgie d'urgence est de 2 heures pour chaque salle-jour, et la capacité en heures régulières est égale à 8 heures pour chaque salle-jour.

On peut observer qu'un usage d'heures supplémentaires est prévu pour les jours 1 et 2, à cause d'une grande demande pour la chirurgie programmable en premier jour (jour 1). On remarque aussi que la charge planifiée dans les salles décroît en fonction du temps. Il est à noter aussi que, pour le problème avec salles identiques, la différence entre la charge maximale et minimale planifiée dans les salles opératoires est relativement faible, en un chaque jour. Autrement, la charge est bien équilibrée entre les salles. Cependant, pour le problème avec salles non-identiques, la charge n'est pas aussi bien équilibrée. Signalons aussi que le coût du planning correspondant aux salles non-identiques est 1.5% plus élevé que celui correspondant aux salles identiques.

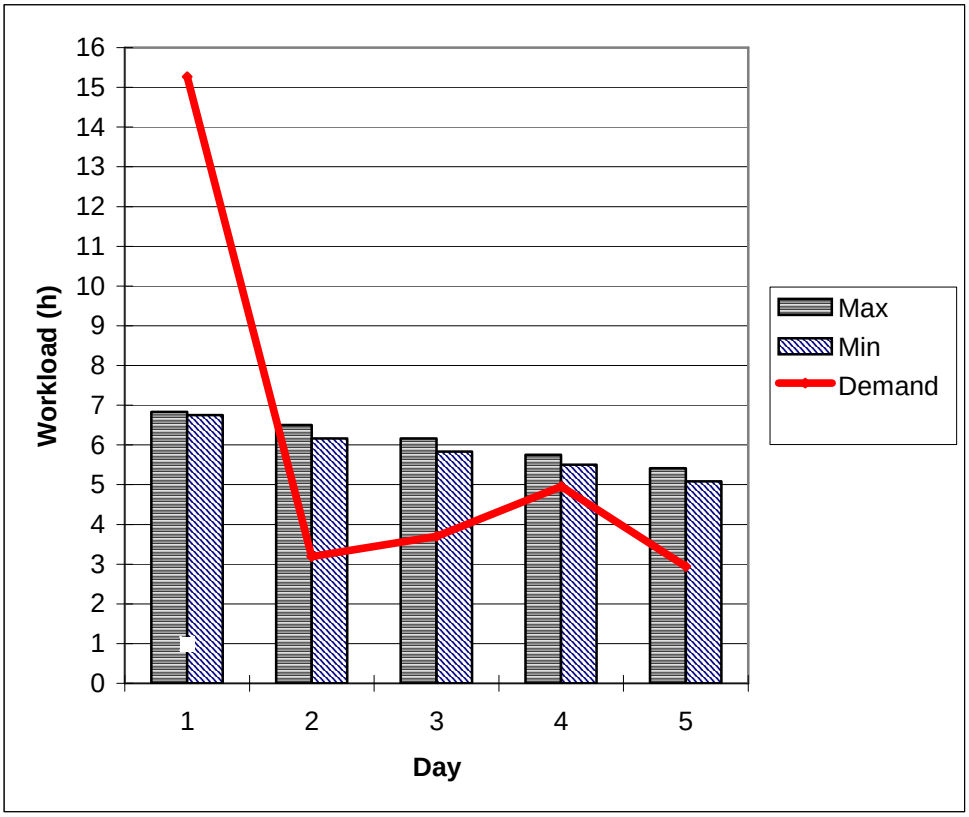

FIG. 4.1 - Répartition des charges pour un problème avec 12 salles identiques

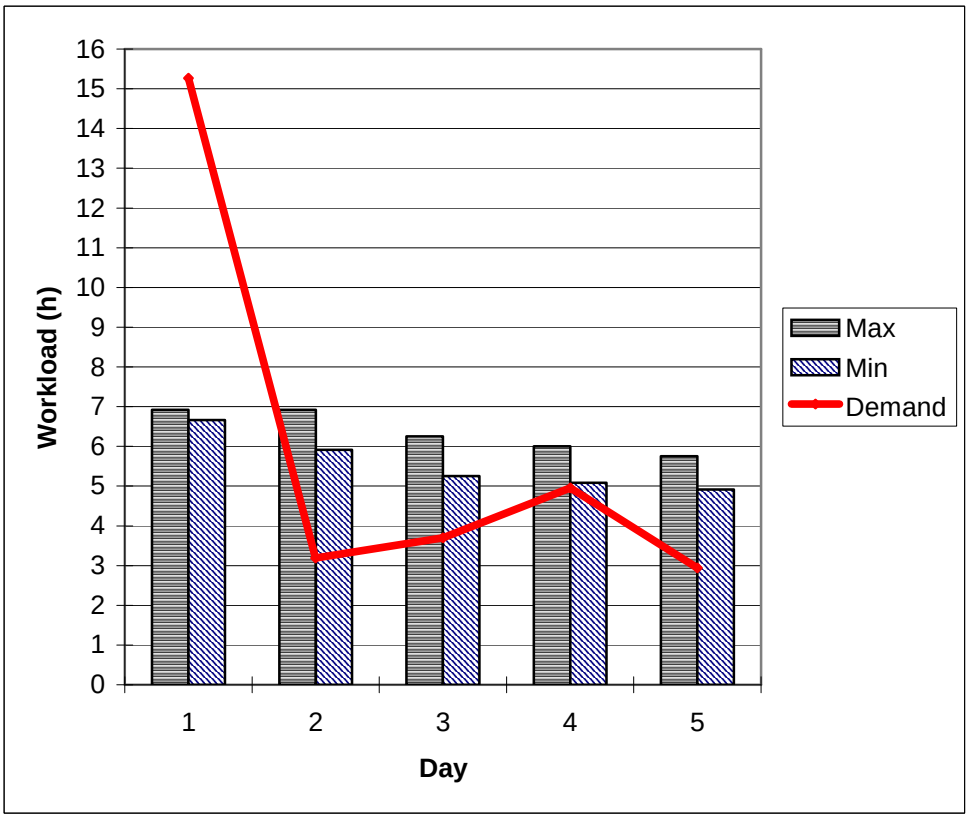

FIG. 4.2 - Répartition des charges pour un problème avec 12 salles non-identiques (R = 400)

4. 5 Conclusion

Dans ce chapitre, nous avons proposé un modèle stochastique pour la planification de la chirurgie élective dans un bloc opératoire, et une méthode de résolution basée sur la génération de colonnes.

La méthode de génération de colonnes permet d'obtenir rapidement des bonnes solutions ainsi qu'une borne inférieure qui permet d'évaluer leur qualité. Selon les tests numériques, la méthode proposée fournit des solutions proches de l'optimum, avec un gap de dualité inférieure à 2%, pour des problèmes de tailles réalistes (12 salles opératoires et à peu près 210 patients) en un temps de calcul assez court ; moins que deux minutes.

Notre expérimentation des différentes stratégies de génération de colonnes fait clairement apparaître une stratégie gagnante : la stratégie « all-negative ». Elle permet une amélioration significative en temps de calcul.

Notons aussi que la méthode de génération de colonnes présente un avantage important : elle permet de traiter des contraintes internes à un planning élémentaire au sein du sous-problème de génération de colonnes et non au sein du problème global ; ce qui facilite l'intégration de nombreuses contraintes additionnelles.

Jusqu'à maintenant, les durées d'interventions pour les patients électifs sont considérées comme déterministes. Cependant, en réalité ces durées sont incertaines et peuvent avoir un grand impact sur la qualité du planning. Dans le chapitre suivant, nous considérons le problème de planification avec durées d'interventions aléatoires et en présence de la chirurgie d'urgence.

Chapitre 5

Planification avec durées d'interventions aléatoires

Dans ce chapitre, nous considérons le problème de planification du bloc opératoire avec prise en compte des aléas dus à la chirurgie d'urgence ainsi qu'aux durées des interventions électives. Les durées des interventions et les capacités utilisées par la chirurgie d'urgence sont considérées comme des variables aléatoires. Le problème de planification consiste à déterminer l'ensemble des patients électifs qui seront opérés dans chaque salle opératoire et pour chaque jour sur un horizon de planification donné. L'objectif est de minimiser la somme des coûts relatifs aux patients électifs et des coûts de sur-utilisation des salles opératoires. Nous formulons le problème sous la forme d'un programme stochastique et nous proposons une approche de résolution qui combine la simulation Monte Carlo et la génération de colonnes. Dans un premier temps, nous utilisons la simulation Monte Carlo pour approximer le problème stochastique par un problème d'optimisation déterministe. Ensuite, nous utilisons une approche de génération de colonnes pour la résolution de ce dernier. Afin d'améliorer les performances de cette approche, nous proposons diverses stratégies de génération de colonnes qui permettent d'exploiter la structure du problème. Nous comparons les performances de ces différentes stratégies et nous évaluons les performances globales de l'approche.

(Lamiri *et al*., 2007a)

La variabilité des durées d'interventions engendre très souvent des modifications dans le planning des activités du bloc opératoire, et par conséquent un bouleversement de tout le

fonctionnement de ce dernier. Par exemple, des interventions plus longues que prévu peuvent conduire à un fort usage des heures supplémentaires et/ou le report des interventions déjà programmées. Ce qui ce traduit par une mauvaise qualité de service vis-à-vis des patients et des coûts supplémentaires pour l'hôpital. Dans ce contexte, la planification du bloc opératoire doit intégrer l'aspect aléatoire des durées d'interventions, afin de proposer des plannings robustes à ce genre d'aléas.

Dans ce chapitre, nous étendons le modèle de planification pour tenir compte des incertitudes liées aux durées des interventions électives. Ces dernières sont maintenant considérées comme des variables aléatoires. A cause de cette nouvelle extension, l'approche de génération de colonnes présentée au chapitre précédent n'est plus applicable. Une nouvelle approche combinant la simulation Monte Carlo et la génération de colonnes est présentée.

5. 1 Modèle et formulation mathématique

5. 1. 1 Modèle proposé

Nous nous intéressons toujours à la planification des interventions électives dans un bloc opératoire sur un horizon de planification donné. Les différentes notations utilisées sont similaires à celles introduites au chapitre précédent. Un rappel de ces dernières est présenté ci-dessous.

H : Horizon de planification,

t : (1… H) Indice de période (jour),

S : Nombre de salles opératoires,

s : (1…S) Indice de salle opératoire,

N : Nombre de patients programmables,

i : (1… N) Indice de patient programmable,

d_i : Durée d'intervention du patient i, variable aléatoire

e_i : Date au plus tôt pour opérer le patient i,

a_{its} : Coût d'affectation du patient i à la salle-jour (s, t),

W_{ts} : Variable aléatoire représentant la durée totale des chirurgies d'urgence réalisées en salle-jour (s, t),

$g_{ts}(.)$: Coût moyen d'utilisation de la salle opératoire s en jour t (coût de sur-utilisation).

Les principales différences par rapport au modèle de planification introduit au chapitre précédent concernent : les durées opératoires des interventions et les coûts d'utilisation des salles opératoires.

Durées opératoires

A chaque patient électif *i* est associée une durée d'intervention (ou *durée opératoire*) d_i. Cette durée regroupe la durée de chirurgie, le temps de préparation et de set-up, et la durée de nettoyage et de remise en état de la salle opératoire.

Dans ce chapitre, nous supposons que les durées opératoires d_i sont des variables aléatoires avec des lois de distributions $f_{d_i}(.)$. Ces lois de distribution peuvent être estimées en exploitant l'historique des données.

Dans la littérature, plusieurs travaux essayent de déterminer les lois de distribution des durées opératoires. Certains auteurs suggèrent la loi log-normale (Zhou et Dexter 1998; Strum *et al*., 1998, 2000), alors que d'autres optent pour la loi de Pearson III (Combes *et al*., 2007). Toutefois, dans notre modèle, nous ne faisons pas d'hypothèse concernant ces lois de distribution.

Coûts d'utilisation des salles opératoires

Dans ce chapitre, le coût d'utilisation d'une salle opératoire est seulement le coût des heures supplémentaires. Chaque salle-jour (*s*, *t*) dispose d'une capacité régulière T_{ts} représentant la durée d'ouverture en heures normales de la salle opératoire *s* en jour *t*. Si cette capacité n'est pas suffisante on peut avoir recours à un usage des heures supplémentaires. Si la durée totale des interventions programmées plus les chirurgies d'urgence dépasse la capacité régulière, alors des heures supplémentaires sont nécessaires. Nous désignons par c_{ts} le coût par unité de temps des heures supplémentaires en salle-jour (*s*, *t*).

Soit $y = [y_1, \ldots, y_N]$ un vecteur binaire à *N* composantes, spécifiant les patients planifiés dans la salle-jour (*s*, *t*) ; $y_i = 1$ si le patient *i* est planifié dans la salle-jour (*s*, *t*), 0 sinon. Le coût moyen d'utilisation de la salle *s* en jour *t* est alors :

$$c_{ts} \times E\left[\left(W_{ts} + \sum_{i=1}^{N} y_i d_i - T_{ts}\right)^+\right]$$

L'espérance mathématique *E*[.] est par rapport à la distribution de la variable aléatoire $W_{ts} + \sum_{i=1}^{N} y_i d_i$.

Remarquons que lorsqu'il n'y a pas de patients électifs planifiés en salle-jour (*s*, *t*), i.e. $y = 0$, le coût d'utilisation est $c_{ts} \times E\left[\left(W_{ts} - T_{ts}\right)^+\right]$. Ce coût d'utilisation peut être non nul. En effet, même en absence des interventions électives planifiées des dépassements horaires peuvent avoir lieu dus à la chirurgie d'urgence.

5. 1. 2 Formulation mathématique

À nouveau, le problème de planification consiste à affecter les patients électifs aux différentes salles-jours de telle sorte que la somme des coûts relatifs aux patients et des coûts des dépassements horaires des salles opératoires soit minimale.

Définissons les variables de décisions $x_{its} \in \{0,1\}$, avec x_{its} égale 1 si le patient électif i est affecté à la salle-jour (s, t) et 0 sinon. Par convention, $x_{i,H+1,s} = 1$ signifie que le patient i est rejeté du planning courant.

Le problème de planification est formulé par le programme mathématique suivant, appelé problème général (PG) :

$$J^* = \text{Minimiser } J(X) = \sum_{i=1}^{N} \sum_{t=e_i}^{H+1} \sum_{s=1}^{S} a_{its} x_{its} + \sum_{t=1}^{H} \sum_{s=1}^{S} c_{ts} O_{ts} \tag{5.1}$$

sous contraintes :

$$O_{ts} = E\left[\left(W_{ts} + \sum_{i=1}^{N} x_{its} d_i - T_{ts} \right)^+ \right], \ \forall\, t \in \{1,\ldots, H\}, s \in \{1,..., S\} \tag{5.2}$$

$$\sum_{t=e_i}^{H+1} \sum_{s=1}^{S} x_{its} = 1, \ \ \forall\, i \in \{1,\ldots, N\} \tag{5.3}$$

$$x_{its} \in \{0,1\}, \forall\, i \in \{0, 1\}, t \in \{1,\ldots, H+1\}, s \in \{1,\ldots, S\}$$

L'espérance mathématique $E[.]$ est par rapport à la distribution de la variable aléatoire, à savoir $W_{ts} + \sum_{i=1}^{N} d_i x_{its}$, et $(y)^+ = \max\{0, y\}$.

La fonction objectif (5.1) est la somme des coûts relatifs aux patients électifs et les coûts des dépassements horaires moyens des salles opératoires. Les contraintes (5.2) déterminent les dépassements horaires moyens. Les contraintes (5.3) assurent que chaque patient électif est planifié exactement une fois. Le problème de planification (PG) est un problème combinatoire stochastique.

5. 2 Approche de résolution : Monte Carlo et génération de colonnes

Dans cette section, nous proposons une approche de résolution qui combine la simulation Monte Carlo et la génération de colonnes pour le problème général (PG). Les principales étapes de cette approche sont les suivantes. Dans un premier temps, nous utilisons la simulation Monte Carlo pour approximer le problème stochastique (PG) par un problème d'optimisation déterministe. Ce dernier, sera ensuite résolu par une approche de génération de

colonnes. Pour ce faire, (i) nous reformulons le problème déterministe sous la forme d'un problème mathématique comportant un nombre très élevé de variables (problème maître), (ii) nous résolvons la relaxation linéaire du problème maître en utilisant la génération de colonnes, (iii) nous construisons un planning réalisable à partir de la solution optimale du problème relâché, (iv) et enfin nous essayons d'améliorer le planning réalisable par optimisation locale.

Le schéma général de l'approche de génération de colonnes est similaire à celui utilisé au chapitre précédent. Cependant, nous signalons des différences majeures qui concernent :

- Les sous-problèmes de générations de colonnes : ces sous-problèmes représentent une nouvelle extension du problème de sac-à-dos multidimensionnel classique. Nous étudions la structure des solutions optimales de ces sous-problèmes et nous identifions des propriétés qui permettent de fixer certaines variables et ainsi réduire le nombre des variables de décisions. Ce qui permet d'accélérer leur résolution.

- Les stratégies de génération de colonnes : nous utilisons des stratégies de génération de colonnes « non classiques » qui exploitent la structure du problème et qui permettent d'améliorer les performances de l'approche.

- La construction du planning réalisable : nous utilisons une heuristique d'arrondissement ; heuristique différente de celles proposées dans le chapitre précédent.

5. 2. 1 Approximation Monte Carlo

L'idée de base de cette méthode est d'utiliser la simulation Monte Carlo pour approximer le problème de planification stochastique (PG) par un problème déterministe. Plus précisément, on génère K échantillons aléatoires indépendants et identiquement distribués (i.i.d) pour chaque variable aléatoire. Soit

- pour chaque salle-jour (s, t), K échantillons aléatoires i.i.d W_{ts}^1, …, W_{ts}^K pour la variable aléatoire W_{ts},
- Et, pour chaque patient i, K échantillons aléatoires i.i.d d_i^1, …, d_i^K pour la variable aléatoire d_i.

Un vecteur $[W_{1,1}^k, \ldots, W_{HS}^k, d_1^k, \ldots, d_N^k]$ représente un scénario, $k \in \{1,\ldots, K\}$. C'est aussi un échantillon aléatoire (i.i.d) du vecteur aléatoire $[W, d] = [W_{1,1}, \ldots, W_{HS}, d_1, \ldots, d_N]$.

En utilisant les K scénarios générés, les espérances mathématiques (5.2) évaluant les dépassements horaires peuvent être approximées par des moyennes empiriques. Avec cette

estimation le problème de planification (PG) peut être approximé par le problème déterministe suivant (PG_K):

$$\text{Minimiser } J_K(X) = \sum_{i=1}^{N}\sum_{t=e_i}^{H+1}\sum_{s=1}^{S} a_{its}x_{its} + \sum_{t=1}^{H}\sum_{s=1}^{S} c_{ts}O_{tsK} \tag{5.4}$$

sous contraintes :

$$O_{tsK} = \frac{1}{K}\sum_{k=1}^{K}\left(W_{ts}^{k} + \sum_{i=1}^{N} d_i^k x_{its} - T_{ts} \right)^{+}, \ \forall\, t\in\{1,\ldots, H\}, s\in\{1,..., S\} \tag{5.5}$$

$$\sum_{t=e_i}^{H+1}\sum_{s=1}^{S} x_{its} = 1, \ \forall\, i\in\{1,\ldots, N\} \tag{5.6}$$

$$x_{its}\in\{0,1\}, \forall\, i\in\{0, 1\}, t\in\{1,\ldots, H+1\}, s\in\{1,\ldots, S\}$$

Le critère $J_K(X)$ à minimiser est maintenant un *estimateur* du critère exact $J(X)$. Les contraintes (5.5) fournissent une estimation des dépassements horaires en chaque salle-jour (s, t), en se basant sur les scénarios générés.

5. 2. 2 Approche de génération de colonnes

Dans cette section, nous proposons une formulation en colonnes pour le problème de planification. Similairement au chapitre précédent, nous utilisons le concept de *planning élémentaire*. Un planning élémentaire p est un planning relatif à une salle opératoire donnée en un jour donné ; et il est défini par les variables binaires suivantes :

$$y_{ip} = \begin{cases} 1 & \text{si le patient } i \text{ est affecté au planning élementaire } p, \\ 0 & \text{sinon.} \end{cases}$$

$$z_{tsp} = \begin{cases} 1 & \text{si le planning élémentaire } p \text{ est affecté à la salle - jour } (s,t), \\ 0 & \text{sinon.} \end{cases}$$

Le planning élémentaire p peut être représenté par le vecteur binaire (colonne) $[y_p, z_p] = [(y_{1p},\ldots, y_{Np}), (z_{1p},\ldots,z_{(H\times S)p})]$.

Soit Ω l'ensemble de tous les plannings élémentaires faisables. Un planning élémentaire p est faisable si les deux conditions suivantes sont vérifiées :

- (i) le planning est affecté à une et une seule salle-jour, c.-à-d., $\sum_{t,s} z_{tsp} = 1$,
- (ii) le planning respecte les dates au plus tôt des patients, c.-à-d., $y_{ip} \times z_{tsp} = 0$, $\forall\, t < e_i$,

L'ensemble des plannings élémentaires, défini ci-dessus, ne concerne que les salles-jours comprises dans l'horizon de planification. Aucun planning élémentaire n'est associé à la période « virtuelle » H+1.

Le coût d'un planning élémentaire p peut être exprimé par :

$$C_p = \sum_{t,s} \left[z_{tsp} \sum_i y_{ip} a_{its} + g_{ts}\left(y_p\right) \right], \tag{5.7}$$

où $g_{ts}\left(y_p\right)$ est le coût moyen du dépassement horaire de la salle s en jour t ; estimé en se basant sur les scénarios aléatoires générées à la phase de Monte Carlo. Il est défini par

$$g_{ts}\left(y_p\right) = \left(c_{ts}/K\right) \sum_{k=1}^{K} \left(W_{ts}^k + \sum_{i=1}^{N} d_i^k y_{ip} - T_{ts} \right)^+ \tag{5.8}$$

Le coût du planning est à nouveau constitué de deux parties : les coûts relatifs aux patients électifs et le coût moyen du dépassement horaire de la salle opératoire.

Le problème de planification consiste donc à sélectionner un sous-ensemble des plannings élémentaires. Soit λ_p une variable de décision binaire indiquant si le planning élémentaire p est sélectionné (λ_p = 1) ou non (λ_p = 0). Le problème de planification est maintenant formulé comme suit :

$$\text{Min} \sum_{p\in\Omega} C_p \lambda_p + \sum_i a_{iH+1}\left(1 - \sum_{p\in\Omega} y_{ip}\lambda_p\right) + \sum_{t,s} g_{ts}(0) \times \left(1 - \sum_{p\in\Omega} z_{tsp}\lambda_p\right) \tag{5.9}$$

sous contraintes :

$$\sum_{p\in\Omega} y_{ip}\lambda_p \le 1, \ \forall\, i = 1,\ldots,N \tag{5.10}$$

$$\sum_{p\in\Omega} z_{tsp}\lambda_p \le 1, \ \forall\, t=1,\ldots,H \ , \ \forall\, s=1,\ldots,S \tag{5.11}$$

$$\lambda_p \in \{0,1\}, \ \forall p \in \Omega$$

La fonction objectif (5.9) représente le coût d'un planning global ; elle est constituée de trois parties :

- le coût des plannings élémentaires sélectionnés : $\sum_{p\in\Omega} C_p\lambda_p$
- les coûts engendrés par les patients non planifiés: $\sum_i a_{iH+1}\left(1 - \sum_{p\in\Omega} y_{ip}\lambda_p\right)$
- les coûts d'utilisation des salles opératoires n'ayant pas reçues de plannings : $\sum_{t,s} g_{ts}(0) \times \left(1 - \sum_{p\in\Omega} z_{tsp}\lambda_p\right)$

Les contraintes (5.10) assurent que chaque patient est inclus au plus dans un planning sélectionné, c'est-à-dire qu'il est affecté au plus à une salle-jour sur l'horizon de planification. Les contraintes (5.11) garantissent qu'il y a au plus un planning élémentaire sélectionné pour chaque salle-jour.

Soit $\tilde{C}_p = C_p - \sum_i a_{iH+1}\, y_{ip} - \sum_{t,s} g_{ts}(0)\, z_{tsp}$, le coût *modifié* du planning élémentaire *p*. Le problème de planification peut être maintenant re-exprimé sous la forme du problème mathématique suivant (problème maître, PM) :

$$\text{(PM):}\quad \text{Min} \sum_i a_{iH+1} + \sum_{t,s} g_{ts}(0) + \sum_{p\in\Omega} \tilde{C}_p \lambda_p \tag{5.12}$$

sous contraintes :

$$\sum_{p\in\Omega} y_{ip}\lambda_p \le 1,\ \forall\, i \tag{5.13}$$

$$\sum_{p\in\Omega} z_{tsp}\lambda_p \le 1,\ \forall\, t, s \tag{5.14}$$

$$\lambda_p \in \{0,1\},\ \forall p \in \Omega$$

Similairement au chapitre précédent, nous allons résoudre, par génération de colonnes, la relaxation linéaire du problème maître, le problème maître linéaire (PML). Pour ce faire, nous commençons par résoudre un problème maître restreint (PMR) en considérant seulement un sous-ensemble $\Omega' \subset \Omega$ de colonnes. Ensuite, des colonnes supplémentaires sont ajoutées seulement si le problème de *génération de colonnes* arrive à identifier des colonnes avec un coût réduit négatif.

5. 2. 3 Problème de génération de colonnes

Soient π_i et π_{ts} les solutions duales optimales d'un PMR, associées respectivement aux contraintes (5.13) et (5.14). Le problème de *génération de colonnes* (ou *problème générateur*)

$$\sigma_p = \min_{\{y_{ip}, z_{tsp}\}} \tilde{C}_p - \sum_i \pi_i y_{ip} - \sum_{t,s} \pi_{ts} z_{tsp} \quad \text{tel que } p\in\Omega$$

consiste à identifier la colonne (le planning élémentaire) qui a le coût réduit minimal parmi les colonnes faisables. Il est formulé comme suit :

$$\text{Minimiser} \quad \tilde{C}_p - \sum_{i=1}^{N} \pi_i y_{ip} - \sum_{t=1}^{H} \sum_{s=1}^{S} \pi_{ts} z_{tsp} \tag{5.15}$$

sous contraintes :

$$y_{ip} \times z_{tsp} = 0, \ \forall\ t < e_i, \tag{5.16}$$

$$\sum_{t,s} z_{tsp} = 1 \tag{5.17}$$

$$y_{ip} \in \{0, 1\}, \ z_{tsp} \in \{0, 1\}, \ \forall\ i, t, s$$

avec $\tilde{C}_p = \sum_{t,s} \left[z_{tsp} \sum_i y_{ip} a_{its} + g_{ts}\left(y_p\right) \right] - \sum_i a_{iH+1}\, y_{ip} - \sum_{t,s} g_{ts}(0)\, z_{tsp}$.

Étant donné les solutions duales optimales π_i et π_{ts}, du problème maître restreint actuel, la fonction objectif (5.15) détermine le coût réduit d'une colonne. Les contraintes (5.16) garantissent qu'aucun patient n'est planifié avant sa date au plus tôt. La contrainte (5.17) assure que la colonne est associée à une et une seule salle-jour.

Le problème de génération de colonnes (5.15)-(5.17) peut être décomposé en $H \times S$ sous-problèmes, un par salle-jour. Ainsi, il peut être résolu en résolvant le sous-problème de génération de colonnes (GC_{ts}) pour chaque salle-jour :

$$(GC_{ts}) : \text{Minimiser} \ \sum_{i=1}^{N} \tilde{a}_{its} y_{ip} + g_{ts}\left(y_p\right) - g_{ts}\left(0\right) - \pi_{ts} \tag{5.18}$$

sous contraintes :

$$y_{ip} = 0, \ \forall i \text{ avec } t < e_i \tag{5.19}$$

$$y_{ip} \in \{0, 1\}, \ \forall i = 1, \ldots, N$$

où $\tilde{a}_{its} = a_{its} - a_{iH+1} - \pi_i$.

Donc, le problème générateur peut être résolu en résolvant $H \times S$ sous-problèmes de génération de colonnes (GC_{ts}), et ensuite choisir la solution (colonne) qui a le coût réduit le plus petit.

Définissons pour chaque jour t, l'ensemble $I_t = \{i \ / \ e_i \leq t\}$ de patients électifs qui peuvent être planifiés pour ce jour. Les patients n'appartenant pas à l'ensemble I_t ne peuvent pas être opérés en jour t ; par conséquent, $y_{ip} = 0, \ \forall i \notin I_t$. Le sous-problème (GC_{ts}) peut être formulé comme suit :

$$\sigma_{ts}^{*} = \text{Min} \quad \sigma_{ts}\left(y_p\right) = -\pi_{ts} - g_{ts}\left(0\right) + \sum_{i \in I_t} \tilde{a}_{its} y_{ip} + \left(c_{ts}/K\right) \sum_{k=1}^{K} \left(W_{ts}^{k} + \sum_{i \in I_t} d_i^k y_{ip} \ - \ T_{ts} \right)^{+} \tag{5.20}$$

sous contraintes : $\quad y_{ip} \in \{0, 1\}, \ \forall\ i \in I_t$

Le sous-problème (GC_{ts}) peut être transformé en un programme linéaire à variable mixte :

$$\text{Minimiser } -\pi_{ts} - g_{ts}(0) + \sum_{i \in I_t} \tilde{a}_{its} y_{ip} + (c_{ts}/K) \sum_{k=1}^{K} O_k$$

sous contraintes :

$$O_k \geq W_{ts}^k + \sum_{i=1}^{N} d_i^k y_{ip} - T_{ts}, \ \forall k \in \{1, \ldots, K\}$$

$$O_k \geq 0, \ \forall k \in \{1, \ldots, K\}, \ y_{ip} \in \{0, 1\}, \ \forall i \in I_t$$

(GC_{ts}) est un problème de programmation linéaire en nombres entiers et peut donc être résolu en utilisant une des méthodes classiques tel que le *branch-and-bound* ou *branch-and-cut*.

Le sous-problème (GC_{ts}) est une nouvelle extension du problème de sac-à-dos multidimensionnel classique (Fréville, 2004); on autorise la violation des contraintes de capacité (dépassement de capacité) et on pénalise ces violations. Les O_k représentent les violations et le terme $(c_{ts}/K)\sum_{k=1}^{K} O_k$ représente la pénalité associée à ces violations.

La proposition suivante identifie des propriétés de la solution optimale du sous-problème (GC_{ts}). Désignons par y_p^* une solution optimale du sous-problème (GC_{ts}).

Proposition 5.1 :

- Si $\tilde{a}_{its} \geq 0$, alors $y_{ip}^* = 0$
- Si $\tilde{a}_{its} + (c_{ts}/K)\sum_{k=1}^{K} d_i^k \leq 0$, alors $y_{ip}^* = 1$.

Preuve : Soient y_p une solution quelconque du sous-problème (GC_{ts}) avec $y_{ip} = 1$, et $\bar{y}_p$ une solution telle que $\bar{y}_{jp} = y_{jp}, \forall j \in I_t \setminus \{i\}$ et $\bar{y}_{ip} = 0$. La seule différence entre les deux solutions est au niveau de la $i^{\text{ème}}$ composante.

Nous montrons que :

- si $\tilde{a}_{its} \geq 0$, alors pour toutes solutions y_p et $\bar{y}_p$ on a $\sigma_{ts}(\bar{y}_p) \leq \sigma_{ts}(y_p)$,
- et si $\tilde{a}_{its} + (c_{ts}/K)\sum_{k=1}^{K} d_i^k \leq 0$, alors pour toutes solutions y_p et $\bar{y}_p$ on a $\sigma_{ts}(y_p) \leq \sigma_{ts}(\bar{y}_p)$.

À partir de l'expression (5.20), pour toutes solutions y_p et $\bar{y}_p$, nous avons :

$$\sigma_{ts}(y_p) - \sigma_{ts}(\bar{y}_p) = \tilde{a}_{its} + (c_{ts}/K)\sum_k \left(W_{ts}^k + \sum_{j \in I_t \setminus \{i\}} d_j^k y_{jp} + d_i^k - T_{ts} \right)^+$$
$$-(c_{ts}/K)\sum_k \left(W_{ts}^k + \sum_{j \in I_t \setminus \{i\}} d_j^k y_{jp} - T_{ts} \right)^+$$

Il en découle alors, que $\tilde{a}_{its} \le \sigma_{ts}(y_p) - \sigma_{ts}(\bar{y}_p) \le \tilde{a}_{its} + (c_{ts}/K)\sum_k d_i^k$

Par conséquent, si $\tilde{a}_{its} \ge 0$, alors $\sigma_{ts}(\bar{y}_p) \le \sigma_{ts}(y_p)$. Ce qui signifie qu'il est toujours plus avantageux d'avoir la $i^{\text{ème}}$ composante égale à 0. Donc, la solution optimale vérifie $y_{ip}^* = 0$.

Similairement, si $\tilde{a}_{its} + (c_{ts}/K)\sum_{k=1}^{K} d_i^k \le 0$, alors $\sigma_{ts}(y_p) \le \sigma_{ts}(\bar{y}_p)$. Et par conséquent $y_{ip}^* = 1$. □

La proposition 5.1 permet de réduire le nombre de variables de décision en fixant quelques unes à 0 ou à 1.

Dans ce chapitre, deux méthodes sont utilisées pour la résolution du sous problème de génération de colonnes (CG_{ts}). Une méthode exacte, branch-and-cut, et une méthode heuristique. Dans ce qui suit, nous présentons l'heuristique utilisée.

Selon la valeur de π_{ts}, l'heuristique procède différemment. Rappelons que les π_{ts} sont inférieurs ou égaux à 0, parce qu'ils sont associés aux contraintes (5.14). Si π_{ts} = 0, l'heuristique construit une colonne avec un coût réduit négatif (s'il existe au moins une). Si π_{ts}<0, l'heuristique sélectionne une colonne qui est déjà dans la base du PMR, ensuite elle essaye de l'améliorer pour obtenir une colonne avec un coût réduit négatif. Avant de détailler les différentes étapes de l'heuristique, nous commençons par présenter les deux résultats suivants.

Proposition 5.2 : Si π_{ts} = 0, alors σ_{ts}^* < 0 si et seulement si $\exists\ i \in I_t$ tel que $\tilde{a}_{its} + (c_{ts}/K)\sum_{k=1}^{K}(W_{ts}^k + d_i^k - T_{ts})^+ - g_{ts}(0) < 0$.

Preuve :

Condition nécessaire : Supposons que σ_{ts}^* est strictement inférieure à 0. Soit y_p une solution optimale, c-à-d, $\sigma_{ts}(y_p) = \sigma_{ts}^* < 0$.

À partir de l'expression de $\sigma_{ts}(.)$, donnée par (5.20), nous avons

$$\begin{aligned}\sigma_{ts}(0) &= -\pi_{ts} - g_{ts}(0) + (c_{ts}/K)\sum_{k=1}^{K}(W_{ts}^k - T_{ts})^+ \\ &= -\pi_{ts} - g_{ts}(0) + g_{ts}(0) \quad \text{(Par définition de } g_{ts}(0)\text{, expression (5.8))} \\ &= 0 \qquad \text{(Parce que } \pi_{ts} = 0\text{)}\end{aligned}$$

Comme $\sigma_{ts}(y_p) < 0$, donc y_p est différent du vecteur nul. Par conséquent, il existe $i \in I_t$ tel que $y_{ip} = 1$.

En plus, il existe au moins $i \in I_t$ tel que $y_{ip} = 1$ et tel que la solution $\bar{y}_p$ ($\bar{y}_{jp} = y_{jp}, \forall j \in I_t \setminus \{i\}$ et $\bar{y}_{ip} = 0$) n'est pas optimale, c-à-d,

$$\sigma_{ts}(y_p) - \sigma_{ts}(\bar{y}_p) < 0 \tag{5.21}$$

En utilisant l'expression (5.20), on a

$$\begin{aligned}
\sigma_{ts}(y_p) - \sigma_{ts}(\bar{y}_p) &= \tilde{a}_{its} + (c_{ts}/K)\sum_k \left(W_{ts}^k + \sum_{j \in I_t \setminus \{i\}} d_j^k y_{jp} + d_i^k - T_{ts}\right)^+ \\
&\quad - (c_{ts}/K)\sum_k \left(W_{ts}^k + \sum_{j \in I_t \setminus \{i\}} d_j^k y_{jp} - T_{ts}\right)^+ \\
&\geq \tilde{a}_{its} + (c_{ts}/K)\sum_k \left(W_{ts}^k + d_i^k - T_{ts}\right)^+ \\
&\quad - (c_{ts}/K)\sum_k \left(W_{ts}^k - T_{ts}\right)^+ \\
&= \tilde{a}_{its} + (c_{ts}/K)\sum_k \left(W_{ts}^k + d_i^k - T_{ts}\right)^+ - g_{ts}(0)
\end{aligned} \tag{5.22}$$

À partir de (5.21) et (5.22), on obtient alors

$$\tilde{a}_{its} + (c_{ts}/K)\sum_k \left(W_{ts}^k + d_i^k - T_{ts}\right)^+ - g_{ts}(0) < 0$$

Condition suffisante : Soit $i \in I_t$ tel que $\tilde{a}_{its} + (c_{ts}/K)\sum_{k=1}^{K}(W_{ts}^k + d_i^k - T_{ts})^+ - g_{ts}(0) < 0$.

Désignons par y_p la solution vérifiant $y_{ip} = 1$, et $y_{jp} = 0$, $\forall j \neq i$. On a

$$\begin{aligned}
\sigma_{ts}(y_p) &= \tilde{a}_{its} + (c_{ts}/K)\sum_{k=1}^{K}(W_{ts}^k + d_i^k - T_{ts})^+ - g_{ts}(0) \quad \text{(Parce que } \pi_{ts} = 0\text{)} \\
&< 0
\end{aligned}$$

Par conséquent, $\sigma_{ts}^* \leq \sigma_{ts}(y_p) < 0$ (Par définition de σ_{ts}^*). □

Proposition 5.3 : Si $\pi_{ts}<0$, alors il existe un planning élémentaire $p \in \Omega$' dans le problème PMR tel que $z_{tsp} = 1$ et $\sigma_{ts}(y_p) = 0$.

Preuve : La variable duale π_{ts} est associée à la contrainte $\sum_{p \in \Omega'} z_{tsp}\lambda_p \leq 1$. Donc, selon le théorème de complémentarité en programmation linéaire (Minoux, 1983), on a $\pi_{ts} \times (1 - \sum_{p \in \Omega'} z_{tsp}\lambda_p) = 0$.

Comme $\pi_{ts}<0$, alors $\sum_{p \in \Omega'} z_{tsp}\lambda_p = 1$. Par conséquent, il existe au moins un planning $p \in \Omega$' avec $z_{tsp} = 1$ et $\lambda_p > 0$. Le planning p est associée à la salle-jour (s, t) (car $z_{tsp} = 1$) et

représente une variable de base ($\lambda_p > 0$) pour le problème PMR actuel, et par conséquent son coût réduit $\sigma_{ts}(y_p)$ est égal à 0. □

La proposition 5.3 signifie que, si $\pi_{ts} < 0$ alors il existe une colonne associée à la salle-jour (s, t) qui est déjà dans la base actuelle du PMR.

Les principales étapes de l'heuristique sont les suivantes (plus de détails sont fournis en Annexe 1) :

- Cas $\pi_{ts} = 0$: On commence par identifier le patient i^* qui a la quantité $\tilde{a}_{its} + (c_{ts}/K)\sum_{k=1}^{K}(W_{ts}^k + d_i^k - T_{ts})^+ - g_{ts}(0)$ la plus négative, et on fixe $y_{i^*p} = 1$ (si aucun n'existe, alors il n'y a pas de colonne améliorante, c-à-d $\sigma_{ts}^* > 0$, selon Proposition 5.2) et $y_{ip} = 0$, $\forall i \neq i^*$. Donc y_p représente une colonne améliorante, puisque $\sigma_{ts}(y_p) < 0$. Ensuite, on essaye d'améliorer itérativement la colonne. A chaque itération, on sélectionne un patient parmi les restants et on l'insère dans la colonne y_p. Le patient sélectionné est celui qui permet d'améliorer le plus la colonne y_p. Les itérations continuent jusqu'à ce que la solution "y_p" ne puisse plus être améliorée.

- Cas $\pi_{ts} < 0$: On sélectionne une colonne de la base courante $p \in \Omega'$ et qui est associée à la salle-jour (s, t), ensuite on essaye de l'améliorer de manière itérative en ajoutant ou en enlevant un patient. L'existence d'une telle colonne est garantie grâce à la proposition 5.3.

5. 2. 4 Différentes stratégies de génération de colonnes

Dans ce travail nous utilisons trois stratégies différentes pour la génération des nouvelles colonnes.

"***All-negative strategy***" : Cette stratégie consiste à résoudre jusqu'à l'optimalité tous les sous-problèmes (CG_{ts}), ensuite toutes les colonnes avec coût réduit négatif sont ajoutées au problème maître restreint (PMR).

"***Two-phase strategy***" : Cette stratégie fait appel à la méthode heuristique pour la résolution du sous-problème de génération de colonnes. Afin de garantir l'optimalité de la solution du problème maître linéaire (PML), une approche à deux phases est utilisée. Dans un premier temps, les sous-problèmes (CG_{ts}) sont résolus en utilisant la méthode heuristique. Si cette dernière n'arrive plus à trouver des colonnes améliorantes (avec un coût réduit négatif), on fait alors appel à la méthode exacte qui va identifier des nouvelles colonnes améliorantes ou qui va prouver l'optimalité de la solution courante du PML (si il n'y a plus de colonnes

avec coût réduit négatif). Ce processus est répété jusqu'à ce que le PML soit résolu à l'optimalité.

"***Cyclic strategy***" : Cette stratégie consiste à résoudre seulement un seul sous-problème (CG_{ts}) par itération de génération de colonnes, quand c'est possible. Les sous-problèmes sont considérés un par un et résolus à l'optimalité jusqu'à ce qu'on trouve une colonne $[y_p, z_p]$ avec un coût réduit négatif. Soit (s, t) la salle-jour associée à cette colonne, c-àd, $z_{tsp} = 1$ et $\sigma^*_{ts} = \sigma_{ts}(y_p)$.

(y_p) est ensuite testée comme solution candidate pour les autres sous-problèmes ($CG_{t's'}$), pour tout $(t', s') \neq (t, s)$. Si $\sigma_{t's'}(y_p) < 0$, alors la colonne $[y_p, z'_p]$ représente une colonne améliorante, et par conséquent elle est ajoutée au problème maître restreint (PMR). (z'_p) est un vecteur d'éléments nuls, sauf $z_{t's'p} = 1$.

Les salles-jours sont considérées de manière cyclique. Ce processus s'arrête si à une itération donnée tous les sous-problèmes sont résolus sans identifier aucune colonne améliorante.

5. 2. 5 Construction d'un planning réalisable

Comme les contraintes d'intégrités ont été relâchées, la solution optimale du problème maître linéaire (PML) peut être fractionnaire et par conséquent non faisable pour le problème général (PG), Dans cette section, nous présentons deux méthodes heuristiques pour la construction d'une solution entière. L'une utilise les variables x_{its} du problème général (PG), alors que l'autre utilise les variables λ_p du problème maître (PM).

Réaffectation progressive

Cette heuristique est la même que celle utilisée au chapitre précèdent. Cependant, nous rappelons ici son fonctionnement pour faciliter la lecture.

Désignons par x_{its} la solution du problème général correspondant à la solution optimale du problème maître linéaire PML, c'est-à-dire, $x_{its} = \sum_{p\in\Omega} y_{ip} z_{tsp} \lambda_p$. Dans un premier temps, on sélectionne un patient électif i dont la matrice d'affectations $[x_{its}]_{ts}$ contient des fractions, c'est-à-dire affecté de manière fractionnaire à plusieurs salles-jours. Ensuite, considérons les solutions $X^{(s', t')}$ obtenues à partir de $X = [x_{its}]_{its}$ en réaffectant le patient i seulement à la salle-jour (s', t'), et en gardant les affectations d'autres patients même si elles sont fractionnaires.

Le patient i est finalement affecté à la salle-jour (s_i, t_i) qui minimise $J(X^{(s', t')})$, c'est-à-dire $(s_i, t_i) = \arg\min_{\{(s', t')\}} J(X^{(s', t')})$. Bien évidemment, afin que la solution soit réalisable. La

solution X est ensuite remplacée par $X^{(s_i, t_i)}$, et le processus continue pour les autres patients ayant des affectations fractionnaires. L'ordre de traitement de ces patients est arbitraire.

Heuristique d'arrondissement

Dans un premier temps, le problème maître linéaire (PML) est résolu par génération de colonnes ; soit $\{\lambda_p\}$ la solution optimale. Ensuite, toutes les variables avec λ_p =1 (si il en existe) sont fixées à 1, et la variable avec la plus grande valeur fractionnaire est arrondie (fixée) à 1. Les variables (colonnes) fixées forment une solution partielle du problème maître (PM). Notons que si $\{\lambda_p\}$ satisfait les contraintes d'intégrité, alors elle représente une solution optimale du (PM) ; sinon elle fournit une borne inférieure.

Les patients et les salles-jours non inclus dans la solution partielle forment un problème *résiduel*. Ce dernier est à nouveau résolu par la génération de colonnes et des nouvelles variables sont fixées de la même manière que précédemment. Ce processus est répété jusqu'à ce que la solution du problème résiduel soit entière, et dans ce cas on a plus besoin d'arrondissement. À ce stade, la solution partielle représente un solution réalisable pour le problème maître. Cette heuristique d'arrondissement, peut être considérée comme une heuristique « *dive-and-fix* » basée sur l'arbre du *branch-and-price* (Wolsey, 1998).

5. 2. 6 Heuristiques d'amélioration

Afin d'améliorer la qualité de la solution faisable (un planning réalisable), nous utilisons deux heuristiques à recherche locale. Elles sont identiques à celles proposées au chapitre précédent.

Optimisation locale par réaffectation

La solution est itérativement améliorée en réaffectant les patients électifs. À chaque itération, on détermine pour chaque patient le meilleur gain qui peut être réalisé en le réaffectant à une autre salle-jour (ou à la période H+1) et en gardant les affectations des autres patients. Bien évidemment, la nouvelle affectation doit satisfaire la date au plus tôt du patient ainsi que les capacités totales des salles-jours. Ensuite, on choisit le patient dont la réaffectation apporte le gain le plus important, et on le réaffecte. Ce processus est répété jusqu'à ce qu'on ne puisse plus améliorer la solution.

Optimisation locale par permutation

Le principe de cette méthode est d'itérativement améliorer la solution en permutant à chaque fois les affectations d'une paire de patients électifs.

À chaque itération, on considère un patient *i*, et on détermine un patient *j* pour permuter leurs affectations. Le patient *j* est celui qui apporte le gain le plus important lorsqu'il est permuté avec le patient *i*, sans toucher aux affectations des autres patients et tout en satisfaisant les dates au plus tôt des patients et les capacités totales des salles-jours. Ensuite, les deux patients *i* et *j* sont permutés. À l'itération suivante, deux autres patients sont considérés et ainsi de suite. Ce processus est répété jusqu'à ce qu'il n'y ait plus de permutations améliorantes.

5. 2. 7 Combinaisons des différentes heuristiques

La méthodologie de résolution proposée pour le problème de planification approximé (PG_K) est composée de trois phases. Dans la première phase, le problème maître linéaire (PLM) est résolu par génération de colonnes ; le problème maître restreint (PMR) résolu par CPLEX LP et les sous-problèmes de génération de colonnes (GC_{ts}) résolus soit par une méthode heuristique (présentée en section 5.2.3) soit par un algorithme de branch-and-cut comme CPLEX IP. Dans la deuxième phase, une solution faisable est construite à partir de la solution optimale du PML. Cette phase fait appel soit à l'heuristique de réaffectation progressive (**RP**) soit à l'heuristique d'arrondissement (**HA**). Rappelons que cette dernière fait appel encore à la génération de colonnes pour la résolution des problèmes résiduels. La troisième phase a pour objectif d'améliorer la solution faisable en utilisant l'heuristique d'optimisation locale (**OL**) et/ou d'optimisation par permutation (**OP**).

Dans la section suivante nous testons les deux combinaisons suivantes (Méthode 1 et Méthode 2) :

- Méthode 1 : Génération de colonnes, RP, OL, et OP.
- Méthode 2 : Génération de colonnes, HA, OL, et OP.

5. 3 Expérimentations numériques

Les expérimentations numériques présentées dans cette section ont été réalisées sur un PC avec un processeur Pentium 4 à 3 GHz avec une mémoire de 512 Mo, utilisant un système d'exploitation Windows XP. Les algorithmes ont été programmés en MS Visual C++ et faisant appel à la bibliothèque d'optimisation CPLEX 9.0.

5. 3. 1 Génération des instances

Nous considérons un horizon de planification de 5 jours ($H = 5$) et des problèmes avec 3 et 6 salles opératoires ($S = 3, 6$). Les capacités régulières T_{ts} des salles-jours sont fixées à 8 heures. Les coûts des heures supplémentaires sont de 500 €/heure.

Les capacités aléatoires utilisées par les urgences W_{ts} suivent une loi log-normale avec une moyenne de 2 heures ($E[W_{ts}]$ = 2 heures). Les écarts types sont générés de manière aléatoire et uniforme à partir de l'ensemble $\{0.1 \times E[W_{ts}] ,\ldots, 0.5 \times E[W_{ts}]\}$.

Les durées opératoires d_i sont supposées suivre aussi une loi log-normale. Pour chaque patient, l'espérance E[d_i] est générée de manière aléatoire et uniforme de l'intervalle [0.5 heure, 3heures] , et l'écart type de d_i est généré aléatoirement à partir de l'ensemble $\{0.1 \times E[d_i] ,\ldots, 0.5 \times E[d_i]\}$.

Les dates au plus-tôt e_i et les coûts d'affectation a_{its} des patients électifs sont générés de la même manière que dans le chapitre précédent (Section 4.4.1).

Similairement au chapitre précédent, nous considérons deux classes de problèmes. Une classe où les salles opératoires sont identiques, et dans ce cas pour chaque patient électif *i* les coûts d'affectation a_{its} dépendent seulement du jour *t*. Pour la seconde classe, les salles sont de différents types et sont allouées à différentes spécialités. Un patient électif appartenant à une spécialité donnée peut être affecté soit à une salle allouée à la spécialité en question, soit à une salle allouée à une autre spécialité, mais dans ce cas il y a un coût supplémentaire, une pénalité.

Les patients électifs sont générés un par un jusqu'à ce que la somme des durées opératoires moyennes dépasse 75% de la capacité régulière totale disponible sur tout l'horizon de planification. Rappelons qu'avec $E[W_{ts}]$= 2 heures et T_{ts}= 8 heures, la charge moyenne due à la chirurgie d'urgence est 25% de la capacité totale. De cette manière, la demande pour la chirurgie élective et la chirurgie d'urgence représente 100% de la capacité totale disponible.

Le nombre de scénarios aléatoires *K* utilisés pour approximer le problème général (PG) par un problème déterministe (PG_K) est égal à 100. Comme nous l'avons constaté au chapitre 3, un tel nombre de scénarios est suffisant pour avoir une estimation assez précise du problème « exact » (PG).

5. 3. 2 Comparaison des stratégies de génération de colonnes

Nous commençons par comparer l'impact des différentes stratégies de génération de colonnes (all-negative, two-phase et cyclic strategy) sur la résolution du problème maître linéaire (PML). La construction de solutions faisables n'est pas considérée à ce stade. Le problème maître restreint (PMR) est initialisé avec un ensemble de colonnes Ω' générées aléatoirement (formant une solution faisable).

Les résultats numériques représentent les moyennes basées sur 10 instances générées aléatoirement pour chaque taille et la classe de problèmes. Ces résultats sont présentés en

tableau 5.1. Pour chaque stratégie, nous présentons le nombre d'itérations de la procédure de génération de colonnes (Nb Itérations), le nombre de colonnes générées (Nb Colonnes), le temps de calcul nécessaire pour la résolution du problème générateur (CPU Générateur) et le temps calcul total nécessaire pour la résolution du PML (CPU).

Pour les instances générées, le nombre moyen de patients électifs à planifier est égal à 53,9 pour des problèmes avec 3 salles et à 106,9 pour des problèmes avec 6 salles.

Notons que pour une instance donnée le même ensemble de scénarios est utilisé avec les différentes stratégies de génération de colonnes.

		Stratégie	Nb Itérations	Nb Colonnes	CPU Générateur	CPU
Salles identiques	3 Salles	All-negative	59.8	744.9	76.9	77.3
		Two-phase	68.6	665.8	64.9	65.2
		Cyclic	282.6	1333.4	55.7	56.6
	6 Salles	All-negative	104.4	2748.1	594.2	596.4
		Two-phase	112.4	2205.5	548.3	550.3
		Cyclic	689.6	6105.0	406.2	415.3
Salles non-identiques	3 Salles	All-negative	50.1	617.6	36.4	36.6
		Two-phase	72.1	605.4	22.6	23.0
		Cyclic	311.6	1202.5	27.9	28.7
	6 Salles	All-negative	75.7	1941.2	191.6	193.1
		Two-phase	93.3	1714.9	128.1	129.4
		Cyclic	658.4	5266.0	128.9	136.2

TAB. 5.1 - Comparaison des différentes stratégies de génération de colonnes

À partir du tableau 5.1, on peut remarquer que les stratégies « two-phase » et « cyclic » sont meilleures que la stratégie « all-negative » ; elles permettent une réduction considérable en temps de calcul.

Avec la stratégie « all-negative », les $H \times S$ sous-problèmes sont résolus de manière exacte à chaque itération de génération de colonnes ; ce qui est gourmand en temps de calcul.

Par rapport à la stratégie « all-negative », les stratégies « two-phase » et « cyclic » augmentent le nombre d'itérations de génération de colonnes ainsi que le nombre de colonnes générées. Mais elles permettent de diminuer le temps de calcul total. En effet, le nombre d'itérations

augmente parce que des colonnes de moins bonne qualité sont ajoutées au PMR. Cependant, comme les sous problèmes ne sont pas toujours résolus de manière exacte, le temps de calcul est réduit.

Nous constatons aussi que pour les problèmes de salles identiques la stratégie « cyclic » est meilleure que la stratégie « two-phase ». Mais le contraire est vrai avec les problèmes de salles non identiques. Pour des problèmes de salles identiques, les sous-problèmes sont similaires et par conséquent une colonne améliorante pour une salle-jour a de forte chance d'être améliorante pour plusieurs autres salles-jours.

On remarque aussi que la plus grande partie du temps de calcul est consommée pour la résolution des sous-problèmes de génération de colonnes, et que les problèmes de salles non identiques sont plus faciles à résoudre (temps de calcul plus court).

5. 3. 3 Comparaison des différentes combinaisons

Nous évaluons et comparons maintenant les performances de Méthode 1 et Méthode 2 (présentées en section 5.2.7). Nous utilisons la stratégie « cyclic » pour la génération de colonnes. Les critères de performances sont le temps de calcul et le « Gap » par rapport à la borne inférieure fournie par la génération de colonnes.

Afin de montrer l'intérêt d'utiliser une approche de génération de colonnes pour la résolution du problème approximé (PG_K), nous résolvons ce dernier en utilisant directement un algorithme de *branch-and-cut* de CPLEX MIP ; nous désignons cette méthode par « **MIP** ». Pour cette méthode nous fixons un budget en temps de calcul égal à une heure.

Les résultats numériques sont présentés en tableau 5.2 ; ils sont basés sur 10 instances générées aléatoirement pour chaque taille et classe de problème en considération. Pour une instance donnée, le même ensemble de scénarios est utilisé avec les différentes méthodes. Pour chaque méthode, nous fournissons Gap_max, Gap_av, Gap_min – respectivement le Gap maximal, moyen et minimal ; CPU_max, CPU_av, CPU_min – respectivement le temps de calcul maximal, moyen et minimal. Les temps de calcul sont en secondes et les Gap en pourcentage (%).

Le Gap est défini, de la même manière qu'au chapitre précédent, par (*UB* - *LB*) / *UB*. Où *UB* est le coût de la solution faisable, et *LB* est la borne inférieure fournie par la génération de colonnes, c-à-d le coût optimal du problème maître linéaire (PML).

		Salles Identiques		Salles Non-Identiques	
		3 Salles	6 Salles	3 Salles	6 Salles
Méthode 1	Gap_max	8.3	10.3	3.5	5.7
	Gap_av	5.8	6.0	1.0	4.5
	Gap_min	3.1	2.5	0.0	2.6
	CPU_max	94.0	650.5	54.1	276.0
	CPU_av	57.0	419.1	28.9	139.7
	CPU_min	27.2	295.6	16.0	91.2
Méthode 2	Gap_max	2.6	1.8	0.8	2.6
	Gap_av	1.5	1.0	0.2	1.6
	Gap_min	0.5	0.3	0.0	0.4
	CPU_max	149.6	1332.3	61.7	405.2
	CPU_av	94.5	967.4	33.7	292.1
	CPU_min	45.9	623.4	17.6	223.3
MIP	Gap_max	6.6	12.1	3.5	8.5
	Gap_av	5.0	10.9	1.4	6.6
	Gap_min	3.7	8.5	0.3	4.7
	CPU_max	3600	3600	3600	3600
	CPU_av	3600	3600	3600	3600
	CPU_min	3600	3600	3600	3600

TAB. 5.2 - Performances des Méthode 1 & Méthode 2

À partir des résultats numériques, on peut remarquer que la Méthode 2 utilisant l'heuristique d'arrondissement (HA) pour construire une solution faisable fournit des solutions de bonne qualité avec un Gap moyen inférieur à 2%. Cette méthode est nettement meilleure que la Méthode 1 utilisant la réaffectation progressive (RP) pour la construction des solutions faisables. Cependant, elle est plus gourmande en temps de calcul que la Méthode 1.

On peut aussi constater que les problèmes avec salles identiques nécessitent un temps de calcul plus élevé et ont des Gap légèrement supérieurs à ceux des problèmes avec salles non identiques.

À partir du tableau 5.2, on peut aussi remarquer que la méthode MIP est lourde en temps de calcul et les solutions fournies au bout d'une heure de calcul sont plus mauvaises que celles trouvées par Méthode 1 et Méthode 2. Ceci montre bien l'intérêt d'utiliser une approche de génération de colonnes pour la résolution du problème au lieu d'une simple utilisation du solveur CPLEX MIP.

En conclusion, Méthode 2 - heuristique basée sur la génération de colonnes - est une méthode efficace pour la résolution du problème de planification approximé (PG_K) ; elle fournit des solutions proches de l'optimum en un temps de calcul raisonnable.

5. 3. 4 Bénéfices d'une modélisation stochastique

Afin d'évaluer les bénéfices d'une modélisation stochastique du problème de planification, nous considérons aussi une version déterministe du problème.

La version déterministe consiste à remplacer les variables aléatoires W_{ts} ($t \in \{1,\ldots, H+1\}$, $s \in \{1,\ldots, S\}$) et d_i ($i \in \{1,\ldots, N\}$) par leurs moyennes respectives $E[W_{ts}]$ et $E[d_i]$. Le problème de planification (PG) est dans ce cas un problème d'optimisation déterministe (problème linéaire en variables mixtes). Désignons par X^{det} la solution optimale du problème déterministe.

Pour une instance donnée du problème,

- (i) nous résolvons la version déterministe en utilisant le solveur CPLEX MIP, et nous obtenons la solution X^{det}.

- (ii) en plus nous résolvons la version stochastique en utilisant notre approche (Méthode 2). Soit X^{sto} la solution obtenue. Rappelons que 100 scénarios (K=100) sont utilisés par l'approximation Monte Carlo pour obtenir le problème approximé (PG_K).

Afin de comparer les deux solutions X^{det} et X^{sto}, nous évaluons leurs coûts $J(X^{det})$ et $J(X^{sto})$. Pour ce faire, nous utilisons la simulation Monte Carlo avec un nombre très élevé de scénarios (10^6 scénarios), c-à-d, $J\left(X^{det}\right) \cong J_K\left(X^{det}\right)$ et $J\left(X^{sto}\right) \cong J_K\left(X^{sto}\right)$ avec $K = 10^6$ (Le critère $J_K(.)$ est donné par l'expression (5.4)).

La quantité $J(X^{det})$ - $J(X^{sto})$ représente le gain réalisé en utilisant une modélisation stochastique (*Value of the stochastic solution*, (Ruszczyński et Shapiro, 2003)).

Nous évaluons le gain en pourcentage ($J(X^{det})$ - $J(X^{sto})$) / $J(X^{det})$ pour des problèmes de planification avec 3 salles opératoires. Sur la base de 10 instances pour chaque classe de problèmes, le gain moyen est de 7,96% pour des problèmes avec salles identiques, et de 7,08% pour des problèmes avec salles non-identiques.

Nous notons que le gain est positif pour toutes les instances considérées. Nous signalons aussi que la résolution des versions déterministes des problèmes est très coûteuse en temps calcul. Le temps de calcul nécessaire à CPLEX MIP pour trouver la solution X^{det} est supérieur à 3 heures (en moyenne).

À partir de ces résultats, nous remarquons que les solutions obtenues par notre approche sont meilleures que les solutions déterministes. Tout en utilisant un modeste nombre de scénarios

(K= 100), les solutions fournies permettent une réduction du coût considérable par rapport à la solution déterministe.

5. 4 Conclusion

Dans ce chapitre, nous nous sommes intéressés à la planification du bloc opératoire tout en considérant les incertitudes liées aux durées d'interventions électives ainsi qu'à la chirurgie d'urgence. Nous avons proposé un modèle de planification stochastique et nous avons développé une approche de résolution qui combine la simulation Monte Carlo et la génération de colonnes.

Cette approche de résolution fournit des solutions approchées de bonne qualité, à moins de 2% de l'optimum, et permet de résoudre des problèmes de taille réaliste en un temps de calcul relativement court.

Nous avons proposé deux stratégies de génération de colonnes élaborées qui permettent d'exploiter la structure de problème pour accélérer la résolution : génération cyclique (*cyclic strategy*) et génération en deux phases (*two-phase strategy*). Les tests numériques ont montré que ces deux stratégies ont des performances similaires, et qu'elles sont efficaces et apportent des améliorations significatives en temps de calcul.

Moyennant des tests numériques, nous avons aussi montré qu'une réduction considérable des coûts peut être réalisée par une modélisation stochastique du problème. Tout en utilisant un modeste nombre de scénarios, l'approche proposée fournit des solutions de meilleure qualité que les solutions déterministes.

Enfin, nous notons que le modèle proposé dans ce chapitre peut être facilement étendu pour prendre en compte les coûts de sous-utilisation des salles opératoires. Avec cette extension, les propositions 5.1 et 5.2 ne sont plus valides, mais l'approche globale de résolution reste toujours applicable.

Conclusion générale

L'objectif de ce travail de thèse était de développer des modèles et outils d'optimisation pour la planification des blocs opératoires avec prise en compte des aléas dus essentiellement à la chirurgie d'urgence et aux durées des interventions chirurgicales. Dans le premier chapitre, nous avons décrit les différentes formes d'aléas qui caractérisent l'environnement du bloc opératoire, et nous avons souligné l'importance de les prendre en compte lors de la planification. Dans le chapitre 2, nous avons exposé un état de l'art sur les méthodes et approches développées pour la gestion des blocs opératoires. Nous avons noté ainsi l'utilisation prépondérante des approches déterministes pour la planification des blocs opératoires ; des approches qui font abstraction de toutes formes d'aléas et qui ne permettent pas de les intégrer lors de la planification. Dans les chapitres 3, 4 et 5, nous avons proposé différents modèles de planification stochastiques qui permettent de modéliser de manière explicite différentes formes d'aléas, et nous avons développé différentes approches de résolution.

Dans le troisième chapitre nous avons proposé un modèle stochastique pour la planification du bloc opératoire avec prise en compte de la chirurgie d'urgence, mais qui ne tient pas compte des incertitudes liées aux durées des interventions programmées. Dans ce chapitre, nous avons supposé que les salles opératoires sont polyvalentes, et seule la capacité globale de l'ensemble des salles opératoires est prise en compte. Dans ce modèle de planification, nous nous sommes intéressés à déterminer les dates d'intervention des patients électifs. L'affectation des patients à des salles opératoires spécifiques n'a pas été prise en compte. Nous avons formulé le problème de planification sous la forme d'un programme mathématique stochastique. Nous avons proposé une méthode d'optimisation qui combine la simulation Monte Carlo et la programmation en nombres mixtes. Nous avons étudié les propriétés de convergence de cette méthode et nous avons montré que les solutions obtenues convergent en exponentielle vers des vraies solutions optimales du problème. Cependant, l'inconvénient de cette méthode est qu'elle n'est efficace qu'avec des problèmes de faible taille. Pour palier à cet inconvénient, nous avons développé une méthode approchée basée sur la technique de relaxation Lagrangienne.

Dans le quatrième chapitre, nous avons apporté des extensions au modèle de planification pour permettre une affectation des patients aux salles opératoires, et la prise en compte d'une structure coût plus élaborée pour ces dernières. Nous avons proposé une approche de résolution basée sur la génération de colonnes. Nous avons testé différentes stratégies de génération de colonnes et nous avons évalué les performances de cette approche. Les tests numériques ont montré que cette approche fournit des solutions approximées très proches de l'optimum en un temps de calcul très court, pour des problèmes de tailles réalistes.

Dans le cinquième chapitre, nous avons considéré la planification du bloc opératoire avec prise en compte des aléas dus à la chirurgie d'urgence et aux durées des interventions électives. Cette extension a rendu l'approche de génération de colonnes présentée au chapitre 4 inapplicable. Nous avons alors développé une nouvelle approche qui combine la simulation Monte Carlo et la génération de colonnes. Nous avons exploité la structure des sous-problèmes de génération de colonnes pour accélérer leur résolution et pour développer des stratégies de génération de colonnes « non classiques ». Moyennant des expérimentations numériques, nous avons illustré l'efficacité de cette approche et nous avons montré que des gains significatifs peuvent être réalisés en utilisant une approche stochastique pour la planification du bloc opératoire.

Pour récapituler, nous avons considéré, dans le cadre de cette thèse, le problème de planification sous incertitudes du bloc opératoire. Nous avons proposé des modèles stochastiques qui (i) capturent les éléments essentiels à considérer lors de la planification des activités chirurgicales, (ii) permettent de modéliser de manière explicite différentes formes d'aléas tels que les incertitudes relatives aux durées des interventions et à la chirurgie d'urgence, et (iii) peuvent être facilement étendus pour modéliser d'autres contraintes dues aux pratiques du terrain.

Nous avons développé plusieurs approches et méthodes d'optimisation complètes et complémentaires pour la planification stochastique du bloc opératoire. Moyennant des expérimentations numériques, nous avons évalué les performances de ces différentes approches et nous avons montré leur efficacité. En particulier, nous avons montré que des gains considérables peuvent être réalisés moyennant une modélisation stochastique du problème de planification du bloc opératoire.

Ce travail de thèse ouvre la voie à des nombreuses perspectives de recherche qui se situent essentiellement sur deux plans : l'amélioration des méthodes d'optimisation et l'enrichissement des modèles de planification.

Les performances de l'approche de génération de colonnes peuvent être améliorées en accélérant la convergence du problème maître linéaire. Des techniques de stabilisation

remplissant ce rôle existent, et nous pensons qu'il serait intéressant de les mettre en œuvre. Une autre piste d'amélioration consiste à résoudre les sous-problèmes de génération de colonnes de manière efficace ; particulièrement les sous-problèmes rencontrés dans le chapitre 5. Ceci peut être réalisé en exploitant la structure des sous-problèmes pour développer des inégalités valides (tel que les *inégalités couvrantes* et/ou *couvrantes inverses*) pour accélérer la résolution exacte, ou développer des heuristiques performantes pour une résolution approchée. Il serait également intéressant d'exploiter la génération de colonnes pour développer une méthode de résolution exacte, *branch-and-price*.

Dans les différents modèles de planification présentés dans ce mémoire, nous avons supposé que les coûts relatifs aux patients représentent plusieurs facteurs : coûts d'hospitalisation, préférences des chirurgiens et des patients, adéquation des salles opératoires, des contraintes médicales, etc. Dans ce travail, nous avons supposé qu'on dispose de cet ensemble de coûts. Cependant, la détermination de ces coûts est une tâche qui n'est pas facile et qui nécessite la collaboration de plusieurs acteurs. L'élaboration d'une procédure pour estimer ces coûts est une direction intéressante pour nos recherches futures.

Nous notons que certaines contraintes (concernant la disponibilité des chirurgiens, l'adéquation des salles opératoires, ou des dates limites pour les interventions) peuvent être facilement prises en compte de manière explicite, en définissant pour chaque patient un ensemble explicit de salles-jours dans lesquelles le patient peut être planifié. Les modèles et approches de résolution proposés dans ce mémoire peuvent facilement intégrer ce genre de restrictions. Toutefois, il serait intéressant de tester la sensibilité des résultats par rapport à ces extensions.

Dans ce travail, nous avons supposé que les salles opératoires représentent les seules ressources critiques. Mais, il serait aussi intéressant d'inclure d'autres ressources telles que la salle de réveil, les lits d'hospitalisation, la salle des soins intensifs, etc.

Annexe 1: Résolution heuristique du sous-problème de génération de colonnes

Le sous-problème de génération de colonnes (GC_{ts}) (5.20)

$$\sigma_{ts}^{*} = \text{Min} \quad \sigma_{ts}(y_p) = -\pi_{ts} + \sum_{i \in I_t} \tilde{a}_{its} y_{ip} + (c_{ts}/K) \sum_{k=1}^{K} \left(W_{ts}^{k} + \sum_{i \in I_t} d_i^k y_{ip} - T_{ts} \right)^{+} - g_{ts}(0)$$

$$\text{sous contraintes :} \qquad y_{ip} \in \{0, 1\}, \ \forall\, i \in I_t$$

est résolu heuristiquement de la manière suivante :

Si $\pi_{ts} = 0$ alors construire une solution (colonne) avec un coût réduit négatif (PROCEDURE 1).

Sinon sélectionner et améliorer une colonne *p* déjà existante dans le PMR (PROCEDURE 2).

Nous rappelons que le multiplicateur de simplex π_{ts} est inférieur ou égal à 0.

PROCEDURE 1 : Construire une colonne (améliorante) avec un coût réduit négatif

Étape 1 : Initialiser

$y_{ip} = 0, \ \forall\, i \in I_t$

l'ensemble de patients candidats : $I \leftarrow I_t$

l'ensemble de patients inclus dans la colonne : $J \leftarrow \varnothing$

le coût réduit de la colonne : $\sigma_{ts}^{*} = 0$

Étape 2 : Déterminer $i^{*} = \underset{i \in I}{arg\,min} \quad \Delta_i = \tilde{a}_{its} + (c_{ts}/K) \sum_{k=1}^{K} (W_{ts}^{k} + d_i^k - T_{ts})^{+} - g_{ts}(0)$

Étape 3 : Si $\Delta_{i^*} \geq 0$ alors STOP ; il n'existe pas de colonnes avec coût réduit négatif.

Étape 4 : Insérer le patient i^{*} dans la colonne, $y_{i^*p} = 1$, et actualiser

l'ensemble de patients candidats : $I \leftarrow I \setminus \{i^{*}\}$

l'ensemble de patients inclus dans la colonne : $J \leftarrow J \cup \{i^{*}\}$

le coût réduit de la colonne : $\sigma_{ts}^{*} \leftarrow \sigma_{ts}^{*} + \Delta_{i*}$

Étape 5 : Déterminer $i^{*} = \underset{i \in I}{arg\,min} \quad \Delta_i = \tilde{a}_{its} + (c_{ts}/K)\sum_{k=1}^{K}\left(W_{ts}^{k} + \sum_{j \in J} d_j^k + d_i^k - T_{ts}\right)^{+}$

$$- (c_{ts}/K)\sum_{k=1}^{K}\left(W_{ts}^{k} + \sum_{j \in J} d_j^k - T_{ts}\right)^{+}$$

Étape 6 : Si $\Delta_{i*} \geq 0$ alors STOP ; aucun autre patient ne peut être inséré dans la colonne sans dégrader le coût réduit de cette dernière.

Sinon aller à l'Étape 4.

PROCEDURE 2 : Sélectionner et améliorer une colonne existante

Étape 1 : Sélectionner une colonne [y_p, z_p] existante dans le PMR et associée à la salle-jour(s, t) ; $p = \underset{p' \in \Omega' \text{ et } z_{tsp'}=1}{arg\,max}\ \lambda_{p'}\tilde{C}_{p'}$.

Étape 2 : Améliorer la solution y_p en utilisant la recherche locale. Le voisinage d'une solution y_p est défini de la manière suivante :

$V(y_p) = \{y_{p'}$ tel que $\exists\ i \in I_t$ avec $y_{ip'} = 1 - y_{ip}$ et $y_{jp} = y_{jp'}\ \ \forall\ j \in I_t \setminus \{i\}\ \}$

Annexe 2 : Principe de résolution par génération de colonnes

La génération de colonnes permet de résoudre certains programmes linéaires contenant un nombre de variables tel qu'il interdit l'application de l'algorithme du simplexe au problème dans sa globalité. Le principe de cette technique consiste à ne manipuler qu'un petit nombre de variables à la fois et à identifier les variables entrant en base au cours de la résolution sans les énumérer explicitement (Lübbecke et Desrosiers, 2005 ; Minoux, 1983).

Considérons (P), le programme linéaire sous forme standard suivant :

$$(P)\begin{cases} Min & z = c.x \\ s.c. & A.x = b \\ & x \geq 0 \end{cases}$$

avec :

– n : nombre de variables ($x = (x_1,..., x_n)^T$),

– m : nombre de contraintes ($m \leq n$),

– A = (a_{ij}) : Matrice des contraintes ($m \times n$, $rang(A) = m$),

– c : vecteur ligne des profits (ou gains),

– b : vecteur colonne des seconds membres.

On suppose que (P) comporte un nombre de contraintes "raisonnable" et un nombre de variables très largement supérieur ($n \gg m$), tel que la matrice A ne peut pas être explicitée. On suppose cependant que A est connue *implicitement*, c'est-à-dire que ses colonnes correspondent aux éléments d'un ensemble que l'on sait caractériser.

On suppose également que l'on dispose d'un algorithme efficace (appelé *algorithme générateur*) pour exhiber une colonne de A minimisant une fonction linéaire de la forme $z(A_j) = c_j - \pi.A_j$, où $\pi = (\pi_1,\ldots, \pi_m)$ est un vecteur-ligne quelconque.

Soit Ω un sous-ensemble des variables de (P) et A' la sous-matrice de A correspondante (on suppose que A' est de rang m). On note (PR) la restriction de (P) à Ω.

La résolution par le simplexe de (PR) fournit une solution de base optimale ($B^{-1}.b \mid 0$) associée à une base B issue de A', de rang m.

Cette solution est admissible mais pas nécessairement optimale pour (P). Pour l'améliorer, en suivant la méthode du simplexe, on cherche donc à faire entrer en base une variable de (P) de profit marginal (appelé aussi *coût réduit*) strictement négatif (en minimisation). On sait qu'une telle variable n'existe pas dans Ω (sinon ($B^{-1}.b \mid 0$) n'est pas optimale pour (PR)) ; il faut donc exhiber une des variables non explicitées de (P) et générer la colonne de A correspondante.

Or, on sait que le coût réduit d'une variable x_j (hors base) est défini par :

$$\overline{c}_j = c_j - \pi.A_j$$

où π est le vecteur des multiplicateurs du simplexe associés à B ($\pi = c_B.B^{-1}$), c'est-à-dire le vecteur des valeurs duales de (PR).

On peut donc déterminer, en utilisant l'*algorithme générateur*, la colonne A_r de A telle que :

$$\overline{c}_r = Min_j(z = c_j - \pi.A_j)$$

Le problème présenté ci-dessus est appelé *oracle*, ou *sous-problème de génération de colonnes*, ou encore *problème générateur*.

Tant que $\overline{c}_r$ est strictement négatif, on ajoute la variable x_r à Ω et la colonne A_r à A' et on itère le processus en résolvant le nouveau programme linéaire (PR). Lorsque l'algorithme générateur détermine $\overline{c}_r \geq 0$, l'algorithme s'arrête ; en effet, toutes les variables hors base ont alors un coût réduit positif ou nul, la solution courante est donc minimale pour (P).

Comme le simplexe, la méthode de résolution par génération de colonnes converge (chaque colonne de A est générée au plus une fois) et est exacte.

Bien entendu, l'hypothèse faite sur l'existence d'un algorithme générateur efficace est une hypothèse forte et peut donner le sentiment que la difficulté du problème a seulement été masquée. Cependant, dans de nombreuses applications, les colonnes de la matrice A ont une structure naturelle particulière que l'on peut exploiter avantageusement dans un algorithme générateur.

Bibliographie

Aickelin, U. and Dowsland, K. A. (2000). Exploiting problem structure in a genetic algorithms approach to a nurse rostering problem, *Journal of Scheduling*, 31, 139-153.

Aickelin, U. and White, P. (2004). Building better nurse scheduling algorithms, *Annals of Operations Research*, 128, 159-177.

Albert, F., Trilling, L. et Marcon, E. (2007). Hospital reorganization: how to help decision makers?. *The 33 Euro Working Group ORAHS*, Saint-Etienne, France.

Amladi, P. (1984). Outpatient health care facility planning and sizing via computer simulation. *Proceedings of the 1984 Winter Simulation Conference*, Dallas, Texas, USA, 28-30 November, 705-711.

Augusto, V., Xie, X. and Perdomo, V. (2007). Operating theatre scheduling with limited Recovery beds and patients recovery in operating rooms. *Proceedings of the International Conference on Industrial Engineering and System Management*, May 30-June 2, Beijing, China.

Badri, M. and Hollingsworth, J. (1993). A simulation model for scheduling in the emergency room. *Int J Opns and Prod Mgmt*, 13, 13-24.

Bard, J. F. and Purnomo, H. W. (2005a). A column generation-based approach to solve the preference scheduling problem for nurses with downgrading, *Socio-Economic Planning Sciences*, 39, 193-213.

Bard, J. F. and Purnomo, H. W. (2005b). Preference scheduling for nurses using column generation, *European Journal of Operational Research*, 164, 510-534.

Barnhart, C., Johnson, E.L., Nemhauser, G.L., Savelsbergh, M.W.P. and Vance P.H. (1998) Branch-and-price: Column generation for solving huge integer programs. *Operations Research*, **46**, 316-329.

Beliën, J. and Demeulemeester, E. (2006). Building cyclic master surgery schedules with leveled resulting bed occupancy, *European Journal of Operational Research*, 176, 1185-1204.

Beliën, J. and Demeulemeester, E. (2007). A branch-and-price approach for integrating nurse and surgery scheduling, *European Journal of Operational Research*, In Press.

Bellanti, F., Carello, G., Della Croce, F. and Tadei, R. (2004). A greedy-based neighborhood search approach to a nurse rostering problem, *European Journal of Operational Research*, 153, 28-40.

Besombes, B., Marcon, E., Albert, F., Merchier, L. et Bernaud, M. (2006). Quels indicateurs de performance pour piloter le regroupement de plateaux médico-techniques ? Retour d'expérience au CHU de St-Etienne. *GISEH 2006*, Luxembourg.

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

Blake, J.T. and Donald, J. (2002). Mount Sinai hospital uses integer programming to allocate operating room time, *Interface*, 32(2), 63-73.

Blake, J.T., Carter, M., O'Brien-Pallas, L. and McGillis-Hall L.(1995). A surgical process management tool, *Proceeding of the 8th World Congress on Medical Informatics MEDINFO 95*, Greenes R., Vancouver B.C.(Ed.): International Medical Informatics Association.

Bleakley, A., Hobbs, A., Boyden, J. and Walsh, L. (2004). Safety in operating theatres: Improving teamwork trough team resource management, *Journal of Workplace Learning*, 16, 83-91.

Bharadwaj, A., Sen, A. and Vinze, A, (1999). Scheduling cardiac procedures: A knowledge-based approach, *IEEE Transactions on Engineering Management*, 46(3), 322-334.

Bosi, F. and Milano, M. (2001). Enhancing constraint logic programming branch and bound techniques for scheduling problems, *Software Practice and Experience*, 31, 17-42.

Bounekkar, A., Deslandres, V., Magny, D. L. et Trilling, L. (2006). Étude des facteurs influençant le taux d'occupation des salles dans le contexte du regroupement de plateaux médico-techniques. *GISEH 2006*, Luxembourg.

Broun, G. (2002). Le plateau technique médical à l'hôpital, *Edition Eska*. Paris.

Brusco, M. J. and Jacobs, L. W. (1993). A simulated annealing approach to the cyclic staff-scheduling problem, *Naval Research Logistics*, 40, 69-84.

Burke, E. K., De Causmaecker, P., Petrovic, S. and Vanden Berghe, G. (2003). Chapter 7: Variable neighborhood search for nurse rostering problems, in M. G. C. Resende & J. P. de Sousa (eds), *METAHEURISTICS: Computer Decision-Making*, Kluwer (Combinatorial Optimiza-tion Book Series), 153-172.

Burke, E.K., De Causmaecker, P., Vanden Berghe, G. and Van Landeghem, H. (2004). The state of the art of nurse rostering. *The Journal of Scheduling*, 7, 441–499.

Calos, H. and Whitlock, A. (1986). Monte Carlo Methods, *vol 1: Basics, John Wiley*, New York.

Cardoen, B., Demeulemeester, E. and Beliën, J. (2006). Optimizing a multiple objective surgical case scheduling problem. *FETEW Research Report KBI_0625, K.U.Leuven*.

Cardoen, B., Demeulemeester, E. and Beliën, J. (2007). Determining surgery schedules on the operational level through column generation, *The ORAHS'2007 Conference*, Jul. 15-20, Saint-Etienne, France.

Carter, M. (2006). A Case Study of a Simulation-Based Decision Support Tool. *The ORAHS'2006 Conference*, July 23-28, Wroclow, Poland.

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

Chaabane, S. (2004). Gestion prédictive des blocs opératoires, *Mémoire de Thèse*, Institut National des Sciences Appliqués de Lyon.

Chaabane, S., Guinet, A., et Trilling, L. (2004). Pilotage conjoint de ressources hospitalière humaines et matérielles : un problème d'ordonnancement avec cycles. *Actes de GISEH*

2004, pages 14-23, Mons, Belgique.

Cheang, B., Li, H., Lim, A. and Rodrigues, B. (2003). Nurse rostering problems – a bibliographic survey. *European Journal of Operational Research*, 151, 447–460.

Cheng, B., Lee, J. and Wu, J. (1997). A nurse rostering system using constraint programming and redundant modeling, *IEEE Transactions on Information Technology in Biomedicine*, 1(1), 44-54.

Chen, J. G. and Yeung, T. W. (1993). Hybrid expert-system approach to nurse scheduling, *Computers in Nursing*, 11(4), 183-190.

Combes, C., Dussauchoy, A., Meskens, N. and Vandamme, J.P. (2007). Identification of bivariate Gamma probability distribution model to fit different periods of time related to surgical activity, *Proceedings of the International Conference on Industrial Engineering and System Management*, May 30- June 2, Beijing, China.

Currie, K., Iskander, W., Michael, L. and Coberly, C. (1984). Simulation modeling in health care facilities. *Proceedings of the 1984 Winter Simulation Conference*, Dallas, Texas, USA, 28-30 November, 713-717.

Dai, L. and Chen, C.H. (1997). Rates of convergence of ordinal comparison for dependent discrete event dynamic systems. *Journal of Optimization Theory and Applications*, **94**, 29-54.

Darmoni, S. J., Fajner, A., Mahe, N., Leforestier, A., Vondracek, M., O., S. and Baldenweck, M. (1995). Horoplan: computer-assisted nurse scheduling using constraint-based programming, *Journal of the Society for Health Systems*, 5(1),: 41-54.

Decker, K. M. (1991). The Monte Carlo method: Theory and application. *Computer Methods in applied Mechanics and Engineering*, 89: 463-483.

Delesie, L. (1998). Bridging the gap between clinicians and health managers, *European Journal of Operational Research*, 105, 248-256.

Dembo, A. and Zeitouni, O. (1993). Large deviations techniques and applications. *Jones and Barlett Publishers*, Boston.

Denton, B., Gupta, D. (2003) A sequential bounding approach for optimal appointment scheduling. *IIE Transactions*, **35**, 1003-1016.

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

Denton, B., Rahman, A., Nelson, H. and Bailey, A. (2006). Simulation of a Multiple Operating Room Surgical Suite. *Proceedings of the 2006 Winter Simulation Conference*.

Dexter, F., Macario, A., Qian, F. and Traub, R.D. (1999a). Forecasting surgical groups' total hours of elective cases for allocation of bloc time: application of time series analysis to operating room management. *Anesthesiology*, 91, 1501-1508.

Dexter, F., Macario, A. and Traub, RD., (1999b). Which algorithm for scheduling add-on elective cases to maximizes operating room utilization? Use of bin packing algorithms and fuzzy constraints in operating room management, *Anesthesiology*, 91, 1491-1500.

Dexter, F. and Traub, R.D. (2000). Sequencing cases in the operating room: predicting whether one surgical case will last longer than another. *Anesthesia & Analgesia* 90:975-979.

Dexter, F. and Macario, A. (2002). Changing allocations of operating room time from a

system based on historical utilization to one where the aim is to schedule as many surgical cases as possible, *Anesthesia & Analgesia*, 94, 1272-1279.

Dexter, F., Blake, J.T., Penning, D.H., Solan, B., Chung, P. and Lubarsky, D.A.(2002). Use of Linear programming to estimate impact of changes in a hospital's operating room time allocation on perioperative variable costs, *Anesthesiology*, 96(3), 718-724.

Dowsland, K. (1998). Nurse scheduling with tabu search and strategic oscillation, *European Journal of Operational Research*, 106, 393-407.

Dussauchoy, A., Combes, C., Gouin, F. et Botti, G. (2003). Simulation de l'activité d'un bloc opératoire en utilisant des données recueillies au niveau d'un département d'anesthésie. *Conférence GISEH 2003*, Lyon, France.

Everett, J.E. (2002). A decision support simulation model for the management of an elective surgery waiting system, *Health Care Management Science*, 5, 89-95.

Fei, H., Chu, C., Meskens, N. and Artiba, A. (2006). Solving surgical cases assignment problem by a Branch-and-Price approach. *International Journal of Production Economics*, to appear.

Fei, H. (2006). Vers un outil d'aide à la planification et l'ordonnancement des blocs opératoires. *Mémoire de Thèse*, Facultés Universitaires Catholiques de Mons.

Fei, H., Meskens, N. and Chu, C. (2007). An operating theatre planning and scheduling problem in the case of an open scheduling strategy. *Proceedings of the International Conference on Industrial Engineering and System Management*, May 30- June 2, Beijing, China.

Féniès, P., Gourgand, M. et Tchernev, N. (2004). Une contribution à la mesure de le performance dans la supply chain hospitalière : L'exemple du processus opératoire. *Actes de GISEH 2004*, Mons, Belgique.

Féniès, P. et Rodier, S. (2006). Un modèle décisionnel générique pour l'évaluation de la performance du processus logistique : Application en contexte hospitalier. *GISEH 2006*, Luxembourg.

Fisher, M.L. (1981). The Lagrangian relaxation method for solving integer programming problems, *Management Science*, 27(1), 1-18.

Fréville, A. (2004). The multidimensional 0–1 knapsack problem: An overview. *European Journal of Operational Research*, 155, 1–21.

Gallivan, S. (1998). Evaluation of priority strategies for hospital admissions. *The ORAHS'1998 Conference*, July 19-24, Rome, Italy.

Garey, M.R. and Johnson, M.R. (1979). Computers and Intractability: A Guide to the Theory of NP-Completeness, *Freeman*, San Francisco.

Garnett, J. (1998). Modelling an operating theatre suite. *The 12th European Simulation Multiconference*, Manchester, UK.

Gerchak, Y., Diwakar, G. and Mordechai, H. (1996). Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Science*, 42, 321-334.

Goffin, J.L. (1977). On the convergence rate of subgradient optimization methods. *Mathematical Programming*. 13, 329–347.

Goldman, J. and Knappenberger, H.A. (1968). How to determine the optimum number of operating rooms. *Modern Hospital*, 111(3), 114-116.

Gourgand, M., Mebrek, F. and Tanguy, A. (2005) Hospital Logistic Modelling and simulation, *European Simulation and Modelling Conference*, October 24-26, Porto, Portugal.

Guinet, A. and Chaabane, S. (2003). Operating theatre planning. *International Journal of Production Economics*, 85, 69-81.

Gürkan, G., Özge, A. Y. and Robinson, S. M. (1999). Sample-path solutions of stochastic variational inequalities. *Mathematical Programming*, 84:313-334.

Hammami, S. (2006). Aide à la décision dans le pilotage des flux matériels et patients d'un plateau médico-technique. *Thèse de doctorat*, Institut National Polytechnique de Grenoble.

Hammersley, J. M. and Handscomb, D. C. (1964). Monte Carlo Method. *Methuem & Co Ltd*, London.

Hans, E. W., Wullink, G., Houdenhoven, M. V. and Kazemier, G. (2006). Robust surgery loading. *European Journal of Operational Research*, Inpress.

Held, M., Wolfe, P. and Crowder, H.D. (1974). Validation of subgradient optimization. *Mathematical Programming*, 6 :62–88.

Hsu, V. N., Matta, R. and Lee, C. Y. (2003). Scheduling patients in an ambulatory surgical center, *Naval Research Logistics*, 50(3), 218-238.

Ho, C. and Lau, H. (1992). Minimizing total cost in scheduling outpatient appointments, *Management science*, 38(12), 1750-1764.

Hopkins, D.S.P., Gerson, A., Levin, P.J. and Merchant, R.S. (1982). A model for optimizing the number of operating rooms in a hospital surgical suite. *Health Care Management Review*, 49-64.

Ikegami, A. and Niwa, A. (2003). A subproblem-centric model and approach to the nurse scheduling problem, *Mathematical Programming*, 97(3), 517-541.

Iskander, W.H. and Carter, M.W. (1991). A simulation model for a same day care facility at a university hospital. *Proceedings of the 1991 Winter Simulation Conference*, Phoenix, Arizona, USA, 8-11 December, 846-853.

Isken, M. and Hancock, W. (1991). A heuristic approach to nurse scheduling in hospital units with non-stationary, urgent demand, and a fixed staff size, *Journal of the Society for Health Systems*, 2(2), 24-41.

Jan, A., Yamamoto, M. and Ohuchi, A. (2002). Search algorithms for nurse scheduling with genetic algorithms, *Operations Research/Management Science at Work, the International Series in Operations Research & Management Science*, Vol. 43, Kluwer Academic Publishers, 149-161.

Jaumard, B., Semet, F. and Vovor, T. (1998). A generalized linear programming model for nurse scheduling, *European Journal of Operational Research*, 107-118.

Jebali, A. (2004). Vers un outil d'aide à la planification et à l'ordonnancement des ressources dans les services de soins, *Mémoire de Thèse*, Institut National Polytechnique de Grenoble.

Jebali, A., Hadjalouane, A. B. and Ladet, P. (2006). Operating rooms scheduling.

International Journal of Production Economics, 99, 52-62.

Jun, J.B., Jacobson, S.H. and Swisher, J.R. (1999). Application of discrete-event simulation in health care clinics: A survey. *Journal of the Operational Research Society*, 50, 109-123.

Jones, A. W., Sahney, V. K. and Kurtoglu, A. (1983). A discrete event simulation for the management of surgical suite scheduling. *Proceedings of the 16th annual symposium on Simulation*, Tampa, Florida, United States, 263 – 278.

Kao, E.P.C. and Tung, G.G. (1981). Aggregate nursing requirement planning in a public health care delivery system. *Socio-Economic Planning Sciences*, 15(3), 119-127.

Kharraja, S., Chaabane, S. et Marcon, E. (2002). Évaluation de performances pour deux stratégies de programmation opératoire de bloc. Conférence Internationale Francophone d'Automatique CIFA, Nantes, France.

Kharraja, S. (2003). Outils d'aide à la planification et l'ordonnancement des plateaux médico-techniques, *Mémoire de Thèse*, Université Jean Monnet.

Kharraja, S. and Marcon, E. (2003). Construction automatique du plan directeur d'allocation des plages horaires. *Proceedings de GISEH'03*, 17-18 janvier, Lyon, France.

Klafehn, K.A. and Owens, D. (1987). A simulation model designed to investigate resource utilization in a hospital emergency room. *Proceedings of the 11th Annual Symposium on Computer Applications in Medical Care*, Washington DC, USA, 1-4 November, 676-679.

Klafehn, K.A., Owens, D., Felter, R., Vonneman, N. and McKinnon, C. (1989). Evaluating the linkage between emergency medical services and the provision of scarce resources through simulation. *Proceedings of the 13th Annual Symposium on Computer Applications in Medical Care*, Washington DC, USA, 5-8 November, 335-339.

Klassen, K.J. and Rohelder, T.R. (1996). Scheduling outpatient appointments in a dynamic environment. *Journal of operations Management*, 14, 83-110.

Kleywegt, A., Shapiro, A. and Homem-de-Mello, T. (2001). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2): 479-502.

Kontak-Forsyth, M. and Grant, A.E. (1995). OR booking policy: development and implementation, *Canadian Nursing Journal*, 13(1).

Kusters, R.J. and Groot, P.M.A. (1996). Modelling resource availability in general hospitals: Design and implementation of a decision support model. *European Journal of Operational Research*, 88, 428-445.

Kuzdrall, P.J., Kwak, N.K. and Schmitz, H.H. (1974). A technical note on the operating-room and recovery room usage. *Operations Research*, 22, 434-440;

Kuzdrall, P.J., Kwak, N.K. and Schmitz, H.H. (1981). Simulating space requirements and scheduling policies in a hospital surgical suite. *Simulation*, 27: 163-171.

Kwak, N., Kuzdrall, P.J. and Schmitz, H. (1975). Simulating the use of space in a hospital surgical suite. *Simulation*, 24, 147-152.

Kwak, N.K., Kuzdrall, P.J. and Schmitz, H.H. (1976). The GPSS Simulation of Scheduling Policies for Surgical Patients, *Management Science*, 22(9), 982-989.

Lafon, N. et Landry, S. (2001). Gérer plus efficacement les stocks du bloc opératoire à partir de la programmation des interventions chirurgicales. *Gestion hospitalière*, Avril 2001 No 405.

Lamiri, M. and Xie, X. (2006). Operating rooms planning using Lagrangian relaxation technique. *Proceedings of the 2006 IEEE Conference on Automation Science and Engineering*, October 8-10, 2006, Shanghai Pudong, China, IEEE Catalog Number: 06EX1338C, ISBN: 1-4244-0311-1, p. 186 – 191.

Lamiri, M., Grimaud, F. and Xie, X. (2006a). Optimization methods for surgery planning under uncertain demand for emergency surgery. *Information Control Problems In Manufacturing 2006: A Proceedings volume from the 12th IFAC International Symposium*, St Etienne, France, 17-19 May 2006, A. Dolgui, G. Morel, C. Pereira (Eds.), Elsevier Science, 2006, ISBN: 978-0-08-044654-7, vol. 3, p. 633- 638.

Lamiri, M., Xie, X., Dolgui, A. and Grimaud, F. (2006b). A stochastic model for operating room planning with elective and emergency demand for surgery, *European Journal of Operational Research*, (accepted, In Press).

Lamiri, M. and Xie, X. (2007). Operating room planning with elective and emergency patients. *Proceedings of the International Conference on Industrial Engineering and Systems Management*, Beijing, China, May 30- June 2, 2007.

Lamiri, M., Dréo, J., and Xie, X. (2007a). Operating room planning with random surgery times. *Proceedings of IEEE Conference on Automation Science and Engineering*, Scottsdale, USA, Sept 22-25, 2007.

Lamiri, M., Grimaud, F. and Xie, X. (2007b). Optimization methods for a stochastic surgery planning problem, *International Journal of Production Economics*, (Under review).

Lamiri, M., Xie, X. and Zhang, S. (2007c). Column generation for operating theatre planning with elective and emergency patients, *IIE Transactions*, (accepted).

Lans, M.V.D., Hans, E. W., Hurink, J. L., Wullink, G., Houdenhoven, M. V. and Kazemier, G. (2006). Anticipating Urgent Surgery in Operating Room Departments. *BETA working paper WP-158, ISSN: 1386-9213*, University of Twente.

Lapierre, S.D., Batson, C. and McCaskey, S. (1999). Improving on-time performance in health care organizations: a case study, *Health care Management Science*, 2, 27-34.

Law, A.M. and Kelton, W.D. (1991), *Simulation Modeling and Analysis*, McGraw-Hill, New York.

Lebowitz, P. (2003). Schedule the short procedure first to improve OR efficiency. *AORN Journal* 78:651-659.

Levecq, P., Meskens, N., et Artiba, A. (2003). Utilisation d'une approche Data Mining pour la spécification des durées en milieu hospitalier, *Santé et systémique*, 7, 191-203.

Liu, L. and Liu, X. (1998a) Dynamic and static job allocation for multi-server systems. *IIE transactions*, **30**, 845-854.

Liu, L. and Liu, X. (1998b) Block appointment systems for outpatient clinics with multiple doctors. *Journal of the Operational Research Society*, **49**, 1254-1259.

Liyanage, L. and Gale, M. (1995). Quality improvement for the Campbelltown hospital emergency service. *IEEE International Conference on Systems, Man, and Cybernetics*, Vancouver, British Columbia, Canada, 22-25 October, 1997- 2002.

Lovejoy, W. and Li, Y. (2002). Hospital operating room capacity expansion. *Management Science*, 48(11), 1369-1387.

Lübbecke, M. E. and Desrosiers, J. (2005). Selected Topics in Column Generation. *Operations Research*, **53**, 1007–1023.

Macario, A., Dexter, F. and Traub, R.D. (2001). Hospital profitability per hour of operating room time can vary among surgeons, *Anesthesia & Analgesia*, 93, 669-675.

Magerlein, J., and Martin, J. (1978). Surgical demand scheduling: A review. *Health Services Research*, 31(4), 418-433.

Marcon, E., Kharraja, S. and Simmonet, G. (2003a). The operating theatre scheduling: an approach centered on the follow-up of the risk of no realization of the planning. *International Journal of Production Economics*, **85**, 83- 90.

Marcon, E., Kharraja, S., Smolski N. and Luquet, B. (2003b) Determining the number of beds in the post-anesthesia care unit: a computer simulation approach. Journal of the Operational Anesthesia Research Society, *Anesthesia and Analgesia*, 96, 1415-1423.

Martello, S. and Toth, P. (1990). Knapsack Problems: Algorithms and Computer Implementations, *Wiley*, 36-38.

Marty, J. (2001). Organisation des sites opératoires. *Conférences d'actualisation*, 203-224.

Mason, A. J. and Smith, M. C. (1998). A nested column generator for solving rostering problems with integer programming, *International Conference on Optimisation: Techniques and Applications*, 827-834.

McHugh, M.L. (1989). Computer simulation as a method for selecting nurse staffing levels in hospitals. *Proceedings of the 1989 Winter Simulation Conference*, Washington DC, USA, 4-6 December, 1121-1129.

Mcleod, H., Ham, C. and Kipping, R. (2003). Booking patients for hospital admissions: evaluation of a pilot programme for day cases, *British Medical Journal*, 327, 1147-1151.

Meier, L., Sigal, E. and Vitale, F.R. (1985). The use of simulation model for planning ambulatory surgery. *Proceedings of the 1985 Winter Simulation Conference*, San Francisco, California, USA, 11-13 December, 558-563.

Minoux, M. (1983). Programmation mathématique : Théorie et algorithmes. *Collection technique et scientifique des télécommunications DUNOD*, Paris.

Murphy, D.R. and Sigal, E. (1985). Evaluating surgical block schedules using computer simulation, *Winter Simulation Conference Proceeding*, Gantz D., Blais G., Solomon S. (Eds.), 551-557.

Ogulata, S.N. and 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.

Okada, M. (1992). An approach to the generalized nurse scheduling problem-generation of a declarative program to represent institution-specific knowledge, *Computers and Biomedical Research*, 25(5), 417-434.

Olson, E. and Dux, L.E. (1994). Computer model targets best route for expanding hospital surgicenter. Indust Engng, 26: 24-26.

O'Neil, L. and Dexter, F. (2004). Evaluating the efficiency of hospital's perioperative service using DEA. Operation research and health care: A hand book of methods and

application, Chapter 6. (Eds) Brandeau, M.L., Sainfort, F. and Pierskalla, Kluwer, London.

Overdyk, F.J., Harvey, S.C., Fishman, R.L., and Shippey, F. (1998). Successful strategies for improving operating room efficiency at academic institutions, *Anesthesia & Analgesia*, 86, 896-906.

Ozkarahan, I. (1989). Flexible nurse scheduling support systems, *Computer Methods and Programs in Biomedicine*, 30,145-153.

Ozkarahan, I. (2000). Allocation of surgeries to operating rooms by goal programming, *Journal of Medical Systems*, 24(6), 339-378.

Patterson, P. (1996). What makes a well-oiled scheduling system, *Journal of OR Manager*, 12(9), 19-23.

Perdomo, V., Augusto, V. and Xie, X. (2006). Operating theatre scheduling using Lagrangian relaxation. *Proceedings of the 2006 International Conference on Service Systems and Service Management*, Troyes, France.

Persson, M. and Persson, J.A. (2005). Optimization modeling of hospital operating room planning using a logistic perspective, *The 16th Annual Conference of POMS*, 29 April-2 Mai, 2005, Chicago, USA.

Petrovic, S., Beddoe, G. and Vanden-Berghe, G. (2003). Storing and adapting repair experiences in personnel rostering, Practice and Theory of Automated Timetabling, *Fourth International Conference*, Gent, Vol. 2740, Springer, 185-186.

Pham, D.N. and Klinkert, A. (2006). Surgical case scheduling as a generalized job shop scheduling problem. *European Journal of Operational Research*, In press.

Plambeck, E.L., Fu, B.R., Robinson, S.M. and Suri, R. (1996). Sample-path optimization of convex stochastic performance functions. *Mathematical Programming*, Series B, 75:137-176.

Polyak, B.T. (1976). Minimization of unsmooth functionals. *Soviet Mathematics*, 8 :593–597.

Polyak, B.T. (1969). Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9 :14–29.

Ramis, F.J., Palma, G.L. and Baesler, F.F. (2001). The use of simulation for process improvement at an ambulatory surgery center. *Proceedings of the 2001 Winter Simulation Conference*, ed. B.A. Peters, J.S. Smith, D.J. Medeiros ans M.W. Rohrer, 1401-1407.

Reymondon, F., Pellet, B. and Marcon, E. (2006). Methodology for designing medical device packages based on sterilisation costs. *INCOM 2006*, Volume 3, 701-706, Saint-Etienne, France.

Roland, B., Martinelly, C.D., Riane, F. and Pochet, Y. (2007). Scheduling operating theatre under human resources constraints. *Proceedings of the International Conference on Industrial Engineering and System Management*, May 30- June 2, Beijing, China.

Rossi-Turk, D. (2002). Comment garantir la qualité et la séqurité au bloc opératoire par une programmation et logistique innovante?. *Santé et Systémique*, 6, 1-3.

Ruszczyński, A. and Shapiro, A. (2003). Stochastic Programming. *Handbooks in Operations Research and Management Science*, v10, Elsevier, Amsterdam, The Netherlands.

Sahney, V.K., Knappenberger, H.A. and Purohit, H.C. (1976). Elective admission scheduling

through evolutionary policy determination, *Proceedings of the 1st International conference of systems science in health care*, 5-9 July 1976, Paris, France.

Santibanez, P., Begen, M. and Atkins, D. (2004). Managing surgical waitlists for a British Columbia health authority. *Working paper, Sauder School of Business, University of British Columbia: Vancouver*, Canada.

Schmitz, H.H. and Kwak, N.K. (1972). Monte Carlo simulation of operating-room and recovery room usage. *Operations Research*, 20, 1171-1180;

Scott, S. and Simpson, R.M. (1998). Case-bases incorporating scheduling constraint dimensions: Experiences in nurse rostering, *Advances in Case-Based Reasoning*, Vol. 1488, Springer, 392-401.

Shukla, R.K., Ketcham, J.S. and Ozcan, Y.A. (1990). Comparison of subjective versus data base approaches for improving efficiency of operating room scheduling. *Health Services Management Research*, 3, 74-81.

Sier, D., Tobin, P. and McGurk, C. (1997). Scheduling Surgical Procedures, *Journal of the Operational Research Society*, 48, 884-891.

Smolski, N., Marcon, E., Chaabane, S., Luquet, B. et Viale, J.P. (2002). Impact des stratégies de brancardage sur l'occupation de la SSPI : Etude par simulation », *44ème Congrès national d'Anesthésie et de Réanimation*, 19-22 septembre 2002.

Strum, D.P., Vargas, L.G., May, J.H. and Bashein, G. (1997). Surgical suite utilization and capacity planning: a minimal cost analysis model, *Journal of Medical Systems*, 21(5), 309-322.

Strum, D.P., May, J.H. and Vargas, L.G. (1998). Surgical procedure times are well modeled by the log normal distribution. *Anesthesia & Analgesia*, 86, S47.

Strum, D.P, Vargas, L.G. and May, J.H. (1999). Surgical subspecialty block utilization and capacity planning: a minimal cost analysis model. *Anesthesiology*, 90, 1176-1185.

Strum, D.P., May, J.H. and Vargas, L.G. (2000). Modeling the uncertainty of surgical procedure times: comparison of the log-normal and normal models. *Anesthesiology*, 92, 1160-1167.

Swisher, J.R., Jun, J.B, Jacobson, S.H and Balci, O. (1997). Simulation of the Queston physician network. *Proceedings of the 1997 Winter Simulation Conference*, Atlanta, Georgia, USA, 7-10 December, 1146-1154.

Tanomaru, J. (1995). Staff scheduling by a genetic algorithm with heuristic operators, *Proceedings of the IEEE Conference on Evolutionary Computation*, New York, 456-461.

Trilling, L. (2006). Aide à la décision pour le dimensionnement et le pilotage de ressources humaines mutualisées en milieu hospitalier. *Mémoire de Thèse*, Institut National des Sciences Appliqués de Lyon.

Trivedi, V.M. and Warner, D.M. (1976). A branch and bound algorithm for optimum allocation of float nurses, *Management Science,* 22(9), 972-981.

Trivedi, V.M. (1981). A mixed-integer goal programming model for nursing service budgeting. *Operation Research*, 29, 1019-1034.

Van Oostrum, J.M., Houdenhoven, M.V., Hurink, J.L., Hans, E.W., Wullink, G. and Kazemier, G. (2005). A master surgical scheduling approach for cyclic scheduling in

operating room departments. *Memorandum n° 1789, ISSN 0169-2690, To appear in: OR Spectrum*, University of Twente.

Velasquez, R. and Melo, M.T. (2005). Set Packing Approach for Scheduling Elective Surgical Procedures. *The Annual International Conference of the German Operations Research Society*, September 7–9, Bremen, Germany.

Venkataraman, R. and Brusco, M.J. (1996). An integrated Analysis of Nurse and Scheduling Policies. *OMEGA International Journal of Management Science,* 24, 57-71.

Vissers, J.M.H., Adan, I.J.B.F. and Bekkers, J.A. (2005). Patient mix optimization in tactical cardiothoracic surgery planning: a case study. *IMA Journal of Management Mathematics*, 16, 281-304.

Vissers, J.M.H. (1998). Patient flow-based allocation of inpatient resources: a case study. *European Journal of Operational Research,* 105, 356-370.

Warner, M. (1976). Nurse staffing, scheduling and reallocation in the hospital. *Hospital and Health Services Administration*, 21(3), 77-90.

Weil, G., Heus, K., Francois, P. and Poujade, M. (1995). Constraint programming for nurse scheduling, *Engineering in Medicine and Biology Magazine, IEEE,* 14, 417-422.

Weiss, E.N. (1990). Models for determining started start times and case orderings in hospital operating rooms. *IIE Transactions*, 22, 143-150.

Wilt, A. and Goddin, D. (1989). Health care case study: Simulating staffing needs and work flow in an outpatient diagnostic center. *Indust Engng*, 21, 22-26.

Wolsey, L.A. (1998). Integer programming, *Wiley-Interscience*, New York.

Wright, I.H, Kooperberg, C., Bonar, A.B. and Bashein, G. (1996). Statistical modeling to predict elective surgery time: Comparison with a computer scheduling system and surgeon-provided estimates, *Anesthesiology*, 85(6), 1235-1245.

Xie, X. (1997). Dynamics and Convergence Rate of Ordinal Comparison of Stochastic Discrete-Event Systems. *IEEE Transactions on Automatic Control,* 42, 586-590.

Zhou, J. and Dexter, F. (1998). Method to assist in the scheduling of add-on surgical cases, Upper prediction bounds for surgical case durations based on the log-normal distribution, *Anesthesiology*, 89 (5), 1228-1232.

École Nationale Supérieure des Mines de Saint-Étienne

N° d'ordre : 446 GI

Prénom Nom : Mehdi Lamiri

Titre de la thèse : Operating Rooms Planning Under uncertainties

Spécialité : Industrial Engineering

Mots clefs : Planning under uncertainties, Operating rooms, Stochastic programming, Column generation, Lagrangian relaxation, Monte Carlo optimization

Résumé

Facing ever increasing health care demand, limited government support and increasing competition, hospitals are more and more aware of the need to use their resources as efficiently as possible. Operating Rooms (ORs) are among the most critical resources that generate highest costs for a hospital. For these reasons, planning OR activities has become one of the major priorities for hospitals.

The planning problem consists of determining a plan that specifies the set of elective patients that would be operated in each OR in each period over a planning horizon. This problem has been addressed in the health care literature and several approaches for OR planning have been proposed. However, most existing approaches are based on deterministic models that do not consider uncertainty related to the ORs' environment. Yet, uncertainty is inherent to the world of health care; it concerns essentially emergency patients' arrivals, surgery durations, and equipments and medical staff availability. The goal of our research is to develop optimization models and solution approaches (algorithms) for ORs planning under uncertainties.

In this thesis, we proposed several OR planning models that (i) capture essential factors relevant to surgical activities planning, (ii) explicitly take into account several kinds of uncertainties such as random demand for emergency surgery and random surgery durations, and (iii) can be easily extended to consider other real world constraints. We have also developed several solution methods and approaches for the stochastic ORs planning problem.

Using numerical experiments, we have evaluated the performances of the different solution approaches and showed their efficiency. Numerical experiments also show the importance of explicit modeling of uncertainties. Compared with deterministic OR planning models, those neglect uncertainty, our stochastic planning methods yield to significant cost reductions.

École Nationale Supérieure des Mines de Saint-Étienne

N° d'ordre : 446 GI

Prénom Nom : Mehdi Lamiri

Titre de la thèse : Planification des blocs opératoires avec prise en compte des aléas

Spécialité : Génie Industriel

Mots clefs : Planification sous incertitudes, Bloc opératoire, Programmation stochastique, Génération de colonnes, Relaxation lagrangienne, Optimisation Monte Carlo

Résumé

Le bloc opératoire constitue l'un des secteurs les plus coûteux et les plus importants dans un établissement hospitalier. Afin, d'utiliser de manière efficace et rationnelle les ressources (humaines et matérielles) disponibles tout en assurant une bonne qualité de service vis-à-vis des patients, la planification du bloc opératoire est devenue l'une des premières préoccupations des établissements hospitaliers.

Le problème de planification du bloc opératoire consiste à déterminer pour un horizon de plusieurs jours, l'ensemble d'interventions qui seront réalisées dans chaque salle opératoire. Ce problème a été traité dans la littérature et plusieurs approches de planification ont été proposées. Toutefois, les approches existantes sont essentiellement basées sur des modèles déterministes qui font abstraction de toutes sortes d'aléas. Or, le bloc opératoire est sujet à nombreuses formes d'aléas qui concernent essentiellement la chirurgie d'urgence et les durées d'interventions. À ce titre, l'objectif de notre travail de recherche est de développer des modèles et méthodes pour la planification des activités chirurgicales dans le bloc opératoire tout en tenant compte des aléas importants.

Dans le cadre de cette thèse, nous avons proposé des modèles stochastiques qui (i) capturent les éléments essentiels à considérer lors de la planification des activités chirurgicales, (ii) permettent de modéliser de manière explicite différentes formes d'aléas tels que les incertitudes relatives aux durées des interventions et à la chirurgie d'urgence, et (iii) peuvent être facilement étendus pour modéliser d'autres contraintes dues aux pratiques du terrain. Nous avons aussi développé plusieurs approches et méthodes d'optimisation complètes et complémentaires pour la planification stochastique du bloc opératoire. Moyennant des expérimentations numériques, nous avons évalué les performances de ces différentes approches et nous avons montré leurs efficacités. En particulier, nous avons montré que des gains considérables peuvent être réalisés moyennant une modélisation stochastique du problème de planification du bloc opératoire.