Design and flow control of stochastic health care networks without waiting rooms : A perinatal application

Canan Pehlivan

Ecole Nationale Supérieure des Mines
SAINT-ETIENNE

NNT : *2014 EMSE 0731*

THÈSE

présentée par

Canan PEHLIVAN

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

Spécialité : Génie Industriel

DESIGN AND FLOW CONTROL OF A STOCHASTIC HEALTHCARE NETWORK WITHOUT WAITING ROOMS: A PERINATAL APPLICATION

soutenue à Saint-Etienne, le 23 Janvier 2014

Membres du jury

Président :	Sylvie NORRE	Professeur, LIMOS, Montluçon
Rapporteurs :	Ger KOOLE	Professeur, VU University Amsterdam, Amsterdam
	Christian TAHON	Professeur, L'Université de Valenciennes, Valenciennes
Examinateur(s):	Jean-Philippe GAYON	Maitre de conférences HDR, Laboratoire G-SCOP Grenoble INP, Grenoble
	Sylvie NORRE	Professeur, LIMOS, Montluçon
Directeur de thèse :	Xiaolan XIE	Professeur, École Nationale Supérieure des Mines de St-Étienne, Saint-Etienne
Co-Encadrant :	Vincent AUGUSTO	Chargé de Recherche, École Nationale Supérieure des Mines de St-Étienne, Saint-Etienne
Invité(s) éventuel(s):	Catherine CRENN HEBERT	Docteur, Hôpital Louis Mourier, Paris

Spécialités doctorales :	Responsables :
SCIENCES ET GENIE DES MATERIAUX	K. Wolski Directeur de recherche
MECANIQUE ET INGENIERIE	S. Drapier, professeur
GENIE DES PROCEDES	F. Gruy, Maître de recherche
SCIENCES DE LA TERRE	B. Guy, Directeur de recherche
SCIENCES ET GENIE DE L'ENVIRONNEMENT	D. Graillot, Directeur de recherche
MATHEMATIQUES APPLIQUEES	O. Roustant, Maître-assistant
INFORMATIQUE	O. Boissier, Professeur
IMAGE, VISION, SIGNAL	JC. Pinoli, Professeur
GENIE INDUSTRIEL	A. Dolgui, Professeur
MICROELECTRONIQUE	

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

AVRIL	Stéphane	PR2	Mécanique et ingénierie	CIS
BATTON-HUBERT	Mireille	PR2	Sciences et génie de l'environnement	FAYOL
BENABEN	Patrick	PR1	Sciences et génie des matériaux	CMP
BERNACHE-ASSOLLANT	Didier	PR0	Génie des Procédés	CIS
BIGOT	Jean Pierre	MR(DR2)	Génie des Procédés	SPIN
BILAL	Essaid	DR	Sciences de la Terre	SPIN
BOISSIER	Olivier	PR1	Informatique	FAYOL
BORBELY	Andras	MR(DR2)	Sciences et génie de l'environnement	SMS
BOUCHER	Xavier	PR2	Génie Industriel	FAYOL
BRODHAG	Christian	DR	Sciences et génie de l'environnement	FAYOL
BURLAT	Patrick	PR2	Génie Industriel	FAYOL
COURNIL	Michel	PR0	Génie des Procédés	DIR
DARRIEULAT	Michel	IGM	Sciences et génie des matériaux	SMS
DAUZERE-PERES	Stéphane	PR1	Génie Industriel	CMP
DEBAYLE	Johan	CR	Image Vision Signal	CIS
DELAFOSSE	David	PR1	Sciences et génie des matériaux	SMS
DESRAYAUD	Christophe	PR2	Mécanique et ingénierie	SMS
DOLGUI	Alexandre	PR0	Génie Industriel	FAYOL
DRAPIER	Sylvain	PR1	Mécanique et ingénierie	SMS
FEILLET	Dominique	PR2	Génie Industriel	CMP
FOREST	Bernard	PR1	Sciences et génie des matériaux	CIS
FORMISYN	Pascal	PR0	Sciences et génie de l'environnement	DIR
FRACZKIEWICZ	Anna	DR	Sciences et génie des matériaux	SMS
GARCIA	Daniel	MR(DR2)	Génie des Procédés	SPIN
GERINGER	Jean	MA(MDC)	Sciences et génie des matériaux	CIS
GIRARDOT	Jean-jacques	MR(DR2)	Informatique	FAYOL
GOEURIOT	Dominique	DR	Sciences et génie des matériaux	SMS
GRAILLOT	Didier	DR	Sciences et génie de l'environnement	SPIN
GROSSEAU	Philippe	DR	Génie des Procédés	SPIN
GRUY	Frédéric	PR1	Génie des Procédés	SPIN
GUY	Bernard	DR	Sciences de la Terre	SPIN
GUYONNET	René	DR	Génie des Procédés	SPIN
HAN	Woo-Suck	CR	Mécanique et ingénierie	SMS
HERRI	Jean Michel	PR1	Génie des Procédés	SPIN
INAL	Karim	PR2	Microélectronique	CMP
KERMOUCHE	Guillaume	PR2	Mécanique et Ingénierie	SMS
KLOCKER	Helmut	DR	Sciences et génie des matériaux	SMS
LAFOREST	Valérie	MR(DR2)	Sciences et génie de l'environnement	FAYOL
LERICHE	Rodolphe	CR	Mécanique et ingénierie	FAYOL
LI	Jean Michel		Microélectronique	CMP
MALLIARAS	Georges	PR1	Microélectronique	CMP
MOLIMARD	Jérôme	PR2	Mécanique et ingénierie	CIS
MONTHEILLET	Franck	DR	Sciences et génie des matériaux	SMS
PERIER-CAMBY	Laurent	PR2	Génie des Procédés	DFG
PIJOLAT	Christophe	PR0	Génie des Procédés	SPIN
PIJOLAT	Michèle	PR1	Génie des Procédés	SPIN
PINOLI	Jean Charles	PR0	Image Vision Signal	CIS
POURCHEZ	Jérémy	CR	Génie des Procédés	CIS
ROUSTANT	Olivier	MA(MDC)		FAYOL
STOLARZ	Jacques	CR	Sciences et génie des matériaux	SMS
SZAFNICKI	Konrad	MR(DR2)	Sciences et génie de l'environnement	CMP
TRIA	Assia		Microélectronique	CMP
VALDIVIESO	François	MA(MDC)	Sciences et génie des matériaux	SMS
VIRICELLE	Jean Paul	MR(DR2)	Génie des Procédés	SPIN
WOLSKI	Krzystof	DR	Sciences et génie des matériaux	SMS
XIE	Xiaolan	PR0	Génie industriel	CIS

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

BERGHEAU	Jean-Michel	PU	Mécanique et Ingénierie	ENISE
BERTRAND	Philippe	MCF	Génie des procédés	ENISE
DUBUJET	Philippe	PU	Mécanique et Ingénierie	ENISE
FORTUNIER	Roland	PR	Sciences et Génie des matériaux	ENISE
GUSSAROV	Andrey	Enseignant contractuel	Génie des procédés	ENISE
HAMDI	Hédi	MCF	Mécanique et Ingénierie	ENISE
LYONNET	Patrick	PU	Mécanique et Ingénierie	ENISE
RECH	Joël	MCF	Mécanique et Ingénierie	ENISE
SMUROV	Igor	PU	Mécanique et Ingénierie	ENISE
TOSCANO	Rosario	MCF	Mécanique et Ingénierie	ENISE
ZAHOUANI	Hassan	PU	Mécanique et Ingénierie	ENISE

Mise à jour : 07/04/2013

PR 0	Professeur classe exceptionnelle
PR 1	Professeur 1ère classe
PR 2	Professeur 2ème classe
PU	Professeur des Universités
MA (MDC)	Maître assistant
DR	Directeur de recherche

Ing.	Ingénieur
MCF	Maître de conférences
MR (DR2)	Maître de recherche
CR	Chargé de recherche
EC	Enseignant-chercheur
IGM	Ingénieur général des mines

SMS	Sciences des Matériaux et des Structures
SPIN	Sciences des Processus Industriels et Naturels
FAYOL	Institut Henri Fayol
CMP	Centre de Microélectronique de Provence
CIS	Centre Ingénierie et Santé

Acknowledgments

First and foremost, I would like to express my sincere gratitude and very great appreciation to Prof. Xiaolan Xie, my thesis director and supervisor, for his valuable guidance and immense knowledge that enlightened my way throughout the thesis.

I would also like to thank Dr. Vincent Augusto, my co-advisor, who launched this thesis study with a great enthusiasm and offered his assistance and patience whenever they are needed.

My grateful thanks are also extended to Dr. Catherine Crenn Hebert and her team who provided us essential information and data about perinatal networks in France. I am also very grateful to the remaining members of my dissertation committee. Their academic support is greatly appreciated.

My sincere thanks go to the members of our research lab, particularly Dr. Thierry Garaix for each time being available to offer his valuable help.

Besides, there are several other people that I would like to acknowledge with much appreciation.

First, I owe special thanks to my parents for their everlasting patience, belief and all of the sacrifices that they have made on my behalf throughout my life.

I would like to offer my keen appreciation to Dr. Thomas G. Yeung, who had a big influence on me in pursuing a PhD study. I am truly thankful for his company, inspiration and contribution in opening new doors in my life.

I would like to offer my gratitude to Dr. Carlos Rodriguez, quite a friend, who has been a big help while struggling in a completely new culture, a relief in difficult moments, most importantly who sees through and understands.

I owe many thanks to Aurelie Royon, who has always been there for me to ease my life and cheer my days. Her pure presence is enough to make everything better. Additionally, many thanks to Damien Perrier, a special person, who made my day whenever he is around and provided his full support at the most needed times. Their great support and friendship were treasure.

Finally, I would like to express my deep appreciation to Nevzat Kaya for his unconditional belief in me and his patience in my long time being away. I am grateful for his presence in my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Several recent worldwide demographical and sociological changes have led to important new challenges in healthcare industry. With the democratization of health related information, population has become ever more aware and active in health related activities. Because of this increased awareness, the expectations regarding the health services have also increased. Furthermore, the recent technological and scientific advancements have led to a general increase in life expectancy which have created a population shift.

On the other side, inherit randomness, complexity of operations, highly expensive and scarce resources are most commonly known problems encountered in health care systems which makes the system more difficult to manage. Those difficulties have also given rise to a shortage of available resources. In particular, "hospital networks" have emerged as a new organizational form designed to face those challenges; specifically by better linking the different entities of the health system to surge cooperation, improve knowledge sharing, assure effective access to health and better stand against the inefficiency and inadequacy of healthcare resources ([Bravi 2012]). The main objective of this study is to develop innovative tools that take into account all players of the network and improve a network of hospitals from various perspectives.

In this thesis, by being motivated from the challenges in perinatal networks, we address design and flow control of a stochastic healthcare network where there exist multiple levels of hospitals, resources and different types of patients. Patients are supposed urgent; thus can be rejected and overflow to another facility in the same network if no service capacity is available at their arrival. Rejection of patients due to lack of service capacity is a common phenomenon in overflow networks. We approach the problem from both strategic and operational perspectives and develop state of the art methodologies in order to evaluate the performance of such a complex system and improve the efficiency (rejection rates) in network. The developed methodologies are being synthesized under the project COVER, an ongoing project stems from the collaboration with perinatal network of Nord Hauts-de-Seine, with an ultimate aim of being applied to the real perinatal network.

Scientific Challenges and Objectives

1. Improve the Design of a Stochastic Network: Regarding the continuous change, high uncertainty and variability of healthcare networks, health authorities constantly face important and difficult strategic decisions of how to design or adjust healthcare networks. In this thesis, this challenging issue is addressed via developing a decision-aid tool to determine the locations and capacities of healthcare facilities that ensures a minimum service level in the network.

3. Better Control the Patient Flow of a Stochastic Overflow Network: Strategic level perception mainly ensures enough capacity to allow patients to be treated in their preferred hospitals. However at the operational level in loss networks, a patient rejected from a facility may overflow to another facility in the same network. Therefore, a hospital may receive both its own patients and overflow patients from other hospitals. In such a healthcare network where there exist different types of arrivals, employing the dynamic admission control strategies increases flexibility in the allocation of resources among different arrival types. In this thesis, we investigate optimal patient admission control policies in order to improve the control of patient flow and the control of rejection rates in the network.

2. Evaluate the Performance of a Stochastic Overflow Network: In an overflow loss network, the most important performance measures are considered as the rejection probabilities in each hospital and overflow probabilities among hospitals. In order to evaluate the performance of a stochastic overflow network and propose to-the-point solutions regarding the bottlenecks, it is important to be able to quantify the performance measures, determine the most loaded hospitals and the highest rejected patient groups. In this thesis, we developed approximation methodologies to quantify dispatching probabilities between hospitals. Furthermore, a realistic performance evaluation is achieved via a simulation model constructed to accurately represent a perinatal network.

This thesis is composed of seven chapters. The research strategy adopted in this thesis is described in the following as represented in Figure 1.1.

Chapter 2 presents background information about healthcare networks in France, particularly on our primary application area: perinatal networks, along with discussion of current challenging issues of perinatal networks. Furthermore, a comprehensive review of recent studies on healthcare networks is provided. Finally, the global contribution of this study is stated in relation to the body of work in literature.

Chapter 3 focuses on strategic location and capacity planning of a pure-loss hierarchical network. A nonlinear location and capacity planning model is developed to minimize the total cost of changes while keeping the service level (admission rate) of all service units above some given level. Different linearization models of Erlang-loss

function and their properties are proposed. Linearization transforms the nonlinear model into compact mixed integer programs solvable to optimality with standard solver. Application to a real-life perinatal network is then presented.

Chapter 4 presents some approximation methodologies in order to evaluate the performance of an overflow-loss network considering the presence of overflows from one hospital to another. Several approximation methods are proposed for different possible overflow structures in a network; forward routing and feedback routing. Diversified numerical examples are provided to evaluate the effectiveness of the approximation methodologies.

Chapter 5 focuses on operational management of an overflow-loss network. Considering the presence of overflows, a hospital may receive both its own patients and overflow patients from other hospitals. In such systems where there exist different types of arrivals, there might also exist different priorities. In this chapter we study optimal admission control policies via Markov Decision Processes on different size networks. For smaller size networks, we provide the structural properties of optimal admission control policy. For big scale networks, we propose a near-optimal heuristic policy whose performance is evaluated by an LP model developed to compute a tight upper-bound on the problem.

Chapter 6 presents a joint "agent-based" "discrete-event-system" simulation model of a stochastic hierarchical overflow-loss perinatal network (Hauts-de-Seine). Our main objective is to represent a perinatal network accurately with an adequate detail level in order to evaluate the performance measures (rejection probabilities) and more importantly evaluate the strength of the optimal results of our analytical models considering the numerous underlying assumptions.

Finally, Chapter 7 concludes the thesis by summarizing the key results and dis-

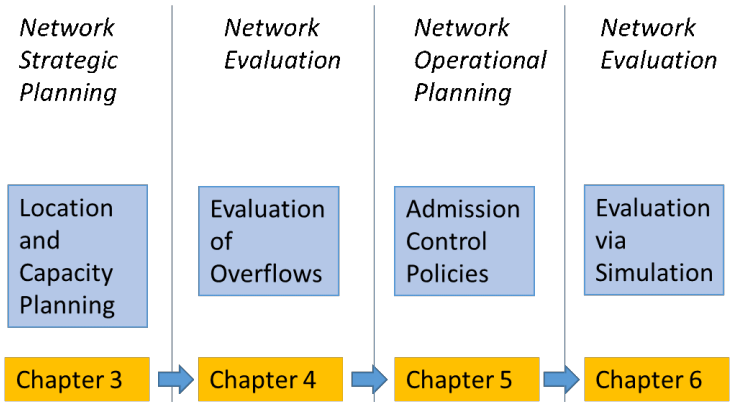

Figure 1.1: Research Strategy

cussing the final output of this thesis, the project COVER along with the prospective research directions.

Introduction

Plusieurs changements démographiques et sociologiques récents à travers le monde ont conduit à de nouveaux défis importants dans le secteur de la santé. Avec la démocratisation des technologies de l'information liées à la santé, la population en général est devenue de plus en plus consciente et active dans les activités liées à la santé. En raison de cette prise de conscience, les attentes concernant les services de santé ont également augmenté. En outre, les récents progrès technologiques et scientifiques ont conduit à une augmentation générale de l'espérance de vie, entraînant ainsi une augmentation globale de la demande en soins.

D'autre part, la complexité des traitements médicaux, les flux de patients, la mise en oeuvre de ressources coûteuses et rares sont les problèmes les plus couramment rencontrés dans les systèmes de soins, et rendent l'organisation plus difficile à gérer. Ces difficultés ont également donné lieu à une pénurie de ressources disponibles. En particulier, "les réseaux hospitaliers" ont émergé comme une nouvelle forme d'organisation concue pour relever ces défis, en particulier par une meilleure articulation entre les différents organismes du système de santé et une coopération entre professionnels de santé permettant également d'améliorer le partage des connaissances, d'assurer un accès effectif à la santé et de mieux résister aux inefficacité et l'insuffisance des ressources de soins de santé ([Bravi 2012]). Dans cette thèse, notre objectif principal est de développer des outils innovants qui prennent en compte tous les acteurs du réseau et d'améliorer l'organisation général d'un réseau d'hôpitaux sous différentes perspectives.

Cette thèse est motivée par une collaboration avec le réseau de périnatalité 92 Nord (Île de France) et vise à améliorer l'organisation global du réseau de maternité du territoire pour une meilleure prise en charge des femmes enceintes ou ayant récemment accouché. Nous nous intéressons en particulier à un réseau de santé stochastique hiérarchique sans attente où plusieurs types d'hôpitaux, des ressources multiples et différents types de patients entrent en interaction. Les patients sont considérés comme urgents et peuvent donc être rejetés vers un autre établissement du même réseau si la capacité d'accueil du service de soins est insuffisante à leur arrivée. Le rejet des patients en raison d'une capacité insuffisante est un phénomène courant dans le réseaux de soins. Nous abordons le problème sous deux angles, stratégique et opérationnel, et nous développons des méthodes innovatrices afin d'évaluer la performance d'un système aussi complexe et d'améliorer l'efficacité (taux de rejet) du réseau.

Cette thèse se décompose en sept chapitres. La stratégie de recherche adoptée est décrite ci-après et illustrée Figure 1.

Dans le chapitre 2, nous proposons une présentation générale des réseaux de soins en France, en particulier sur notre domaine d'application principal: les réseaux de

périnatalité. Une discussion est également proposée sur les problèmes récurrents liés à ce type de réseaux. En outre, un examen complet des études récentes sur les réseaux de soins est fournie. Enfin, la contribution globale de cette étude est mise en avant par rapport à l'ensemble des travaux de la littérature.

Dans le chapitre 3, nous nous concentrons sur la localisation stratégique et la planification de la capacité d'un réseau hiérarchique de soins sans attente. Un modèle de localisation non linéaire et un modèle de planification de capacité sont développés pour minimiser le coût total de changements tout en gardant le niveau de service (taux d'admission) de toutes les unités de service ci-dessus d'un certain niveau donné. Différents modèles de linéarisation de la fonction Erlang-loss et leurs propriétés sont proposés. La linéarisation transforme le modèle non-linéaire en programmes mixtes en nombres entiers que nous pouvons résoudre optimalement avec les solveurs standards. Une application concrète au réseau périnatal est également présentée.

Dans le chapitre 4, nous présentons plusieurs méthodes d'approximation pour évaluer la performance d'un réseau avec perte et débordements d'un hôpital à l'autre. Différentes structures de débordement possibles dans un réseau sont examinées. Plusieurs exemples numériques sont fournis pour évaluer l'efficacité des méthodes d'approximation.

Dans le chapitre 5, nous nous concentrons sur la gestion opérationnelle d'un réseau de soins avec perte sans attente. Compte tenu de la présence de débordements, un hôpital peut recevoir à la fois ses propres patients et les patients provenant d'autres hôpitaux. Dans ces systèmes où il existe différents types d'arrivées, il pourrait également exister des priorités différentes. Dans ce chapitre, nous étudions les politiques de contrôle d'admission optimales via les processus de décision de Markov sur des réseaux de tailles différentes. Pour les réseaux de petite taille, nous fournissons les propriétés structurelles de la politique de contrôle d'admission optimale. Pour les réseaux d'envergure, nous proposons une politique heuristique pseudo-optimale dont la performance est évaluée par un modèle de programmation linéaire développé pour calculer une borne supérieure pour ce problème.

Dans le chapitre 6, nous présentons un modèle de simulation mixte multi-agents et à événements discrets d'un réseau périnatal stochastique hiérarchique sans attente (Hauts-de-Seine). Notre objectif principal est de modéliser un réseau périnatal avec un niveau de détail suffisant pour évaluer les indicateurs de performance (probabilité de rejet, coût de prise en charge), et surtout d'évaluer la robustesse des résultats optimaux de nos modèles analytiques présentés dans les chapitres précédents, en tenant compte des nombreuses hypothèses sous-jacentes ignorées jusqu'à maintenant.

Enfin, une conclusion générale de la thèse est présentée dans le chapitre 7 avec un résumé des principaux résultats et une discussion sur les apports de cette recherche

et de son application via le projet COVER. Plusieurs perspectives de recherche futures sont également proposées.

Chapter 2

Healthcare Networks & Literature Review

Contents

2.1 Healthcare Networks in France

Healthcare networks bring together a wide range of public and private actors (hospitals, specific service units, healthcare professionals and managers) at local or regional level that seek for management effectiveness, decisional efficacy and coordination. The healthcare network system encourages and improves knowledge sharing (working experience) among healthcare professionals and clinicians. The system has also been proved to increase quality of care, the spread of innovations and the adoption of new clinical practices ([Grenier 2004]).

Healthcare networks in France are composed of multiple private and public hospitals. They provide tailored support for individuals in terms of health education, prevention, diagnosis and treatments. Preserved by law since 2002, the networks are one of the main mechanisms for coordinating actors (health, social and medico-social) involved in the patient pathway.([www.sante.gouv.fr])

In metropolitan France, there are 643 networks funded by an amount of 167 million euros by Regional Intervention Fund (FIR) representing approximately 2500 professionals. Table 2.1 presents the categories of Healthcare Networks in France.

Table 2.1: Category of Health Care Networks in France (Activity Report of FIQCS 2010)

Name of the Field	Total Number of HC Networks
Gerontology (Gérontologie)	120
Palliative care/nursing (Soins Palliatifs)	95
Diabetology (Diabétologie)	65
Addiction (Addictologie)	42
Support for Teens (Prise en charge des adolescents)	38
Cancer (Cancérologie)	37
Handicap	35
Perinatal (Périnatalité)	32
Pathology (Poly pathologie)	29
Cardiovascular (Cardiovasculaire)	23
Obesity (l'obésité)	21
Respiratory Diseases (Pathologies respiratoires)	21
Neurology (Neurologie)	18
Infectious Diseases (Maladies infectieuses)	17
Distress (Douleur)	10
Precariousness (Précarité)	9
Nephrology (Néphrologie)	6
Rare Diseases (Maladies rares)	5
Mental Health (Santé Mentale)	4
Rheumatology (Rhumatologie)	1
Others	15
TOTAL	643

The law of 21 July 2009 on the reform of the "hospital, patients, health and territories" (loi HPST, Hôpital, Patients, Santé, Territoire) created the Regional Health Agencies-ARS (Agences Régionales de Santé) as pillars to reform the health system. The Regional Health Agencies are responsible for ensuring a unified control of regional health, better meet the needs of the population and improve the efficiency of the system ([www.ars.fr]). They are also responsible for prevention activities conducted in the regions and of providing care based on the needs of the population. Furthermore, the agencies ensure a more coherent and effective approach to health policies pursued in an area, organize coordination, and allow greater fluidity of care pathway to meet the needs of patients. However, constantly changing environment and demographics, inherit randomness in health care systems (random, seasonal

arrivals of patients and service times), complexity of operations and diverse patient flows, lack of coordination and cooperation among hospitals and lack of integration between different components of the system make the regional health control quite challenging.

This thesis is stimulated of the challenges that the healthcare networks and subsequently ARS encounters (detailed in the following sections). This study is motivated notably from the "Perinatal Network of Hauts-de-Seine" in France and it is realized under collaboration with ARS of Ile-de-France and partially supported by Agence National de Recherche (ANR-11-TECS-010-04) and National Natural Science Foundation of China (No.71131005). Our ultimate objective is to provide innovative tools that take into account all players of the network and improve the system as a whole.

In the following section, we provide descriptive information about perinatal networks (perinatal care, structure of maternity facilities and patient flows) along with the specific characteristics of Perinatal Network of Hauts-de-Seine. In section 2.3 the critical points and difficulties observed in the network are summarized. Section 2.4 presents a comprehensive literature review of studies developed with an aim to improve health care delivery networks. Finally section 2.5 sets our study in relation to the body of work in literature and clarifies our global contribution to the area.

2.2 Description of Perinatal Networks and Maternity Facilities

Perinatal care consists of the period immediately before and after birth. The target patient groups are pregnant women in delivery phase and newborn infants. In perinatal networks, maternity facilities provide different services to different group of patients with different criticality levels.

Obstetrics Care Unit (OB) provides labor services to women. Childbirth is a process achieved through broadly either a caesarean-section (C-section) or a vaginal delivery. Although there might exist numerous complications and relatively different procedures, i.e., induced vaginal deliveries, roughly it can be assumed two distinctive patient flows for pregnant women in an OB. In an OB, there exist Antepartum, Labor&Delivery Rooms (LDR), Operating Rooms (OR), Post Anesthesia Care Unit (PACU) and Mother Baby Rooms (MB). Women receives vaginal delivery in a Labor&Delivery Room (LDR) whereas C-sections are conducted in OR rooms. Women recover from anesthesia after their C-sections in the Post Anesthesia Care Unit (PACU).

Neonatal units are part of perinatal networks. Different neonatal service units are established according to the level of complexity of care required.

Basic Neonatal Care Unit provides medical care to newborn infants who have mild illness expected to resolve quickly or who are recovering after intensive care. The principal patient group is commonly newborn infants with a gestational age greater than 34 weeks or weight greater than 1800 g, or with chronic lung disease needing long-term oxygen and monitoring.

Special Neonatal Care Unit provides medical care to newborn infants who are moderately ill with problems expected to resolve quickly or who are recovering after intensive care. The principal patient group is newborn infants with a gestational age of 32 weeks (or greater), or a weight of 1500 g (or greater) or with chronic lung disease needing long-term oxygen and monitoring. Resuscitation (artificial respiration, cardiac massage) and stabilization of ill infants before transfer to an appropriate care facility is also provided in this unit.

Neonatal Intensive Care Unit (NICU) provides intensive care for extremely ill newborn infants with $<$1500g and $<$32 weeks' gestation. NICUs have the capabilities to provide mechanical ventilation and advanced respiratory support, access to a full range of pediatric medical subspecialists, advanced imaging (including computed tomography, magnetic resonance imaging and echocardiography), pediatric surgical specialists and pediatric anesthesiologists (capable of surgical repair of serious cardiac malformations).

Remark 1: In this study, for simplification reasons, we consider basic and special neonatal care unit as one unit, and assume that there exist 2 type of neonatal care unit in a network; basic+special and NICU.

In a perinatal network, maternity facilities differ according to the type of neonatal service they provide. In a maternity facility, all service units interact inherently with each other as women in labor who need obstetrics services may subsequently require basic neonatal or NICU services for their newborn babies.

A total of 602 public and private maternity hospitals (of which 37% are run privately) cover the French metropolitan territory, with the majority of them being a branch of a public hospital or private clinic. In 2007, only 4% of this total was performing fewer than 300 births a year ([Baray 2012]). Maternity facilities show a nested hierarchical network structure (a system of different types of interacting facilities). In a nested hierarchy, a higher-level facility provides all the services provided by a lower-level facility. Three levels of maternity facilities can be distinguished among the country's total number of 602 facilities:

Level 1 Maternity Facilities: Small clinics composed of Obstetrics Unit (OB) (without neonatal units) where labor is held but no special neonatal service is provided. They provide service to pregnant women at low risk, defined as those

with a singleton pregnancy, not hospitalized during pregnancy, anticipating a safe delivery with birth weight to be 2.5 kg or more, a gestational age of 37 weeks or more, generally women who had no medical risk factors. Level 1 type maternity facilities account for the majority of admissions. In France, there exist 303 level-1 type facilities.

Level 2 Maternity Facilities: Middle-sized clinics contain Obstetrics Unit and basic and/or special neonatal care unit where both obstetrics and neonatal services are provided. Pregnant women at medium risk and babies born premature at 33 weeks requiring special care but not tremendously expensive treatment are accommodated in those facilities. Level 2 facilities can be classified as 2A and 2B according to the type of neonatal care provided. Level 2A maternity facilities provide basic neonatal care whereas Level 2B facility provide both special and basic care. In France, there exist 223 Level 2 type facilities.

Remark 2: Based on previous remark, since in this study we didn't consider basic and special neonatal care separately, we neither break down level 2 facilities into level 2A and 2B. Instead we work with general level 2 maternity facilities.

Level 3 Maternity Facilities: Maternity hospitals located mostly in big cities and centralized locations offer all type of services: Obstetrics, basic+special neonatal care and neonatal intensive care services. Those hospitals are specialized in monitoring pathological pregnancies (e.g. severe hypertension, diabetes). Pregnant women at medium-high risk and premature infants with gestational age smaller than 33 weeks are cared in that type of facilities. NICUs where those highly critical newborns are cared, are only located at Level 3 Maternity Hospitals. In France, there exist 76 Level 3 type hospitals.

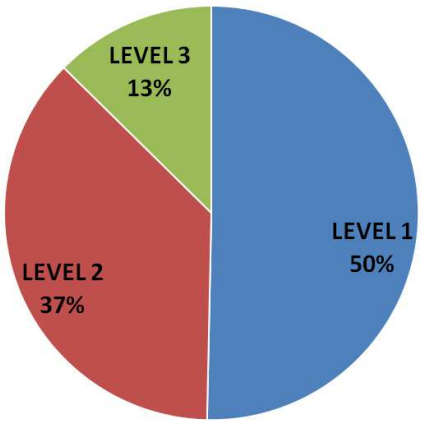

Figure 2.1: Breakdown of French maternity hospitals by type (from [Baray 2012])

Patients may have various flows in the maternity facilities according to different needs they may have. In an OB unit, there exists broadly two patient flows based on the type of delivery, vaginal (labor) or Caesarian, with significantly different service times. Patient flow can be summarized as below:

(i) Vaginal deliveries are sent to a "labor and delivery room (LDR)" to give birth, and then to "postpartum" to recover from the delivery.
(ii) Women receiving a C-section are sent to an "operating room (OR)". After C-section procedure, women recover from anesthesia in the "post anesthesia recovery unit (PACU)". Finally, they are taken to "postpartum/mother baby rooms" to recover fully before leaving the hospital. Women having C-section are expected to stay longer in the hospital compared to the women having vaginal delivery.
(iii) The newborn infants who need special care are transferred to either basic or intensive care neonatal units according to the criticality of their health status.

2.2.1 Perinatal Network of Hauts-de-Seine

Hauts-de-Seine Perinatal Network is one of the regional healthcare networks in Ile-de-France (IDF) in France. It is located in the health area 92, French administrative district (département) Hauts-de-Seine. It is the most crowded area after the central Paris in IDF (75). Hauts-de-Seine Perinatal Network is divided into two networks; Nord and Sud. In this study we mainly focus on the "Nord" part which is composed of 14 communes (lowest level of administrative division in France). The approximate number of pregnant women calculated in 2012 for Nord and Sud are 13000 and 10000 respectively. The representation of the network is presented in Figures 2.2 and 2.3.

In Hauts-de-Seine Nord there exist 6 public, 4 private maternity facilities. In this study, we focus on only public facilities due to their non-lucrative nature. It is a nested hierarchical network where there exist one type-3 (Hopital Louis Mourier), three type-2 (CMC Foch, CH Neuilly Courbevoie, Hopital Franco-Britannique) and two type-1 (Cash de Nanterre, Hopital Beaujon) public maternity facilities providing at most 3 level of services to different class of patients. Even though CMC Foch is a type 2 hospital, due to its proximity to highly populated areas, the bed occupancy is higher in this hospital compared to the type 3 regional hospital (Louis Mourier) which is located in fairly more rural part of the department. It is also important to mention that Nord Hauts-de-Seine Perinatal Network receives high number of patients originated from the neighboring departments such as central Paris (77), Hauts-de-Seine Sud (92 Sud), Seine-Saint-Denis (93), Val-d'Oise (95) and Yvelines(78).

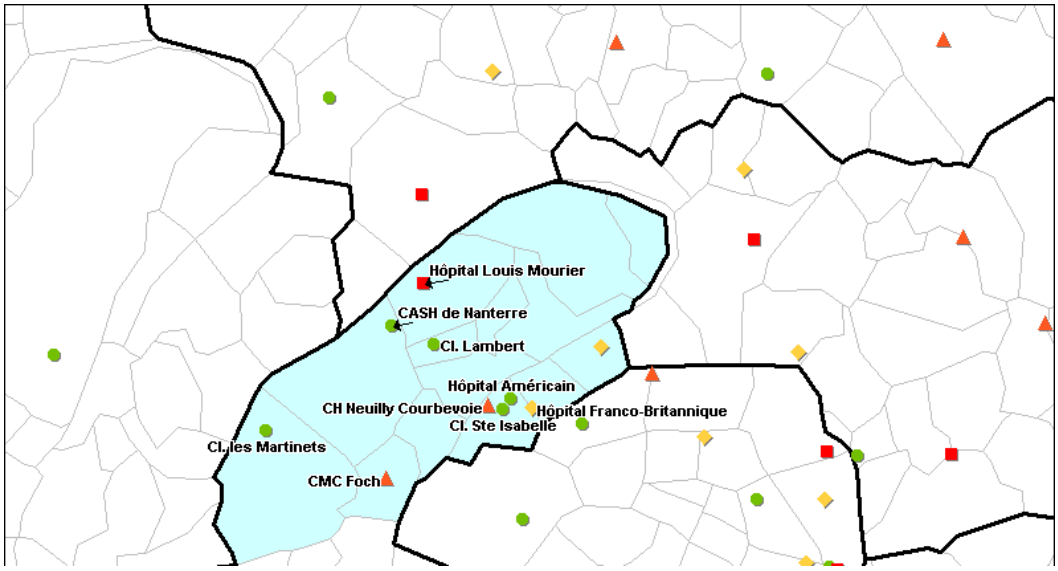

Figure 2.2: Perinatal Network Hauts-de-Seine Nord

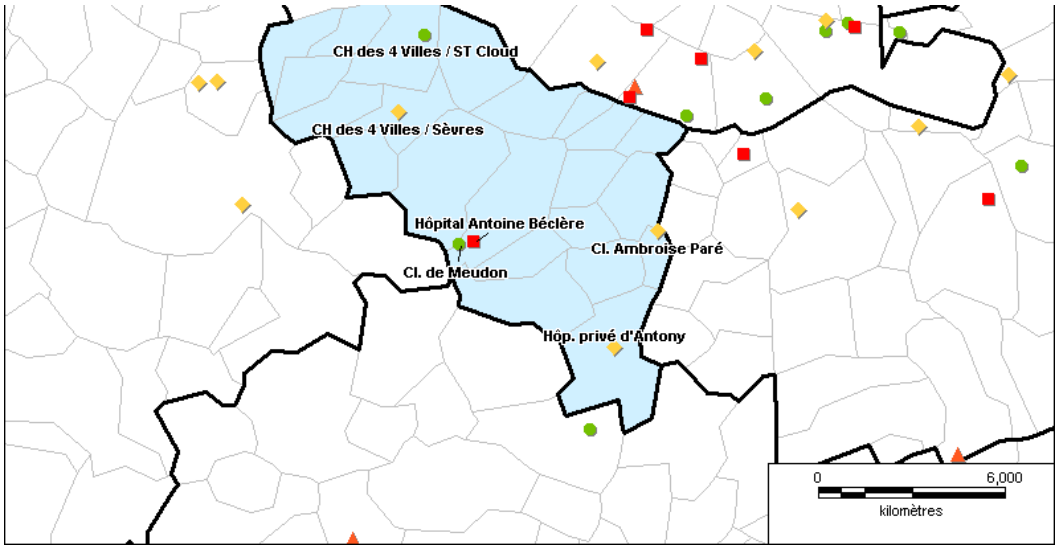

Figure 2.3: Perinatal Network Hauts-de-Seine Sud

2.3 Difficulties and Challenges in Perinatal Networks

As in many other countries, the number of maternity units in France has been reduced regularly since the 1970s ([Bréart 2003]). This trend is associated with several factors. One of the main motives of closures is the common belief that delivery in small or less specialized units with a low delivery volume is considered to be less safe (even though not a proved fact), thus the government merges resources in more centralized units by closing the small structures. Another provoking factor is the shortage of resources (a common problem in most healthcare networks) such as insufficient number of gynecologists, obstetricians and midwives, inadequate levels of equipment. Furthermore, government policies aiming to reduce the costs of health service provision can be named as another reason to closures ([Pilkington 2008]).

The closure rates of maternity facilities differ according to the region and its characteristics (urban or rural area, population demographics, migration etc.). Figure 2.4 presents the geographic variations in closure rates during 1998-2003. The closures are not uniform across France and they varies with a rate between 0-36%.

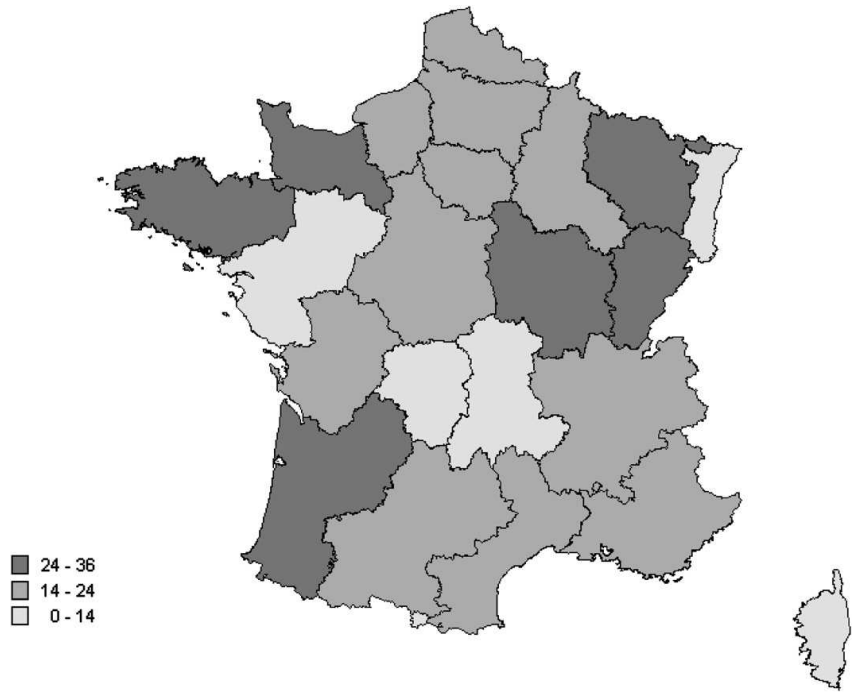

Figure 2.4: Rate of closure of maternity facilities in France by administrative regions 1998-2003 (from [Pilkington 2008])

Table 2.2 presents the existing number of health care facilities in France in years 2001-2010 (INSEE) while pointing out the different closure tendency associated with the type of facilities. It is observed that most of the hospitals that are chosen to be

Table 2.2: Change in number of HC facilities in France 2001-2010

Hospital Types	**Number of Healthcare Facilities**									
	2001	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**
Regional Hospitals (CHR/CHU)	29	29	29	29	29	29	29	29	29	29
Central Hospitals	530	523	518	515	504	499	498	490	485	498
Central Hospitals specialized on psychiatry	89	89	87	87	87	88	86	87	87	87
Local Hospitals	344	345	342	341	347	343	340	331	319	290
Other Facilities	19	22	21	22	20	18	19	17	17	14
Total Public	**1011**	**1008**	**997**	**994**	**987**	**977**	**972**	**954**	**937**	**918**

closed are local and central hospitals while the number of regional hospitals stays still. Along with hospital closures, the number of hospital beds also diminished during those years. Table 2.3 presents the amount of downsize in hospital beds at different type of hospitals. Regardless of the type of hospital, there is a tendency of downsizing in France hospitals.

Table 2.3: Change in number of hospital beds in France 2001-2010

Types of Hospitals	**NUMBER OF HOSPITAL BEDS**									
	2001	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**
Regional Hospitals (CHR/CHU)	89,107	87,947	86,723	86,179	84,337	84,639	84,771	84,991	83,143	81,108
Central Hospitals	178,290	176,092	173,426	171,537	168,649	166,397	167,451	164,731	157,567	153,031
Central Hospitals specialized on psychiatry	45,554	44,605	43,607	43,120	42,881	42,641	40,903	41,515	41,148	40,283
Local Hospitals	24,111	23,619	23,184	22,563	22,293	21,559	21,360	19,729	16,648	12,876
Other Facilities	1,486	2,045	2,140	2,069	1,944	1,798	2,066	1,807	1,843	1,562
Total public	**338,548**	**334,308**	**329,080**	**325,468**	**320,104**	**317,034**	**316,551**	**312,773**	**300,349**	**288,860**

While hospitals are being closed and hospital beds are downsized, there has been a quadratic increase of the bed occupancy rates in hospitals with the effect of rising population and increased usage of healthcare resources in time. It is studied in [Royston 2009] that hospital admission refusals (rejection rates) rise gradually with bed occupancy. Running a hospital close to 100% bed occupancy is not possible without turning patients away.

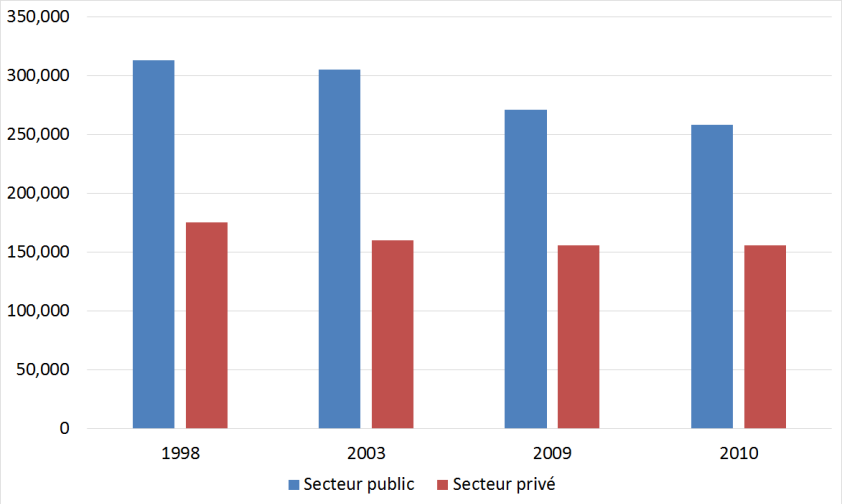

Figure 2.5: Evaluation of Number of Hospital Beds in France

Those closures or mergers improved perinatal care in terms of quality of service. However they also had negative consequences in the sense that accessibility decreased, especially in rural areas, and increased distance to hospitals limit pregnant women's access to maternity facilities and danger the life of both women and baby([Pilkington 2008]).

The biggest risk associated with closures can be defined as the reduction in the accessibility of maternity services. This risk is higher for the women from socially disadvantage backgrounds which may contribute to social inequalities ([Attar 2006]). Moreover, as in all medical specialties that involve management of life-threatening emergencies, rapid access is crucial also in perinatal and neonatal care in order to avoid accidental deliveries outside of hospitals, and avoid the risk of maternal and neonatal morbidity ([Gulliford 2002]). A recent study [Zeitlin 2004] shows that only about half of all babies hospitalized in large NICUs were born in the adjoining maternity unit. In other words, a large proportion of babies were born in lower-level maternity units in France and transferred to the NICUs by ambulance after birth ([Zeitlin 2004]). Transportation after birth is known to increase the risk of mortality and morbidity. One possible reason of this problem is the lack of evaluation of pregnancy and the risk level of baby before birth, and not being able to refer the woman to the appropriate facility before the time of delivery. The other reason can be defined as lacking of efficient resource planning and efficient distribution of scarce resources among NICUs according to the needs of the region.

In networks, lack of organization, cooperation and coordination may lead to inefficiently allocated resources and unproductive policies. In perinatal networks, considering that most of the pregnant women and newborn babies needing special

care are urgent, unavailability or inadequacy of necessary resources, inefficient distribution of capacity may lead to rejection and diversion of patients to other facilities even to other regions. This diversion may cause long travel distances which might give rise to fatal occurrences. Rejection can seriously affect the health state of the person and increase the risk of morbidity, and needed to be kept under control. Therefore, "rejection probability" is considered to be an important performance measure in this work as it is in most of the loss systems in literature. An improvement in rejection rates means better accessibility and increased effectiveness in the network.

Independent of closures and mergers, in most of developed countries neonatal care units suffer from extremely high rejection rates. In WHO report 2010, rejection rates in NICU units in Europe is approximated around 25-30%. Random arrivals, highly variable and long length of stays and insufficient number of available resources are the fundamental reasons to this crisis. Moreover, in NICUs the necessary equipment (neonatal cots) are very expensive and management wants to keep them highly utilized, which also increases the rejection probability ([Asaduzzaman 2010]). Therefore in a system where resources are very expensive thus scarce, it is very important how we manage them, how we distribute them among hospitals and which policies we employ.

"Rejection" problem described for perinatal networks are not only specific to perinatal networks but also a very common problem of most of the health care networks whose target patient group require urgent care, and cannot wait. Those problems and challenges in perinatal networks created the motivation to launch this research study. "Hauts-de-Seine Perinatal Network" is chosen as the main application field for demonstrating the results of the developed methodologies. However, the proposed methodologies are kept sufficiently general so that they can also be applied to any other stochastic hierarchical loss service networks.

2.4 Literature Review on Managing Healthcare Networks

This section reviews the body of work in operations research which is devoted to health care delivery networks from a broad perspective. The difficulty in measuring quality and value of outcomes due to the highly stochastic and complex nature of health care services, the multi-objective nature of decisions due to the existence of several decision makers (multi-hospitals, multi-departments, physicians, nurses, administrators,etc.), third party payment mechanisms for diagnoses and treatments led to the development of many different methodologies, interdisciplinary studies, quantitative and also qualitative studies focused on improving the efficiency of healthcare networks.

We scoped various topics likely to be relevant to healthcare networks to be used as part of the literature search strategy. In addition, we identified further cross cutting issues such as cooperation, coordination, integration as being potentially relevant to the study. In our final search strategy we looked for articles in the subject areas of:

- Health Care Networks
- Multi-Hospitals/Units
- Cooperation/coordination/integration
- Delivery/Logistic
- Decision support tools
- Operations Research
- Optimization
- Resource management

In literature many studies have been conducted and many tools have been developed and applied in order to improve the quality and access on health care delivery networks. The body of work can be classified in 3 major sections.

2.4.1 Strategic Location of Healthcare Facilities

One of the most crucial strategic decision that can be given in order to improve quality and efficiency of care in a network is the location of facilities. Location of healthcare delivery facilities and services mostly focuses on objectives such as ease of access and/or speed of access, as well as a need for achieving low cost. They can be applied in national (territorial), regional or local network levels.

Research on facility location field is abundant. Some main stream models (P-median models ([Hakimi 1964]) , covering models ([Schilling 1993]) , maximal covering models ([ReVelle 1986]) developed for solving facility location are frequently used in healthcare applications. Facility location problems are mostly NP-hard. Therefore heuristic models and lagrangian relaxation techniques are widely used when the algorithms built are not efficient or unable to compute an optimal solution in reasonable time, which is commonly the case in big complex multi-facility location problems. Hierarchical models capture variations in the facilities network, deploy a complex formulation and often require heuristic models to obtain a solution ([Şahin 2007]).

Location models are reviewed by many authors by being classified regarding different aspects. Pierskalla ([Pierskalla 1994]) reviewed applications of operations research in healthcare delivery by focusing on location selection and capacity planning. Daskin ([Daskin 2005]) reviewed the location problem as a critical element of strategic planning. He summarized those mainstream models and their recent variations according to static, dynamic and stochastic nature of the time component. For respecting the broad perspective of the analysis, a comprehensive review of facility location models is not presented here. In the following, we introduce the new challenges and improvements in the field by presenting some recent interesting location-allocation studies in healthcare.

Griffin ([Griffin 2008]) addressed the problem of improving the effectiveness of Community Health Centers (CHCs) which are built to provide family-oriented healthcare services for people living in rural and urban medically underserved communities. The authors developed an optimization model to determine the best location and number of new CHCs in a geographical network, what services each CHC should offer at which capacity level. In the model, the weighted demand coverage of the needy population is maximized subject to budget and capacity of each facility and service. The model application to the state Georgia demonstrated that optimizing the overall network can result in improvements of 20% in several measures such as the number of encounters, service of uninsured people, and coverage of rural counties. The proposed model is used to analyze policy questions such as how to serve the uninsured.

Siddhartha and Cote ([Syam 2010]) developed a cost minimization model to or-

ganize the network of specialized health care services such as traumatic brain injury (TBI) treatment. The model is applied to one of the Department of Veterans Affairs' integrated service networks. In the application, they analyzed the effect of some critical factors: (1) degree of centralization of services, (2) the role of patient retention as a function of distance to a treatment unit, and (3) the geographic density of the patient population. Those factors are investigated with respect to the trade-off between the cost of providing service and the need to provide such service. They showed that all three factors are useful to decision-makers in selecting locations.

Zhang ([Zhang 2010]) presented a methodology to design the network of Preventive healthcare whose aim is to reduce the frequency of life-threatening illnesses by protection and early detection. The level of participation in preventive healthcare programs is a critical determinant in terms of their effectiveness and efficiency. In order to improve accessibility and maximize participation, authors developed a bilevel nonlinear optimization model and solved the convex problem by the gradient projection method and a Tabu search algorithm. The model is used to analyze an illustrative case, a network of mammography centers in Montreal, and a number of interesting results and managerial insights are discussed, especially about capacity pooling.

Essoussi ([Essoussi 2009]) studied the case of the centralization of medical supplies in a French health care network. He analyzed the potential gains of such cooperative scheme through multi criteria optimization approach associated to an integrated inventory-location-allocation problem. A procedure of three steps is suggested to solve the whole problem.

Rodriguez et al.([Rodriguez-Verjan 2012]) studied the healthcare at home (HAH) facility location-allocation problem considering the decisions of locating facilities in a region, assignment of demand to HAH companies, authorization for delivering some specific care and allocation of resources. To tackle this problem authors proposed a multi-period, multi-resource, multi-facility location-allocation problem. Instances derived from real case study (HAH network of Rhone-Alps region) are solved using branch and bound procedure.

Locating public services for moving (nomadic) population groups is a difficult challenge as the locations of the targeted populations seasonally change. In [Ndiaye 2008], the population groups are assumed to occupy different locations according to the time of the year, i.e., winter and summer. A binary integer programming model is formulated to determine the optimal number and locations of primary health units for satisfying a seasonally varying demand. This model is successfully applied to the actual locations of 17 seasonally varying nomadic groups in the Middle East. Computational tests are performed on different versions of the model in order analyze the tradeoffs among different performance measures.

2.4.2 Capacity Planning in Healthcare Delivery Networks

In a scarce resource environment, determining necessary capacity in order to deal with the fluctuating demand in the region or optimal allocation of the existing capacity among hospitals, services, different class of patients, etc. is clearly significant for dealing with healthcare networks.

In a health care network structure, capacity planning models usually focuses on decisions such as determining the number of resources (beds, nurses, staffing levels, staff skill mix, etc.) necessary to meet a certain demand ([Ahmed 2009], [Lavieri 2009]), resource allocation for a particular service within a multi-hospital network for control of specific diseases ([Earnshaw 2002]), infectious diseases and immunization programs ([Brandeau 2003]), resource allocation among different levels of the health services ([Santibáñez 2009]), different disease types in a hospital network ([Govind 2008], [Flessa 2000]), resource/budget allocation among geographic areas ([Horev 2004]); different regions of a country, urban versus rural areas, or developing countries ([Flessa 2000]).

The most fundamental measure of hospital capacity is the number of inpatient beds. Recently Govind ([Govind 2008]) examined a network of hospitals in a predefined geographical area to determine the bed capacity that each hospital in the network should devote to different disease classes to maximize speed of access to care. They considered the spatio-temporal pattern of disease incidence in the area of focus. Their model incorporated the driving distance, rather than crow-flight distance, to a healthcare facility, as well as driving speeds on different types of roads in determining the speed of access. Santibanez ([Santibáñez 2009]) developed a multi-period mathematical programming model to provide options for configuring the system, specifically the location of clinical services and allocation of bed capacity across the services in hospitals. The decisions in the model are based on population access, critical mass standards, and clinical adjacencies. They applied their model on a real large-scale network with 12 hospitals with multiple services.

The other major component of hospital capacity is workforce, particularly nurses. Nurses are the chief caregivers and have a significant impact on clinical outcomes ([Aiken 2002]). In addition, nursing costs consist of a big portion of hospital budgets. There have been many articles on the use of optimization models to determine nurse staffing ([Kwak 1997], [Green 2002]). In terms of strategic workforce planning, Lavieri ([Lavieri 2009]) proposed a linear programming hierarchical planning model that determines the optimal number of nurses to train, promote to management and recruit over a 20 year planning horizon to achieve specified workforce levels.

Another important element of hospital capacity is planning extremely expensive machines such as the Magnetic Resonance Imaging Devices (MRIs). For such specialized services 100% utilization is tried to be achieved in the operating policies.

Very recently Mahar et al. ([Mahar 2011]) addressed the problem of allocating such specialized medical services in a multi-hospital network. The authors developed a nonlinear optimization model to determine where to locate specialized capacity across a capacitated hospital network taking into account both cost and patient service levels. The model minimizes total cost of fixed operating and variable cost of maintaining each specialized capacity, penalty cost of unmet demand at a location, and transportation cost of diverting demand between facilities while also enforcing a target level of patient service. Their results showed that aggregating demand across hospitals and allocating specialized services to a subset of hospitals rather than all hospitals in the system can yield cost savings due to enabling reduction of the necessary number of MRI machines, however it could also lead to increased travel distances for patients.

From an analytical perspective, the capacity planning models involve complex dynamics therefore mathematical optimization models are used quite frequently where the models attempt to explicitly represent the functioning of the system, resulting in large linear and integer models with many variables and constraints. Simulation models are also widely used in capacity planning in order to gain insights to guide strategies and decisions. Simulation is quite useful to mimic the behavior of a health care network in order to evaluate its performance and analyze the outcome of different scenarios. A combined use of simulation and optimization is also proposed in [De Angelis 2003].

2.4.3 Decision Support Tools for managing Healthcare Delivery Networks

The quantitative models developed with operational research tools are strong but challenging to be understood and used by decision makers. Due to the complexity of healthcare operations, distinctiveness of patient pathways and differences among structures, it is challenging to come up with some generic models which can be used by all players in the healthcare network; though necessary in order to standardize the operations, improve information flow and cooperation, and increase the quality of service; moreover critical in public-health emergencies (bioterrorist attacks, disease outbreaks, pandemics, etc.) where the decisions should be given urgently or frequently.

There have been some efforts in literature to develop electronic web-based integrated decision support tools aiming to help decision makers, managers, health professionals on operational or medical decision making, standardizing operations and managing information flow among different units in healthcare systems. Some of them focus on managing integrated care delivery networks that are developed to be implemented in multiple sites/hospitals to facilitate information flow and strategic decision making.

Bhargava et al. ([Bhargava 1999]) classified Decision Support Systems (DSS) broadly according to their specialized areas. Data driven DSS organize and analyze large volumes of data using database queries and online analytical processing (OLAP) techniques. Model-driven DSS represent decision models through optimization, stochastic modeling, simulation, statistics and logic modeling and provide

Table 2.4: Decision Support Tools in existing Literature

Authors/ Developers	Pub.Date	Name of the Tool	Purpose	Field	Methodology
Eva Lee et al.	2013	**RealOpt**	Strategic, Operational Planning	Disaster Management	Mathematical programming, Simulation, Graph-drawing tools
Shuurman et al.	2008	**wGUI**	Location-Allocation	Managing Rural Areas	-
Stein et al.	2012	**AsiaFluCap**	Strategic Planning	Pandemic Management	Simulation
Montreuil et al.	2010	**Agent-based Framework**	Strategic Planning	Chronic Diseases	Agent-based Simulation
Cunnich et al.	2011	**ALProst**	Decision Making	Screening Chronic Diseases	MCDA
Javitt et al.	1994	**PROPHET**	Decision Making	Screening Chronic Diseases	Decision trees, Markov processes, Monte-Carlo Simulation
Roth et al.	2009	**PrepLink**	Communication (Information System)	Disaster Management	Documents, Inventory, GIS
Dr Mike Stein	2006	**The Map of Medicine**	knowledge management	Decision Making	-

analysis support. Communication driven DSS link multiple decision makers over space and time. Knowledge driven DSS assist decision makers in the selection of alternatives, such as medical expert systems and scenario generators. Document driven DSS integrate a variety of storage and processing technologies to provide document retrieval and analysis.

For University of Pittsburgh Medical Center in partnership with the Pennsylvania National Guard, Roth et al. ([Roth 2009]) developed a comprehensive web-based healthcare-related electronic disaster management system called **PrepLink** which can be deployed in multiple sites in a region or more broadly. With PrepLink, emergency and healthcare workers can access timely information related to public health, safety, planning, preparation, and can communicate rapidly, share team plans across disparate locations through password-protected private pages. The system stores multiple key documents and contains asset inventories, a GIS, patient tracking, and a command-and-control module.

Recently, for tactical and strategic operational planning of a catastrophe including biological, chemical, radiological incidents, natural disasters or a disease outbreak, Eva Lee et al. ([Lee 2013]) designed and implemented a decision support tool called **RealOpt**. It combines mathematical modeling, large-scale simulation and linked them with automatic graph-drawing tools and a user-friendly interface. Operational research technology is integrated into a powerful decision support and data management tool. For this work, an OR team worked together with public health experts at the US Centers for Disease Control and Prevention (CDC) in order to address the fundamental challenges:

- distribution of medical supply
- determining locations of dispensing facilities
- optimal facility staffing and resource allocation
- routing of the population
- dispensing methods

RealOpt has been used by over 6,500 public health and emergency directors in US covering all states, plus many international users. RealOpt has been applied in hundreds of trainings and vaccination events, including anthrax preparedness, as well as seasonal flu and H1N1 vaccination events. CDC experts stated that "RealOpt is the first system that looks at the design of strategic planning and operational response on the ground. It gives policy makers a tool to assess their capabilities for handling large-scale medical emergencies and how they might handle scenarios ranging from local public health emergencies to a situation of national magnitude" (Media Newswire 2008).

For disaster management, several simulation tools have been developed and made publicly available. Recently, Stein et al. ([Stein 2012]) developed a user-friendly,

comprehensive, flexible resource modeling tool, the **AsiaFluCap** Simulator, to support decision makers involved in a pandemic management to test and compare different pandemic scenarios and assess their impact on health care resource capacity. The tool consists of an epidemiological model combined with a resource model containing 28 health care resources, a graphical user interface, and a link to export simulation results to GIS software for creating illustrative maps for geographic analysis of distribution of resources. The tool is developed in Microsoft Excel Â© (Microsoft Corporation, Redmond, WA) and using the programming language VBA (Visual Basic for Applications). The simulator is freely available at www.cdprg.org and especially useful for developing countries where resources are limited and management is needed on reallocation of regional resources.

Shuurman et.al. ([Schuurman 2008]) created a web-based graphical user interface (**wGUI**) in order to improve evidence-based decision making of health policy makers about service reallocations and hospital closures in rural areas in Canadian province of British Columbia. It provides information about the geographical area around a service, the total number of people in each existing or hypothetical service area as well as percentages of the population not served within the specified road travel time. The tool supports decisions about hospital closures, openings and service reallocations in rural and remote regions based primarily on the criterion of serving the largest possible population within certain (maximum acceptable) travel time.

Charfeddine et al. ([Charfeddine 2010]) proposed a framework for integrated agent-simulation modeling of the population with specific chronic disease (chronic obstructive pulmonary disease (COPD)) in a large region and the network (healthcare delivery network in Quebec, Canada). There are two main differentiating facets of the integrated model they proposed: characterization of population demand and modeling healthcare delivery network. Demand for healthcare services is expressed deeply through the stochastic modeling of the health state evolution of each person (represented as an agent) in a population of potential and actual patients, where the implications of this evolution generates the demand in terms of patient needs for healthcare and their frequency. In modeling healthcare delivery network part, the organization and functioning of the healthcare delivery network is captured with an adequate detail level. Objective is to assess the current organization of networks and the impact of possible changes.

Réseau Santé is a live, open, voluntary and collaborative mapping tool of healthcare delivery networks in the Quebec province (Canada) proposed by [Montreuil 2011]. Such platform provides (i) organizational mapping (overview of actual organizations, identification of organizational issues), (ii) epidemiological mapping (helping to identify and explain the relations between viral or bacterial communities, their human hosts), and (iii) logistical mapping (highlighting patient and resource flows through the network, providing an overview of actual flow patterns, identifying

Figure 2.6: Platform "Réseau Santé"

and analyzing constraints and logistical issues). Target users are healthcare professionals, managers of healthcare structures, patients and public, and researchers. The platform is available and provides a wide range of information to the public.

For managing chronic diseases, Cunnich et al. ([Cunich 2011]) introduced a new decision-support tool for specifically prostate cancer screening (**ALProst**) which is grounded in multi-criteria decision analysis (MCDA) framework. With interviews, individual weights for different criteria are set and performance output of each decision on each criterion is analyzed. For each decision option, an expected value algorithm calculates a score. Given the uncertainty and tradeoffs between benefits and harms for prostate cancer screening, this tool seeks for a risk reduction in prostate cancer-specific mortality through helping general practitioners to consider multiple factors while giving complex decisions.

Javitt et al. ([Javitt 1994]) used a computer-modeling system, called **PROPHET** (PROspective Population Health Event Tabulation), which analyzes events and costs incurred during the lifetime course of irreversible chronic diseases such as diabetes. They proposed a cohort model which combines features of decision trees, Markov processes and Monte Carlo simulation techniques. The model describes individuals whose progress is updated in two-monthly time steps; disease progression and mortality rates depend on age and disease severity. It uses data from several databases and studies and incorporates them.

In England, Dr. Mike Stein ([Stein 2006]) created an innovative communication tool, **The Map of Medicine**, for improving the knowledge management and improve quality in clinics network. Map of Medicine provides access to detailed clinical information for proper monitoring of clinical pathway of patients and the recovery of the most detailed information instantly as needed. It also facilitates the standard-

ization of practices, minimization of risk and variation in the Treatment of patients across the health system, efficient use of resources and resource planning across a healthcare system. Being deployed within the Connecting for Health program, it is the largest healthcare IT program in the world. The Map of Medicine is currently utilized in 79 National Health Service (NHS) organizations in England.

2.5 Positioning and Scientific Contribution of this study in service network literature

In this thesis, various fundamental challenges in a stochastic overflow-loss hierarchical network are addressed. We focused on problems from both strategic and operational levels. In chapter 3, a strategic facility location and capacity problem is addressed. We developed a nonlinear location & capacity planning model that captures the stochasticity in each hospital. Linearization of this model is achieved in such a way that its LP version becomes capable to solve big scale territorial problems in a reasonable time which is a practice rarely achieved in location planning literature. After setting the necessary capacity in chapter 3 without considering overflows, in chapter 4 we developed new approximation methodologies to evaluate the performance of the existing network in terms of the rejection and overflow probabilities between hospitals. In chapter 5, we proposed optimal admission control policies for different size of networks in order to better utilize the scarce resources in hospitals such that total profit is maximized. Finally, we generated a comprehensive simulation model that allows us to examine the proposed analytical models and evaluate the performance of the system more realistically.

In parallel to those research studies, we aimed to combine all methodologies developed in this thesis in a decision support tool foreseen under the project named "COVER". The main objective of COVER project consists in proposing a collaborative decision aid platform in order to assist health system managers to effectively plan strategic and operational decisions of a healthcare network and evaluate the performance of their decisions. This project was founded by Saint-Etienne Métropole in 2010 to allow the development of a web platform and its connection with decision aid and simulation tools developed in this thesis. In this way, the COVER platform allows the health care professionals and hospital managers to access scientific tools designed to improve the efficiency of the health care network that they are working in.

To summarize, we approached the problems of a "stochastic overflow-loss service network" from various possible perspectives. We developed state-of-the-art methodologies by merging operational research tools such as queuing theory, mathematical programming, markov decision processes, agent-based and discrete-event-simulation that have been mostly used disjointedly in the literature. To the best of our knowledge, managing overflow-loss hierarchical networks considering dependencies, cooperation, as well as overflows among units has not been considered in the existing literature.

2.6 Résumé du chapitre

Dans ce chapitre, nous proposons une revue de littérature sur les réseaux périnatals (périnatalité, la structure des services de maternité et les flux de patients) ainsi que les caractéristiques spécifiques du réseau périnatal des Hauts-de-Seine. Ensuite, les points critiques et les difficultés observées dans le réseau sont résumées. La Section 2.4 présente une revue exhaustive de la littérature d'études développés dans le but d'améliorer les réseaux de prestation de soins de santé. Enfin, la section 2.5 définit notre étude par rapport à l'ensemble des travaux dans la littérature et clarifie notre contribution globale à la région.

Dans de nombreux pays, dont la France, les réseaux périnatals ont subi de profonds changements concernant la centralisation des naissances dans les grandes unités [Pilkington 2008]. De nombreux services de maternité avec moins de naissances par an sont fermés et beaucoup d'autres ont opté pour une fermeture de leur bloc opératoire pour éviter les problèmes de qualité et de sécurité dus à un faible volume d'activités de chirurgie. Afin d'assurer la qualité et la sécurité des soins, il est recommandé dans [Vallancien 2006] de suivre des consignes de sécurité pour fermer les petits services, et notamment un seuil d'activités (2000 chirurgies par an pour une salle d'opération et 500 naissances par an pour un service de maternité). En conséquence, le nombre de services de maternité en France passe de 1369 en 1975, à 1010 en 1985, 814 en 1996, 779 en 1997, 576 en 2007, 520 en 2011 [IGAS 2013]. Le nombre de lits d'obstétrique est divisé par 2. De 1997 à 2008, on comptait 196 services de maternité fermés et 568 opérations de réorganisation des services de maternité (fusion, la réduction des effectifs). Cependant, une telle échelle de réduire les effectifs ont donné lieu à de nombreux problèmes tels que le manque d'accès aux soins obstétriques [Nesbitt 1997], les grossesses faiblement suivies principalement dans les régions rurales du pays, l'augmentation des taux de natalité à l'hôpital, et le risque finalement plus élevé de mortalité néonatale que la mortalité et la morbidité [Blondel 2011]. La réduction des effectifs des unités de service et ses effets associés est encore un débat en cours aujourd'hui. Les chiffres ci-dessus montrent que les autorités sanitaires françaises font constamment face à des décisions stratégiques et opérationnelles importantes et difficiles sur la façon de régler et de gérer les réseaux de périnatalité.

Dans les réseaux, le manque d'organisation, la coopération et la coordination peut conduire à une utilisation sous-optimale des ressources allouées et à des politiques improductives. Dans les réseaux de périnatalité, étant donné que la plupart des femmes enceintes et des nouveau-nés nécessitant des soins spéciaux sont urgents, l'absence ou l'insuffisance des ressources nécessaires, la répartition inefficace des capacités peut entraîner le rejet et le détournement des patients vers d'autres établissements, même dans d'autres régions. Cette déviation peut causer de longues distances à parcourir et pourrait donner lieu à des accidents mortels. Le rejet peut sérieusement affecter l'état de santé de la personne et accroître le risque de morbidité

qui doit être maintenu sous contrôle. Par conséquent, "la probabilité de rejet" est considéré comme une mesure de performance importante dans ce travail car il est utilisé dans la plupart des systèmes sans attente de la littérature. Une amélioration du taux de rejet signifie une meilleure accessibilité et une plus grande efficacité dans le réseau.

Dans cette thèse, divers défis fondamentaux dans un réseau hiérarchisé stochastique sans attente sont abordés. Nous nous concentrons sur les problèmes de niveau stratégique et opérationnel. Dans le chapitre 3, la partie stratégique, nous avons développé un modèle de planification de capacité & de localisation non linéaire qui tient compte des phénomènes aléatoires dans chaque hôpital. La linéarisation de ce modèle est réalisé afin d'obtenir une version linéaire capable de résoudre des problèmes de grande taille (échelle d'un département) en temps raisonnable. Dans le chapitre 4 nous avons développé de nouvelles méthodes d'approximation pour évaluer les performances du réseau existant en termes de rejet et de transferts entre les hôpitaux. Dans le chapitre 5, nous avons proposé des politiques de contrôle d'admission optimales pour différentes tailles de réseaux afin de mieux utiliser les ressources limitées dans les hôpitaux au moyen du profit total qui doit être maximisé. Finalement, nous générons un modèle de simulation complet qui nous permet de tester des modèles analytiques proposés et d'évaluer la performance du système de façon plus réaliste.

Parallèlement à ces études, nous avons cherché à combiner toutes les méthodes développées dans cette thèse dans un outil d'aide à la décision prévu dans le cadre du projet intitulé "COVER". L'objectif principal du projet COVER consiste à proposer une plate-forme collaborative d'aide à la décision afin d' aider les gestionnaires du système de santé afin de planifier efficacement les décisions stratégiques et opérationnelles d'un réseau de soins de santé et d'évaluer la performance de leurs décisions. Ce projet a été financé par Saint-Étienne MÉtropolé en 2010 pour permettre le développement d'une plateforme web et son lien avec les outils d'aide à la décision et les outils de simulation développés dans cette thèse. De cette façon, la plate-forme COVER permet aux professionnels des soins de santé et les gestionnaires d'hôpitaux d'accéder aux outils scientifiques visant à améliorer l'efficacité du réseau de soins de santé.

Pour résumer, nous avons abordé les problèmes d'un "réseau de service débordement perte stochastique" à partir de différents points de vue possibles. Nous avons développé des méthodologies state-of-the-art de la fusion des outils de recherche opérationnelle comme la théorie des files d'attente, la programmation mathématique, les processus de décision de Markov, la simulation multi-agents et à événements discrets, le plus souvent utilisé de manière disjointe dans la littérature. D'après état de l'art, la gestion des rejets dans les réseaux sans attentes stochastiques hiérarchiques avec dépendances et coopération n'ont pas été pris en compte dans la littérature existante.

Chapter 3

Dynamic Location and Capacity Planning

Contents

3.1 Introduction

In many countries including France, perinatal networks have undergone various dramatic changes regarding centralization of births in larger units [Pilkington 2008]. Many maternity services with fewer child-births per year are closed and many others have their operating room closed to avoid quality and safety problems due to small volume of surgery activities. In order to ensure the quality and safety of care, it is recommended in [Vallancien 2006] to close small services below some safety-threshold of activities (2000 surgeries per year for an operating theatre and 500 births for a maternity service). As a result, the number of maternity services in France changes from 1369 in 1975, to 1010 in 1985, 814 in 1996, 779 in 1997, 576 in 2007, 520 in 2011 [IGAS 2013]. The number of obstetric beds is divided by 2. From

1997 to 2008, there were 196 maternity services closed and 568 maternity service reorganization operations (merging, downsizing). However, such scale of downsize gave rise to many problems such as poor access to obstetric care [Nesbitt 1997], weakly monitored pregnancies mainly in the rural parts of the country, increase of birth rate out of hospital, and eventually higher risk of neonatal mortality and morbidity [Blondel 2011]. Downsizing service units and its associated effects is still an ongoing discussion today.

The above numbers show that French health authorities constantly face important and difficult strategic decisions of how to adjust the perinatal network, i.e. location of maternity services and capacity/demand allocation. This work aims at developing a decision-aid tool to match locations and capacity of maternity services with demographic changes.

Given the continuous change, high uncertainty and variability of perinatal networks [Harper 2002], location and capacity planning are essential but challenging activities that need to be dealt with from many different perspectives under a rigorous analysis. For example, in health care networks, conventionally there is a strong relationship among the hospitals located in the same network in terms of resource transfer and distribution of the demand. Nevertheless, this is often ignored in most cases as capacity decisions are taken independently of the other hospitals in the network. In addition to that, health care networks often face random arrivals and stochastic service times which make the network a stochastic system and capacity planning a delicate and complex issue.

Nevertheless, most health authorities are using methods based on ratios and target bed occupancy rates, often using the same target occupancy rates for different sized units. However, these methods fail to consider the variability in hospitalization demands over time [Nguyen 2005]. A preliminary work [Pehlivan 2012] has been developed for planning capacity in a stochastic hierarchical network of hospitals where we studied the ways to determine capacity planning without changing the existing infrastructure in the network. However, capacity decisions in network are long-term strategic decisions which should be considered along with the network design decisions. During a long planning horizon, network may require new facilities or it may be necessary to close some others due to demographic changes. With that perspective, in this paper we are motivated to extend our preliminary work in order to incorporate network planning decisions such as opening new/closing existing service units and eventually hospitals.

This paper proposes a general framework of nested hierarchical service networks to capture the specific features of perinatal networks. In a nested hierarchy, a higher-level facility provides all the services provided by a lower-level facility. A customer requiring a certain level of service will additionally require all lower-level services. Customers arrive randomly from different demand zones and are assigned to differ-

ent facilities to be served. Customers are supposed urgent and are lost if no service capacity is available at their arrival. Rejection of customers due to the lack of service capacity is the common phenomenon in overflow networks. Pure loss queues are used and each service unit in a facility is modeled as either an M/G/c/c or M/M/c/c pure loss queuing system from which the analytical relationship between service capacity and customer acceptance rate is obtained by Erlang Loss formula.

We propose a multi-period, multi-facility, multi-service mathematical optimization model for the periodic service unit location and capacity decisions in order to ensure a minimum desired customer acceptance rate in each service. Service unit location and service capacity decisions are adjusted dynamically to match the forecasted demand changes over space and over time. The mathematical model includes service level constraints expressed in nonlinear Erlang loss formula. Various linearization approaches of the nonlinear constraints are proposed using a pointwise representation, maximum offered load functions and minimum service capacity functions. Linear regressions of these latter functions are used for efficient approximation of large-scale problems. Structural properties of these linearization models are proved.

The above mathematical models are applied to the perinatal network in Hauts-de-Seine in Paris, France. The mathematical models and sensitivity analysis show how to match demographic changes by adjusting neonatal services of each hospital and their staffed-beds.

The original contributions of the paper can be summarized as follows: (i) a new nested hierarchical service network model covering our real perinatal network as a special case, (ii) a mathematical programming model for service unit location and capacity planning under service level constraints, (iii) new linearization approaches and their properties of the Erlang-loss function which defines the service level, (iv) application to a real-life perinatal network to best meet demographic changes.

The remainder of this paper is organized as follows. Section 3.2 reviews the related capacity planning and facility location literature and defines our contribution. Section 3.3 formally defines hierarchical service networks. Section 3.4 proposes a mathematical model for service unit location and service capacity planning. Section 3.5 proposes different approaches for linearization of the customer acceptance rate constraints and establishes their properties. Section 3.6 presents the application to the perinatal network. Finally, conclusion and future research directions are given in Section 3.7.

3.2 Literature review

This section is a review of service facility location and capacity planning models most relevant to our study. As this chapter is mainly motivated by healthcare applications, a special attention is given to the models in healthcare literature.

Within the area of facility location, the problem of maximizing the use of or access to service networks has been often addressed by considering travel distances or capacity utilization. Verter and Lapierre [Verter 2002] examined the optimal number and locations of preventative health facilities by using the maximal covering location problem to maximize the number of potential patients who can utilize the facilities. Griffin et al. [Griffin 2008] investigated a variation of the maximal covering location problem whereby the location and services offered at each location are determined. Stummer et al. [Stummer 2004] determined the location and size of medical departments. Galvao et al. [Galvão 2006] studied hierarchical location of perinatal facilities in municipality of Rio de Janeiro by incorporating capacity constraints. Tanonkou et al. [Tanonkou 2008] proposed a Lagrangian relaxation approach for a stochastic distribution network with random demand and random supply lead times. The problem consists in determining distribution center (DC) opening/closing and customer zones to DC assignment in order to minimize the total cost related to DC opening, transportation, running inventory and safety stocks. Recently Mahar et al. [Mahar 2011] developed a nonlinear optimization model to determine where to locate specialized capacity across a medical network's capacitated hospitals and take into account both cost and patient service levels. For a review of service facility location planning, the reader is referred to [Daskin 2005]. Facility location models are big size models working on regional level where it is common to develop heuristics in order to solve the problems in a reasonable amount of time. In our work we were able to introduce some modifications in the model and solve the model to optimality in a reasonable time.

Capacity decisions in a healthcare setting can be considered at various levels: Unit ([Ridge 1998, Kim 1999, Gorunescu 2002, Asaduzzaman 2010]), Hospital ([Harper 2002, Cochran 2006], [Li 2008]) and Regional Network ([Flessa 2000, Stummer 2004, Govind 2008, Santibáñez 2009, Günes 2009, Syam 2010]). In this work, we focus on location and capacity decisions for regional network, by considering more uncertainty than previous models. There have been some works taking into account stochastic variability in long-term demographics [Stummer 2004, Santibáñez 2009]. We consider this long-term uncertainty as well as daily stochastic arrivals and service times at the unit level. Furthermore, existing regional models primarily focus on location of facilities and allocation of monetary resources whereas we give more detailed operational capacity decisions such as staffed-beds.

There are various methodologies used in the literature in order to optimize these capacity planning decisions. At the unit/hospital level, the most common method is

simulation-based optimization ([Kao 1981, Ridge 1998, Kim 1999, Harper 2002]) due to its ability to capture the complex nature of the system. However, a tremendous amount of data is required to create a meaningful simulation. This data generally is not obtainable for a multi-facility environment.

The other most commonly used method is queuing theory which is particularly useful for modeling patient flows and determining the minimum capacity requirements in stochastic systems. It also allows capturing the stochastic nature of arrivals and service times that is typical in healthcare systems [Bretthauer 1998]. There are several works that use queuing theory, but all consider only either the unit or hospital level. They also focus only on operational aspects and do not consider strategic long-term decisions such as changing demographics through the years as we do. In one early work combining queuing and optimization, Bretthauer et al. [Bretthauer 1998] proposed a methodology to model a blood bank and an outpatient clinic as a network of queuing stations. An optimization/queuing framework is used to minimize the capacity cost required to achieve a target level of customer service (waiting time). The resulting nonlinear model can only be solved using branch and bound and other approximation techniques. In this paper, we use the refused admission probability as the service level indicator and we are able to linearize our problem to a mixed integer program that can be solved efficiently. Gorunescu et al. [Gorunescu 2002] introduced a queuing model (M/PH/c) where the mean bed occupancy and lost patient probabilities are obtained and used to determine the number of beds required to achieve a target acceptance probability. We use the same derivation for lost patient probabilities, but we use a mathematical optimization model to optimize the number of staffed-beds in the network whereas Gorunescu et al. [Gorunescu 2002] focused on the bed capacity of a single hospital. Li et al. [Li 2008] integrated the model of Gorunescu et al. [Gorunescu 2002] with a multi-objective bed allocation model. The results from queuing model are used to construct a multi-objective model within a goal programming framework that takes into account the resource conflicts between departments of a hospital.

There are several other queuing models in health care; their focus is on the analysis of the queuing model itself considering special properties such as overflow networks or the blocking property. Asaduzzaman et al. [Asaduzzaman 2010] proposed a queuing model to determine the number of beds at all care units for any desired overflow and rejection probability in a neonatal unit, but does not consider the interaction with obstetrics units. Cochran and Bharti [Cochran 2006] combined queuing theory and simulation to balance bed utilization across an obstetrics hospital and minimize the blocking of beds. They considered obstetrics and neonatal unit within a hospital and passage probabilities between them but, did not consider the interaction between the maternity facilities at the network level. We believe that these interactions between the units and among the hospitals within the network should not be ignored. Therefore, we focus on the dependency between service units in a hospital and also the interaction among hospitals within a network structure.

3.3 Hierarchical Service Networks

Consider a hierarchical service network offering a set of $\overline{S}$ **services** to **customers** within specialized service units located in a set $I = \{1, 2, ..., \overline{I}\}$ of **facilities**. The set S of services is indexed by integer s with $s \in \{1, 2, ..., \overline{S}\}$. Service s is said to be of higher level than another service s' with $s' < s$. For simplicity, we will call a specialized unit for service s in a facility i the service s of facility i and denote it as (i, s). Customers are all from a given region partitioned into some demand zones to be precised later.

Assumption 1: There are $\overline{S}$ types of customers with nested hierarchy defined as follows. A customer requiring service s also requires all lower-level services. A customer of type-s is a customer requiring services $1, 2, \ldots, s$. Hence a type-1 customer requires only service 1 while a type-$\overline{S}$ customer requires all services. Further, there is a one-to-one correspondence between the set of services and the set of customer types and the same index will be used for simplicity.

Assumption 2: All services of a customer are met by service units of the same facility. As a result, a type-s customer can only be assigned to facilities with all service units s' with $s' = 1, 2, \ldots, s$. This assumption also implicitly assumes the nested hierarchy of facilities.

Remark 1: In our perinatal application, facilities are hospitals, service units are obstetric units (OB), neonatal units and neonatal intensive care units (NICU). Customers are pregnant women where the newborns are considered as the extension of woman. Services include obstetric care ($s = 1$), neonatal care ($s = 2$) and neonatal intensive care ($s = 3$). A type-3 woman needs type-1 obstetric care for herself and both type-2 and type-3 neonatal cares for her baby. Commonly a newborn treated in NICU goes through neonatal unit as well after an initial critical phase. Assigning each woman and her baby to the same hospital is a reasonable assumption. The strict hierarchy among different services can be replaced by a less restrictive assumption of a customer-dependent subset of services for each customer. The model can be easily extended and the results still hold.

Remark 2: Assumption 1 also implies that one pregnant woman gives birth to one baby. Although it holds for most pregnant women, it is just an approximation of the real case and extension to multiple births needs to bulk arrivals that are not considered in this paper.

Assumption 3: Each service unit s of a facility i has a number c_{is} of identical servers called service capacity. It can be modified over time by opening/closing of the unit, by purchasing additional servers, by downsizing unneeded servers or by transferring some servers to another facility.

Remark 3: For the strategic capacity planning of our perinatal network, rather than determining the optimum number of each type of resources, we utilize the concept of **staffed-beds**. Pre-defined ratios among resources are usually required in healthcare. A staffed-bed is the combination of physical equipment (bed) with an appropriate coverage of nurses and physicians [Dexter 2001]. In our applications, typical ratios are 5 beds per nurse in an obstetric unit and 2 beds per nurse in NICU.

Remark 4: The location and capacity planning decisions (closing/opening, transfer, etc.) are realistic in our perinatal application. As mentioned in Section I, there have been regularly maternity services closed or restructured in France since 1975. Most often, staff personnel are not fired but transferred as many French maternity services are public. While some regions in France have their population reduced, other regions have seen significant increase of their population and opening or upgrading maternity services becomes necessary to ensure equal access to cares of the entire French population.

Assumption 4: Each facility i has some existing and potential service units s. Each existing service unit (i, s) can be closed during the planning horizon but cannot be reopened. Each potential service unit (i, s) can be opened during the planning horizon but cannot be closed again. Let IS be the set of all services units (i, s) in all facilities, IS_o the set of existing ones and IS_c the set of potential ones. Further, $(i, s+1) \in IS$ implies $(i, s) \in IS$ and $(i, s+1) \in IS_o$ implies $(i, s) \in IS_o$.

Remark 5: We consider in this paper a planning horizon of 5 to 10 years. For such a time horizon, it is not reasonable to close an existing (open a new) maternity service and reopen (close) it later. If an existing type-1 hospital i can be upgraded by opening type-2 and type-3, then service units $(i, 1), (i, 2), (i, 3)$ are all in IS. If the hospital cannot be upgraded, then IS contains only $(i, 1)$. Similar restrictions apply for potential hospitals. The last part of Assumption 4 is a natural consequence of Assumption 1 and 2. It defines different types of facilities.

Assumption 5: For each customer type, the service region is partitioned into different **demand zones**. Each demand zone-k of a customer type-s corresponds to a geographic territory for a customer type-s and all customers (of type-s) from the demand zone are assigned to exactly one facility. Let K be the set of all demand zones for all customer types and K_s be the set of demand zones of type-s customers with $K = U_{s \in S} K_s$. As a result, two customers of different types from the same geographic location belong to different demand zones.

Remark 6: In the perinatal application, the motivation of different zone-partitions for different types of patients is due to significant difference in arrival rate and number of service units. The number of type-1 patients is significantly higher than that of type-3 patients. Further there are fewer type-3 NICU while each facility has a type-1 obstetric unit. As a result, each type-3 NICU is expected to cover a much

larger territory than a basic type-3 obstetric unit. For these reasons, different territory partitions are used for type-3 and type-1 patients such that each demand zone has a meaningful annual demand. This partition into different demand zones for different services allows easy modeling of demand demographic changes in space and in time.

Assumption 6: Customers of each demand zone-k arrive according to an independent Poisson process of rate d_k. Upon the arrival of a zone-k customer corresponding to type-s and assigned to facility i, s new demands each corresponding to a service $s' \leq s$ are generated simultaneously. Each new demand s' is accepted if a server in service unit (i, s') is available and it is rejected if all servers in (i, s') are busy. The service time of any demand in a service unit is a random variable of arbitrary distribution which depends on the type of the customer generating the demand.

Remark 7: Poisson arrivals are often observed in systems without appointments but with arrivals controlled by nature such as birth delivery [Brandeau 2004]. Furthermore, in stochastic systems, the coefficient of variation (CV) of length of stay (LOS), which is defined as the ratio of the standard deviation to the mean, is typically close to one, satisfying the usage of negative exponential distribution as a service time assumption.

Remark 8: Since patients of perinatal networks are urgent, they should not be allowed to wait. Therefore, inadequate capacity in a hospital results in the patient being deferred to another facility which may cause long travel distances to a non-preferred hospital. In other words inadequate capacity in a unit results in rejection of the patient which might cause fatal occurrences so that **rejection probability** is considered to be an important performance measure in this work as it is in most of the pure loss systems in literature.

Remark 9: In the perinatal application, assumption 6 implies that the three demands (OB, NICU and basic neonatal service)all needed for a type-3 woman can be accepted or rejected independently. This assumption is reasonable if “rejected" demands can still be accommodated in the same hospital by undesirable over capacity such as adding a third bed in a double room. This assumption does not hold if rejected demand cannot be handled in the same hospital. In this case the woman should be transferred to a less preferred hospital and all her demands be met in this new hospital. Extension to this latter case is not addressed in this paper.

Remark 10: In our perinatal application, the service time for the type-1 obstetric care is not the same for all women. Service times in OB of women can change a lot depending on if the birth is caesarian /vaginal. NICU babies from type-3 women stay longer than basic neonatal.

Remark 11: Under assumption 6, for a service unit with two types of customers

needing its service with arrival rate λ_i, service times X_i and service rate $\mu_i = 1/E[X_i]$, the service unit is an M/G/c/c queue with total arrival rate $\lambda = \lambda_1 + \lambda_2$, service time X equal to X_i with probability $\lambda_i/(\lambda_1 + \lambda_2)$ and average service rate $\mu = 1/E[X]$. The total offered load $a = \lambda/\mu$ is given by $a = a_1 + a_2$.

Since the system considered in this paper is stochastic, capacity requirements cannot be set deterministically. It is known from the literature that queuing theory is a suitable modeling technique when the system considered is stochastic [Asaduzzaman 2010]. Thus, in order to better evaluate the uncertainty in capacity requirements, we utilize queuing theory to obtain a performance measure which sets an analytical relationship between the capacity and service level in each hospital. This performance measure is set as “rejection rate" originated from Erlang loss function which is commonly interpreted as a good service level indicator and important for quality of care [De Bruin 2010].

Under on-going assumptions, each service unit (i, s) is an Erlang loss system, i.e. an M/G/c/c queue with Poisson arrivals, general service time distribution, c identical servers and no waiting room. The arrival rate is the total rate of demands generated by customers assigned to facility i and requiring service s. The service time distribution is the probability combination of the service times of all these customers. The rejection probability at service unit is insensitive to service time distribution [Cochran 2006, De Bruin 2010] and is given by the Erlang loss function:

$$B(c, a) = \frac{a^c/c!}{\sum_{k=0}^{c} a^k/k!}$$

where λ is the total arrival rate of demands for the service unit, μ the average service rate of the service unit, $a = \lambda/\mu$ the so-called offered load which is the sum of loads offered by each type of customers, and c the number of servers of the service unit. By convention, $B(c, 0) = 0, \forall c \geq 0$. This Erlang loss function holds even if service time at a service unit depends on the type of customer trigger the demand, i.e. even if the last part of Assumption 6 does not hold. However, in this case, the average service rate depends on the demand zone-to-facility assignment.

3.4 Capacity Planning And Service Units location

In this section, we consider the problem of selecting the set of service units to open or keep open and the service capacity of each open service unit in order to meet demand changes of different demand zones subject to service level constraints. A multi-period model is proposed in this section.

The decision options available over time include: opening a potential service unit in a facility, closing an existing service unit, purchasing new servers for a service unit, firing servers of a service unit, transferring servers from one service unit to another. Service re-location and capacity planning decisions are not taken consecutively but simultaneously according to the changing demand demographics in the network.

Each service unit is associated with the following costs: fixed opening cost, fixed closing cost, fixed operation cost which correspond respectively to construction cost, inconvenience cost of closure and maintenance/infrastructure cost. Capacity planning is associated with the following costs: cost of purchasing new servers, cost of firing existing servers, cost of transferring a server from a service unit to another, holding cost of a server or equivalently salary cost, patient assignment costs which depend on customers' preferences.

The problem consists in determining service unit relocation and capacity planning decisions over a multi-period time horizon in order to minimize the total costs over the planning horizon such that the rejection probability for each service unit and for each period is below a given service-dependent target and each demand zone is served by a facility within a limited distance.

The problem is formally defined by the following notation.

Sets and Indices

T	set of time periods indexed by $t = \{0, 1, ..., \overline{T}\}$
I	set of facilities indexed by $i, j = \{1, 2, ..., \overline{I}\}$
S	set of services indexed by $s = \{1, 2, ..., \overline{S}\}$
IS	set of all existing and potential service units indexed by $(i, s) \subseteq I \times S$
IS_o	set of existing service units initially open $IS_o \subseteq IS$
IS_c	set of potential service units initially closed $IS_c = IS - IS_o$
K_s	set of demand zones of type-s customers
K	set of all demand zones indexed by $k = U_{s \in S} K_s$
I_k	set of facilities that can serve zone-k customers $I_k \subseteq I$

d_{kt}	demand rate of zone k in period t
μ_{ks}	service rate of any service unit (i, s) for each demand generated by zone-k customers
b_{kst}	service-s load offered by zone-k customers in period t, i.e. $b_{kst} = d_{kt}/\mu_{ks}$
c_{is0}	initial number of servers of service unit (i, s)
LB_{is}	minimum number of servers of service unit (i, s)
UB_{is}	maximum number of servers of service unit (i, s)
α_s	target rejection probability of service s

Costs

OC_{is}	Fixed Opening Cost of service unit (i, s)
SC_{is}	Shutdown Cost of service unit (i, s)
FC_{is}	Fixed operation cost per time period of service unit (i, s)
PC_s	Unit purchasing cost of a type-s server
HC_s	Per period holding cost or salary cost of a type-s server
DC_s	Downsizing cost for firing a type-s server
TC_{ijs}	Transfer cost of server s from facility i to facility j
e_{ki}	unit assignment cost of a zone-k customer to facility i

Decision Variables

δ_{ist}	binary variable equal to 1 if service unit (i, s) is open in period t
x_{kit}	binary variable equal to 1 if demand zone k is assigned to facility i in period t
n_{ist}	number of servers of service unit (i, s) purchased in period t
z_{ist}	number of servers of service unit (i, s) fired in period t
y_{ijst}	number of servers transferred in period t from service unit (i, s) to (j, s)

Auxiliary Variables

c_{ist}	number of servers of service unit (i, s) in period t
a_{ist}	total offered load of of service unit (i, s) in period t

Objective function to minimize

$$\begin{aligned}\min & \sum_{(i,s)\in IS_c} OC_{is}\delta_{is\overline{T}} + \sum_{(i,s)\in IS_0} SC_{is}(1-\delta_{is\overline{T}}) \\ & + \sum_{t\in T}\sum_{(i,s)\in IS}(PC_s n_{ist} + DC_s z_{ist} + HC_s c_{ist} + FC_{is}\delta_{ist}) \\ & + \sum_{t\in T}\sum_{i,j\in I|(i,s)\in IS\&(j,s)\in IS} TC_{ijs}y_{ijst} + \sum_{k\in K}\sum_{i\in I} e_{ki}x_{kit}\end{aligned}$$

Constraints

$$\sum_{i \in I_k} x_{kit} = 1, \qquad \forall k \in K, t \in T \tag{3.1}$$

$$x_{kit} \leq \delta_{ist}, \qquad \forall (i,s) \in IS, t \in T, k \in K_s \tag{3.2}$$

$$a_{ist} = \sum_{k \in K_s \cup K_{s+1} \ldots \cup K_{\overline{S}}} b_{kst} x_{kit}, \qquad \forall i \in I, t \in T \tag{3.3}$$

$$c_{ist} = c_{is(t-1)} + n_{ist} - z_{ist} - \sum_j y_{ijst} + \sum_j y_{jist}, \qquad \forall (i,s) \in IS, t \in T \tag{3.4}$$

$$z_{ist} + \sum_j y_{ijst} \leq c_{is(t-1)}, \qquad \forall (i,s) \in IS, t \in T \tag{3.5}$$

$$c_{is(t-1)} + n_{ist} + \sum_j y_{jist} \leq UB_{is}, \qquad \forall (i,s) \in IS, t \in T \tag{3.6}$$

$$LB_{is}\delta_{ist} \leq c_{ist} \leq UB_{is}\delta_{ist}, \quad \forall (i,s) \in IS, t \in T \tag{3.7}$$

$$\delta_{i(s+1)t} \leq \delta_{ist}, \qquad \forall (i,s) \in IS, s \neq \overline{S}, t \in T \tag{3.8}$$

$$\delta_{ist} \leq \delta_{is(t-1)}, \qquad \forall (i,s) \in IS_0, t \in T \tag{3.9}$$

$$\delta_{ist} \geq \delta_{is(t-1)}, \qquad \forall (i,s) \in IS_c, t \in T \tag{3.10}$$

$$B(c_{ist}, a_{ist}) \leq a_s, \qquad \forall (i,s) \in IS, t \in T \tag{3.11}$$

$$n_{ist}, z_{ist}, y_{ijst}, c_{ist} \in N, \delta_{ist}, x_{kit} \in \{0,1\} \tag{3.12}$$

Constraints 3.1-3.3 are demand “assignment" constraints. Constraint 3.1 ensures that each demand zone is assigned to only one facility covering it. Recall that the service region is partitioned into demand zones for each customer type and hence each demand zone corresponds to a combination of a geographic zone and a customer type. With constraint 3.2, a demand zone is assigned to a facility having appropriate service unit. According to assumption 1 and 2, constraints 3.3 define the total offered load of each service unit (i,s) to be the sum of loads of all demands generated customers assigned to facility i and of types equal to or higher than s. Constraint 3.4 is the flow balance equation for the number of servers of each service unit. It takes into account purchased servers, fired servers, servers transferred out and servers transferred in. Constraint 3.7 sets upper and lower bounds for the capacity of each service unit depending on whether it is opened. Constraint 3.8 is a direct consequence of assumption 2 on the nested hierarchy of facilities and ensures that opening a service s+1 in a facility implies the opening of service s in the same facility. Constraints 3.9-3.10 are direct consequences of Assumption 4 and guarantee that if a service is closed, it will not re-open again throughout the planning horizon and vice versa. Constraint 3.11 sets an upper bound for the rejection probability for each service unit in the network.

Constraint 3.5 and 3.6 provide stronger formulation and are satisfied for optimal solutions of the problem. Constraint 3.5 ensures that the total number of fired

servers and servers transferred out does not exceed the number of servers in the previous period as, otherwise, some servers will be fired or transferred out immediately after they are purchased or transferred in, leading to higher total cost. Constraint 3.6 holds as the cost structure ensures that either the capacity increases through purchasing and transfer in or the capacity decreases through firing and transfer out, i.e. either $n_{ist} + \sum_j y_{jist} = 0$ or $z_{ist} + \sum_j y_{ijst} = 0$.

Constraints 3.8 is redundant as it is ensured by constraints 3.1-3.3 and the rejection probability constraints 3.11.

The problem under consideration is NP-hard as the classical capacitated facility location problem is a special case with one type of service and one time period and fixed capacities, i.e. $c_i = LB_i = UB_i$. In this case, the rejection probability constraints reduce to classical capacity constraints $\lambda_i \leq A_i$ for some A_i such that $B(c_i, A_i) = \alpha$.

Remark 12: In this paper, the service rate μ_{ks} does not depend on the facility i to which a demand is assigned to. All results of this paper still hold if this rate depends on the facility by using the service rate μ_{kis} and the offered load $b_{kist} = d_{kt}/\mu_{kis}$.

Remark 13: If $\mu_{ks} = \mu_s \quad \forall k, s$ which holds for our perinatal application, the offered load constraint 3.3 can be replaced by the following more compact ones:

$$D_{i\overline{S}t} = \sum_{k \in K_{\overline{S}}} d_{kt} x_{kit}, \qquad \forall i \in I, t \in T$$

$$D_{ist} = \sum_{k \in K_s} d_{kt} x_{kit} + D_{i(s+1)t}, \qquad \forall (i,s) \in IS, s \neq \overline{S}, t \in T$$

$$a_{ist} = D_{ist}/\mu_s \qquad \forall (i,s) \in IS, t \in T$$

where D_{ist} is the total arrival rate of (i, s) in period t.

3.5 Linearization of the mathematical model

This section presents different methods for linearization of the mathematical model of the previous section. The main nonlinear constraint is the rejection probability constraint $B(c_{ist}, \lambda_{ist}/\mu_{is}) \leq \alpha_s$. The first method relies on a point-wise representation of the rejection probability formula $B(c, a)$ for all possible values of capacity c and offered load a. The second method uses the maximum admissible offered load for any given capacity c. The third method relies on the minimum required capacity for any given offered load a. Some linear approximations will also be given.

Note that the second and third methods lead to more compact and stronger linearized models than the point-wise representation. Nevertheless we still introduce the point-wise representation linearization as it serves the basis for other methods and can be used if the service level constraint is replaced by a service-level-dependent criterion.

3.5.1 Linearization by point-wise representation

Rejection probability formulation (Erlang loss function) utilized in the constraint 3.11 gave rise to a highly nonlinear mathematical model. The first linearization requires the following assumption.

Assumption 7: All offered loads b_{kst} from all demand zones for all services and for all periods are integer multiple of some base offered load Δ. Let $L = \sum_{k,s,t} b_{kst}/\Delta$.

The first linearization uses the following additional variables:

q_{clist} binary variable equal to 1 if $c_{ist} = c$ and $a_{ist} = l\Delta$.

Constraint 3.11 is equivalently replaced by the following constraints:

$$c_{ist} = \sum_{c=LB_{is}}^{UB_{is}} \sum_{l=0}^{L} c q_{clist}$$

$$a_{ist} = \sum_{c=LB_{is}}^{UB_{is}} \sum_{l=0}^{L} l\Delta q_{clist}$$

$$\sum_{c=LB_{is}}^{UB_{is}} \sum_{l=0}^{L} q_{clist} \leq 1$$

$$\sum_{c=LB_{is}}^{UB_{is}} B(c, l\Delta) q_{clist} \leq \alpha_s$$

With the rejection probability constraint being replaced by their linearization constraints, the capacity planning and service unit location model is converted into

a linear mix integer program called **Model 1**. However the four sets of constraints above makes the whole enumeration of the rejection probabilities at all values of possible arrivals and capacity which is computationally demanding. Furthermore, certain combinations of c and l resulting in rejection probabilities greater than α_s are not feasible for the problem and can be eliminated. This awareness prompts a room for improvement and leads us to redefine the linearization constraints by eliminating the redundant parts.

3.5.2 Linearization by maximum admissible offered load

The second linearization method leading to **Model 2** relies on the concept of maximum admissible offered load $\overline{a}(c, \alpha)$ with respect to a given capacity c and a given rejection probability target α that is defined as follows:

$$B(c, \overline{a}(c, \alpha)) = \alpha$$

Since $B(c, a)$ is strictly monotone in c and in $a, \overline{a}(c, \alpha)$ is well defined. Further the following holds:

$$B(c, a) < \alpha, \qquad \forall a < \overline{a}(c, \alpha)$$

The second linearization uses the following additional variables:

q_{cist} binary variable equal to 1 if $c_{ist} = c$

and replaces the rejection probability constraint 3.11 by the followings:

$$c_{ist} = \sum_{c=LB_{is}}^{UB_{is}} c q_{cist}$$

$$\sum_{c=LB_{is}}^{UB_{is}} q_{cist} \leq 1$$

$$a_{ist} \leq \sum_{c=LB_{is}}^{UB_{is}} \overline{a}(c, \alpha_s) q_{cist}$$

With these modifications in rejection probability constraints, the model has gained some important desirable qualifications. Due to the fact that we do not enumerate all of the rejection probabilities defined with all possible arrivals and capacity, the number of necessary computations and memory held have been decreased drastically. The model is still linear and less demanding in terms of number of decision variables. Furthermore the assumption of integer arrivals is no more needed.

In our healthcare application, capacity c (number of staffed-beds in each service in each hospital) is upper bounded and desired rejection probability is commonly

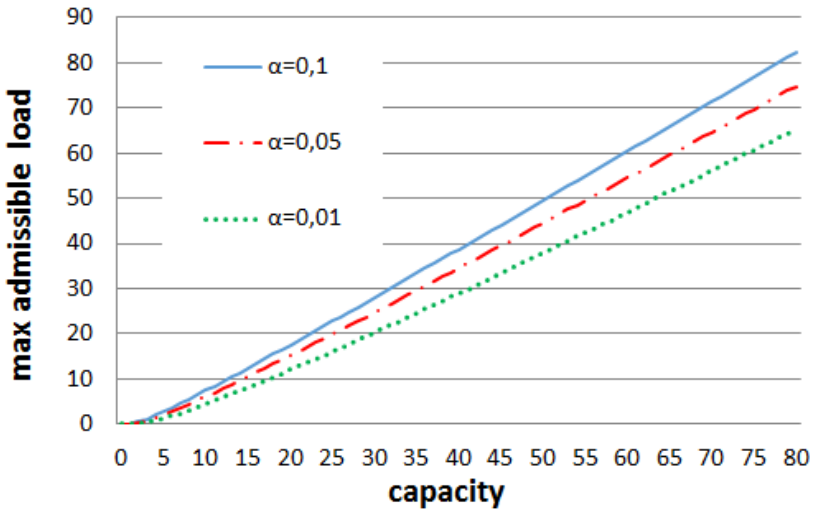

Figure 3.1: Maximum admissible offered load vs. capacity

smaller than 10%. Maximum admissible offered load $\overline{a}(c, \alpha)$ is one-dimension vector for each α. Figure 3.1 gives the maximum admissible offered load for different rejection probability of 10%, 5% and 1%.
Each offered load function can be closely approximated with a linear regression with $R^2 \cong 1$:

$$\overline{a}(c, 0.1) \approx 1.0572 \times c - 4.1648$$
$$\overline{a}(c, 0.05) \approx 0.9674 \times c - 4.5263$$
$$\overline{a}(c, 0.01) \approx 0.8531 \times c - 5.1443$$

It is also observed that the function $\overline{a}(c, \alpha)$ is convex and the linear approximation oscillates between underestimation and overestimation and the error is bigger for smaller c and α. Another interesting remark is that $\overline{a}(c, \alpha)$ tends to be linear in c. This result will be confirmed in our asymptotic analysis which proves the following asymptotes:

$$\overline{a}(c, 0.1) = 1.111 \times c - 10 + o(1/c)$$
$$\overline{a}(c, 0.05) = 1.053 \times c - 20 + o(1/c)$$
$$\overline{a}(c, 0.01) = 1.010 \times c - 100 + o(1/c)$$

Asymptotic behavior allows us to manage quite big size problems such as the capacity or location planning of hospitals in entire city or even in a country. In addition to that, asymptotic expansions may also provide us the upper and lower bounds for the capacity planning in medium size problems.

Asymptotic behavior of the function is not suited for our capacitated healthcare network where the capacity has both lower and upper bounds *(L, U)*. On the other side, linear fit to a curve may over/underestimate the maximum admissible offered loads, where the rejection probability constraints may not hold exactly.

In this paper we use instead an optimal linear approximation $\widehat{a}(c,\alpha)$ of maximum admissible offered load $\overline{a}(c,\alpha)$ in order to minimize the total deviation in the range of interest *(L, U)* while ensuring the rejection probability target α i.e. $B(c,\widehat{a}(c,\alpha)) \leq \alpha$. More specifically,

$$\min_{x,y} \sum_{c=L}^{U} \overline{a}(c,\alpha) - \widehat{a}(c,\alpha)$$

subject to

$$\widehat{a}(c,\alpha) \leq \overline{a}(c,\alpha), \qquad \forall c \in \{L, L+1, ..., U\}$$
$$\widehat{a}(c,\alpha) = x * c + y$$

With a range $(L,U) = (10,80)$ and a service-level target $\alpha = 0.05$, the following linear approximation is obtained:

$$a_{ist} \leq 0.9829 c_{ist} - 4.5461 \cdot \delta_{ist}$$

This modification in the constraints allows us to eliminate all additional decision variables. We will call this linearization method **linear approximation of maximum admissible offered load**. Note that this linear approximation can also be considered as a tangent line of the maximal admissible load function.

3.5.3 Linearization by minimum admissible capacity

The third linearization method leading to **Model 3** relies on the concept of minimum admissible capacity $\underline{c}(a,\alpha)$ with respect to a given offered load a and a given rejection probability target α that is defined as follows:

$$\underline{c}(a,\alpha) = \min\{c | B(c,a) \leq \alpha\}$$

Since $B(c,a)$ is strictly monotone in c and in a, the following holds:

$$B(c,a) < \alpha, \qquad \forall c > \underline{c}(a,\alpha)$$

The third linearization needs assumption 7 and uses the following additional variables:

q_{list} binary variable equal to 1 if $a_{ist} = l\Delta$.

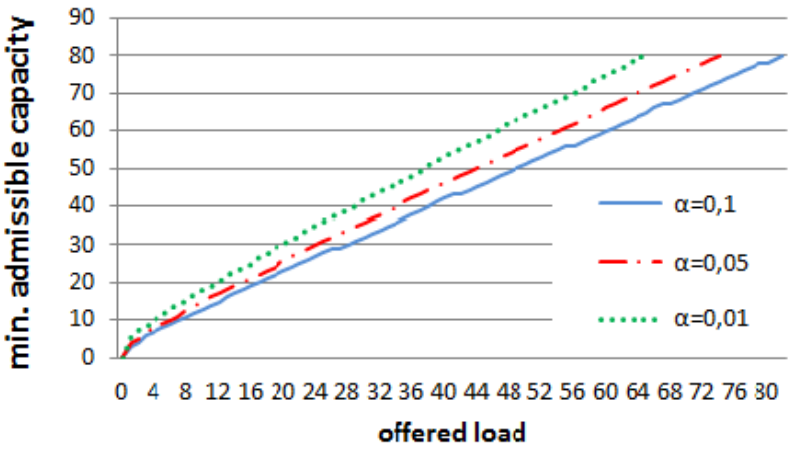

Figure 3.2: Minimum admissible capacity vs. offered load

and replaces the rejection probability constraint 3.11 by the followings:

$$a_{ist} = \sum_{l=0}^{L} l\Delta q_{list}$$

$$\sum_{l=0}^{L} q_{list} \leq 1$$

$$c_{ist} \geq \sum_{l=0}^{L} \underline{c}(l\Delta, \alpha_s) q_{list}$$

Figure 3.2 gives the minimum admissible capacity for different rejection probability of 10%, 5% and 1%. The following observations are made: $\underline{c}(a, \alpha)$ is almost concave in a and asymptotically linear with the following asymptotes:

$$\underline{c}(a, 0.1) = 0.9a + 9 + o(1/a)$$
$$\underline{c}(a, 0.05) = 0.95a + 19 + o(1/a)$$
$$\underline{c}(a, 0.01) = 0.99a + 99 + o(1/a)$$

The previous linear approximation approach is extended to determine **the linear approximation $\widehat{c}(a, \alpha)$ of minimal admissible capacity** that ensures the service level target. As for the maximum admissible offered load, for the practical range of healthcare application, linear approximation is preferred to these asymptotes. This leads to the following linear approximation of the rejection probability constraints for α=0.05:

$$c_{ist} \geq 1.0246 a_{ist} + 4.7009 \cdot \delta_{ist}$$

3.5.4 Properties of maximum offered load and minimum capacity

The purpose of this section is to investigate the properties of the maximum admissible offered load function $\overline{a}(c,\alpha)$ and the minimum admissible capacity function $\underline{c}(a,\alpha)$. Note that $\underline{c}(a,\alpha)$ takes integer values and is not even continuous in a and $\overline{a}(c,\alpha)$ is defined on integer points of c. To avoid these inconveniences, we adopt the usual extension of the Erlang loss function to real capacity c called **continued Erlang loss function** defined as follows [Jagerman 1974]:

$$B(c,a)^{-1} = a\int_0^\infty e^{-ay}(1+y)^c dy \tag{3.13}$$

The continued Erlang loss function $B(c,a)$ has the following properties.

Theorem 1. (i) The continued Erlang loss function $B(c,a)$ is equal to the usual Erlang loss function at integer point of c, (ii) it is strictly decreasing in c and strictly increasing in a, (iii) $B(c,a)^{-1}$ is jointly log-convex and hence jointly convex in c and a.

As a result, $\overline{a}(c,\alpha)$ and $\underline{c}(a,\alpha)$ can be extended to all positive real c as follows:

$$B(c,\overline{a}(c,\alpha)) = \alpha$$
$$B(\underline{c}(a,\alpha),a) = \alpha$$

The definition of $\overline{a}(c,\alpha)$ is the same as above but that of $\underline{c}(a,\alpha)$ differs. $\underline{c}(a,\alpha)$ can now be any positive real and the minimal admissible capacity of Section 3.5.3 is just the integer greater or equal to the real $\underline{c}(a,\alpha)$.

Theorem 2. $\overline{a}(c,\alpha)$ is increasing in c and in α and $\underline{c}(a,\alpha)$ is increasing in a and decreasing in α.

Theorem 3. $\overline{a}(c,\alpha)$ is convex in c and $\underline{c}(a,\alpha)$ is concave in a for any given α.

Theorem 4. $\underline{c}(a,\alpha) = (1-\alpha)a + (1-\alpha)/\alpha + o(1/a), \quad$ for all $\alpha > 0$.

Theorem 5. $\overline{a}(c,\alpha) = (1-\alpha)^{-1}c - 1/\alpha + o(1/c), \quad$ for all $\alpha > 0$.

Theorem 6. $B((1-\alpha)a,a) > \alpha, \quad \forall a > 0, \quad \forall \alpha < 1$.

From the asymptotes, the concavity, convexity and Theorem 6, the following bounds of maximum offered load and minimum admissible capacity can be derived:

$$(1-\alpha)a < \underline{c}(a,\alpha) < (1-\alpha)a + (1-\alpha)/\alpha$$
$$(1-\alpha)^{-1}c > \overline{a}(c,\alpha) > (1-\alpha)^{-1}c - 1/\alpha$$
$$\overline{a}(c,\alpha) \geq \overline{a}(c_0,\alpha) + (\overline{a}(c_0+1,\alpha) - \overline{a}(c_0,\alpha))(c-c_0)$$

for all $c \in N, c_0 \in N$. The bounds of $\underline{c}(a, \alpha)$ determined with continued Erlang loss function can be converted into integer-valued bounds and hence can be used to bound the minimum admissible capacity defined with usual Erlang loss function.

This section ends with the formal proofs of all above results.

Proof of Theorem 1. (i) is from Theorem 3 and (iii) from Theorem 19 of [Jagerman 1974]. From 3.13, $B(c, a)$ is strictly decreasing in c for $c \geq 0$ and $B(c, a) < B(0, a) = 1$. The proof of the monotonicity in a needs Theorem 15 of [Jagerman 1974] which gives:

$$\frac{\partial B(c,a)}{\partial a} = \{\frac{c}{a} - 1 + B(c,a)\}B(c,a), \qquad \forall a > 0$$

which implies that $\partial B(c, a)/\partial a > 0$ if $c \geq a$. If $c < a$, from Lemma 2 of [Jagers 1986], $aB(c, a) > a - c$ which implies $c/a - 1 + B(c, a) > 0$ and $\partial B(c, a)/\partial a > 0$. Q.E.D.

Proof of Theorem 2. Obvious as $B(c, a)$ is strictly decreasing in c and strictly increasing in a. Q.E.D.

Proof of Theorem 3. Since α is given, we use notation $\overline{a}(c)$ and $\underline{c}(a)$ for simplicity. For all c, $c' > 0$ and $\theta \in [0, 1]$, from the joint convexity of $B(c, a)^{-1}$:

$$\begin{aligned} &B(\theta c + (1-\theta)c', \theta\overline{a}(c) + (1-\theta)\overline{a}(c'))^{-1} \\ &\quad \leq \theta B(c, \overline{a}(c))^{-1} + (1-\theta)B(c', \overline{a}(c'))^{-1} = 1/\alpha \end{aligned}$$

which, together with the monotonicity of $B(c, a)^{-1}$ in a, implies $\overline{a}(\theta c + (1-\theta)c') \leq \theta\overline{a}(c) + (1-\theta)\overline{a}(c')$ and proves the convexity of $\overline{a}(c)$. Consider now any $a, a' > 0$ and $\theta \in [0, 1]$. Similarly

$$\begin{aligned} &B(\theta\underline{c}(a) + (1-\theta)\underline{c}(a'), \theta a + (1-\theta)a')^{-1} \\ &\quad \leq \theta B(\underline{c}(a), a)^{-1} + (1-\theta)B(\underline{c}(a'), a')^{-1} = 1/\alpha \end{aligned}$$

which, together with the monotonicity of $B(c, a)^{-1}$ in c, implies $\underline{c}(\theta a + (1-\theta)a') \geq \theta\underline{c}(a) + (1-\theta)\underline{c}(a')$ and proves the concavity of $\underline{c}(a)$. Q.E.D.

Proof of Theorem 4. For notation simplicity, let $\overline{\alpha} = 1 - \alpha$. Consider the capacity $c = \overline{\alpha}a + w$. Let $x = 1/a$. Hence,

$$q \equiv a/c = (\overline{\alpha} + wx)^{-1} \tag{3.14}$$

$$1/c = x(\overline{\alpha} + wx)^{-1} \tag{3.15}$$

For all $a > 2, w/\alpha, q > 1/(1-\alpha/2) > 1$, from Theorem 13 of [Jagerman 1974],

$$B(c,a)^{-1} = B(c,qc)^{-1} = \frac{q}{(q-1)} - \frac{q}{(q-1)^3}\frac{1}{c} + \frac{(2q^2+q)}{(q-1)^5}\frac{1}{c^2} + o(1/c^2) \tag{3.16}$$

Substituting 3.14 and 3.15 into 3.16 and then taking Taylor expansion lead to:

$$\begin{aligned}
\frac{q}{q-1} &= \frac{1}{\alpha} + \frac{w}{\alpha^2}x + \frac{w^2}{\alpha^3}x^2 + o(x^2) \\
\frac{q}{(q-1)^3}\frac{1}{c} &= \frac{\overline{\alpha}}{\alpha^3}x + \left(\frac{w}{\alpha^3} + \frac{3\overline{\alpha}w}{\alpha^4}\right)x^2 + o(x^2) \\
\frac{2q^2+q}{(q-1)^5}\frac{1}{c^2} &= \frac{2\overline{\alpha}+\overline{\alpha}^2}{\alpha^5}x^2 + o(x^2)
\end{aligned}$$

As a result,

$$\begin{aligned}
B(c,a)^{-1} = \frac{1}{\alpha} &+ \left(\frac{w}{\alpha^2} - \frac{\overline{\alpha}}{\alpha^3}\right)x \\
&+ \left(\frac{2\overline{\alpha}+\overline{\alpha}^2}{\alpha^5} + \frac{w^2}{\alpha^3} - \frac{w}{\alpha^3} - \frac{3\overline{\alpha}w}{\alpha^4}\right)x^2 + o(x^2)
\end{aligned}$$

This leads to:

$$\begin{aligned}
B(\overline{\alpha}a + w, a)^{-1} &= \frac{1}{\alpha} + \left(w - \frac{\overline{\alpha}}{\alpha}\right)\frac{1}{\alpha^2 a} + o(1/a) \\
B(\overline{\alpha}a + w_0, a)^{-1} &= \frac{1}{\alpha} + \frac{\overline{\alpha}}{\alpha^5 a^2} + o(1/a^2)
\end{aligned}$$

$\forall w \neq w_0 \equiv \overline{\alpha}/\alpha$. As a result, for large enough $a, B(\overline{\alpha}a + w, a) > \alpha$ for all $w < w_0$ and $B(\overline{\alpha}a + w_0, a) < \alpha$. Since $B(\overline{\alpha}a + w, a)$ is strictly decreasing in $w, \underline{c}(a,\alpha) = \overline{\alpha}a + w_0 + o(1/a)$ which completes the proof. Q.E.D.

Proof of Theorem 5. Let $a = c/\overline{\alpha} - w/\overline{\alpha}$. Hence, $B(c, c/\overline{\alpha} - w/\overline{\alpha}) = B(\overline{\alpha}a + w, a)$. From the proof of Theorem 4, for large enough c, i.e. large enough $a, B(\overline{\alpha}a + w, a) > \alpha$ for all $w < w_0 \equiv \overline{\alpha}/\alpha$ and $B(\overline{\alpha}a + w_0, a) < \alpha$ which implies $B(c, c/\overline{\alpha} - w/\overline{\alpha}) > \alpha$ for all $w < w_0$ and $B(c, c/\overline{\alpha} - w_0/\overline{\alpha}) < \alpha$. Since $B(c, c/\overline{\alpha} - w/\overline{\alpha})$ is strictly decreasing in $w, \overline{\alpha}(c,\alpha) = c/\overline{\alpha} - w_0/\overline{\alpha} + o(1/c)$ which completes the proof. Q.E.D.

Proof of Theorem 6. From Lemma 2 of [Jagers 1986], $aB(c,a) > a - c$ which implies $B((1-\alpha)a, a) > \alpha$. Q.E.D.

3.6 Optimization of a perinatal network

This section presents an application of the previous results to the optimization of the perinatal network of maternity facilities in the department North Hauts-de-Seine, France which is known as the most populated region in Ile-de-France after central Paris. The perinatal network is represented in Figure 3.3 where light (dark) color symbolizes population rise (decrease) in the corresponding commune. The network contains 6 public facilities ($H0, H3, H4, H5, H8, H9$). Four potential facilities ($H1, H2, H6, H7$) are considered.

This case study is partially supported by Agence Régionale de Santé (ARS), the regional health authority which provided us certain information and data. Unfortunately, data required for the study is tremendous due to the large-scale of the work, and sometimes hard to quantify (i.e. cost structure). Therefore, we must make certain assumptions or conduct some sensitivity analysis in order to determine the reasonable parameter settings for our case study.

Initially, key characteristics about the network and key assumptions in the case study are determined. Afterwards we discuss the results of DOE (design of experiments) for cost parameters. Finally, the results are presented.

3.6.1 Key Characteristics and Assumptions

In this application, servers are staffed-beds. Bed capacities, c_{is0} (current number of beds), lower bounds (LB) and upper bounds (UB) of the 10 hospitals are given in Table 3.1.

North Hauts-de-Seine department is composed of 14 population centers. In this study, we consider 21 demand zones for different patient types. Type 1 patients are higher in number and can be treated in all hospitals. All 14 real population centers are considered as the demand zones for type-1 patients (needing only OB). However, type-2 and 3 patients are lower in number and are covered by fewer hospitals. We gathered the patients in neighboring population centers who are covered by the same hospital set and generate demand zones: demand zones 15-19 for type-2 patients (needing OB and neonatal) and demand zones 20-21 for type-3 patients (needing OB, neonatal, and NICU). Different demand zones for different patient types are represented in Figure 3.4.

Each demand zone is covered by a set of hospitals but has different preferences for each hospital in this set. The order of preference of each hospital is extracted from the arrival data of previous years. The demand zones, associated patient type, the hospitals covering them and the demand scenario used in the case study are presented in Table 3.2.

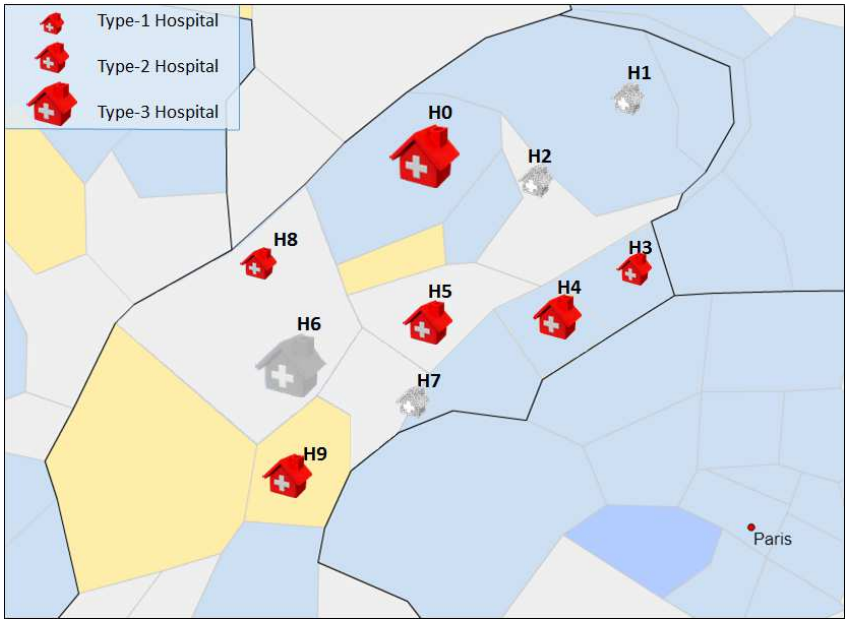

Figure 3.3: North Hauts-de-Seine Perinatal Network

Table 3.1: Maternity Facilities and Bed Capacities

	Type	OB			Neonatal			NICU		
		c_{i10}	LB	UB	c_{i20}	LB	UB	c_{i30}	LB	UB
H0	**3**	**65**	**10**	**80**	**20**	**5**	**25**	**13**	**5**	**20**
H1	*1*		*5*	*30*						
H2	*1*		*5*	*30*						
H3	**1**	**40**	**10**	**50**						
H4	**2**	**30**	**10**	**50**	**15**	**5**	**25**			
H5	**2**	**30**	**10**	**50**	**10**	**5**	**25**			
H6	*3*		*10*	*80*		*5*	*25*		*5*	*20*
H7	*1*		*5*	*30*						
H8	**1**	**25**	**10**	**30**						
H9	**2**	**50**	**10**	**60**	**20**	**5**	**25**			

For all OB units in each hospital, ARS provided us yearly arrival and LOS data of each pregnant woman and their domicile information. By assuming no seasonality within the year, we were able to compute daily offered load of each population center and each hospital in the network. ARS also provided the ratio of neonates requiring basic and intensive care, from where the offered load of neonates requiring basic and intensive care are computed.

In order to reach a realistic demand scenario, we analyzed the studies conducted by national statistics authorities such as INSEE (National Institute of Statistics and Economic Studies) and ATIH (Technical Agency of Information on Hospitalization) in France and come up with a realistic prospective demand scenario that represent the changing demographics in Hauts-de-Seine. In the forecasted data of

Table 3.2: Demand Zones and Demand Scenario

Demand Zones		Patient type	Hospitals most to least preferred	Demand in each period t										
				Current period	t=1	t=2	t=3	t=4	t=5	t=6	t=7	t=8	t=9	t=10
1	VILLENEUVE LA G.	1	0,1,3	4	4	4	3	3	3	3	3	2	2	2
2	GENNEVILLIERS	1	0,1,2,3,8	8	8	7	7	7	6	6	5	5	4	4
3	COLOMBES	1	0,1,2,3,5,8	16	15	15	14	13	13	12	11	11	10	10
4	ASNIERES SUR S.	1	2,3, 0,1,4	14	14	13	12	11	11	10	9	9	8	8
5	BOIS COLOMBES	1	0,2,3,4,5	4	4	4	4	3	3	3	3	3	2	2
6	CLICHY	1	0,2,3,4,5	11	11	10	9	9	8	8	7	7	6	5
7	LEVALLOIS PERRET	1	4, 3, 0, 5	10	10	10	9	9	8	8	8	7	7	6
8	NEUILLY SUR SEINE	1	4,5, 9,6, 8, 0	4	4	4	4	3	3	3	3	3	3	3
9	COURBEVOIE	1	5,9,4,6,0,7,8	12	11	11	11	11	12	12	12	11	11	11
10	LA GARENNE C.	1	0,4,5,7,8,9,6	3	3	4	4	4	5	5	6	7	7	8
11	PUTEAUX	1	9,6,8,7,5,0	6	7	7	8	9	10	11	12	13	14	16
12	NANTERRE	1	8,9,6,7,0	16	17	18	19	20	21	22	24	27	30	35
13	SURESNES	1	9,6,8,7,0	8	9	9	10	11	12	14	15	16	16	17
14	RUEIL MALMAISON	1	9,8,7,6,0	6	6	7	7	8	9	11	11	13	15	16
15	1, 2	2	0	3	3	3	3	3	2	2	2	2	2	2
16	3, 4, 5	2	0,4,5	9	8	8	8	7	7	7	6	6	6	5
17	6, 7	2	4,0,5	6	6	5	5	5	5	4	4	4	4	4
18	8, 9, 10	2	5,4,9,0,6	5	5	5	5	5	5	6	7	9	12	12
19	11, ..., 14	2	9,6,4,5,0	10	10	11	12	13	14	14	15	15	16	16
20	1, ..., 7	3	0,6	7	7	7	7	6	6	6	6	6	6	6
21	8, ..., 14	3	6,0	6	6	6	6	7	7	8	9	10	11	12

the department, a prospective migration is observed at some certain communes and additionally it should be noted that 58% of the population in the department is between the ages of 20-59. Therefore, in our demand scenario an increase of 2% for some communes and a decrease of 1% for some other communes are considered to model the migration in population while we let the total demand increase gradually. This leads to a demand evolution over a time horizon of 10 years with a yearly time period. This prospective demand evolution forecasted for each demand zone throughout the planning horizon is presented in Table 3.2.

In related literature [Gorunescu 2002, Asaduzzaman 2010, Li 2008], the desired value of acceptance probability in maternity facilities is defined as 95%. Therefore, we set the upper bound for rejection rate as 0.05 for all patient types (both women and neonates).

3.6.2 Setting Cost Parameters with DOE

Cost parameters are one of the most important parameters to set. The objective function of our model is composed of different cost items. This makes the model highly sensitive to the cost structure.

We obtained costs of the beds (PC_s) themselves from the average price of each bed type on the market. Holding cost for a staffed-bed (HC_s) is based on actual staff salaries and the pre-defined relationship between human resources and a bed (ratios of human resources (i.e., midwives, obstetricians, nurses, and pediatricians) for each bed is provided by ARS). However, true costs for downsizing (DC), transfer (TR), fixed cost of keeping a structure open (FC), and closing costs (SC) are

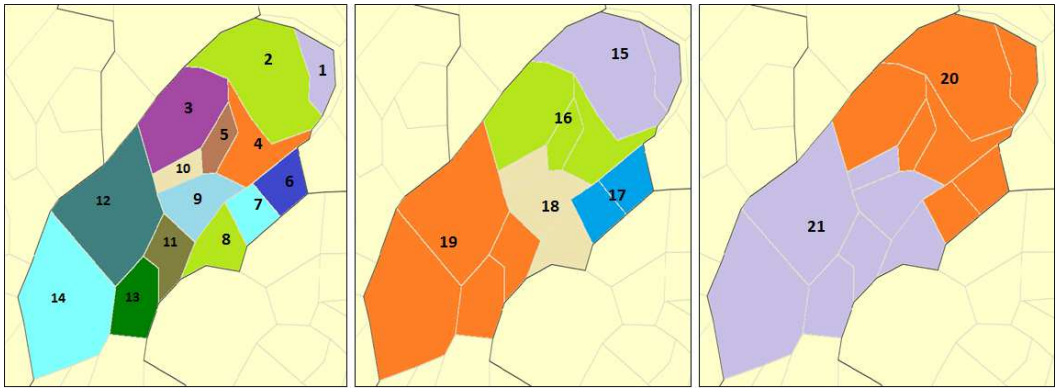

Figure 3.4: Representation of demand zones for type 1, type 2 and type 3 patients, respectively from left to right

unknown. They are mostly related with the inconvenience of different stakeholders. They may take on different values depending on the perspective and preferences of different stakeholders, eventually result in different outputs. In order to have a rigorous understanding of the model behavior in terms of the cost structure, we conducted design of experiments for the most difficult to set cost items.

The details of DOE are given in Appendix A. In this paper we only present the main results. When FC is low and DC is high, more services are kept open whereas when FC is high and DC is low more services get downsized and closed. Other than those intuitive results, as FC and DC are both low, even though it is cheap to keep services open due to small FC, it is observed that the model tends to close services and downsize resources in order to decrease holding cost which constitutes the biggest portion of the total cost. On capacity planning side, there is often a tradeoff between downsizing and transferring resources. FC does not have a crucial effect on the number beds downsized neither has a direct effect on the number of beds transferred.

In the remaining of our perinatal network case study, the cost structure presented in Table 3.3 is used. All costs are given in some standardized money units and we assume that all maternity facilities have identical cost structure while cost values are differentiated according to the type of services. The motivation for this cost structure is explained below.

Setting DC: In healthcare networks, downsizing resources is not a good strategy considering that most of the human resources are highly educated, experienced, scarce, thus too expensive to downsize. DC is one time cost and directly linked with holding cost such that when a staffed-bed is cut-off, holding cost of this resource is also eliminated for the rest of the planning horizon. Therefore, in order not to let

any downsize, we set DC regarding the relation $DC_s \geq HC_s \cdot T$.

Setting TR: Transfer is possible in the network only if the following relation holds: $TR_s \leq DC_s + PC_s$. So TR is set to its medium level in DOE setting.

Setting FC: It was observed from DOE study that setting FC on the extremities (too high or too low) leads some unrealistic configurations. So it is kept at its medium level.

Setting SC: Even though downsizing is not desirable, it is possible to close structures in healthcare. It is a one-time cost incurs for each unit closed and works in the opposite direction of FC which incurs periodically for every unit kept open. SC should be set much bigger than FC for a feasible solution. Once FC is defined it is easy to define SC, therefore it is set in relation with FC.

Setting OC: Opening cost stems from the construction costs of a facility which should be set quite high for not letting the model open a structure for a demand that can be handled within the existing structure.

Unit assignment costs (e_{ki}) are set according to the preference ranking of each demand zone. While the most preferred hospitals have no assignment cost, as the preference ranking increase, the cost increases exponentially (0, 10, 40, 90,...). This cost is set after a sensitivity analysis where we compared the percentage contribution of assignment costs compared to the total cost, and the unit assignment costs are set to keep this percentage contribution in a meaningful range.

Table 3.3: Cost Data

Costs	**Obstetrics (s=0)**	**Neonatal(s=1)**	**NICU(s=2)**
Purchase Cost (PC)	1	3	7
Holding Cost (HC)	7	10	20
Decrement Cost (DC)	100	150	200
Transfer Cost (TR)	15	20	40
Fix Cost (FC)	30	30	30
Shutdown Cost (SC)	200	200	200
Opening Cost (OC)	1000	1000	1000

Remark 14: Holding cost incurs at each period for each staffed-bed hold in each hospital. Compared to other costs, the total holding cost is enormously big which makes other capacity decisions insensitive to the decision process. In order to balance the effect of cost items, we choose to work with a modified RHC (relative holding cost) where the emphasis is mostly on the new invested resources. For this purpose, the lower bound of Theorem 6 for the capacity is subtracted from the total

holding cost and introduced in the objective function instead of regular holding cost:

$$RHC = \sum_{t \in T} \left(\sum_{(i,s) \in IS} HC_s \cdot c_{ist} - \sum_{s \in S} HC_s \cdot \alpha_s \sum_{k \in K_s} b_{kst} \right)$$

3.6.3 Comparison of linearization models

A test instance is developed with the demand scenario of Table 3.2 and the cost structure of Table 3.3. Five different capacity planning and service unit relocation models where constraint 3.11 is replaced by the linearization models presented in Section 3.5 are coded in C++. For model comparison purposes, the proposed test instance is solved with all 5 models using CPLEX 12.5 on a computer supporting at most 2.67 Ghz processor with 6GB RAM.

Due to huge memory requirements, piecewise linearization model (Model 1) could not reach any solution whereas all other linearization models are able to find optimum solutions. Comparisons of models are presented in Table 3.4. The total cost of the exact optimum solution is 16340. Even though linearization models 2 and 3 (maximum offered load $\overline{a}(c, \alpha)$ and minimum capacity $\underline{c}(a, \alpha)$) are able to find the optimum value, the proposed capacity planning (demand assignments and capacity decisions) differs slightly in that they point out alternative solutions. In terms of CPU time, linearization model 2 performs slightly better than model 3. On the other side, both of the linear regression approximations of the two models reach their optimum solution though in quite long time and the optimum values they found overestimate the real optimum as expected. This is consistent with the observation of Section 3.5 that linear approximation “Max Load" is a tangent line to the max load curve and at some parts underestimate the values of “maximum admissible load for a certain capacity". Likewise, linear approximation “Min Cap" is a tangent line to the min required capacity curve and overestimates the values of “minimum capacity required for a certain amount of arrival". Note that the opti-

Table 3.4: Linearization Model Comparison

Model	**Total Cost**	**Optimality Gap**	**CPU time**
Linearization (Max Load)	16340	0%	19.53 min
Linearization (Min Cap)	16340	0%	20.82 min
Linear Approximation (Max Load)	16455.8	0%	136 min
	16455.9	0.30%	113 sec
	16458.9	0.50%	28 sec
	16505	1%	20 sec
Linear Approximation (Min Cap)	16465.8	0%	259 min
	16465.9	0.30%	113 sec
	16475.9	0.50%	30 sec
	16481	1%	11 sec

mality gap is with respect to the optimal solution of the linear approximation model.

Further, in linear approximation models, near optimum solutions with optimality gap of less than 0.5% are reached very fast, however it takes quite long time to reach the optimum. To better understand this convergence, Table 3.4 presents CPU time needed to reach various optimality gaps. Even in a very small gap (0.3%), both approximations are able to get quite close to their optimum in reasonable CPU time.

To summarize, linearization by maximum offered load proves itself to be the most appropriate model and is used in remainder of our case study. It is able to find the optimum solution in a reasonable CPU time. On the other side, linear approximation models are quite fast however they overestimate the optimum. Those approximation models are appropriate in bigger instances where they can even reach the optimum.

3.6.4 Sensitivity analysis

This subsection considers three scenarios: (i) the base scenario with demand scenario of Table 3.2 and cost structure of Table 3.3, (ii) the base scenario but with very high transfer cost TC, (iii) the base scenario but with higher acceptance rate of 99% for neonatals and NICUs. The goal of the base scenario is to better understand the detailed capacity planning decisions. Scenario (ii) is considered to understand the value of collaboration between different hospitals through bed transfer. Scenario (iii) is considered to investigate the impact of rejection probabilities.

The optimum capacity planning for NICUs, Basic Neonatals, and OBs of the base scenario is presented in Figures 3.5, 3.6 and 3.7 respectively where the dashed bars represent number of staffed-beds transferred and black bars represent number of staffed-beds purchased at each period.

In the current system there is only one NICU providing service in the network and it has limited capacity. As the total demand increases gradually, at some point in the planning horizon, demand for NICU exceeds the total capacity that is possible to be kept in the existing hospital. Therefore in order to maintain the required service level, capacity increment by new investments becomes an inevitable action to take. Thus, a new NICU unit (H6) is launched at period 8. Since NICU units imply other lower level services, the decisions on NICUs force some other decisions on OBs and basic neonatal units in the network. Consequently, model proposes to launch a type-3 hospital (H6) in the increasing population area (south of the region).

In basic neonatal units, with some minor transfers between units, the capacity is balanced and existing neonatal units are able to meet their own demand until period 8 where a new basic neonatal unit is forced to be opened along with the opening of H6. Furthermore, in the beginning of the time period, the model proposes to close

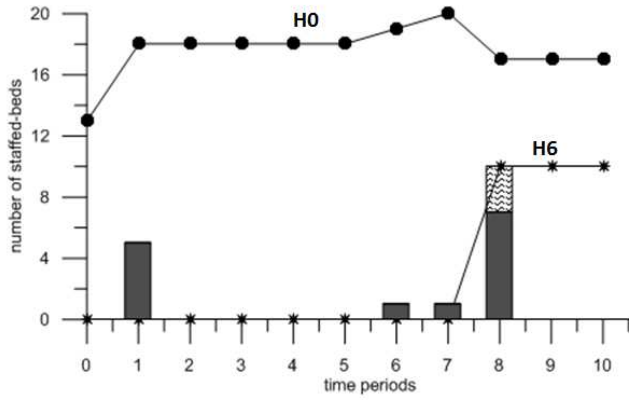

Figure 3.5: Optimal Capacity Planning for NICU

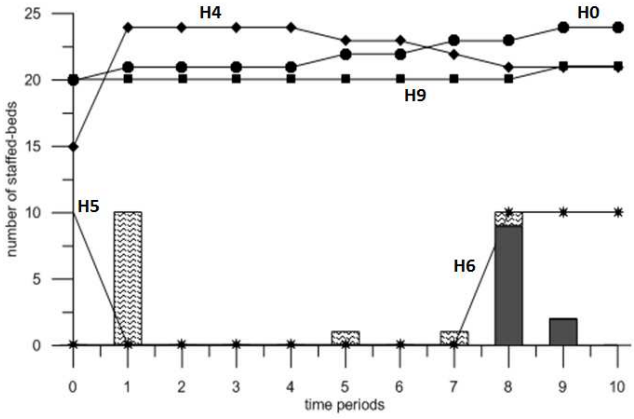

Figure 3.6: Optimal Capacity Planning for Basic Neonatal Units

the basic neonatal unit of H5 (which consists the lowest number of staffed-beds and is located next to a decreasing population area) and transfer its resources for the sake of centralization.

As expected, the biggest changes in capacity planning and investments are proposed for basic neonatal and NICUs rather than OBs. Bed capacity in OBs is almost sufficient for the demand changes in the first half of the planning horizon. With the cooperation in the network (transfer of resources) existing OBs meet their service level requirements without having the need to make new investments almost until the last period of the planning horizon. The migration in department is clearly observable in the capacity changes in the network.

With the presence of the randomness, it is well-known that bigger structures and centralized resources perform better. If hospitals in perinatal network were not capacitated, our model would choose to increase the capacities of the most preferred existing units while closing the least preferred ones and making necessary resource

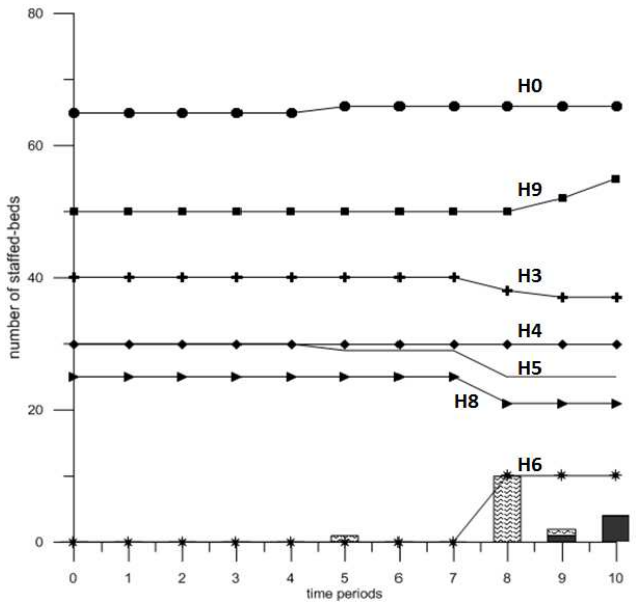

Figure 3.7: Optimal Capacity Planning for Obstetric Units

transfers instead of opening new structures and making new investments. However when the facilities are capacitated, opening new structures is an indispensable decision.

Scenario 2 with very large transfer cost TC is considered to quantify the impact of network cooperation via resource mobility or transfer. Big TC eliminates the

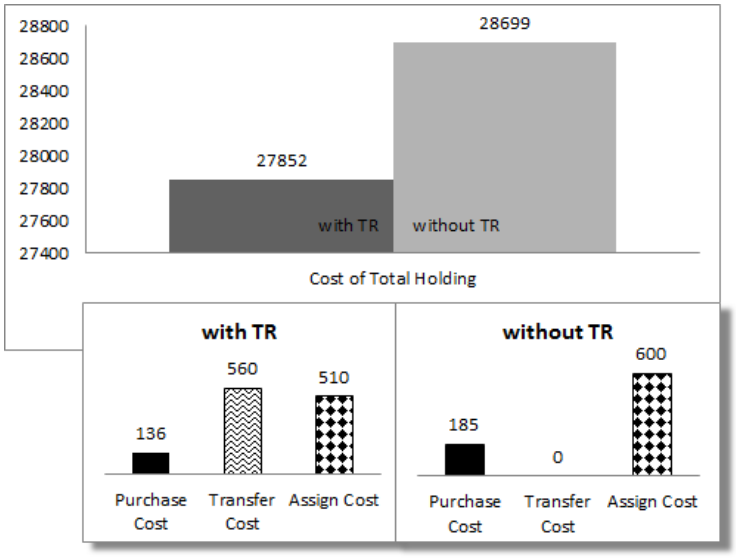

Figure 3.8: Network with cooperation vs. without cooperation

possibility of sharing (transferring) resources among hospitals. With large TC, it is no more profitable to close an existing facility since its cost cannot be compensated by other decisions such as downsizing or transferring. The model has to make new investments in order to ensure the desired service level, which results in increasing holding cost, presented in Fig 3.8. Detailed results and plots can be found in Appendix B.

Scenario 3 with higher service level of 99% acceptance rate for neonatal cares and NICUs is used to investigate the impact of higher service level. Seeking higher service level for neonates whose health conditions are generally highly critical is natural and a higher service level would provide an exponential improvement in their health states. It is a known fact that NICUs are the bottlenecks of the system due to their lack of capacities. With the increased service level, a new NICU unit is opened in scenario 3 at the beginning of the planning horizon. Higher investment and more transfers are needed to ensure a higher service level, leading to higher increment and holding costs. The much higher service level of 99% acceptance rate for the most critical patients is obtained with an additional 18% expenses in the network. Fig. 3.9 compares different cost items. Detailed results and supporting plots can be found in Appendix C.

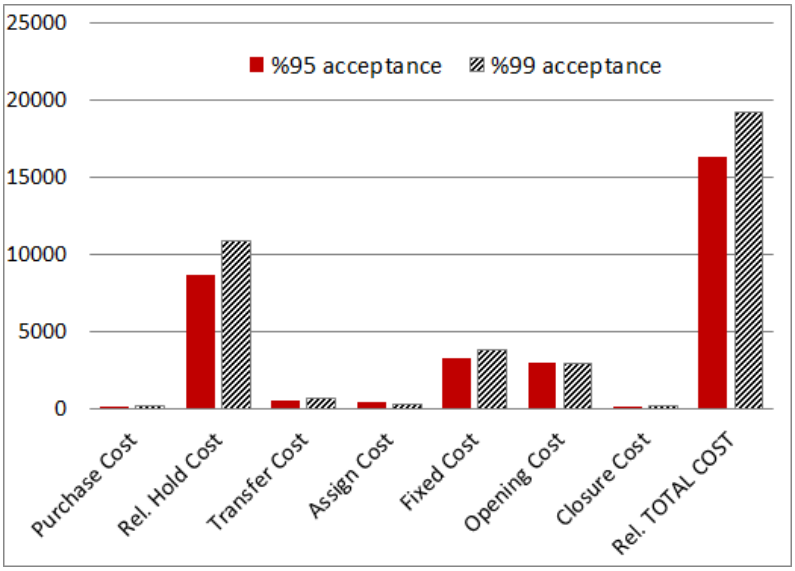

Figure 3.9: Costs vs. Service Level

3.7 Conclusion

This paper proposed a dynamic capacity planning and service unit location of a stochastic hierarchical service network where we consider simultaneously location (opening/closing) and service capacity decisions (increase, decrease, transfer) in order to ensure a minimum desired customer acceptance rate for each service unit in each facility. Poisson arrivals and random service times are handled by modeling each service unit in each facility as an Erlang loss system. A multi-period, multi-facility, multi-service optimization model is proposed where service levels are given by nonlinear Erlang loss formula.

We were able to linearize the nonlinear model by various linearization models which allow us to solve big size problems to optimality in a reasonable time. Structural properties of these linearization models are proved.

The best performed linearized model is applied to a real-life perinatal network. Different decision scenarios are tested in order to evaluate the importance of the model decisions and their effect on the total cost. Cooperation among players (in this work by transferring resources) in a hierarchical network proves its importance by resulting in significant cost savings. Finally we tested the model to quantify the amount of expense related with necessary changes in the network structures and accordingly capacity reorganization if a higher service level is needed for some services.

The scope of the numerical experiment of this paper is limited to prove the applicability and compare performances of the different linearization models. Realizing extensive numerical experiments on perinatal networks is out of the scope of this work, however quite necessary for a prospective real life application. Future research can be pursued in several directions. Detailed statistical analysis and extensive numerical experiments especially with different regional configurations and different demand scenarios are needed in order to investigate how location and capacity planning decisions depend on these parameters.

In this work, the stochastic aspect is characterized by Poisson arrivals and random service times which were shown in the literature reasonable assumptions for modeling maternity services. Further, we assumed a gradual annual demand changes over the planning horizon encouraged from the fact that demographic changes are often slow on a planning horizon of 5 to 10 years. The extension of our model to longer horizon of 20 to 50 years would lead to significant uncertainties in demand and an interesting future research direction that can be addressed by robust optimization or stochastic programming.

In this work demands are assumed urgent and are lost if they are rejected at their assigned facility. This assumption is reasonable for the purpose of this paper, i.e. for location and capacity planning decisions at strategic level to ensure that each

facility has enough capacity to accommodate its assigned customers at some service level. It is however not acceptable at daily operational level where patients rejected from their own hospital are not lost but overflow to another hospital. Real-time admission and overflow control of patients in a hierarchical network is an interesting and rich research area which we consider as a future work.

3.8 Résumé du chapitre

Ce chapitre traite de la planification de la capacité dynamique et de la localisation des services de soins d'un réseau. Dans cette optique, nous proposons un nouveau modèle de réseau de services hiérarchique dans lequel les deux installations et les clients ont une hiérarchie imbriquée. Une installation de niveau supérieur fournit tous les services fournis par un dispositif de niveau inférieur. Un client exigeant un certain niveau de service aura en outre besoin de services de niveau inférieur. Une loi de Poisson décrit les arrivées des clients, et des durées de service aléatoires sont pris en compte. Chaque unité de service est modélisée comme un système d'Erlang-loss et le niveau de service défini comme étant la probabilité de l'acceptation du client est déterminée par la fonction dite d'Erlang-loss. Le problème de la localisation et de la planification des capacités consiste à déterminer quand et où pour ouvrir ou fermer une unité de service, la capacité de l'unité de service et de la répartition des demandes d'installations . Un modèle de programmation non linéaire est proposé pour la minimisation du coût total tout en maintenant le niveau de service de l'ensemble des unités de service supérieures à un certain niveau donné. Différents modèles de linéarisation de la fonction Erlang-loss et leurs propriétés sont proposées. La linéarisation transforme le modèle non-linéaire en programmes mixtes en nombres entiers pouvant être résolu à l'optimalité avec les solveurs standards en temps raisonnable. Une application à un réseau périnatal est ensuite présenté.

L'étude de cas est motivée par notre collaboration avec le réseau de périnatalé dans les Hauts-de-Seine, près de Paris, France. Il s'agit d'un réseau d'installations de maternités de différents types où au plus trois niveaux de services (obstétrique, réanimation néonatale et néonatalité soins intensifs) sont fournis. Tous les établissements fournissent le niveau le plus bas des soins: services d'obstétrique. Les types de services de maternité diffèrent selon le niveau de service fourni. Le réseau montre une propriété hiérarchique imbriquée où un établissement de niveau supérieur fournit tous les services fournis par un établissement de niveau inférieur. Chaque hôpital a un nombre limité de lits pour chaque service. Du fait des changements démographiques , le réseau périnatal est confronté au problème de la réinstallation de ses services en périnatalité et de l'ajustement de sa capacité pour répondre à la demande changeante. Les modèles mathématiques du présent document tiennent compte des différentes options telles que l'ouverture ou la fermeture d'une unité de service, l'expansion des capacités, la réduction et le transfert des ressources humaines. Les modèles sont utilisés pour aider les manageurs à prendre des décisions afin de minimiser le coût global du réseau sans dégrader le niveau de service.

Chapter 4

Performance Evaluation of an Overflow Loss Network

Contents

4.1 Introduction

In a loss hospital network, due to scarce resources, not enough beds and high variability, there is inevitably rejections and overflows. As discussed in the strategic capacity and location planning of a hospital network presented in chapter 3, in order to ensure the patients to be admitted to their 1^{st} preferred hospitals, it is necessary to make extremely expensive investments on beds which should be supported with experienced human workforce; most of the time even not feasible to provide. Therefore, in such a resource-scarce system there always exist some patients not being accepted to their preferred hospitals and overflow to other hospitals, sometimes even more than once.

In such an environment, it is important to be able to quantify the amount of overflows among hospitals, determine the most loaded hospitals and the highest rejected patient group in order to propose to-the-point solutions regarding the bottlenecks.

For that, one needs to be able to evaluate the rejection rates in each hospital for each arriving patient group in an overflow loss network.

The purpose of our study in this chapter is to develop methods (performance analysis tools) for approximating the blocking probability of **each arrival stream** in each station (hospital) in order to evaluate the performance of an overflow loss network. Blocking probability estimation in a multi hospital loss network where a patient can overflow more than once involves two key issues. First, overflow streams do not follow a Poisson distribution, so Erlang-loss formula cannot be employed for loss calculation. An overflow stream has a higher variance than a Poisson stream, thus one-moment characterization is generally not accurate. Secondly, overflow streams might be statistically correlated. In our approximation method, we address the first issue by characterizing each incoming and outgoing overflow stream as Interrupted Poisson Process (IPP). The correlations among streams in each station are managed by using a binomial moment approximation method which allows us to compute blocking probability for each stream separately.

We consider two systems with different overflow structures. First considered network is denoted as "Acyclic Overflow Loss Network" where the overflows flow along one direction, such that customers overflow from lower-tiers to higher-tiers. For performance evaluation of such a system, our approximation method is based on a binomial moment matching of IPP characterization of non-Poisson arrivals. Second considered system assumes the existence of feedback overflows where two hospitals may overflow to each other. For this network, several efforts are realized to evaluate highly correlated/interdependent blocking probabilities. Several methods are proposed on a small network structure.

The remainder of this chapter is structured as follows. In section 4.2, we provide a comprehensive literature review on performance evaluation in overflow loss networks and discuss our contribution to this field. In section 4.3 and 4.4, the considered systems are described and proposed methodologies are presented. The numerical analysis and comparison with other methods are presented in Section 4.5. Finally, concluding remarks and directions for future research are presented in Section 4.6.

4.2 Literature Review

In literature, there have been numerous efforts to model and analyze blocking probabilities in overflow loss networks; most of the studies on this area mainly belong to telecommunication networks literature. There is also a significant body of literature on overflow models for ambulance location called hypercube models. Small size overflow networks are usually modeled as multi-dimensional Markov processes and analyzed with queuing theory. However, as system grows, the state space increases gradually and the modeling remains unmanageable with Markov processes. Approx-

imations therefore play an important role in estimation of blocking probabilities.

The simplest approach is Erlang's Fixed point approximation (EFPA). The main idea behind, is to combine all streams arriving to a server into a marginal stream and compute the blocking probability by using the well-known Erlang Loss formula. The overflow of each arriving stream is then decomposed imposing the computed blocking probability on each one. EFPA is well-known to be inaccurate for overflow loss networks because it characterizes the traffic offered by any stream as if it were a Poisson process even though an overflow process arrives with a higher variability (a greater peakedness) compared to a Poisson process ([Wong 2007]). Furthermore, it ignores the possible statistical dependence among servers while computing the being busy probability of each server. Various strengthened EFPA are developed and introduced in literature ([Wong 2007]).

Wilkinson [Wilkinson 1956] studied loss systems and described the parameters to characterize an overflow stream as mean and variance. With **"equivalent random method"**, he introduced the concept of peakedness, which is not sufficient for a complete characterization of an overflow stream, yet it has been found to be useful in many applications. Peakedness of a stream is defined as variance-to-mean ratio of the distribution of the number of busy servers on an infinite server group processing this traffic with an exponential service time [Fredericks 1980]. The peakedness of a Poisson stream is 1, while the peakedness of an overflow stream is always greater than 1. With equivalent random method, the parameters (mean load (a') and peakedness z') of an overflow stream which originates from a system where a Poisson stream with mean load $a = \lambda/\mu$ fed to a station with N number of exponential servers can be calculated as

$$a' = a * B(N, a)$$
$$z' = 1 - a' + \frac{a}{N - a + a' + 1} > 1$$

where $B(N, a)$ is the Erlang Loss probability

Hayward in 1959 used the peakedness concept to estimate the blocking probabilities in a system where the arrivals are overflow streams. He developed "**Hayward approximation**" which can be seen as a scaling of Erlang's loss formula taking into account the variability of the incoming overflow stream [Chevalier 2003]. An overflow incoming stream with mean load a and peakedness z arriving to a station with N servers can be approximated by a Poisson traffic stream with intensity a/z and peakedness 1 arriving to a station with N/z servers. That is, Hayward approximation estimates the blocking probability of an overflow incoming stream with $B(N, a, z)$ $B(N/z, a/z, 1)$. [Fredericks 1980] developed a natural extension to the Hayward approximation for estimating blocking in a more general system and determination of quantities of interest other than blocking. They showed that Hayward

approximation underestimate blocking for light load.

Neal Scott ([Neal 1971]) extended the "**equivalent random method**" to take into account the correlation among arriving streams. He considered multiple stations with finite servers and additionally two infinite server stations in parallel. Finite server stations receive various independent groups of customers and the arrivals finding all servers busy in a station overflows to other stations and eventually to the two infinite server stations. This system is modeled as Markov system of equations which consist infinite set of equations. Neal showed that it suffices to know the various moments of the distribution of the number of busy servers on two infinite server stations $(L1, L2)$. He used two-dimensional binomial-moment generating function to calculate the necessary moments from where he was able to assess the correlation between $L1$ and $L2$, thus the correlation between two overflow streams of the system. Our proposed method is motivated from this extension.

Recently Chevalier et al. ([Chevalier 2003]) evaluate the performance of a call center with different class of calls where the customer who finds all lines busy leaves the system. They mainly developed their algorithm around the approximation method proposed by Hayward and extended by Fredericks. The overflow calls arriving to a pool of servers are aggregated and an aggregated blocking probability is computed in each station by using Hayward's approximation of blocking. Finally, they evaluate the system performance by computing the system out rejection probability.

Afterwards, Franx et al. ([Franx 2006]) addressed a multi-class blocking system where different class of jobs flow through stations according to a fixed overflow policy. First time in literature, authors were interested in the individual blocking probabilities of each class. They approximated each inter-overflow time with a hyperexponential distribution and they proposed an approximation method to calculate the higher moments and eventually the blocking probabilities for each class.

Most recently Huang et al. ([Huang 2008]) developed an approximation method (MOA) for loss calculation in hierarchical networks with many arrivals sharing a common server group. Their approximation method is based on blocking probabilities matching to calculate the variances of each arriving stream.

Until here, we summarize the studies considering peakedness and overflow stream characterization. On the other side, there are some studies based on estimating the busy fractions of each server in order to evaluate system performance. In [Larson 1974], Larson studied the behavior of a multi-server queuing system with distinguishable servers and proposed an approximation hypercube (AH) model which estimates the fractions of time a server is busy(ρ_i) and dispatching probabilities (f_{im}) (probability a customer type m is assigned to server i) assuming one server at each station and exponential service times. The approximation is derived by

treating server workloads as being independent and adjusting this error through a correction factor $Q(N, \rho, k-1)$ which is a function of number of servers N, traffic intensity ρ and server preference k. Then, they defined "**dispatching probability**" as the probability of $(k-1)$ busy servers followed by an idle server when sampling at random without replacement from the identical servers:

$$f_{im} = Q(N, \rho, k-1)(1-\rho_i) \prod_{l=1}^{k-1} \rho_{a_{ml}}$$

where a_{mk} is k^{th} preferred server of customer m. Larson's approximation procedure is derived by noting $\rho_i = R_i.\tau$ where R_i is the total arrival rate assigned to server i ($R_i = \sum_m \lambda_m f_{im}$) and τ is service time. Two equations are combined to give an iterative procedure for determining busy fractions ρ_i.

In [Jarvis 1985], Jarvis extended the AH model to allow average service times to depend on the server and customer location. Recently Budge et al. ([Budge 2009]) extended the algorithm by allowing multiple servers at a station. They introduce a new correction factor based on random sampling of stations.

In the first part of the study, we consider an acyclic network with several stations with multiple servers, and several customer groups (Poisson and/or overflow streams) arriving to each station where two customer groups arriving to the same station may split and overflow to different stations. In such an overflow loss system, there exist three key issues. First, as in all overflow loss systems, overflow streams have higher variability than Poisson streams and Erlang-loss formula cannot be used for the blocking probability calculation. To address this issue, we characterize each overflow stream as an Interrupted Poisson Process (IPP) which allows us to assume a Markovian system.

Second, the multiple customer streams arriving to one station share the common server capacity of this station, thus overflow characterization of each of those streams is not evident and resultant overflow streams are statistically correlated. The most of the existing studies ([Chevalier 2003] and [Fredericks 1980]) consider one arrival stream to a station and estimate one blocking probability for that stream. The underlying theory in those studies is the equivalent random theory which is not useful for analyzing correlated overflows. Recently Huang et al. [Huang 2008] addressed this correlation by developing an approximation algorithm to compute variance of each overflow stream separately yet still considering one blocking probability. Moreover, in their approximation all incoming streams are assumed Poisson and Erlang-loss probability is utilized for blocking probability matching. To the best of our knowledge, the only existing study that analyzes each overflow stream individually and calculates a blocking probability for each is the work of [Franx 2006]. Independently from this work, we further address this problem on multi-class systems with more complex overflow routing. We develop an approximation method (BinomIPP)

built upon a markovian system with IPPs and binomial moment matching mainly motivated from Neal's extention ([Neal 1971]). With the proposed methodology, we estimate the blocking probability perceived by each customer group (each arrival stream separately) at each station in an overflow network, from where more accurate overflow stream characterization could be achieved.

Third, there might exist correlation among stations mainly dependent on the overflow structure (routing rules) among stations in the network. In an acyclic network where lower tiers are overflowing to higher ones, this type of correlation is expected to be quite low, thus in the proposed method we focus on each station separately assuming independency.

In the second part of the study, we consider a feedback (cyclic) overflow loss network where customers may overflow to any station in the network (forward and backward routings are allowed). In such a system, in addition to the two key source of errors existing in an acyclic overflow network (described above), there exists also high correlation among stations that has never been considered before in the literature to the best of our knowledge. For a small size instance, we proposed some approximation methodologies and an iterative version of BinomIPP for evaluating the performance of such a system.

4.3 Methodology for Acyclic Overflow Network

In this section, we consider an acyclic overflow loss network where customers may overflow from lower-tiers to higher-tiers regarding a forward routing (no backward overflow is considered). This type of routing is commonly observed in our application field (perinatal networks) where patients overflow from small hospitals to generally bigger and centralized ones.

Consider a loss network which is composed of I number of stations each with N_i-servers ($N_i > 1$). Each station i is fed by an independent Poisson arrival stream (λ_i) which is called "offered stream". A stream rejected from a station (and overflows to another station) is called "overflow stream". An offered stream may overflow several times to different stations. An arriving customer finding all N_i-servers busy in a station i overflows to another station according to a forward routing rule. The structure of an example network with a forward routing rule is illustrated in Figure 4.1.

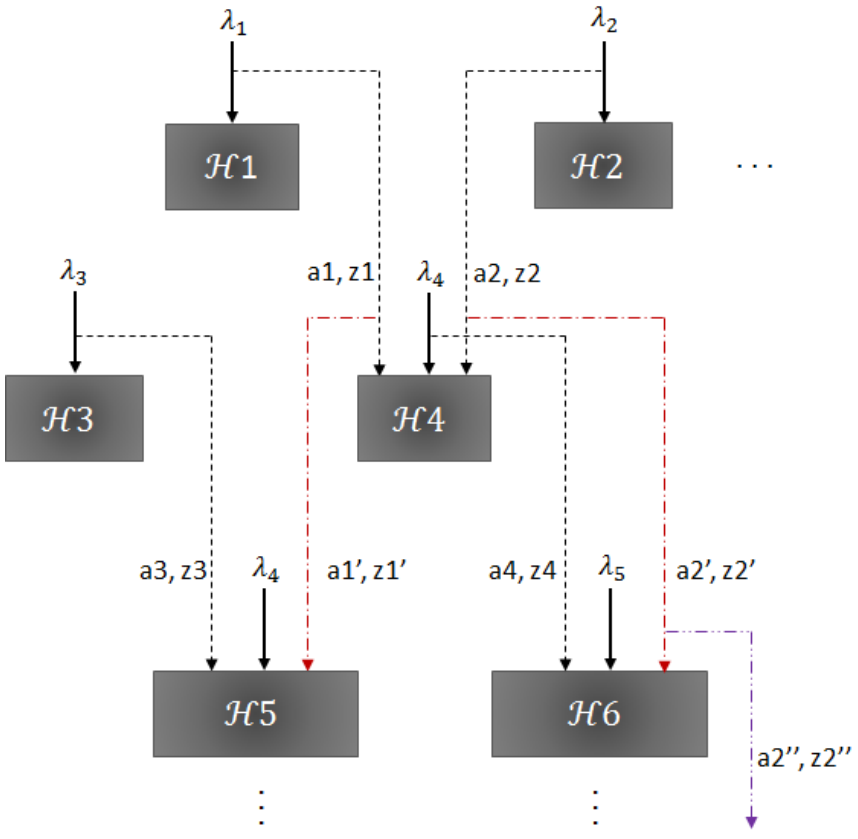

Figure 4.1: Representation of an Acyclic Overflow Loss Network

It is important to note that two customer groups arriving to the same station do

not have to overflow to the same station (see H4 in Fig 4.1). Due to this property, we need to follow each stream separately but in relation with other streams since they share the common server capacity.

In such a system we want to calculate the blocking probability of each stream in each station where there might exist more than one incoming overflow streams which are more variable than Poisson. In the following subsections, we describe the proposed methodology. First, we describe how to approximate an arriving overflow stream as an IPP process. Second, we present our methodology to estimate blocking probabilities and other entities of interest (higher moments) for each stream. And finally, we describe how we use the three moment matching method for characterizing the outgoing overflow stream of each incoming stream.

4.3.1 Approximating overflow streams with IPP

An Interrupted Poisson Process (IPP) can be defined as a Poisson Process with a random on/off switch which is alternately turned on for an exponentially distributed time $(1/\gamma)$ and then turned off for another independent exponentially distributed time $(1/\mathrm{w})$. In literature, in many studies, IPP is considered as a simple and accurate method for approximating overflow traffic. Approximation is obtained by matching either the first two or three moments of an interrupted poisson process to those of an overflow process ([Kuczura 1973]).

Let, λ be the Poisson arrival rate, $(1/\gamma)$ be the mean on-time, $(1/\mathrm{w})$ be the mean off-time of the random switch, $1/\mu$ be the mean service time (it is normalized $(\mu = 1)$ in this study). Let the IPP stream offered to an infinite server station. Let random variable l denotes the number of busy servers in infinite server station. From distribution of l, one can obtain the parameters; mean offered load to the infinite server station (a) and variance of this stream (v) which together characterize an overflow process match of IPP.

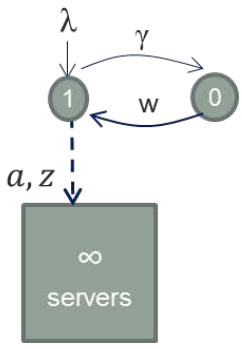

Figure 4.2: Representation of an IPP stream

Let the state of the system described by (l, α) where α is the state of the switch

taking on the value of 1 or 0 according to whether the process is on or off. System is illustrated in Fig 4.2.

The balance equations for the stationary state probabilities are:

$$\begin{aligned}(\lambda+\gamma)p(0,1) &= wp(0,0)+\mu p(1,1)\\ (\lambda+\gamma+l\mu)p(l,1) &= wp(l,0)+(l+1)\mu p(l+1,1)+\lambda p(l-1,1) \quad l\geq 1\\ (w+l\mu)p(l,0) &= \gamma p(l,1)+(l+1)\mu p(l+1,0) \quad l\geq 0\end{aligned}$$

We are interested in the moments of the distribution $f(L)$ of the number of busy servers in the infinite server station. In [Kuczura 1973] these system of equations are solved and the factorial moments of L are given as:

$$\begin{aligned}B(1) &= E(L) = \frac{\lambda w}{(w+\gamma)}\\ B(2) &= E((L(L-1))/2!\,) = \frac{\lambda^2}{2!}\frac{w}{(w+\gamma)}\frac{w+1}{(w+\gamma+1)}\\ B(3) &= E((L(L-1)(L-2))/3!\,) = \frac{\lambda^3}{3!}\frac{w}{(w+\gamma)}\frac{w+1}{(w+\gamma+1)}\frac{w+2}{(w+\gamma+2)}\end{aligned}$$

From the 1st and the 2nd moment of $f(L)$, the overflow stream mean (a) and peakedness (z), that can approximate the input IPP process, can be defined as:

$$\begin{aligned}a &= E(L)\\ z &= \frac{Var(L)}{E(L)} = 1+a\left(\frac{(w+1)}{(w+\gamma+1)}-\frac{w}{(w+\gamma)}\right)\geqslant 1\end{aligned}$$

Above we compute the parameters of an overflow process (a,z) from its IPP process defined with (λ,γ,w). In this study we need to proceed the opposite way such that we need to define IPP parameters of a known overflow stream (a,z). In this, we have three parameters (λ,γ,w) to choose so that IPP gives the best approximation to the overflow process. For that, we used Rapp's approximation given in [Kuczura 1973] as:

$$\begin{aligned}\lambda &= az+3z(z-1)\\ w &= \frac{a}{\lambda}\left(\frac{(\lambda-a)}{(z-1)}-1\right)\\ \gamma &= \left(\frac{\lambda}{a}-1\right)w\end{aligned}$$

4.3.2 Approximation of Blocking Probabilities with BinomIPP

In the proposed approach, each station is treated independently. We call the proposed approximation methodology **BinomIPP** which is applied to each station consecutively starting from the lower-tiers. The developed method is based on a Markov Process where overflow arrival streams are approximated with IPPs. In the following, we describe our proposed methodology focusing on one station regarding its both Poisson and IPP arrivals.

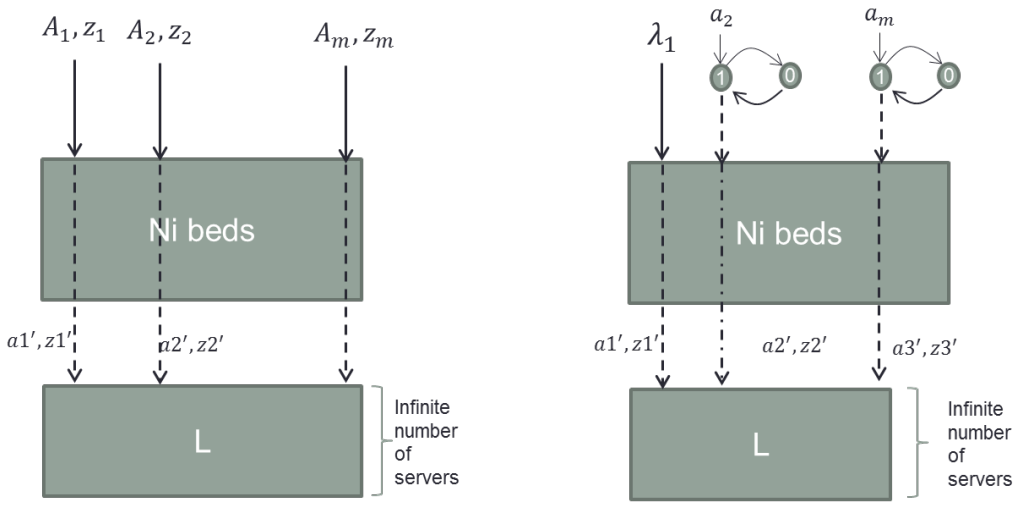

Figure 4.3: Representation of methodology on one station

Consider a service system i which consists a station $\mathcal{N}$ with N_i number of servers and an artificial station $\mathcal{L}$ where there are infinite number of servers. The arrivals to station N is generated by m independent streams with a load intensity of $A_1, A_2, ...A_m$ and peakedness $z_1, z_2, ...z_m$ where $z \geqslant 1$. Arrival streams with $z > 1$ (higher variability than Poisson), are defined as interrupted poisson processes (IPPs) each with an on/off switch (as described in the previous section). A customer arrival can occur only when the switch is on.

If a customer from any group finds all servers busy in station $\mathcal{N}$, then s/he is served by one of the servers in station $\mathcal{L}$. All rejected patients from station $\mathcal{N}$ will be served in station $\mathcal{L}$.

We define this system as a multidimensional Markov Chain with a state space $(p_1, p_2, ..., p_m, n, l)$ where p_m is binary parameter $\{0, 1\}$ denoting the random switch is on(1) / off(0) for each IPP stream (a_m, γ_m, w_m). n and l denotes the number of busy servers in station $\mathcal{N}$ and $\mathcal{L}$ respectively. For simplifying the presentation of the equations, we consider that there are $m = 3$ incoming arrival streams (one Poisson and two IPPs) to the considered station. The balance equations for this system can

be written as:

For $0 \leqslant n < N_i$

$$P(1,1,n,l)(a_1+\gamma_1+a_2+\gamma_2+\lambda+n+l) = P(0,1,n,l)w_1 + P(1,0,n,l)w_2 + P(1,1,n-1,l)(a_1+a_2+\lambda) + P(1,1,n+1,l)(n+1) + P(1,1,n,l+1)(l+1) \tag{4.1}$$

$$P(1,0,n,l)(a_1+\gamma_1+w_2+\lambda+n+l) = P(0,0,n,l)w_1 + P(1,1,n,l)\gamma_2 + P(1,0,n-1,l)(a_1+\lambda) + P(1,0,n+1,l)(n+1) + P(1,0,n,l+1)(l+1) \tag{4.2}$$

$$P(0,1,n,l)(w_1+a_2+\gamma_2+\lambda+n+l) = P(1,1,n,l)\gamma_1 + P(0,0,n,l)w_2 + P(0,1,n-1,l)(a_2+\lambda) + P(0,1,n+1,l)(n+1) + P(0,1,n,l+1)(l+1) \tag{4.3}$$

$$P(0,0,n,l)(w_1+w_2+\lambda+n+l) = P(1,0,n,l)\gamma_1 + P(0,1,n,l)\gamma_2 + P(0,0,n-1,l)(\lambda) + P(0,0,n+1,l)(n+1) + P(0,0,n,l+1)(l+1) \tag{4.4}$$

Similar relations hold on the boundary of the state space

For $n = N$

$$P(1,1,N,l)(a_1+\gamma_1+a_2+\gamma_2+\lambda+N+l) = P(0,1,N,l)w_1 + P(1,0,N,l)w_2 + P(1,1,N-1,l)(a_1+a_2+\lambda) + P(1,1,N,l-1)(a_1+a_2+\lambda) + P(1,1,N,l+1)(l+1) \tag{4.5}$$

$$P(1,0,N,l)(a_1+\gamma_1+w_2+\lambda+N+l) = P(0,0,N,l)w_1 + P(1,1,N,l)\gamma_2 + P(1,0,N-1,l)(a_1+\lambda) + P(1,0,N,l-1)(a_1+\lambda) + P(1,0,N,l+1)(l+1) \tag{4.6}$$

$$P(0,1,N,l)(w_1+a_2+\gamma_2+\lambda+N+l) = P(1,1,N,l)\gamma_1 + P(0,0,N,l)w_2 + P(0,1,N-1,l)(a_2+\lambda) + P(0,1,N,l-1)(a_2+\lambda) + P(0,1,N,l+1)(l+1) \tag{4.7}$$

$$P(0,0,N,l)(w_1+w_2+\lambda+N+l) = P(1,0,N,l)\gamma_1 + P(0,1,N,l)\gamma_2 + P(0,0,N-1,l)(\lambda) + P(0,0,N,l-1)(\lambda) + P(0,0,N,l+1)(l+1) \tag{4.8}$$

where $P(p_1,p_2,n,l)$ is the probability of being in state (p_1,p_2,n,l).

The above balance equations are quite difficult to solve due to the existence of an infinite set L. Furthermore, instead of an explicit solution of above equations, being able to calculate various moments of the random variable l is sufficient for us to proceed and characterize the overflow streams. As proposed in ([Neal 1971]), moments of the random variable l can be obtained by using a binomial-moment

generating function which can be defined as:

$$B(p_1, p_2, ..., p_m, n; x) = \sum_{l=0}^{\infty} P(p_1, p_2, ..., p_m, n, l)(1+x)^l \tag{4.9}$$

and binomial moments are defined as:

$$B_l(p_1, p_2, ..., p_m, n) = \sum_{k=l}^{\infty} \binom{k}{l} P(p_1, p_2, ..., p_m, n, k) \tag{4.10}$$

Thus, l^{th} moment of $f(L)$: distribution of the number of busy servers in the infinite server station can be written as:

$$\begin{aligned} B_l = E\binom{L}{l} &= \sum_{p_1}\sum_{p_2}...\sum_{p_m}\sum_{n} B_l(p_1, p_2, ..., p_m, n) \\ &= \sum_{p_1}\sum_{p_2}...\sum_{p_m}\sum_{n}\sum_{k=l}^{\infty} \binom{k}{l} P(p_1, p_2, ..., p_m, n, k) \end{aligned}$$

For $l = 1$ (1st moment) : $B_1 = E\binom{L}{1} = E(L)$

For $l = 2$ (2nd moment): $B_2 = E\binom{L}{2} = E\left(\frac{(L(L-1))}{2!}\right)$

For $l = 3$ (3rd moment) : $B_3 = E\binom{L}{3} = E\left(\frac{(L(L-1)(L-2))}{3!}\right)$

From moment information, several properties of overflow streams can be achieved. From the 1st and 2nd moment, mean (a) and variance (v) and consequently peakedness (z) of the outgoing overflow streams can be calculated as in the following:

$$\begin{aligned} a &= B_1 \\ v &= 2B_2 + B_1 - B_1^2 \\ z &= \frac{(2B_2 + B_1 - B_1^2)}{B_1} \end{aligned}$$

However, in order to incorporate correlation among streams, we want to characterize the outgoing overflow stream of each arrival stream separately, yet considering their dependency. For that, different moments for each stream is required to be extracted from the above equations.

Relations for binomial moments are obtained by multiplying both sides of above state balance equations by $(1+x)^l$ and summing on l. The complete derivation is given in Appendix D. Eventually the following finite system of equations are achieved:

For $0 \leqslant n < N_i$

$$B_l(1,1,n)(a_1+\gamma_1+a_2+\gamma_2+\lambda+n+\boldsymbol{l}) = B_l(0,1,n)w_1 + B_l(1,0,n)w_2 + B_l(1,1,n-1)(a_1+a_2+\lambda) + B_l(1,1,n+1)(n+1)$$

$$\begin{aligned} B_l(1,0,n)(a_1+\gamma_1+w_2+\lambda+n+\boldsymbol{l}) = {} & B_l(0,0,n)w_1 + B_l(1,1,n)\gamma_2 \\ & + B_l(1,0,n-1)(a_1+\lambda) \\ & + B_l(1,0,n+1)(n+1) \end{aligned}$$

$$\begin{aligned} B_l(0,1,n)(w_1+a_2+\gamma_2+\lambda+n+\boldsymbol{l}) = {} & B_l(1,1,n)\gamma_1 + B_l(0,0,n)w_2 \\ & + B_l(0,1,n-1)(a_2+\lambda) \\ & + B_l(0,1,n+1)(n+1) \end{aligned}$$

$$\begin{aligned} B_l(0,0,n)(w_1+w_2+\lambda+n+\boldsymbol{l}) = {} & B_l(1,0,n)\gamma_1 + B_l(0,1,n)\gamma_2 \\ & + B_l(0,0,n-1)(\lambda) + B_l(0,0,n+1)(n+1) \end{aligned}$$

For $n = N$

$$\begin{aligned} B_l(1,1,N)(\gamma_1+\gamma_2+N+l) = {} & B_l(0,1,N)w_1 + B_l(1,0,N)w_2 \\ & + B_l(1,1,N-1)(a_1+a_2+\lambda) \\ & + B_{l-1}(1,1,N)(a_1+a_2+\lambda) \end{aligned}$$

$$\begin{aligned} B_l(1,0,N)(\gamma_1+w_2+N+l) = {} & B_l(0,0,N)w_1 + B_l(1,1,N)\gamma_2 \\ & + B_l(1,0,N-1)(a_1+\lambda) + B_{l-1}(1,0,N)(a_1+\lambda) \end{aligned}$$

$$\begin{aligned} B_l(0,1,N)(w_1+\gamma_2+N+l) = {} & B_l(1,1,N)\gamma_1 + B_l(0,0,N)w_2 \\ & + B_l(0,1,N-1)(a_2+\lambda) + B_{l-1}(0,1,N)(a_2+\lambda) \end{aligned}$$

$$\begin{aligned} B_l(0,0,N)(w_1+w_2+N+l) = {} & B_l(1,0,N)\gamma_1 + B_l(0,1,N)\gamma_2 + B_l(0,0,N-1)(\lambda) \\ & + B_{l-1}(0,0,N)(\lambda) \end{aligned}$$

For $l = 0$; with the addition of following constraint,

$$\sum_{p_1}\sum_{p_2}\sum_{n} B_0(p_1,p_2,n) = 1$$

$B_0(p_1,p_2,n)$ is the probability of being in state (p_1,p_2,n) and the above finite equations have an exact solution.

From the state probabilities of system being full (p_1,p_2,N), the blocking probability for each arriving stream can be calculated. System rejects a customer from IPP(i) stream only if there is an arrival (switch is on with $p_i = 1$) when the system is full ($n = N$). Whereas a customer from Poisson stream is rejected whenever system is full ($n = N$). The blocking probabilities for each arriving stream can be written as:

- Blocking probability experienced by Poisson stream:
  $B_0 = \sum_{p_1} \sum_{p_2} B_0(p_1, p_2, N)$
- Blocking probability experienced by IPP stream 1:
  $B_0^{IPP1} = \sum_{p_2} B_0(1, p_2, N)$
- Blocking probability experienced by IPP stream 2:
  $B_0^{IPP2} = \sum_{p_1} B_0(p_1, 1, N)$

For $l = 1$, the mean outgoing (overflowed) load of each incoming stream can be written as:

- Mean blocked arrivals of Poisson stream:
  $B_1 = \lambda \sum_{p_1} \sum_{p_2} B_0(p_1, p_2, N) = \lambda B_0$
- Mean blocked arrivals of IPP1 stream:
  $B_1^{IPP1} = a_1 \sum_{p_2} B_0(1, p_2, N) = a_1 B_0^{IPP1}$
- Mean blocked arrivals of IPP2 stream:
  $B_1^{IPP2} = a_2 \sum_{p_1} B_0(p_1, 1, N) = a_2 B_0^{IPP2}$

In addition to the 1^{st} order moments, we can also obtain higher order moments for each stream by using a very useful equation that can be obtained by summing all equations up (through 1-8). From recursive equations, we can obtain higher moments for each stream as:

- For Poisson stream: $lB_l = \lambda \sum_{p_1} \sum_{p_2} B_{l-1}(p_1, p_2, N)$
- For IPP1 stream : $lB_l^{IPP1} = a_1 \sum_{p_2} B_{l-1}(1, p_2, N)$
- For IPP2 stream : $lB_l^{IPP2} = a_2 \sum_{p_1} B_{l-1}(p_1, 1, N)$

Several moments for each arrival stream m at station $\mathcal{N}$ can be computed by solving the presented finite system of linear equations recursively. By using moment relations, outgoing overflow stream of each incoming stream of station $\mathcal{N}$ can be characterized as described.

Remark: Computational complexity is a well-known drawback in Markov modeling. As the state space increases, the required number of computations increases exponentially and the problem easily becomes too complex to solve. Compared to other Markov modulated processes on this field in literature, our modeling has a strong point in considering one station at a time. Furthermore, IPP arrivals are represented in binary states. Except the r.v. n corresponds to the number of servers in the considered station, the rest of the r.v.s take binary values. The computational complexity of this Markov Chain is $O(2^m N)$. Binary representation allows us to obtain a computable MC even though we consider numerous arrival streams (high dimension MC).

4.3.3 Three Moment Matching

After an arrival is rejected from a station, s/he may seek for admission in another station. Formally, a rejected traffic (outgoing overflow stream) from a station $\mathcal{N}$ might be an incoming stream for another station $\mathcal{M}$, thus should be characterized as an IPP process. In previous subsections, the three moments of an IPP process (λ, γ, w) are presented and we described our BinomIPP methodology from where we can compute several moments of outgoing overflow stream. For characterizing this outgoing overflow stream as an IPP process, we match the moments of the two processes (see Fig. 4.4).

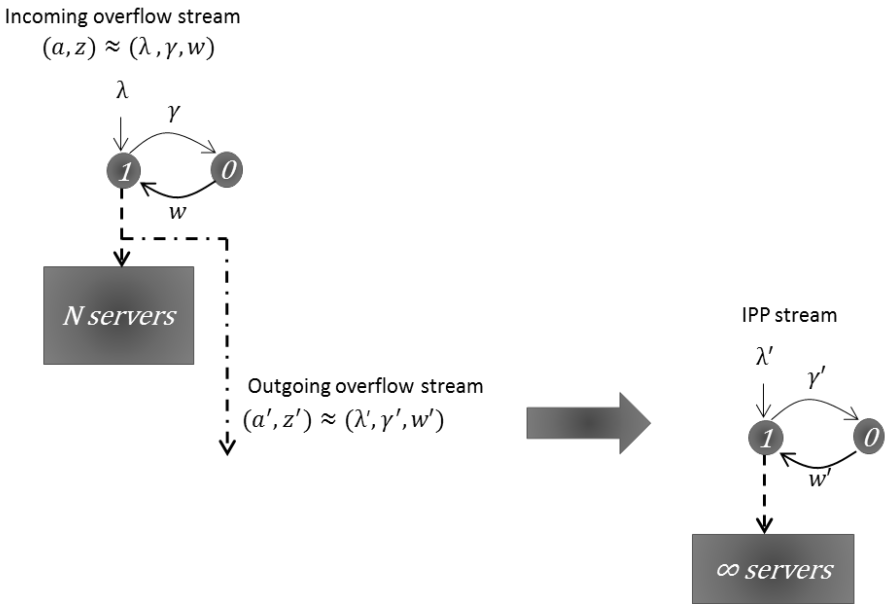

Figure 4.4: Characterization of an outgoing overflow stream

Thus, all incoming and outgoing overflow streams are defined as IPPs whose parameters are computed based on three-moment match. However it is important to note that, it is not always possible to achieve a one-to-one match in the 3rd moment, where it is taken as close as possible.

The proposed methodology is employed repetitively starting from the lower-tier station to higher-tier station following the order of overflows. For each station, the incoming streams are characterized as IPPs (except Poisson arrivals), the blocking probabilities and three moments are calculated via the proposed methodology, the outgoing overflow streams are characterized by using 3MM (three moment matching). We call this proposed methodology BinomIPP.

4.4 Methodology for Feedback Overflow Loss Network

In this section, we consider an overflow loss network where customers may overflow to any hospital such that both forward and backward overflows are allowed. Due to the overwhelming overflow relations among stations, in this section we build the theory and applications on a small network with 3 stations for facilitating the presentation and comprehension of the developed methodologies.

Consider a simple loss network which is composed of three stations each with N_i-servers ($N_i > 1$). Each station i is fed by an independent Poisson arrival stream (λ_i). A customer, finding all N_i-servers busy in station i, overflows to stations $j \in I - i$ within a given routing rule until she is accepted. A customer is rejected to out of network only if all servers in all stations are busy. The structure of an example network designed with a specific routing rule is illustrated in Figure 4.5.

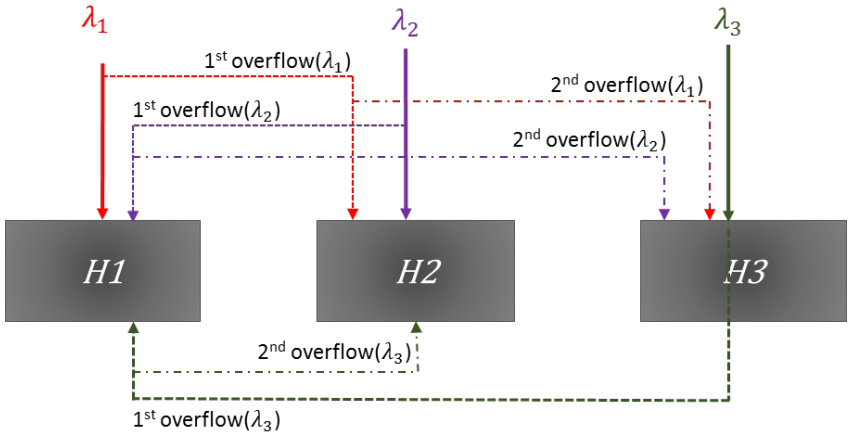

Figure 4.5: Representation of a 3 station Feedback Overflow Loss Network

Such a system where two stations can divert patients to each other involves some key issues which make its analysis highly complex. In the network with acyclic forward routing, it was possible to disregard the dependency among hospitals (each hospital was treated independently) while considering only the dependency among the arrival streams of each station thanks to the one direction routing. However, in the current system there exist a high correlation among both the arrival streams in each station and among the stations themselves.

For the evaluation of such a system and estimation of the rejection rates associated with each stream in each hospital, we worked on several approximation methods ranging from simplistic approaches to sophisticated ones. First, we proposed a simple yet very crude approach assuming Poisson arrivals and one-moment approximation of rejections based on well-known Erlang Loss formula. Second, we came up with a new approach where we aggregate some stations and employ some

probabilistic rules to define the dispatching probabilities between stations. Thirdly, we extended our BinomIPP method to an iterative BinomIPP method. In addition to the developed approximations, we also applied Fredericks and Chevalier approximation in order to assess its performance on such a correlated network.

Remark: The algorithms developed are highly dependent on the overflow routing rules defined among stations. We consider the following routing rule in remaining part of the study.
Station i overflows to station j then station k
Station j overflows to station i then station k
Station k overflows to station i then station j

4.4.1 A simple one-moment iterative approach

This simple method considers one combined offered load arriving to each station and proceeds through iteratively estimating blocking probabilities by using Erlang loss function.

Let a_{it} be the offered load to each station i at iteration t where for $t = 0, a_{i0} = \lambda_i$. B_{it} be the Erlang-Loss blocking probability in station i where

$$B_{it}(a_{it}, N_i) = \frac{(a_{it}^{N_i})/N_i!}{\sum_{n=0}^{N_i}(a_{it}^{n})/n!}$$

Thus, by assuming independency among stations being full, iterative equations are defined as (with normalized service rate $\mu = 1$):

$$\begin{aligned} a_{it} &= \lambda_i \\ &\quad + \lambda_j B_{j(t-1)}(a_{j(t-1)}, N_j) \\ &\quad + \lambda_k B_{k(t-1)}(a_{k(t-1)}, N_k) B_{j(t-1)}(a_{j(t-1)}, N_j) \end{aligned}$$

$$\begin{aligned} a_{jt} &= \lambda_j \\ &\quad + \lambda_i B_{i(t-1)}(a_{i(t-1)}, N_i) \\ &\quad + \lambda_k B_{k(t-1)}(a_{k(t-1)}, N_k) B_{i(t-1)}(a_{i(t-1)}, N_i) \end{aligned}$$

$$\begin{aligned} a_{kt} &= \lambda_k \\ &\quad + \lambda_j B_{j(t-1)}(a_{j(t-1)}, N_j) B_{i(t-1)}(a_{i(t-1)}, N_i) \\ &\quad + \lambda_i B_{i(t-1)}(a_{i(t-1)}, N_i) B_{j(t-1)}(a_{j(t-1)}, N_j) \end{aligned}$$

At each iteration t, the offered load to each hospital and consequently blocking probability estimations are updated till they converge.

This approach is crude in the sense that it considers neither the higher variability of overflow arrivals nor the possible statistical dependence between stations. Furthermore, it works with combined arrival traffic in each hospital therefore we cannot analyze each stream separately.

4.4.2 Aggregation Approach

We consider a three station overflow loss network to describe the aggregation approach. This approach proceeds via an iterative Markov Chain process. Each time, we focus on one station ($i \in I$) by leaving it alone while the other two stations (station $j, k \in I - i$) are merged and are behaved as one aggregated station (see Figure 4.6). While station i is focused, the dispatching probabilities from station j to i (f_{ji}) and from station k to i (f_{ki}) are approximated by imposing some probabilistic selection rules inside the aggregated station.

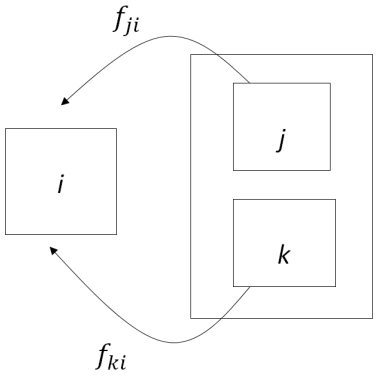

Figure 4.6: Aggregated Approach iteration focusing on hospital i

Three Markov Chain processes are constructed; each focusing on (leaving out) a different station $i = \{1, 2, 3\}$ where the state space is $S = \{(n_{focus}, n_{aggregated})$. For a 3 station network, we obtain three Markov modeling each with state space:

$$S = \{(n_i, n_{jk}) : 0 \leqslant n_i \leqslant N_i, 0 \leqslant n_{jk} \leqslant N_j + N_k\} \quad \forall i = \{1, 2, 3\} \text{ and } j, k \in I - i.$$

where n_i is the number of busy servers in station i and n_{jk} is the number of busy servers in the aggregated station. Finite set of equations of Markov Chain (i) can be written as:

$$(\lambda_i + \lambda_j + \lambda_k)P(0,0) = P(0,1) + P(1,0)$$

$$\begin{aligned}&(\lambda_i + \lambda_j + \lambda_k + n_i)P(n_i, 0)\\ &= \lambda_i P(n_i - 1, 0) + (n_i + 1)P(n_i + 1, 0) + P(n_i, 1) \quad \forall 0 < n_i < N_i\end{aligned}$$

$$\begin{aligned}&(\lambda_i + \lambda_j + \lambda_k + n_{jk})P(0, n_{jk})\\ &= \left[\lambda_j(1 - f_{ji}(n_{jk} - 1)) + (\lambda_k(1 - f_{ki}(n_{jk} - 1))\right] P(0, n_{jk} - 1) + P(1, n_{jk})\\ &+ (n_{jk} + 1)P(0, n_{jk} + 1) \quad \forall 0 < n_{jk} < N_j + N_k\end{aligned}$$

$$
\begin{aligned}
&(\lambda_i + \lambda_j + \lambda_k + n_i + n_{jk})P(n_i, n_{jk}) \\
&= \left[\lambda_j(1 - f_{ji}(n_{jk} - 1)) + (\lambda_k(1 - f_{ki}(n_{jk} - 1))\right] P(n_i, n_{jk} - 1) \\
&+ \left[\lambda_i + \lambda_j f_{ji}(n_{jk}) + \lambda_k f_{ki}(n_{jk})\right] P(n_i - 1, n_{jk}) \\
&+ (n_i + 1)P(n_i + 1, n_{jk}) + (n_{jk} + 1)P(n_i, n_{jk} + 1) \quad \forall 0 < n_{jk} < N_j + N_k
\end{aligned}
$$

$$
(\lambda_i + \lambda_j + \lambda_k + N_i)P(N_i, 0) = P(N_i, 1) + \lambda_i P(N_i - 1, 0)
$$

$$
\begin{aligned}
&(\lambda_i + \lambda_j + \lambda_k + N_{jk})P(0, N_{jk}) \\
&= P(1, N_{jk}) + \left[\lambda_j(1 - f_{ji}(N_{jk} - 1)) + (\lambda_k(1 - f_{ki}(N_{jk} - 1))\right] P(0, N_{jk} - 1)
\end{aligned}
$$

$$
\begin{aligned}
&(\lambda_i + \lambda_j + \lambda_k + n_i + N_{jk})P(n_i, N_{jk}) \\
&= \left[\lambda_j(1 - f_{ji}(N_{jk} - 1))(\lambda_k(1 - f_{ki}(N_{jk} - 1))\right] P(n_i, N_{jk} - 1) \\
&+ (\lambda_i + \lambda_j + \lambda_k)P(n_i - 1, N_{jk}) + (n_i + 1)P(n_i + 1, N_{jk}) \quad \forall 0 < n_i < N_i
\end{aligned}
$$

$$
\begin{aligned}
&(\lambda_i + \lambda_j + \lambda_k + N_i + n_{jk})P(N_i, n_{jk}) \\
&= \left[\lambda_i + \lambda_j(1 - f_{ji}(n_{jk} - 1)) + \lambda_k(1 - f_{ki}(n_{jk} - 1))\right] P(N_i - 1, n_{jk}) \\
&+ (\lambda_i + \lambda_j + \lambda_k)P(N_i, n_{jk} - 1) + (n_{jk} + 1)P(N_i, n_{jk} + 1) \quad \forall 0 < n_{jk} < N_j + N_k
\end{aligned}
$$

$$
\begin{aligned}
&(N_i + N_{jk})P(N_i, N_{jk}) \\
&= (\lambda_i + \lambda_j + \lambda_k)P(N_i - 1, N_{jk}) + (\lambda_i + \lambda_j + \lambda_k)P(N_i, N_{jk} - 1)
\end{aligned}
$$

From the results of each Markov Chain (i), steady state probabilities associated with the focused station i ($P(n_i)$: probability of having n_i number of busy servers in station i) can be obtained.

The key issue in such modeling is determining the rate of overflow from each station (j, k) in aggregated set to station i. We have the higher level information of busy servers in the aggregated set (n_{jk}), but we don't know how many of these busy servers belong to station j or k. In order to better approximate the overflow load arriving to station i from individual stations in aggregated set, we employ a probabilistic selection rule to estimate the being full probability of each station. Therefore, the dispatching (blocking) probabilities from station j to i ($f_{ji}(n_{jk})$) and from station k to i ($f_{ki}(n_{jk})$) are introduced in the MC equations to approximate the flow from these stations. The determination of these probabilities is based on a marginal distribution and a correction factor which is presented in the following.

Calculation of dispatching probability $f_{ji}(n_{jk})$:

The underlying assumption is that when station j has all servers busy, its customers overflow to station i. If the total number of busy servers of aggregated

station does not exceed the total number servers of station j ($n_{jk} < N_j$), station j cannot be full in any case, thus $f_{ji} = 0$.

$$f_{ji}(n_{jk}) = 0 \quad \forall n_{jk} < N_j$$

However, if $n_{jk} \geqslant N_j$, there is a possibility that station j is full. This is possible only if station j has $n_j = N_j$ number of busy servers and station k has the rest of the busy servers ($n_k = n_{jk} - N_j$). We define this possibility as in the following:

$$f_{ji}(n_{jk}) = \frac{P(n_j = N_j)P(n_j = N_j, n_k = n_{jk} - N_j)}{\sum_{\delta=max(n_{jk}-N_j)}^{N_j} P(n_j = \delta)P(n_j = \delta, n_k = n_{jk} - \delta)} \quad \forall n_{jk} \leqslant N_j$$

where $P(n_j)$ is the marginal distribution of having n_j number of beds occupied in station j which is updated at the end of each iteration. The distribution of total number of busy beds to station j and k is highly dependent on the relationship between these two stations. We incorporate this relation to the whole system through a correction factor based on the joint probability of having n_j number of beds occupied in station j while there is $n_k = n_{jk} - n_j$ number of beds occupied in station k which is denoted with $P(n_j, n_k = n_{jk} - n_j)$.

For calculation of the joint probabilities of stations k and j, with other words in order to incorporate the relation between those stations into the dispatching probabilities, a two state MC is constructed whose state space is (n_j, n_k). In this phase, we ignore the existence of station i and consider only the relation between the remaining stations j and k. Therefore, in MC we assume there is a full interaction between station j and k in order to incorporate their routing relations that we couldn't take into account before. Marginal probability of $P(n_j = N_j)$ is initially defined as an Erlang loss probability and updated in the end of each iteration of main Markov Chain (j) where the station j is the focused one.

Calculation of dispatching probability $f_{ki}(n_{jk})$:

The underlying assumption is that when station k has all servers busy, its customers first overflow to station j and then to station i. Therefore, for having an overflow stream from station k to i, the aggregated station needs to be completely full. Thus, the diversion probabilities from station k to station i can be defined as:

$$f_{ki}(n_{jk}) = 0 \quad \forall n_{jk} < N_j + N_k$$

$$f_{ki}(n_{jk}) = 1 \quad \forall n_{jk} = N_j + N_k$$

Algorithm 1 Iteration Algorithm for Aggregation Approach

Step 1. Initialize:
$t := 0$
for all $\forall i = \{1, 2, 3\}$ and $\forall j, k \in I - i$ **do**
$f^0_{ji}(n_{jk})$ and $f^0_{ki}(n_{jk})$ are initialized by using Erlang Loss formulas to calculate $P(n_j = N_j)$ and $P(n_k = N_k)$
$MC(n_i, n_{jk})$ is solved and steady state probabilities $P^0(n_i)$ are obtained.
Step 2. $t := t + 1$
for all $\forall i = \{1, 2, 3\}$ and $\forall j, k \in I - i$ **do**
$f^t_{ji}(n_{jk})$ and $f^t_{ki}(n_{jk})$ are are computed with $P^{t-1}(n_j)$ and $P^{t-1}(n_k)$ steady state probabilities obtained at iteration $(t - 1)$.
$MC(n_i, n_{jk})$ is solved and steady state probabilities $P^t(n_i)$ are obtained.
Step 3.
if steady state probabilities $P^t(n_i), \forall i = \{1, 2, 3\}$ are converged **then**
Stop
else
Go to step 2.

Remark: Numerical results indicate that this method computes very close results to the exact values. However, even though the method performs well for 3 station network, it is not evident how to extend the method in managing the aggregated station and the dispatching probabilities when higher number stations are involved.

4.4.3 Iterative BinomIPP

For 3 hospital fullfeedback overflow implementation, BinomIPP is required to be considered iteratively. BinomIPP has already a recursive algorithm that is to solve for each station consequtively. Due to the consideration of feedback overflows, this recursive algorithm is required to be employed iteratively until the convegence is achieved. The system considered is composed of 3 stations each with a Poisson arrival stream. The algorithm of iterative BinomIPP is presented as in the following:

Algorithm 2 Iteration Algorithm for BinomIPP

Initialization

For each station $i, j, k = \{1, 2, 3\}$ where $i \neq j \neq k$

Equivalent random method is employed independently to each station

Mean and variance of overflow streams in each station are computed.

Overflow streams are characterized with Rapps approximation and defined as IPP_i with (a_i, γ_i, w_i), IPP_j and IPP_k

Iteration Steps

Step 1:

Choose station i

BinomIPP is employed to calculate the moments of each arrival stream λ_i , IPP_j, IPP_k

Overflow streams are characterized with 3MM and defined as IPP_i ,IPP'_j and IPP'_k

Step 2:

Choose station j

BinomIPP is employed to calculate the moments of each stream λ_j, IPP_i, IPP'_k

Overflow streams are characterized with 3MM and defined as IPP_j ,IPP'_i

The third overflow of station k (IPP''_k) is rejected to out of network.

Step 3:

Choose station k

BinomIPP is employed to calculate the moments of each stream λ_k , IPP'_i, IPP'_j

Overflow stream IPP_k is characterized with 3MM.

The third overflows of station i and j (IPP''_i and IPP''_k) are rejected to out of network.

Step 4:

If Blocking probabilities are converged

Stop

Else

Go back to step 1.

4.5 Computational Results

In this section, we evaluate the performance of proposed methodologies and compare them with the performance of existing well performing methodologies over a set of instances.

For the implementation of methodologies, two sets of instances are created regarding different overflow structures; acyclic overflow and feedback overflow respectively. In addition to that, several traffic intensities are considered for each instance set to assess performance alterations of methodologies under different system load.

Acyclic Overflow:
First instance set considers a loss network composed of six hospitals where forward routing is employed (Fig. 4.7). As mentioned before, this type of routing is the most commonly observed overflow structure in healthcare networks such that patients commonly overflow from smaller hospitals to larger centralized ones.

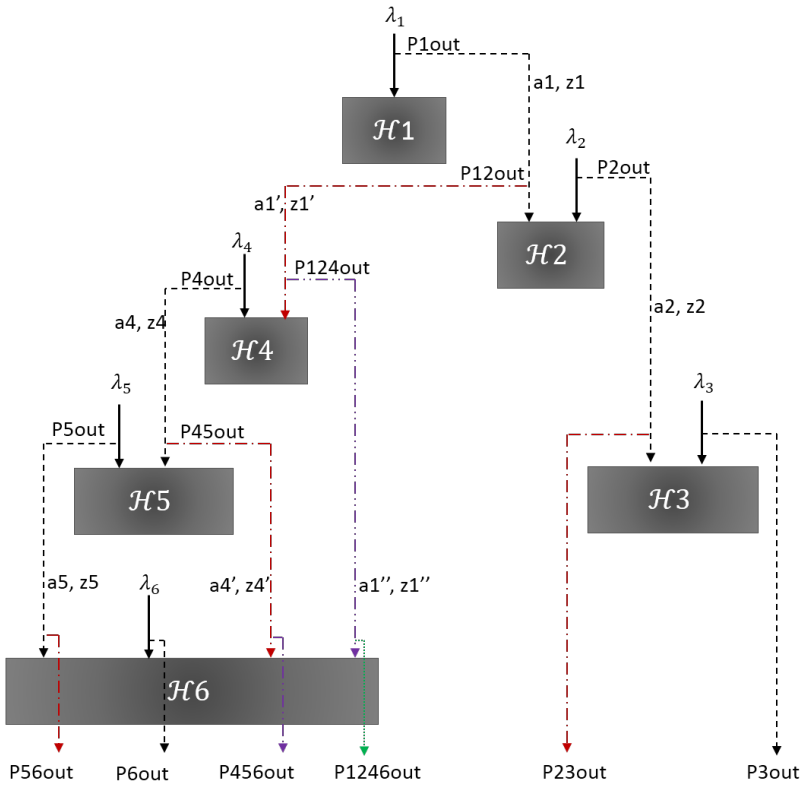

Figure 4.7: Acyclic Network Test Instance

Table 4.1: Test Data for the Acyclic Instance

Station i	Number of Servers (N_i)	Arrival Rates(λ_i)		
		Light Load	Medium Load	Heavy Load
1	5	4	10	15
2	10	9	15	20
3	25	20	28	60
4	20	18	22	40
5	25	22	24	60
6	40	30	35	80

We consider six stations and six different customer groups associated with each station. The routing structure for each customer group is observable in Fig 4.7. Different traffic intensities are introduced to represent light, medium and heavy load. Test Data is presented in Table 4.1. For comparison purposes, both the proposed method BinomIPP and the well-performing existing method Chevalier & Fredericks (C&F) are implemented on the test instances. The results of approximation methods are compared with the results of a discrete event simulation model (SIM) in order to evaluate the accuracy of estimation procedures. Since we are dealing with probabilities (possible to be very small) in order to increase the precision of the simulated model, we consider a 99% CI around loss probabilities. Thus, simulation outputs can be seen as exact results for most of the cases.

Our performance indicators are the blocking probabilities computed for each arrival stream in each station i. There exist 6 stations and 13 blocking probabilities which are represented as in the following:

P_{iout}	denotes the blocking probability of major stream i from station i (Poisson arrival)
P_{jiout}	denotes the blocking probability of stream j from station i
P_{kjiout}	denotes the blocking probability of stream kj from station i (stream rejected from station k and j previously)
$P_{mkjiout}$	denotes the blocking probability of stream mkj from station i (stream rejected from stations m, k and j previously)

The blocking probabilities for each arrival stream in each hospital (all performance indicators) computed with BinomIPP, C&F and SIM are presented in Tables 4.2, 4.3, 4.4 and Figures 4.8, 4.9, 4.10 respectively for heavy, medium and light load. The % relative differences of BinomIPP-SIM and C&F-SIM are presented in Figure 4.11 for all intensity of load.

From all Tables 4.2, 4.3, 4.4, it can be observed that BinomIPP outperforms C&F in all cases and additionally BinomIPP computes very close results to the ones

of simulation (small % rel. errors) for 1st blocking probabilities (P_{iout}) of a system under any intensity of load. However, as the number of overflow experienced by a stream increases, the overflow load tends to decrease while its variability, thus peakedness increases. This disturbs the precision of the approximation and % relative errors start to increase. From the comparison of Table 4.2(heavy load) and Table 4.4 (light load), it can be observed that both BinomIPP and C&F approximates better for heavy load case compared to light load case. This is because, in light load case, an arrival stream already carry a light load, so its overflow load can get very small even it might tend to zero. Considering that in such a light load case one station receives both its normal Poisson load and a very small overflow load with a high variability (which are considered together in blocking probability estimation) deteriorates the precision of the approximation. However this is not the case for heavy load system. Besides, it is very important to note that the comparisons are made with simulation, run under 99% CI setting which gave rise to very small half-length values around (0.005%-0.01%). For heavier load cases, simulation outputs can be considered as exact results whereas under lighter load cases, especially for higher order probabilities the computed half-lengths are not ignorable compared to the very small estimated mean. Thus, even though the % relative errors for higher order probabilities (P_{456}, P_{1246}) are computed very high (see Table 4.4), absolute deviation from simulation results is quite small (the simulation error is also big for such small-probability overflow events). Therefore, we can claim that this situation is a natural consequence of working with very small numbers.

As mentioned before, for stations receiving both a Poisson stream and various incoming streams already overflowed several times (such as station 6), the results tend to degrade more compared to others. For example in station 6, there is a Poisson arriving stream with a relatively high load compared to incoming overflow streams carrying small loads (there is even one overflowed already three times, thus carrying a load almost zero). Due to the fact that loads and variability of streams fed to this station have a large range, the precision of the approximations degrades. The system considered in light load case, gets effected from this situation more than the heavy load system since the loss load tends to zero very quickly in light load case. But still the absolute error remains tiny considering the percentage of rejections. (see Table 4.4).

On the other side, independent from the intensity of traffic, it is observed that C&F method overestimates the first rejection probabilities of a Poisson arrival stream (P_{iout}), while underestimates the higher order overflow stream rejections ($P_{mkjiout}$). This is possibly related with the fact that method combines all arrivals and computes one blocking probability for the aggregated load. Since we are working with very small numbers, this observation is presented better in Figure 4.11 where % relative errors are given. In terms of accurate estimation, BinomIPP seems to better differentiate the arriving streams, thus there is no tendency observed towards either over or under estimation.

Table 4.2: Comparison of Methods in Acyclic Network with Heavy Load

	Heavy Load				
	Blocking Probabilities			% Relative Error	
	BinomIPP	C&F	SIM	SIM-BinomIPP	SIM-C&F
P1out	69.32%	69.32%	69.30%	0.03%	0.03%
P2out	67.56%	68.67%	67.60%	-0.05%	1.58%
P3out	66.36%	66.78%	66.30%	0.09%	0.73%
P4out	58.76%	59.06%	58.70%	0.10%	0.62%
P5out	69.57%	70.66%	69.60%	-0.04%	1.53%
P6out	71.53%	72.30%	70.70%	1.17%	2.26%
P12out	48.95%	47.60%	48.90%	0.10%	-2.65%
P23out	45.95%	45.86%	46.10%	-0.32%	-0.52%
P45out	42.98%	41.74%	42.80%	0.42%	-2.49%
P56out	51.23%	51.09%	51.50%	-0.53%	-0.80%
P124out	30.06%	28.12%	30.30%	-0.79%	-7.21%
P456out	31.57%	30.17%	32.50%	-2.86%	-7.16%
P1246out	21.68%	20.33%	22.30%	-2.77%	-8.84%

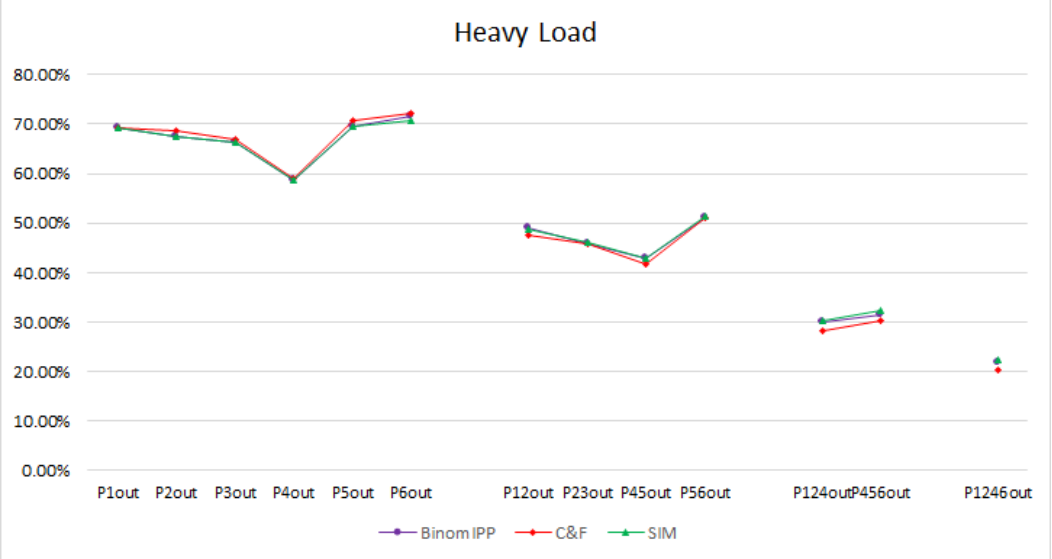

Figure 4.8: Comparison of resulting blocking probabilities obtained from different methods under heavy load acyclic network

Table 4.3: Comparison of Methods in Acyclic Network with Medium Load

	Mix Medium Load				
	Blocking Probabilities			% Relative Error	
	BinomIPP	C&F	SIM	SIM-BinomIPP	SIM-C&F
P1out	56.40%	56.40%	56.40%	-0.01%	-0.01%
P2out	53.91%	55.42%	54.00%	-0.17%	2.62%
P3out	34.41%	35.83%	34.40%	0.03%	4.16%
P4out	28.37%	28.66%	28.30%	0.24%	1.28%
P5out	23.73%	26.29%	23.90%	-0.72%	9.99%
P6out	16.37%	18.11%	16.00%	2.31%	13.21%
P12out	33.22%	31.25%	33.10%	0.35%	-5.58%
P23out	20.84%	19.86%	20.90%	-0.30%	-4.99%
P45out	9.59%	7.54%	9.20%	4.22%	-18.10%
P56out	5.75%	4.76%	6.20%	-7.22%	-23.20%
P124out	10.80%	8.96%	10.90%	-0.96%	-17.82%
P456out	2.08%	1.36%	2.80%	-25.88%	-51.25%
P1246out	1.95%	1.62%	2.40%	-18.78%	-32.39%

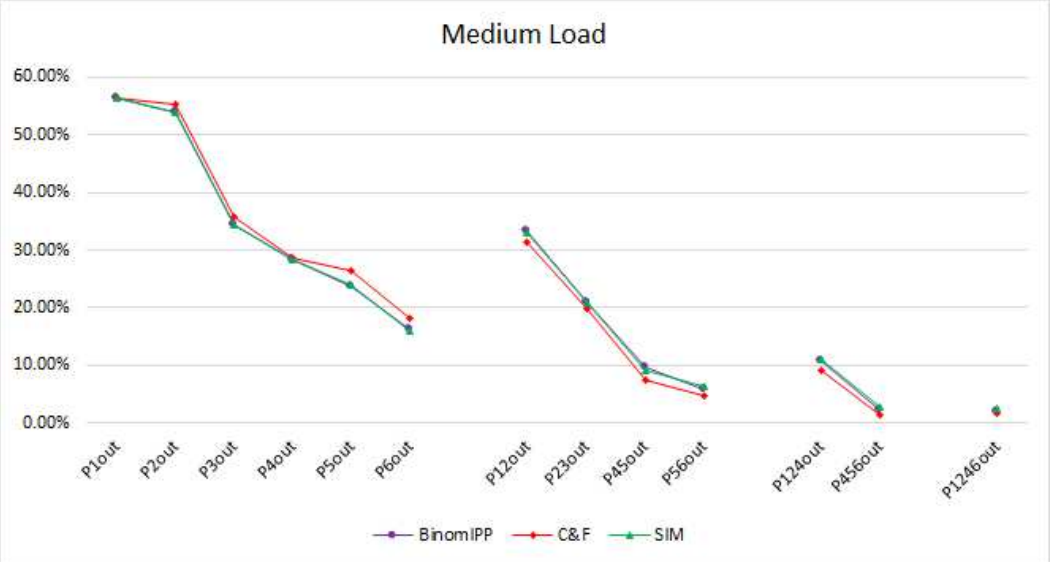

Figure 4.9: Comparison of resulting blocking probabilities obtained from different methods under medium load acyclic network

Table 4.4: Comparison of Methods in Acyclic Network with Light Load

	Light Load				
	Blocking Probabilities			% Relative Error	
	BinomIPP	C&F	SIM	SIM-BinomIPP	SIM-C&F
P1out	19.91%	19.91%	19.90%	0.03%	0.03%
P2out	20.34%	20.98%	20.40%	-0.31%	2.86%
P3out	8.18%	8.75%	8.20%	-0.21%	6.68%
P4out	11.45%	11.37%	11.50%	-0.48%	-1.10%
P5out	12.25%	13.32%	12.30%	-0.38%	8.27%
P6out	3.94%	4.47%	4.00%	-1.48%	11.66%
P12out	5.39%	4.18%	5.30%	1.72%	-21.19%
P23out	2.58%	1.84%	2.50%	3.33%	-26.58%
P45out	2.61%	1.51%	2.40%	8.95%	-36.89%
P56out	1.08%	0.59%	1.10%	-2.27%	-45.93%
P124out	0.74%	0.48%	0.70%	6.17%	-32.13%
P456out	0.17%	0.07%	0.28%	-41.03%	-75.84%
P1246out	0.02%	0.02%	0.01%	90.77%	90.99%

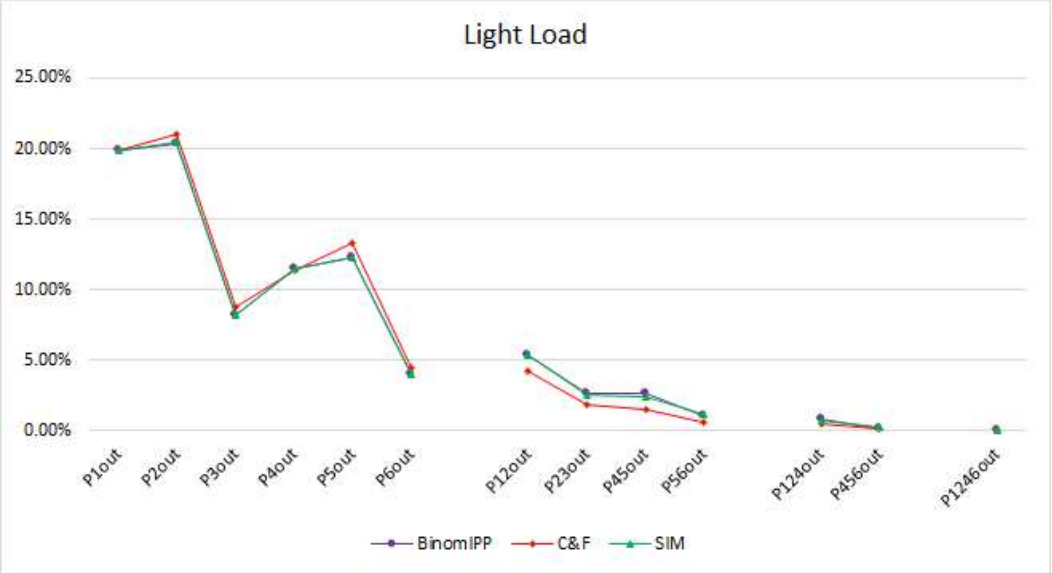

Figure 4.10: Comparison of resulting blocking probabilities obtained from different methods under light load acyclic network

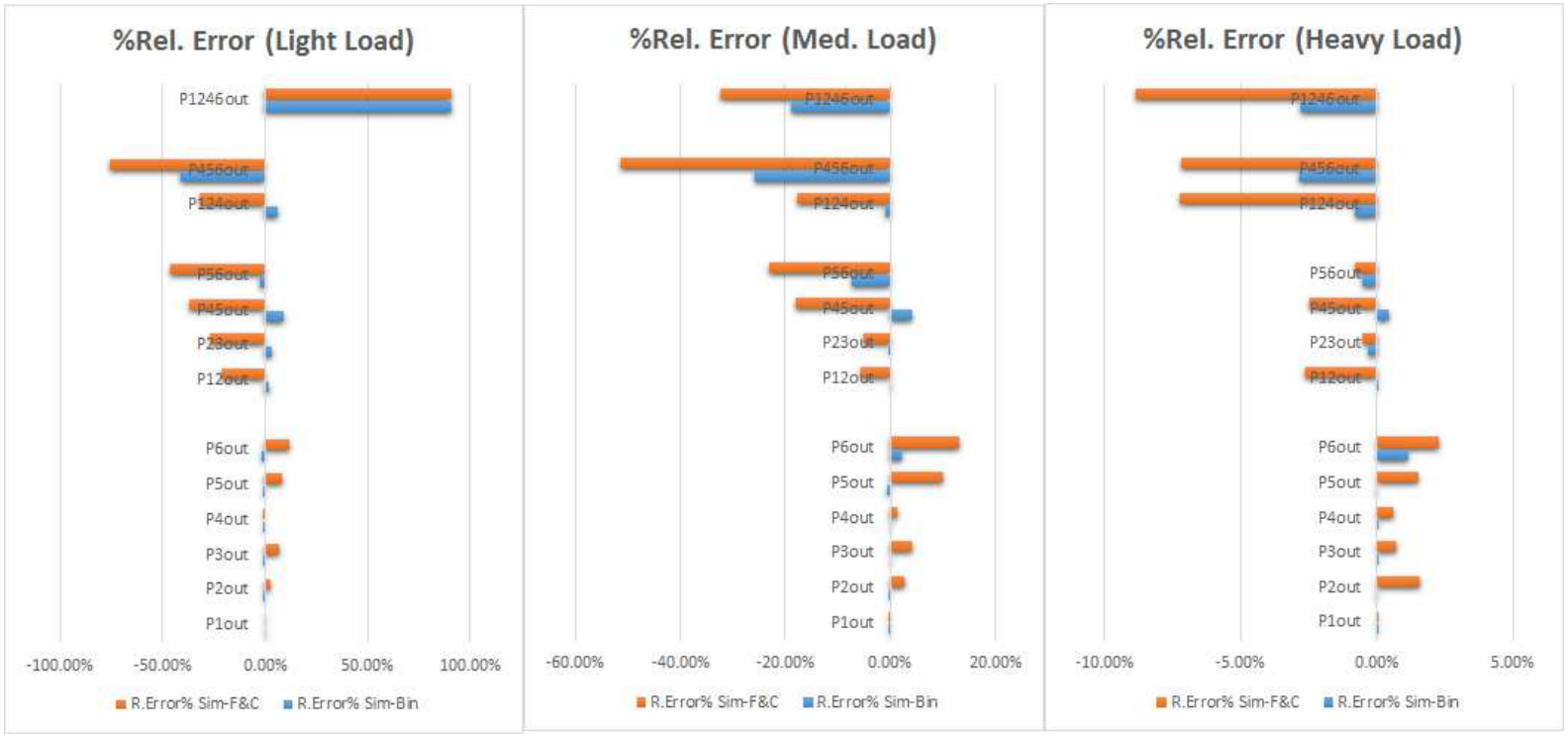

Figure 4.11: Comparison of Percentage Errors of BinomIPP and C&F relative to Simulation results computed for Light, Medium and Heavy load acyclic network respectively

Feedback Overflow:

This instance addresses a smaller network composed of three stations with feedback (backward and forward) overflow where a customer can be rejected out of network only if all stations are full. Test Data (arrival rates, number of servers and routing structure) is presented in Table 4.5. In the test instance we considered a load of medium intensity. On this instance, we evaluate the performance of proposed approximations: simple one-moment iterative approach (Model 1), aggregation approach (Model 2), iterative BinomIPP (Model 3) and the existing approach Chevalier&Fredericks (Model 4). The results of approximation methods are compared with the exact results which are obtained from the three state Markov Chain modeling of the considered instance. The performance indicators are blocking probabilities computed for each arrival stream in each station.

Table 4.5: Test Data for Feedback Overflow Network

	Arrival Rates	Number of Servers	Routing
Station 1	8	10	$\{1,2,3\}$
Station 2	4	5	$\{2,1,3\}$
Station 3	4	5	$\{3,1,2\}$

Exact results, the results obtained from each methodology and % relative errors for each method obtained from the comparison with exact results are presented in Table 4.6

Numerical results for this specific instance indicates that the aggregated approach (Model 2) computes much better estimates compared to the other methodologies. The proposed method performs quite well such that % relative errors are close to zero for most of the blocking probabilities. The highest error is computed as 3.3% for the blocking probability of station 3. Due to the presumed routing rule, station 3 receives the last overflow loads which are very light and highly variable compared to its own Poisson load. This might deteriorate the approximation.

As mentioned before, model 1 is a very crude approach working with the combined load arriving in each station. It works on each station and one aggregated incoming stream to each station. Therefore, it computes the 1^{st} blocking probabilities of each station whereas it is not possible to follow each arrival stream. Furthermore, as observed in Table 4.6, the model overestimates the blocking probability for each station. This error mainly originates from Poisson arrival assumption. On the other side, since the model does not consider each stream separately and higher order overflows, correlation among streams do not affect the results much. As the load gets heavier, the model starts to perform better such that the load of overflow streams gets heavier and deviation from Poisson decreases. Again as the load gets

lighter, the model starts to perform better such that the load of overflow streams tends to zero and for lighter load and in a combined load analysis a very small overflow load remains ignorable, thus Poisson originated error is smaller. Therefore, the approximations for light load case are better.

On the other side, neither model 3 nor model 4 achieved good performance on a feedback overflow routing case. This is mainly because both models perform weak in terms of estimating the correlations among the stations (BinomIPP considers each station independently). It is observed from this particular instance that both methods overestimate the 1st blocking probability and underestimate the higher order blocking probabilities.

Table 4.6: Results obtained from each methodology for feedback overflow network

	Exact Results	Model 1		Model 2		Model 3		Model 4	
	Blocking Prob.	Blocking Prob.	%Rel. Error	Blocking Prob.	%Rel. Error	Blocking Prob.	%Rel. Error	Blocking Prob.	%Rel. Error
P1out	**22.2%**	24.6%	**10.9%**	21.8%	**-1.5%**	23.7%	**6.8%**	26.9%	**21.6%**
P12out	**11.9%**	NA	**NA**	11.9%	**-0.3%**	11.4%	**-4.5%**	11.3%	**-5.6%**
P123out	**6.44%**	6.44%	**0.00%**	6.4%	**-0.5%**	4.9%	**-24.4%**	3.8%	**-41.6%**
P2out	**32.5%**	37.8%	**16.2%**	32.5%	**0.1%**	35.1%	**8.0%**	41.8%	**28.5%**
P21out	**11.9%**	NA	**NA**	11.9%	**-0.3%**	9.9%	**-16.7%**	11.3%	**-5.6%**
P213out	**6.44%**	6.44%	**0.00%**	6.4%	**-0.5%**	3.5%	**-45.2%**	3.8%	**-41.6%**
P3out	**28.5%**	29.4%	**3.2%**	29.4%	**3.3%**	29.6%	**4.1%**	33.4%	**17.1%**
P31out	**10.0%**	NA	**NA**	10.0%	**0.2%**	8.6%	**-13.3%**	9.0%	**-9.7%**
P312out	**6.44%**	6.44%	**0.00%**	6.4%	**-0.5%**	3.5%	**-45.3%**	3.8%	**-41.6%**

4.6 Conclusion

In this chapter we focus on the problem of evaluating blocking probabilities in an overflow loss network, thus performance of the network. We consider two different overflow loss network structure: "Acyclic overflow loss network" where the customers may overflow from lower-tiers to higher tiers following a forward routing rule and "Feedback (cyclic) overflow loss network" where a customer may overflow to any station such that both forward and backward overflows are allowed (a customer leaves the system if all stations are full). For both systems, several approximation methodologies are proposed and tested.

For "Acyclic overflow loss network", we proposed an approximation methodology (BinomIPP) which is based on IPP defined overflow streams and binomial moment matching. The accuracy of the proposed method is verified by the presented simulated data on three numerical instances where three different intensity of offered traffic are considered. Furthermore, the method is compared against an existing well-performing approximation method: Chevalier&Fredericks. The numerical results indicate that our new approximation BinomIPP offers better estimates than C&F and quite close estimates to simulation for each test instance. We claim that the success of the new approximation is due to its ability to discard Poisson error with IPP characterization of overflow streams and consider correlation of streams by estimating moments of each arrival stream separately regarding their dependency.
For "Feedback (cyclic) overflow loss network", we proposed three approximation methodologies: simple one-moment approach, aggregated approach and an iterative BinomIPP. Those methods are tested on a three-station loss network. Numerical results demonstrated that the aggregated approach outperforms the other methods for this specific instance considered. However, the method is based on a two state Markov Chain computation in the aggregated state. As the problem size gets bigger (for higher number of stations), a scalability problem arises, thus it is required to develop a more sophisticated probabilistic selection rule. As a future work, we will focus on this problem and develop an approximation methodology that handles the scalability problem and yields good estimates in such complex networks.

4.7 Résumé du chapitre

Dans ce chapitre, nous nous concentrons sur le problème de l'évaluation des probabilités de blocage dans un réseau sans attente, ainsi ses performances. Nous considérons deux structures de réseau différentes: un "réseau de perte acyclique", où les clients peuvent être rejetés depuis un niveau inférieur vers un niveau supérieur en suivant une règle de routage vers l'avant et "(cyclique) réseau avec perte cyclique et feedback" où un client peut être rejeté vers n'importe quelle station de manière à ce les flux de rejet sont autorisés dans les deux sens (un client quitte le système si toutes les stations sont pleines). Pour les deux systèmes, certaines méthodes d'approximation sont proposées et testées.

Pour les réseaux acycliques, nous avons proposé une méthode d'approximation (BinomIPP) qui est basée sur les flux de débordement définis IPP et "binomial instant" correspondants. La précision de la méthode proposée est vérifiée par les données simulées présentées sur trois exemples numériques où trois différentes intensités de trafic sont jugés. En outre, la méthode est comparée à une méthode d'approximation performante existant: Chevalier & Fredericks 4.2. Les résultats numériques indiquent que notre nouvelle mèthode BinomIPP de rapprochement offre de meilleures estimations que C&F et les estimations assez proches de simulation pour chaque instance de test. Nous avancons l'hypothèse suivante: la qualité de la nouvelle approximation est due à sa capacité à éliminer l'erreur due à l'approximation de Poisson en utilisant la caractérisation IPP des flux de rejet et considère une corrélation de flux grâce à l'estimation des moments de chaque flux d'arrivée séparément, selon leur dépendance.

Pour les réseaux cycliques, nous avons proposé trois méthodes d'approximation: une approche simple, l'approche agrégée et une approche BinomIPP itérative. Les méthodes sont testées sur un réseau de trois stations sans attente. Les résultats numériques ont démontré que l'approche agrégée surpasse les autres méthodes de cette instance spécifique considérée. Cependant, la méthode est basée sur un calcul de la chaine de Markov à deux états dans l'état agrégé. Comme la taille du problème devient de plus en plus importante (pour un plus grand nombre de stations), un problème d'évolutivité se pose, il est donc nécessaire de développer une règle de sélection probabiliste plus sophistiquée. En guise de perspective, nous allons nous concentrer sur ce problème et élaborer une méthodologie d'approximation qui peut résoudre le problème de l'évolutivité et des rendements de bonnes estimations dans des réseaux complexes.

Chapter 5

Admission Control Policies in Overflow Loss Networks

Contents

5.1 Introduction

In the previous chapters a location and capacity planning model is constructed and for our application field, perinatal network, the necessary amount of capacity is defined in each hospital in order to ensure a target minimum rejection probability valid and identical for all patients in the network. While doing that, regarding the strategic level of work, no overflow was considered in the system such that a patient who is rejected once is rejected to outside of network, got lost. Strategic level mainly ensures enough capacity to allow patients to be treated in their most preferred hospitals. However at the operational level in loss networks, it is commonly observed that a patient rejected from a facility may overflow to another facility in the same network. Therefore, a hospital may receive both its own patients and overflow patients from other hospitals. In such systems where there exist different types of arrivals, dynamic admission control strategies are commonly used in order to increase flexibility in the allocation of resources among different arrival types [Lerzan Örmeci 2001]. A dynamic admission policy may decide to admit/reject a patient dependent on the current congestion level of the system upon arrival considering the type of the patient such that by rejecting a patient, system can provide an idle server in anticipation of future arriving prioritized patients who would otherwise be lost.

In this chapter we address an admission control problem in a multi-server loss network of hospitals where each hospital is fed by a major (Poisson) arrival stream of its own patients and non-Poisson arrival stream(s) of patients overflowed from other hospitals. A controller (bed manager) who has complete knowledge of the number of occupied beds at each hospital decides to accept or reject each arrival stream. An amount of reward is gained when a patient is accepted to a hospital depending on the source of the patient. In this study major (Poisson) arrivals are prioritized over overflowed ones. Our objective is to find the optimal policy that maximizes the total discounted rewards. In this work, for simplification purposes, we conduct our study on one type of patient (pregnant women) and consider only bed resources as servers in each unit. As also assumed in the rest of the thesis, each patient is served by one server with an exponential service time, identical for each class of patient and at each hospital.

Our research strategy evolves from simple networks to complicated ones. On simple networks, optimal solutions and proofs of optimality are provided while those results and their implications are used to come up with some near-optimal solution strategies in complicated networks.

In section 5.3, we start with a simple multi-server loss queue of one hospital which receives different types of patients. Patients arrive with a Poisson distribution and their service time is exponential. We studied the admission control problem among different class of arrivals where the optimal policy is proved to be of threshold form.

This problem is one of the first problems that is studied in Markov Decision Process (MDP) literature, yet we present our methodology and our results in here because it generates the basis for methodologies developed for more complicated systems.

In section 5.4, we consider a 2 hospital network where each hospital is fed by their own patients (major Poisson arrival) and an overflow stream originated from the other hospital whenever the hospital gets full. For 2 hospital case, we prove that the optimal control policy is a threshold policy by showing the structural properties of value function. We conduct various numerical studies to analyze the variations of system parameters and their impacts on the output. Finally, we assess the impact of employing an admission control policy, the possible improvements that could be achieved are compared to other policies.

In section 5.5, we consider a hierarchical hospital network with many hospitals (I>2) where all hospitals are overflowing to 2 target hospitals. The structural properties obtained for 2 hospital network in section 5.4 are valid for this case, however additionally there exist overflows from other hospitals that are needed to be considered. We model this system as MDP by assuming that overflow arrivals are Poisson. By using value iteration algorithm, optimal admission policy is computed. In order to assess the performance of MDP optimal policy, we evaluate various scenarios by using discrete-event-simulation. Simulation strengthened the decision making process by incorporating the complexity which cannot be captured by MDP and allow us to assess the impact of Markovian assumption in a complex healthcare setting.

In section 5.6, we consider a big-scale loss hospital network (I>2) where we let all kinds of overflow between hospitals. For this case, we construct the global MDP model, which is recognized as quite complex to solve to optimality due to curse of dimensionality. Therefore, we propose a near-optimal local control policy and a linear programming model for computing a tight upper-bound. We assess the performance of local control policy on two numerical studies.

5.2 Literature Review of Admission Control Policies

In dynamic control problems, Markov Decision Processes (MDP) has been the most commonly used modeling tool which make it possible to characterize the structure of optimal policy and/or numerically solve for its parameters. MDPs are strong sequential decision making tools which can assess the performance of the existing system and propose an improved system considering the prospective actions eventually reward of the system. There is a vast literature on using MDPs on network structures where either the optimal or near-optimal control policies are pursued. Many of the related studies are on telecommunication or production network field. E. Altman [Altman 2002] gives a comprehensive literature review of the models stand in that area. Here, we briefly discuss the studies on admission control in a chronological order.

Early works are mostly focused on admission policies of different class of clients on a single station. Miller [Miller 1969] considers a multi-server queuing system and multi-class of arrivals with ($\mu_1 = \mu_2$) and $r_1 > r_2$. He uses admission control to maximize the expected average reward, and prove the existence of a threshold type policy. Lippmann and Ross [Lippman 1971] analyze the optimal admission rule for a system with one server and no waiting room which receives offers from jobs according to a joint service time and reward probability distribution (this model is usually referred to as the streetwalker's dilemma). Harrison[Harrison 1975] considers a single server queue and two classes of jobs with different Poisson arrival rates, exponential service times and rewards. He analyzed the optimal scheduling rule to maximize the expected service rewards over an infinite horizon and proved the existence of rm rule; meaning the higher reward arrival will be scheduled first on the single server.

Starting from 80s, dynamic control received increased recognition in the control of network of queues where there exists more than one station, involving large-scale systems of interacting components. For smaller networks, researchers examine the structure of optimal control rules by proving some properties of the value function, while for larger networks they develop heuristic rules or near optimal policies.

Hajek [Hajek 1984] considers the control of a general two interacting queue systems that can be either arranged in parallel or in series. In his model, both queues i receive their own Poisson arrivals at rates A_i whereas a third stream of Poisson arrivals (A) can be routed to either queue. In such a system, Hajek studied the routing policies without admission control and used an inductive proof to establish the existence of a monotonic switching curve. Ghoneim and Stidham [Ghoneim 1985] studied two exponential servers in series, each with infinite capacity queue. Using an induction based on value iteration, they established that the optimal value function is concave in each argument and submodular. Hariharan et al. [Hariharan 1990] extended the monotonicity properties derived by Davis [Davis 1977] and Hajek [Hajek 1984] in order to control both admission and routing of customers to two queues

in parallel with infinite capacity and identical server rates. Stidham and Weber [Stidham Jr 1993] provided a comprehensive survey on optimal control of networks of queues. Admission control in a network of facilities with more than 2 queues is commonly approached as a routing problem rather than an admission one (see e.g., Delasay [Delasay 2012], Hordijk et al. [Hordijk 1992]). Few papers are encountered on admission control in networks of more than two queues; however the efforts to obtain some generalized structural results about the optimal admission policy in such systems achieve little success as also mentioned by Stidham and Weber.

In this study, we particularly work on loss networks with multi-stations (in parallel) each with multi-servers, meaning no room for queues (waiting) other than the total number of existing servers. In the vast literature of MDPs, control of admissions to loss queues has been studied by few authors, while control of admissions to loss-network of queues has been studied by fewer.

Carrizosa [Carrizosa 1998] considered a loss system (M/G/c/c) with different class of arrivals. They presented an optimal static admission control policy where there exists a preferred class which is determined by a variation of the $c\mu$ rule.

Ormeci [Lerzan Örmeci 2001] studied dynamic admission control of two class of jobs (with different service rates) to a loss queueing system with c identical parallel servers. They proved the existence of a preferred class and showed that the optimal admission policy is of threshold type. In Ormeci [Örmeci 2006], the authors extended the results of Ormeci [Lerzan Örmeci 2001] by considering arrivals occurring according to a general distribution. We establish the existence of optimal acceptance thresholds for both job classes and show that under certain conditions there exists a preferred class.

Ku Jordan is one of the first authors studied admission control problems in loss networks of queues. In [Ku 1997], they considered a system with two multi-server loss queues in series. Under appropriate conditions, the optimal admission policy is shown to be a switching curve. In [Ku 2002], authors considered a system of multi-server loss-network of queues in parallel and a target loss queue, where customers arrive to target queue after service completion in the network of queues. The target loss queue faces an admission control problem for each arriving stream. The authors prove that the optimal policy is monotonically decreasing thresholds as functions of the occupancy of each queue. In [Ku 2006], Ku and Jordan generalized these results to multi-server loss queues in series with customer classes and proposed a near-optimal heuristic policy which achieves a close performance to the optimal policy.

Chang and Chen [Chang 2003] considered a two-stage no-wait tandem queueing system, in which any customer who finds all servers busy at his destination stage will be lost. They present a feasible admission control policy, called the new never

block the old (NNBO) policy and they compare the loss rate under various admission control policies.

Sheu and Ziedins [Sheu 2010] considered admission and routing controls in a system of N parallel tandem queues with the objective of minimizing loss. They presented an asymptotically optimal policy as N goes to infinity.

Zhang [Zhang 2013] studied a two-station tandem queue loss model where customers arrive to station 1 according to a Poisson process and after service completion in station 1, they arrive to station 2. Admission decision needs to be given for arrival in each station considering the congestion at both stations together. Authors provided the first analytical result on the long-run average reward optimal admission control policy for tandem queues with loss.

Loss-networks has been studied in the literature mostly in terms of tandem queues where a customer visit each queue in an order, and latter queues can be visited only after service-completion in preceding stations, if the customer is accepted. On the other side, a customer rejected in any station is lost. To the best of our knowledge, there is no existing study that considers a network of loss queues in parallel (providing the same service) where any customer rejected overflow to other queues and overflowed (external) customers continue to seek for admission in other stations along with the internal customers of those stations.

5.3 Admission Control Policy in a single Hospital with different class of arrivals

5.3.1 Description of the system

We consider one-hospital loss system with N identical and parallel servers (hospital beds) and multiple arrival streams $m > 1$. Each patient is served by one bed and service rates of all patients are independent and exponentially distributed with μ. If all N servers are occupied, incoming patients are lost. On the other side, admission of a patient from *class-m* brings a reward r_m to the system where $r_1 \geqslant r_i \geqslant 0$. The system representation is pictured in Figure 5.1.

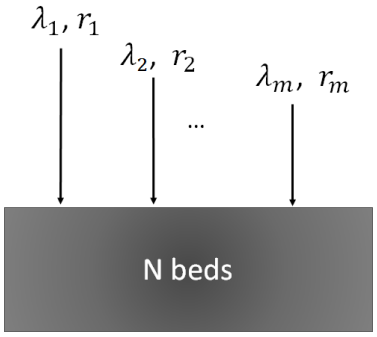

Figure 5.1: A representation of a single hospital with many arrivals

5.3.2 MDP Formulation

This system can be modeled as a continuous-time Markov chain, with states x defined as the number of patients (occupied beds) at the considered hospital where the state space is $\Omega = \{0 \leqslant x \leqslant N\}$. Control action space is defined by $a_m(x) = 1(0)$ indicates that the control action is to admit(1) or reject(0) a *class-m* patient arriving to a single hospital where $m = \{1, 2, ..., M\}$. Therefore, the admission control policy becomes a decision to accept or reject the incoming arrivals, as a function of the state x and the arrival type *class-m*.

Our objective is to maximize the total α-discounted reward $(0 < \alpha < 1)$ over an infinite horizon, over all policies π:

$$V_\pi(X) = \lim_{T\to\infty} E_\pi \left(\sum_t^T R\left(X_t, X_{t+1}\right) \alpha^t | X_0 = X \right)$$

where $R\left(X_t, X_{t+1}\right)$ is the reward obtained from a state change from t to $t+1$. E_π is the conditional expected reward when a certain policy π is employed given an initial state of the system. A control policy π^* is said to be the optimal policy if $V_{\pi^*}(X) = \max_\pi V_\pi(X)$.

In this chapter, we use event-based dynamic programming which focuses on the underlying properties of the value and reward functions, and allows us to study many models at the same time. Eventbased dp is defined with event operators, which can be seen as building blocks of the value function ([Koole 2000]). In this study our event operators are associated with basic events in the system; arrivals and departures of patients. Let us denote $f(x)$ a value function and $Tf(x)$ is the event operator. Operators related to arrivals are $T_{A_m}f(x)$, Operators related to departures are $T_D f(x)$.

In this study we employed well-known uniformization ([Lippman 1975]) giving us an equivalent discrete-time Markov chain by allowing fictitious transitions from a state to itself. The uniformization rate is taken as $\gamma = \sum \lambda_m + N\mu$. Then the equivalent discrete-time system has parameters $p_{A_m} = \lambda_m/\gamma$, $q_D = (N\mu)/\gamma$. corresponding to the probability of real or artificial type-m arrivals and departures.

The optimality equations of event-based dynamic programming can be written as in the following:

$$V(x) = Tf(x) = \sum_m p_{A_m} T_{A_m} f(x) + q_D T_D f(x) \tag{5.1}$$

$$T_{A_m} f(x) = \begin{cases} \max\{r_m + f(x+1), f(x)\} & \text{if } x < N \\ f(x) & \text{if } x = N \end{cases}$$

$$T_D f(x) = \frac{x}{N} f(x-1) + \frac{N-x}{N} f(x),\ \forall x \in \Omega$$

where $\Omega = \{0 \leqslant x \leqslant N\}$ and $r_1 \geqslant r_2 \geqslant ... \geqslant r_i > 0$.

In each state x, the optimal admission control policy decides to accept a patient if the future expected discounted reward is maximized. With other words, a patient can be accepted only if the immediate reward obtained from this patient is bigger than the expected discounted loss caused by future blocking of a patient due to the acceptance of this current patient. The above optimality equations are solved via Value Iteration Algorithm (VI) which convergences to the optimal policy if there exists one. In VI algorithm, we set an initial value to the value function for each state; $V_0(x) = 0$. Then, in every iteration by choosing the actions bringing the max reward, $V_h(x)$ is calculated for each state x using the value function optimality equation 5.1 :

$$\begin{aligned} V_h(x) = \sum_m p_{A_m} \max\{r_m + V_{(h-1)}(x+1), V_{(h-1)}(x)\} \\ + q_D \left(\frac{x}{N} V_{(h-1)}(x-1) + \frac{N-x}{N} V_{(h-1)}(x) \right) \end{aligned}$$

This evaluation is done iteratively until the algorithm converges (V_h converges to V^*) which means optimum policy is reached.

Intuitively we can claim that when a hospital has low occupancy, it is expected that it will accept all arriving patients in order to maximize the immediate reward. However, when the occupancy increase, starting from some threshold level, it is expected to reject arrivals bringing less reward in order to reserve some space for the future arrivals with higher reward. Intuitively, the optimal policy can be interpreted as a threshold policy consisting of several switching curves for each arrival stream m above which the corresponding arrivals are rejected.

5.3.3 Structure properties of the Optimal Admission Control Policy

In this section, we analyze the form of the optimal admission control policy in a series of theorems. Theorem 1 states that, under appropriate conditions, *class-1* arrivals are always accepted and *class-m* ($m > 1$) arrivals need to be controlled. Theorem 2 states that optimal admission policy is of threshold form composed of M-1 switching curves. In a maximization problem in order to show that a threshold policy is optimal, we need to show that value function $Tf(x)$ is monotonically non-increasing and concave. The following properties and lemmas detail the structure of the value function under optimal policy, using value iteration. Proofs of lemmas are given in Appendix E.1.

Properties

A. f is a reward function and its values are always positive. $f(x) \geqslant 0$, $\forall x \in \Omega$

B. f is a monotonically non-increasing function.
$f(x-1) \geqslant f(x)$, $\forall x | ^{\cdot\cdot} x \in \Omega, x-1 \in \Omega$

C. class-1 patients are always preferred over higher class patients.
$r_1 + f(x+1) \geqslant r_i + f(x+1)$, $\forall x \in \Omega | ^{\cdot\cdot} x \in \Omega$, $x+1 \in \Omega$, and $r_1 \geqslant r_i, \forall i$

D. class-1 patients are always accepted to a hospital as long as the hospital is not full.
$r_1 + f(x+1) \geqslant f(x)$, $\forall x \in \Omega | ^{\cdot\cdot} x+1 \in \Omega$

E. f is a concave function.
$\Delta f(x) \geqslant \Delta f(x+1)$, $\forall x \in \Omega | ^{\cdot\cdot} 1 \leqslant x \leqslant N-1$

where $\Delta f(x) = f(x) - f(x-1)$ denotes the difference between two consecutive states. In event based dynamic programming, we need to show that the above properties defined for function f also hold for the operator T.

Lemma 1: A and C hold for $f(x)$, then both hold for $T_{A_m}f(x)$ and $T_Df(x)$ from definitions
Proof. Trivial from definitions. Q.E.D.

Lemma 2: If B and D hold for f(x), then D holds for $T_Df(x)$.
The lemma is proved by establishing the inequality
$T_Df(x+1) - T_Df(x) \geqslant -r_1, \qquad \forall x \in \Omega | \ddot{} 0 \leqslant x \leqslant N-1$

Lemma 3: If E and D hold for f(x), then D holds for $T_{A_i}f(x)$.
It is proved by showing $T_{A_i}f(x+1) - T_{A_i}f(x) \geqslant -r_1, \qquad \forall x \in \Omega | \ddot{} 0 \leqslant x \leqslant N-1$

Lemma 4: If B holds for f(x), then B holds for $T_Df(x)$.
It is proved by showing $T_Df(x+1) - T_Df(x) \leqslant 0 \qquad \forall x \in \Omega | \ddot{} 0 \leqslant x \leqslant N-1$

Lemma 5: If B holds for f(x), then B holds for $T_{A_i}f(x)$.
It is proved by showing $T_{A_i}f(x+1) - T_{A_i}f(x) \leqslant 0 \qquad \forall x \in \Omega | \ddot{} 0 \leqslant x \leqslant N-1$

Lemma 6: If E holds for f(x), then E holds for $T_Df(x)$.
It is proved by showing $\Delta T_Df(x) - \Delta T_Df(x+1) \geqslant 0 \qquad \forall x \in \Omega | \ddot{} 0 \leqslant x \leqslant N-1$

Lemma 7: If E holds for f(x), then E holds for T_{A_i}.
It is proved by showing $\Delta T_{A_i}f(x) - \Delta T_{A_i}f(x+1) \geqslant 0$, $\forall x \in \Omega | \ddot{} 0 \leqslant x \leqslant N-1$

Theorem 1. It is optimal to admit *type-1* patients to the considered hospital in every state unless the hospital is full, i.e., $a_1 = 1$ in all states $x < N$.

Proof is straightforward from Lemma 2 and Lemma 3. Property D holds both for $T_{A_m}f(x)$ and $T_Df(x)$, therefore it also holds for $Tf(x)$.

Theorem 2. The optimal admission control policy for *type-m* patients (*m>1*) is of (M-1) dimensional threshold form. The optimal policy can be characterized by a switching curve for each class of patient, i.e., there exists a threshold C_m for a *class-m* arrival, such that an arrival from this class is accepted if and only if $x < C_m$. Further, $C_m \geq C_{m+1}$.

Proof 2. Theorem is a direct consequence of Lemma 6 and 7. $T_{A_m}f(x)$ and $T_Df(x)$ are proved to be monotonically non-increasing and concave functions, therefore the same properties also hold for $Tf(x)$. Consequently, the optimal policy is of threshold form. The monotonicity of C_m is a direct consequence of the alphabetic reward ordering.

5.4 Admission Control Policy for 2-Hospital

5.4.1 Description of the system

In this section we consider a 2-hospital overflow & loss network. Each hospital is a loss queue with N_i identical and parallel servers (hospital beds). New patients arrive at hospital 1 with a Poisson rate of λ_1 and at hospital 2 with a Poisson rate of λ_2. Each patient is served by one bed and service rates of all patients are independent and exponentially distributed with μ. If all $N_1(N_2)$ servers of hospital 1(2) are occupied, incoming patients to hospital 1(2) are deferred to hospital 2(1). Those deferred patients no longer follow a Poisson distribution and are called overflowed arrivals. Therefore, a hospital may have at most two arrival streams; new and overflowed which from here are called *class-1* and *class-2* arrivals, respectively. A bed manager who has complete knowledge of the system decides when to accept/reject *class-1* and *class-2* patients. The system representation is pictured in Figure 5.2.

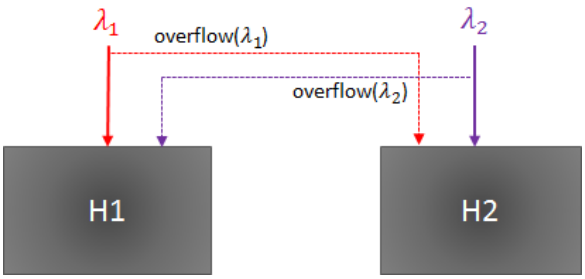

Figure 5.2: Representation of a 2-hospital loss network

Admission of each patient to a hospital brings a reward to the system. In our application field, perinatal network, patients are urgent and rejecting/deferring a patient to another hospital might cause long travel distances which can deteriorate health of both woman and baby. Therefore in this study, *class-1* arrivals are prioritized over *class-2* arrivals by associating a bigger reward to the acceptance of a *class-1* arrival. ($r_{ii} > r_{ij}$)

5.4.2 MDP Formulation

This system can be modeled as a two-dimensional continuous-time Markov chain, with states $x = (x_1, x_2)$ defined as the number of patients (occupied beds) at hospital 1 and 2, respectively where the state space is $\Omega = \{(x_1, x_2) : 0 \leqslant x_1 \leqslant N_1, 0 \leqslant x_2 \leqslant N_2\}$.

Control action space is defined by a 2x2 dimensional mapping matrix with $a_{wj}(X) = 1(0)$ indicates that the control action is to admit(1) or reject(0) a *class-w*

patient arriving to hospital i where $w = \{1, 2\}$ and $i = \{1, 2\}$. Therefore, the admission control policy becomes a decision to accept or reject the incoming arrivals, as a function of the state $x = (x_1, x_2)$, the arrival type (class-w), and the hospital at which the patient is arriving (H1/H2). In some regions of the state space, there is no admission control such that decisions are already defined inherently:

- When both hospitals are not full ($x_1 < N_1$ and $x_2 < N_2$), all patients are accepted to corresponding hospitals
- When both hospitals are full ($x = N$ where $x_1 = N_1$ and $x_2 = N_2$), all patients are rejected.

Therefore in our system, dynamic control of admissions is required when one of the hospitals get full, specifically in the states (x_1, N_2) where $0 \leqslant x_1 < N_1$ and (N_1, x_2) where $0 \leqslant x_2 < N_2$.

An equivalent discrete-time Markov chain allowing fictitious transitions is defined with a uniformization rate taken as $\gamma = \lambda_1 + \lambda_2 + (N_1 + N_2)\mu$. Then the equivalent discrete-time system has corresponding parameters $p_{A_1} = \lambda_1/\gamma$, $p_{A_2} = \lambda_2/\gamma, q_{D_1} = (N_1\mu)/\gamma, q_{D_2} = (N_2\mu)/\gamma$ corresponding to the probability of real or artificial.type-1 arrival, type-2 arrival, type-1 departure or type-2 departure.

Our objective is to maximize the total α-discounted reward ($0 < \alpha < 1$) over an infinite horizon, over all policies π where $V_{\pi^*}(X) = \max_\pi V_\pi(X)$. The optimality equations of event-based dynamic programming can be written as in the following.

$$V(x) = Tf(x) = p_{A_1}T_{A_1}f(x) + p_{A_2}T_{A_2}f(x) + q_{D_1}T_{D_1}f(x) + q_{D_2}T_{D_2}f(x)$$

$$T_{A_i}f(x) = \begin{cases} \max\{r_{ii} + f(x + e_i), r_2 + f(x + e_j), f(x)\} & \text{if } x_i < N_i \\ \max\{r_{ij} + f(x + e_j), f(x)\} & \text{if } x_i = N_i, x_j < N_j, \forall x \in \Omega \\ f(x) & \text{if } X = \boldsymbol{N} \end{cases}$$

$$T_{D_i}f(x) = \frac{x_i}{N_i}f(x - e_i) + \frac{N_i - x_i}{N_i}f(x), \forall x \in \Omega$$

where

e_i denotes the i^{th} unity vector $\Omega = \{(x_1, x_2) : x_i \in \{0, 1, ..., N_i\}\}$ and $N = N_1 + N_2$

r_{ii} : reward of accepting patient i to hospital i
r_{ij} : reward of accepting patient i to hospital j where $i, j \in \{1, 2\}$ and $r_{ii} > r_{ij}$

Remark 1: For facilitating proofs, we consider without loss of generality $p_{A_1} + p_{A_2} + q = 1$ where $q = (N_1 + N_2)\mu/\gamma$. The departure operator $T_{D_i} f(x)$ are consolidated into $T_D f(x)$:

$$T_D f(x) = \frac{N - x_1 - x_2}{N} f(x) + \frac{x_1}{N} f(x - e_1) + \frac{x_2}{N} f(x - e_2),\ \forall x \in \Omega$$

where $T_D f(x) = \frac{N_1}{N} T_{D_1} f(x) + \frac{N_2}{N} T_{D_2} f(x)$.

The optimality equation is modified accordingly:

$$V(x) = Tf(x) = p_{A_1} T_{A_1} f(x) + p_{A_2} T_{A_2} f(x) + q_D T_D f(x)$$

Optimal policy can be obtained from solutions of the optimality equation. In each state X, the optimal admission policy decides to accept a patient if the future expected discounted reward is maximized. With other words, a patient can be accepted only if the immediate reward obtained from this patient is bigger than the expected discounted loss caused by future blocking of a patient due to the acceptance of this current patient. The optimality equation is solved via a value iteration (VI) algorithm which has a fast convergence to the optimal policy. As also mentioned in Section 5.3, the convergence of VI algorithm states that there exists an optimal solution to the admission problem and the form of the optimal policy can be proved through analyzing the structural properties of the given value function.

When a hospital has low occupancy, it is expected that it accepts all arriving patients in order to maximize the immediate reward. However, when the occupancy increases, starting from a certain threshold level, it is expected that class-2 arrivals would be rejected in order to provide idle servers (create more possibility of admission) for class-1 arrivals who will bring more reward in the close future. So the optimal policy can be intuitively defined as a switching curve for each hospital; 2-dimensional threshold policy.

5.4.3 Structure properties of the Optimal Admission Control Policy

In this section, we present a sequence of properties and lemmas in order to show the structural properties of the optimality equation $Tf(X)$, and consequently leading us to a theorem which characterizes the form of the optimal admission control. We would like to show that optimality equation $Tf(X)$ is a monotonically decreasing (non-increasing) and a concave function to be able to conclude that the optimal policy for our problem is of threshold form.Proofs of all properties are given in Appendix E.2.

Properties

A. f is a reward function and its values are always positive.
$f(x) \geqslant 0,\ \forall x \in \Omega$

B. f is a monotonically non-increasing function.
$f(x - e_i) \geqslant f(x)$, $\forall x | \cdot\cdot x \in \Omega, x - e_i \in \Omega$

C. class-1 patients are always preferred over class-2 patients.
$r_{ii} + f(x + e_i) \geqslant r_{ij} + f(x + e_j)$, $\forall x \in \Omega | \cdot\cdot x + e_i \in \Omega$, $x + e_j \in \Omega$, $e_i \neq e_j$

D. class-1 patients are always accepted to a hospital as long as the hospital is not full.
$r_{ii} + f(x + e_i) \geqslant f(x)$, $\forall x \in \Omega | \cdot\cdot x + e_i \in \Omega$

E. f is a concave function.
$\Delta^i f(x) \geqslant \Delta^i f(x + e_i)$, $\forall x \in \Omega | \cdot\cdot 0 \leqslant x_i \leqslant N_i - 1$

where $\Delta^i f(x) = f(x) - f(x - e_i)$ denotes the difference between two consecutive states.

In event based dynamic programming, we need to show that the above properties defined for function f also hold for the operator T.

Lemma 1: If (A) holds for $f(x)$, then (A) holds for $T_{A_i} f(x)$ and $T_{D_i} f(x)$.
Proof is trivial from definitions.

Lemma 2: If B and D hold for $f(x)$, then D holds for $T_{D_i} f(x)$.
The lemma is proved by establishing two inequalities

$$T_{D_i} f(x + e_i) - T_{D_i} f(x) \geqslant -r_{ii}, \forall x \in \Omega | \cdot\cdot 0 \leqslant x_i \leqslant N_i - 1$$
$$T_{D_j} f(x + e_i) - T_{D_j} f(x) \geqslant -r_{ii}, \forall x \in \Omega | \cdot\cdot 0 \leqslant x_i \leqslant N_i - 1$$

Lemma 3: If C and D hold for $f(x)$, then D holds for $T_{A_i} f(x)$.
The lemma is proved by establishing the following inequality

$$T_{A_i} f(x + e_j) - T_{A_i} f(x) \geqslant -r_{ij}, \qquad \forall x \in \Omega | \cdot\cdot 0 \leqslant x_j \leqslant N_j - 1$$

Lemma 4: If C holds for $f(x)$, then C holds for $T_D f(x)$.
The lemma is proved by establishing the following inequality

$$r_{ii} + T_D f(x + e_i) \geqslant r_{ij} + T_D f(x + e_j), \forall x \in \Omega | \cdot\cdot 0 \leqslant x_i \leqslant N_i - 1, 0 \leqslant x_j \leqslant N_j - 1$$

Lemma 5: If C and D hold for $f(x)$, then C holds for $T_{A_i} f(x)$.
The lemma is proved by establishing the following inequalities

$$r_{ii} + T_{A_i} f(x + e_i) \geqslant r_{ij} + T_{A_i} f(x + e_j), \forall x \in \Omega | \cdot\cdot 0 \leqslant x_i \leqslant N_i - 1, 0 \leqslant x_j \leqslant N_j - 1$$
$$r_{ii} + T_{A_j} f(x + e_i) \geqslant r_{ij} + T_{A_j} f(x + e_j), \forall x \in \Omega | \cdot\cdot 0 \leqslant x_i \leqslant N_i - 1, 0 \leqslant x_j \leqslant N_j - 1$$

Lemma 6: If B holds for $f(x)$, then B holds for $T_D f(x)$.
The lemma is proved by establishing the following inequality

$$T_D f(x + e_i) - T_D f(x) \leqslant 0, \qquad \forall x \in \Omega | \cdot\cdot 0 \leqslant x_i \leqslant N_i - 1$$

Lemma 7: If B, C and D hold for $f(x)$, then B holds for $T_{A_i}f(x)$.
The lemma is proved by establishing the two inequalities

$$T_{A_i}f(x+e_i) - T_{A_i}f(x) \leqslant 0,\ \forall x \in \Omega|^{\cdot\cdot}0 \leqslant x_i \leqslant N_i - 1 \tag{5.2}$$
$$T_{A_j}f(x+e_i) - T_{A_j}f(x) \leqslant 0,\ \forall x \in \Omega|^{\cdot\cdot}0 \leqslant x_i \leqslant N_i - 1 \tag{5.3}$$

Lemma 8: If E hold for $f(x)$, then E holds for $T_D f(x)$.
Proof by induction
Lemma 9: If B, C, D, E hold for $f(x)$, then E holds for T_{A_i}.
The lemma is proved by establishing the two inequalities

$$\Delta^i T_{A_i}f(x) - \Delta^i T_{A_i}f(x+e_i) \geqslant 0,\ \forall x \in \Omega|^{\cdot\cdot}1 \leqslant x_i \leqslant N_i - 1$$
$$\Delta^i T_{A_j}f(x) - \Delta^i T_{A_j}f(x+e_i) \geqslant 0,\ \forall x \in \Omega|^{\cdot\cdot}1 \leqslant x_i \leqslant N_i - 1$$

Theorem 1. It is optimal to admit class-1 patients to each hospital in every state unless the considered hospital is full, i.e., $a_{11} = 1\,\forall x_1 < N_1$ and $a_{12} = 1\,\forall x_2 < N_2$.

Proof. Proof is straightforward from Lemma 2 and Lemma 3. Property (D) holds for both operators $T_{A_i}f(x)$ and $T_D f(x)$, therefore it also holds for $Tf(x)$.

Theorem 2. The optimal policy is characterized by a switching curve for each station, i.e., for an overflow arrival to station 2, there exists a threshold C_2 in station 2, such that a class 2 arrival in state (N_1, x_2) is accepted if and only if $x_2 < C_2$. Likewise, there exists a threshold C_1 in station 1 such that a class 2 arrival in state (x_1, N_2) is accepted if and only if $x_1 < C_1$.

Proof. Theorem is a direct consequence of Lemma 6, 7, 8 and 9. $T_{A_i}f(x)$ and $T_D f(x)$ are proved to be monotonically non-increasing and concave functions, therefore the same properties also hold for $Tf(x)$. Consequently, the optimal policy is of threshold form.

5.4.4 Sensitivity Analysis of System Parameters

In the previous section, we characterized the optimal admission policy in a two station multi-server overflow loss system. We found that for fixed parameters $(\lambda_1, \lambda_2, N_1, N_2, \alpha, \mu, r_1, r_2)$, the optimal policy is threshold type in both stations H1 and H2 given by C_1 and C_2 respectively when the other hospital is full. In this section, we investigate the variation of these thresholds with variations in the system parameters.

On a symmetric system where both station has the same number of servers $(N_1 = N_2)$, we measure the change in resulting control points C_1 and C_2 as the intensity of arrivals change. We feed the two identical stations $N_1 = N_2 = 30$ (both with service time $\mu = 1$) with light, medium, heavy arrival rates; both symmetric and asymmetric arrival rates are considered. The other parameters kept unchanged

$(\alpha = 0.99, r_1 = 2, r_2 = 1)$.

Example 1 demonstrates the change in C_1 as the arrival rate λ_1 differs on a symmetric setting where the arrival rate λ_2 is kept fixed. Given a fix medium intensity arrival rate to station 2 ($\lambda_2 = 30$), Fig 5.3 presents the computed control points C_1 when H2 is full, and C_2 when H1 is full as the arrival rate λ_1 changes from light to heavy rates. As intuitive, for light arrival rates, station 1 does not need to employ any control action since there is enough place to serve all class of customers in H1. Meanwhile, H2 reserves small number of servers to favor its own customers since its *class-1* arrival rate ($\lambda_2 = 30$) has medium intensity relative to its capacity $N_2 = 30$. As λ_1 increase, H1 starts to reserve some servers for its own raising customers, it starts to reject more class 2 customers, thus C_1 decreases. Finally when λ_1 gets very high, H1 starts to reject all class 2 arrivals, thus $C_1 = 0$. On the other side, as λ_1 increase, H2 also starts to reserve more, thus C_2 decreases monotonically. However, since arrival rate of its own customers λ_2 is fixed, the increase in λ_1 has an indirect effect on H2 therefore, H2 do not need to reserve as much as H1 does.

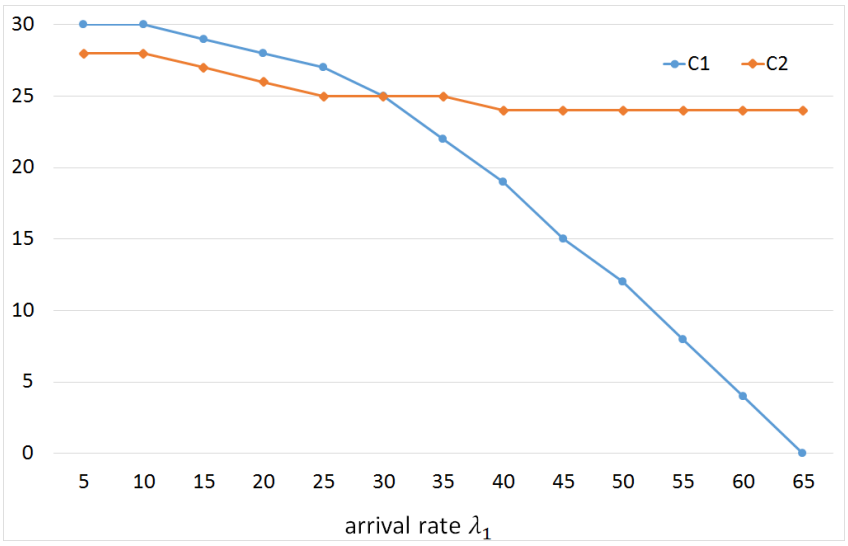

Figure 5.3: Variations in C1 and C2

In Example 2 demonstrates the change in $(C_1 = C_2)$ under various arrival rate scenarios for both symmetric $(N_1 = N_2)$ and asymmetric system $(N_1 \neq N_2)$. Test data used in this study is presented in Table 5.1.

Table 5.1: Test Data For Example 2

		$(\lambda 1, \lambda 2)$			
		light load	**medium load**	**heavy load**	**mix load**
$N1 = N2 = 30$	$\lambda 1 = \lambda 2$	(10,10)	(25, 25)	(50, 50)	-
	$\lambda 1 \neq \lambda 2$	(15,5)	(30, 20)	(60, 40)	(50,10)
$N1 \neq N2$	$\lambda 1 = \lambda 2$	(10,10)	(25, 25)	(50, 50)	-
$N1 = 40, N2 = 20$	$\lambda 1 \neq \lambda 2$	(15,5)	(30, 20)	(60, 40)	(60,10) and (10,60)

For $(N_1 = N_2)$, the control points computed C_1 and C_2 are presented in the Figure 5.4 from where for different arrival rates, the change in C_1 and C_2 can be observed. For symmetric arrival rates, both control points are equal $(C_1 = C_2)$. For light load, C_1 and C_2 is close to N_1 and N_2 whereas as arrival rates increase C_1 and C_2 decreases and tends to zero (observable on the straight line in Figure 5.4). For asymmetric arrival rates, the station receives more arrivals compared to the other station, sets a lower control point; thus rejects more *class-2* customers in order to reserve more space for its own arrivals.

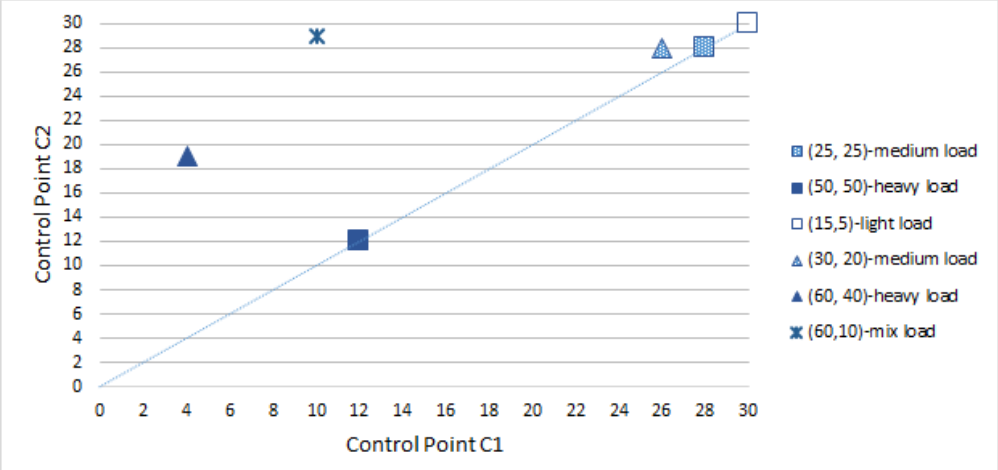

Figure 5.4: Variations in Control Points (C_1, C_2) with respect to various intensity arrival rates to symmetric system $N_1 = N_2$

For $(N_1 \neq N_2)$, the control points C_1 and C_2 computed under arrival rates scenarios defined in Table 5.1 are presented in the Figure 5.5. Under symmetric arrival rate scenarios due to asymmetric servers $(N_1 > N_2)$, station 1 receives lighter load compared to station 2 which receives a heavier load. As a result, station 2 rejects more than station 1, thus C_2 is computed lower than C_1. Therefore the control points (C_1, C_2) are mostly located in the lower triangle of the graph. Under heavy load case, as expected, both stations compute smaller threshold points. Under the

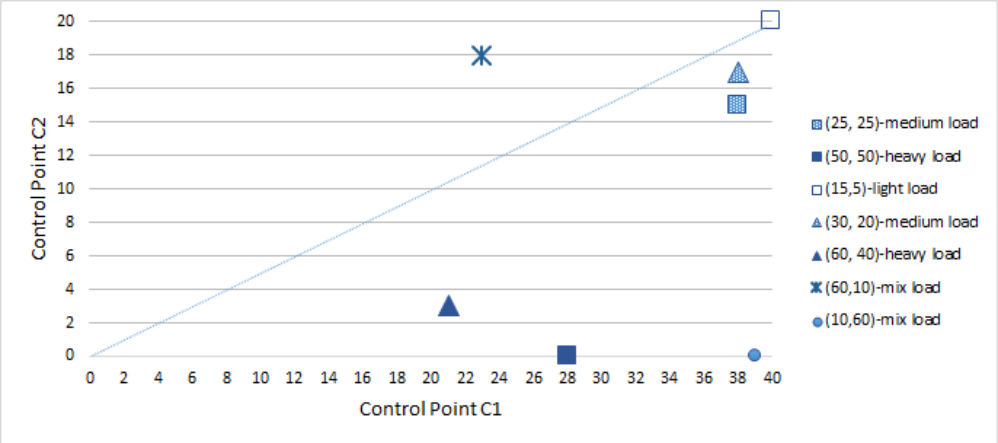

Figure 5.5: Variations in Control Points (C_1, C_2) with respect to various intensity arrival rates to a asymmetric system $N_1 \neq N_2$

considered scenarios, while C_1 does not drop lower than $\frac{N_1}{2}$, C_2 tends to zero. For light load case, class-2 arrivals are all accepted in both stations, thus $(C_1 = N_1)$, $(C_2 = N_2)$.

Example 3 demonstrates the change in control points (C_1, C_2) as unit rewards (R_1, R_2) are altered. In test scenarios, R_2 is fixed to 1 and reward scenarios are generated by altering R_1 values in the range (1 to 6). Two type of arrival rates are considered: symmetric $(\lambda_1 = \lambda_2 = 35)$ and asymmetric $(\lambda_1 = 45, \lambda_2 = 25)$. The other system parameters are defined as $(N_1 = N_2 = 30)$ and $\alpha = 0.999$ and kept unchanged. The optimal threshold points that are computed for each reward scenario (R_1, R_2) are presented in Table 5.2. The corresponding admission probabilities and total reward obtained from each reward scenario is presented in Tables 5.3 and 5.4 respectively for symmetric and asymmetric arrivals.

Table 5.2: Variations in control points with respect to different unit rewards

Test Scenarios		Control Points for N1=N2=30, λ1=λ2=35		Control Points for N1=N2=30, λ1=45, λ2=25	
R1	**R2**	**C1**	**C2**	**C1**	**C2**
1	1	30	30	30	30
2	1	22	22	16	26
3	1	15	15	2	24
5	1	3	3	0	21
6	1	0	0	0	20

From Table 5.2, it is observed that as R_1 increases, the control points (C_1, C_2) decrease. This result is intuitive due to the fact that *class-1* admissions bring more reward to the system compared to *class-2* admissions and as R_1 gets bigger this gap increases. Therefore, it is expected that each station would reserve more capacity

for their own arrivals; thus *class-1* admissions increase whereas *class-2* admissions decrease (Tables 5.3 and 5.4). It is important to note that, R_1 being assigned much higher values than R_2 would quickly end up with rejection of huge percent of *class-2* arrivals. In real life, it is hard to justify the **over** prioritization of a certain group against some others, especially in a healthcare setting. Therefore, in the rest of the study we choose to work with the setting $R_1 = 2$, $R_2 = 1$.

Table 5.3: Total reward obtained under different unit rewards in a symmetric arrival setting

Admission Probabilities		**(R1,R2)**				
		(1,1)	**(2,1)**	**(3,1)**	**(5,1)**	**(6,1)**
Class-1 Admission Pr.	p11in	0.638	0.778	0.780	0.780	0.780
	p22in	0.638	0.778	0.780	0.780	0.780
Class-2 Admission Pr.	p12in	0.170	0.004	0	0	0
	p21in	0.170	0.004	0	0	0
Total Reward		56.58	109.23	163.83	273.05	327.66

Table 5.4: Total reward obtained under different unit rewards in an asymmetric arrival setting

Admission Probabilities		**(R1,R2)**				
		(1,1)	**(2,1)**	**(3,1)**	**(5,1)**	**(6,1)**
Class-1 Admission Pr.	p11in	0.571	0.633	0.633	0.633	0.633
	p22in	0.687	0.898	0.923	0.939	0.942
Class-2 Admission Pr.	p12in	0.238	0.088	0.057	0.026	0.019
	p21in	0.122	0	0	0	0
Total Reward		56.59	105.81	157.18	260.92	312.97

5.4.5 Performance Evaluation of Optimal Admission Control Policy

In the previous section we analyzed the system parameters and their variations individually; explored their effects on the computed optimal control points and their relations with respect to each other. In this section, we aim to discuss the performance of an optimal admission control policy compared to employing no control policy (accepting all *class-2* arrivals) and employing full control policy (rejecting all *class-2* arrivals). Our objective is to measure the total reward difference that we could obtain by controlling the admission of *class-2* customers under light, medium and heavy load setting.

The total rewards obtained under no control policy, full control policy and optimal control policy for light, medium and heavy arrival rates are presented in Table 5.5 Under light arrival rates, both stations are able to serve almost all of their *class-1* patients, thus the number of overflow arrivals (*class-2* customers) for both stations

tends to zero. It can be interpreted that for light load case, since *class-2* customers almost do not exist, there is no observed profit of employing an admission control policy. Under heavy arrivals, the capacity of stations remains insufficient regarding heavy load; thus most of *class-1* customers ends up with being rejected in both stations. An optimal admission control would suggest to reserve big portion of capacity in order to serve more *class-1* patients to increase total reward which would result very close to full control policy (reject all *class-2* customers).

Under medium intensity arrivals where the utilization of servers are not extremely high or extremely low, the optimal control policy performs better compared to the other policies. In comparison with no control and full control policy, optimal policy computes 3% and 1% higher total rewards, respectively. For this specific instance where we consider a network with 2 stations, the % relative rewards are not very distinct. However, as the number of stations interact with each other increase, it is expected from an optimal policy to yield better results. Regarding this issue, the reader is referred to the last Section 5.6 of this chapter where a hierarchical network of hospitals is considered on which we study the near-optimal control policies and its performance evaluation.

Table 5.5: Comparison of Admission Policies under different arrival intensities

		LIGHT LOAD (λ1=25, λ2=20)			**MEDIUM LOAD (λ1=35, λ2=25)**			**HEAVY LOAD (λ1=50, λ2=40)**			**MIX LOAD (λ1=50, λ2=10)**		
		Accept All	*Optimal Control*	*Reject All*	*Accept All*	*Optimal Control*	*Reject All*	*Accept All*	*Optimal Control*	*Reject All*	*Accept All*	*Optimal Control*	*Reject All*
Admission Probabilities		C1=30 C2=30	C1*=28 C2*=28	C1=0 C2=0	C1=30 C2=30	C1*=23 C2*=27	C1=0 C2=0	C1=30 C2=30	C1*=12 C2*=19	C1=0 C2=0	C1=30 C2=30	C1*=18 C2*=29	C1=0 C2=0
Class-1 Admissions	p11in	0.942	0.945	0.947	0.724	0.779	0.780	0.443	0.575	0.575	0.566	0.575	0.575
	p22in	0.980	0.985	0.992	0.810	0.906	0.947	0.472	0.701	0.701	0.853	0.951	1
Class-2 Admissions	p12in	0.053	0.046	0	0.179	0.090	0	0.204	0	0	0.338	0.310	0
	p21in	0.015	0.009	0	0.094	0.003	0	0.176	0	0	0.051	0	0
TOTAL REWARD		87.89	**87.99**	87.03	99.81	**103.04**	101.98	99.32	**113.57**	**113.57**	91.04	**92.02**	77.52

5.5 A Case Study: Admission Control Policy a Hierarchical Overflow Network

5.5.1 Introduction

This section considers admission control policies in a pure-loss hierarchical perinatal network where there are 2 parallel multi-server (target) hospitals fed by new arriving patients and overflowed patients from a set of parallel multi-server hospitals. In this perinatal network setting, we consider the problem of finding an optimal admission policy that recommends how many beds to reserve in two target hospitals for each arriving stream in order to maximize total revenue in the system. At first, we assumed a Markovian system and model the system as a Markov Decision Process (MDP). By using value iteration algorithm, optimal admission policy is computed. Afterwards, we evaluate various policy scenarios (including MDP optimal policy) with a simulation model which strengthens the decision making process by incorporating the complexity which can not be captured by MDP and we assess the impact of Markovian assumption in a complex healthcare setting.

This work is studied on Perinatal Network Haute-de-Seine. In the next section, the contribution aimed with this work is described along with a literature review. In Section 5.5.3, the considered perinatal network is characterized along with necessary assumptions and case study parameters. In Section 5.5.4, an MDP model is constructed. Admission control value function is given, and the optimal policy is found computationally by using the value iteration algorithm. In Section 5.5.5, by using simulation model, several scenarios are generated around the base scenario and evaluated. And finally, conclusion is given.

5.5.2 Literature Review

Healthcare networks are special structures where complexity is quite high due to the fact that "people serve people" meaning people are both the customer and the supply [Roberts 2011]. In such a complex organization, simulation has found widespread application in healthcare literature and is also used as a convenient tool in many studies on healthcare network.[Charfeddine 2010] introduced a global framework for integrated agent-oriented modeling through the Chronic Obstructive Pulmonary Disease (COPD) population and healthcare delivery network in Quebec. [Brailsford 2007] proposed a classification of discrete-event simulation and systems dynamics in healthcare, based on the level of detail on which the model focuses. [Miller 2009] utilized simulation to determine the impact of various patient surge levels on three regional Emergency Departments. Their simulation model pointed out that handling surge depends largely on the percentage of available inpatient beds that a hospital staffs and model also identified the occupancy levels when a hospital should increase in-patient beds.

On the other side MDP models are strong descriptive models that are able to evaluate and predict the performance of existing and proposed systems, and thus are able to improve the design of a system, however requires quite a number of assumptions. There is a vast literature on using Markov decision processes (MDPs) to analyze control policies in a network structure. Many of the related studies are on communication network field. In literature, control of admissions to network of queues has been widely studied however we have encountered few communication papers that have considered the admission control of arrivals to loss queues. [Ku 2002] studied admission control in a multiserver loss queue, where the target queue faces a choice of how many servers to reserve for each arriving stream. They seek the control policy on only one target hospital. [Delasay 2012] focused on optimal routing in a Markovian finite-source, multi-server queueing system with heterogeneous servers, each with a separate queue. They formulate the problem of routing as a Markov Decision Process, they demonstrated that the Shortest Queue policy is optimal when the servers are homogenenous. In our problem, there are two target hospitals (with new patients and overflow patients arriving from other hospitals) overflow to each other when one of them gets full. In such a system, we seek the optimal admission control policies in both hospitals that suggest how many beds to reserve in which hospital.

The problem of finding an optimal admission policy in such a pure loss healthcare network setting is quite interesting, and to our knowledge it has not been studied in the MDP literature. However, MDP models require quite a number of assumptions, therefore it may not be realistic to use MDP results directly on a healthcare problem due to its high complexity. Supporting and improving the decision making process with a simulation model which can better capture the complexity of the healthcare system is the objective of this study.

5.5.3 Perinatal Network Application

This case study is realized in perinatal network of Hauts-de-Seine in Ile-de-France where there are 6 maternity facilities; 2 big-sized ($H1$, $H2$) and 4 small-sized ($H3, H4, H5, H6$). In perinatal network each maternity facility is composed of at most three types of service units $s \in \{1, 2, 3\}$; obstetrics units ($s = 1$) that provide labor services to pregnant women, basic neonatal units ($s = 2$) that provide basic care and neonatal intensive care units NICUs ($s = 3$) that provide special care for newborn babies. Inherently, these units interact with each other as women in labor who need obstetrics unit may subsequently require neonatal or NICU services (this relation is utilized for computing arrival rates of neonatal services). Maternity facilities differ according to the type of service unit included and patient overflow may occur only between the same type of service units.

The perinatal network considered in this study is represented in Figure 5.6 Small-sized hospitals serve only to their own (new) arrivals (*class-1*). Rejected patients

overflow to one of the target hospitals ($H1$ and $H2$) where they are either accepted or rejected out of network. On the other side target hospitals serve both their own (new) arrivals (*class-1*) and overflowed arrivals (*class-2*) from small-sized hospitals. Furthermore, patients of $H1, H2$ overflow to each other whenever one of them is full. The overflow stream of $H1$ ($H2$) is either accepted in hospital $H2$ ($H1$) or rejected out from network. An optimal admission control policy is required and explored for each service unit of $H1$ and $H2$.

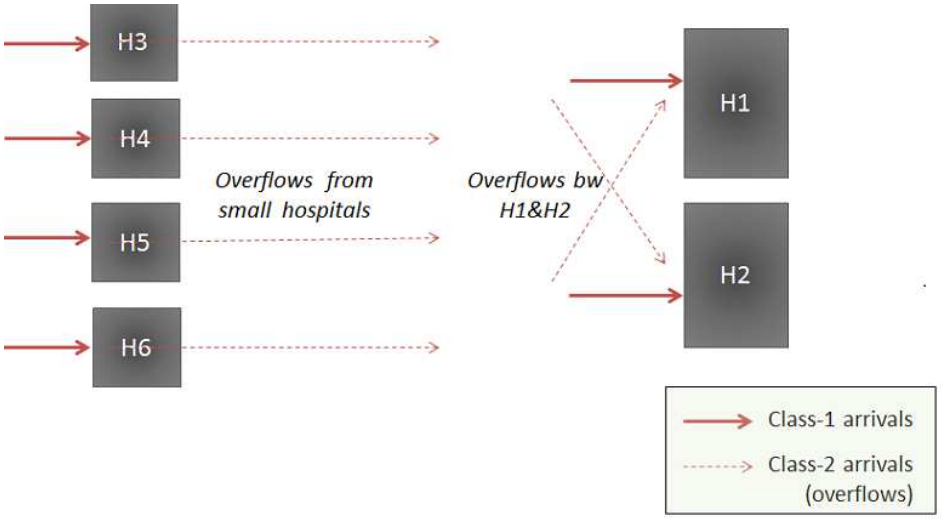

Figure 5.6: Perinatal Network Representation.

5.5.3.1 Parameter Setting of Case Study

This case study is partially supported by Agence Régionale de Santé (ARS), the regional health authority, provided us certain information and data. Unfortunately, data required for a simulation study is tremendous due to the large-scale of the work, therefore we construct the primary simulation model with adequate detail level. Poisson arrivals are often observed in systems without appointments but with arrivals controlled by nature such as birth delivery. Therefore, *class-1* patients are assumed to arrive a service unit s in hospital i with a distribution of Poisson(λ_{si}). For all obstetrics(OB) units in each hospital, ARS provides us yearly arrival data of each pregnant woman, their domicile and preferred hospital information. By assuming no seasonality within the year, we were able to compute daily offered load of each hospital in the network. ARS also provided the ratio of neonates requiring basic and intensive care, from where the arrival rate of neonates requiring basic and intensive care are computed. Furthermore, in heavily stochastic systems, the coefficient of variation (CV) of length of stay (LOS), which is defined as the ratio of the standard deviation to the mean, is typically close to one, satisfying the usage of negative exponential distribution as a service time assumption.

Servers are staffed-beds which is the combination of physical equipment (bed) with an appropriate coverage of nurses and physicians. Each service unit s in each hospital i has multiple identical beds as servers (N_{si}) and each patient requires one bed with an i.i.d exponential service time (μ_{si}) (same for both classes of patients). Arrival rates and existing bed capacity of each service unit in each maternity facility is given in Table 5.6 from where the information about the type of hospitals can be extracted.

Table 5.6: Arrival rates(λ_{si}) and Existing Capacity(N_{si})

	OB		NEO		NICU	
	Arrivals (λ_{1i})	Capacity (N_{1i})	Arrivals (λ_{2i})	Capacity (N_{2i})	Arrivals (λ_{3i})	Capacity (N_{3i})
H1	55	60	24	30	11	8
H2	48	50	12	20	7	5
H3	32	30	15	13	-	-
H4	30	28	12	10	-	-
H5	35	30	-	-	-	-
H6	27	25	-	-	-	-

5.5.4 Markov Decision Process (MDP) Model

In this section we formulate a general MDP model for admission control of patients in any of the service networks. In order to model the overflow system as an MDP process, we assume a Markovian system.

5.5.4.1 Modeling Overflows

Each small-sized hospital can be modeled as an M/M/N/N queue with Poisson arrivals, exponential service time distribution, N identical beds and no waiting room. Then, rejection probability is given by the following Erlang loss function: $B_{si} = \frac{(\lambda/\mu)^N/N!}{\sum_{k=0}^{N}(\lambda/\mu)^k/k!}$. Under this assumption, overflow rate from service unit s of a small hospital i is given as $\vartheta_{si} = B_{si}\lambda_{si},\ \forall s = \{1,2\}, i = \{3,4,5,6\}$ and they are assumed to be Poisson. Therefore, total overflow rate from small hospitals for service s is $\vartheta_s = \sum_{i=3}^{6}\vartheta_{si}$ and is also Poisson.

Remark 1: It is commonly known that overflow arrivals do not follow a Poisson distribution. This is a big but a necessary assumption in order to be able to use MDP since Poisson arrivals is an essential requirement in the application of MDP methods. This approximation signifies the importance of Simulation model to assess and compare the performance of MDP policy with other policies, consequently evaluate the impact of Markovian assumption on resulting output. Markovian approximation and its impact on admission probabilities and total reward is investigated with SIM.

Remark 2: As mentioned before, no patient overflow is assumed to occur among different service units. This fact let us to consider the three service networks independently(OBs, Neonatals, NICUs). Therefore, a general MDP model is formulated in the following section which can be used to describe the flow in each service network s.

5.5.4.2 MDP Formulation

Admission of patients to target hospitals ($H1$, $H2$) is highly dependent on the number of occupied beds in both hospitals. Therefore, this system can be modeled as a two dimensional continuous time Markov Chain model with state $X = (x_1, x_2) \in \Omega$ defined as the number of occupied beds (number of patients) in $H1$ and $H2$ respectively where $\Omega = \{x_1, x_2 : 0 \leqslant x_1 \leqslant N_1 \text{ and } 0 \leqslant x_2 \leqslant N_2\}$ is the state space.

A control action space is defined by a 2x2 dimensional mapping matrix $A(X) = [a_{ki}(X)]$ with $a_{ki}(X) = 1(0)$ indicates the control action is to admit (reject) patient class $k = \{1, 2\}$ at hospital $i = \{1, 2\}$ when the system is in state X. There is a reward of R_k associated with accepting a *class-k* patient where $R_1 > R_2$ so that, admission of a new patient is rewarded more than the overflow patient. Any *class-k* patient rejected from network brings no reward to the system. Some insights:

- $a_{ki}(X) = 0,\ \forall k, \forall i$ at state $X = (x_1 = N_1, x_2 = N_2)$: When both hospitals are full, all arrivals are rejected.
- $a_{1i}(X) = 1,\ \forall i$ at states $x_i < N_i$: *Class-1* patients are always admitted when the corresponding hospital is not full.

Our objective is to find an optimal control policy π^* that maximizes the expected total discounted reward over an infinite horizon. Uniformization results in an equivalent discrete time Markov chain by allowing fictitious transitions from a state to itself. Using uniformization, we express the DP optimality equation below. Therefore, Optimal Value Function for a service network s can be written as:

$$TV(X) = \alpha \sum_{1=1}^{2} \left(\frac{\alpha_i}{\gamma} T_{A_i} V(X) + \frac{\vartheta}{\gamma} T_{OA_i} V(X) + \frac{N_i \mu_i}{\gamma} T_{D_i} V(X) \right)$$

where $\gamma = \sum_{i=1}^{2} (\lambda_i + N_i \mu_i) + \vartheta$ is uniformization rate, α is discount factor $0 < \alpha < 1$.

The Arrival operator T_{A_i} models admission control of *class-1* patients to hospital $i, j = \{1, 2\}$, where $i \neq j$,

$$T_{A_i} V(X) = \begin{cases} r_1 + V(X + e_i) & \text{for } x_i < N_i \\ \max\{r_2 + V(X + e_j), V(X)\} & for\ x_i = N_i \text{ and } x_j < N_j \\ V(X) & \text{for } X = N \end{cases}$$

The Arrival operator T_{OA_i} models admission control of *overflow patients from small hospitals* to hospital i, j

$$T_{OA_i}V(X) = \begin{cases} \max\{r_2 + V(X+e_i), r_2 + V(X+e_j), V(X)\} & \text{for } x_i < N_i \\ \max\{r_2 + V(X+e_j), V(X)\} & \text{for } x_i = N_i \text{ , } x_j < N_j \\ V(X) & \text{for } X = N \end{cases}$$

The Departure operator T_{D_i} models departures from hospital $i = \{1, 2\}$

$$T_{D_i}V(X) = \frac{x_i}{N_i}V(X - e_i) + \frac{N_i - x_i}{N_i}V(X) \qquad \text{for all } X$$

- e_i is the i^{th} unity vector
- r_i is the unif. reward parameter where $r_i = R_i/(\beta + \gamma)$

An optimal control policy should be chosen in each state to maximize the future expected discounted revenue as given by the optimality equation.

5.5.4.3 Computation of an optimal policy with MDP

An optimal policy can be obtained by solving the DP value function analytically however explicit solution of such a function is not trivial. We utilize an efficient computational algorithm, Structured Value Iteration(VI), to reach the optimal policy for our case study.

When both of the hospitals are not full, it is intuitive and also proved in similar studies in literature that optimal action for *class-2* patients is admission to the hospital which has the minimum number of beds occupied (shortest queue) proportional to their *class-1* patient arrival rates. Therefore, in our case study in the states $X = (x_1 < N_1, x_2 < N_2)$ overflow patients will be accepted to hospital i ($a_{2i}(X) = 1$ $a_{2j}(X) = 0$) if $\frac{N_i - x_i}{\lambda_i} < \frac{N_j - x_j}{\lambda_j}$ $\forall i, j = \{1, 2\}$.

Therefore, optimal admission policy for certain states $X = (x_1 < N_1, x_2 < N_2)$ is already determined. In MDP model, we are interested in finding optimal actions for the states $x_i = N_i$ and $x_j < N_j$ (one hospital is full while other is not). With other words, we seek optimal threshold points at $H1$ and $H2$ above which overflowed arrivals will be rejected.
In the computational study, rewards for different class of patients are defined as $R1 = 10$, $R2 = 5$ and the discount factor(α) is taken very close to 1. The proposed VI algorithm is given below:

The stationary policy defined in the value iteration algorithm converges to a 2D threshold based policy which proposes control points x_1^* and x_2^* such that:

Algorithm 3 Value Iteration Algorithm

Require: π^{H1}: the optimal policy for $H1$ when $H2$ is full
π^{H2} : the optimal policy for $H2$ when $H1$ is full
Input: λ, N (given in Table 5.6), $\mu = 1, R1 = 10, R2 = 5, \alpha = 0.97$
1. Initialize:
$st\ \forall X \in \Omega : 0 \leqslant V_0(X) \leqslant V(X)$
$t := 0$
$\forall X, \pi_0^{H1}(X) = 1, \pi_0^{H2}(X) = 1, V_0(X) = 0$
2. $t := t + 1$
for all $X \in \Omega$ **do**
$V_t(X)$
$= \max_{a \in A(X)} \left\{ \sum_{i=1}^{2} \alpha \left(\frac{\lambda_i}{\gamma} T_{A_i} V_{t-1}(X) + \frac{\vartheta}{\gamma} T_{OA_i} V_{t-1}(X) + \frac{N_i \mu_i}{\gamma} T_{D_i} Vt - 1(X) \right) \right\}$
$\pi_t^{H1}(x_1, N_2)$
$= \arg\max_{a \in A(x_1, N_2)}$
$\left\{ \sum_{i=1}^{2} \alpha \left(\frac{\lambda_i}{\gamma} T_{A_i} V_{t-1}(x_1, N_2) + \frac{\vartheta}{\gamma} T_{OA_i} V_{t-1}(x_1, N_2) + \frac{N_i \mu_i}{\gamma} T_{D_i} Vt - 1(x_1, N2) \right) \right\}$
$\pi_t^{H2}(N_1, x_2)$
$= \arg\max_{a \in A(N1, x_2)}$
$\left\{ \sum_{i=1}^{2} \alpha \left(\frac{\lambda_i}{\gamma} T_{A_i} V_{t-1}(N1, x_2) + \frac{\vartheta}{\gamma} T_{OA_i} V_{t-1}(N1, x_2) + \frac{N_i \mu_i}{\gamma} T_{D_i} Vt - 1(N1, x_2) \right) \right\}$
3. Compute:
$dif_{min} = \min_{l \in \Omega} \{V_t(l) - V_{t-1}(l)\}$
$dif_{max} = \max_{l \in \Omega} \{V_t(l) - V_{t-1}(l)\}$
if $dif_{max} - dif_{min} \leqslant \varepsilon \times dif_{min}$ **then**
Policy $\pi_t^{H1}(x_1, N2)$ and $\pi_t^{H2}(N1, x_2)$ are optimal for H1 and H2 respectively
Stop
else
Go to step 2.

- accept *class-2* patients to $H1$ while $(x_1, N2) < (x_1^*, N2)$ and reject while $(x_1, N2) \geqslant (x_1^*, N2)$
- accept *class-2* patients to $H2$ while $(N1, x_2) < (N1, x_2^*)$ and reject while $(N1, x_2) \geqslant (N1, x_2^*)$

The value iteration algorithm is solved independently for three service networks and the optimal policies are found as:

Optimal Policy for Obstetrics (s=1): $x_1^* = 58$ and $x_2^* = 48$
Optimal Policy for Neonatals (s=2): $x_1^* = 29$ and $x_2^* = 19$
Optimal Policy for NICUs (s=3): $x_1^* = 7$ and $x_2^* = 4$

5.5.5 Discrete Event Simulation

In this section we build our simulation model which serves as a practical platform to generate and evaluate various policy scenarios and strengthens the decision making

process by incorporating the complexity in the model which can not be captured by MDP. With the results of those scenarios, Sim model will be able to measure the impact of crucial Markovian assumption posed in MDP model (Poisson distributed overflow streams), the impact of different control policies on admission probabilities and corresponding expected total rewards.

The Sim model is constructed in a commercial software Rockwell Arena 2011. Results are tested in a computer 2.67 GHz and 6 Go. As simulation output, the relevant performance indicators are taken as the admission probabilities of each arriving stream to target hospitals. From admission probabilities, the total reward is calculated by using the unit rewards per patient $R1$ and $R2$. Admission probabilities are defined in Table 5.7.

Table 5.7: Simulation Output

p_i^s	Probability of *class-1* patients admitted in service unit s in hospital $i = \{1, 2\}$
p_{ji}^s	Probability of *class-2* patients rejected from service unit s of hospital $j = \{1, 2, 3, 4, 5, 6\}$ and admitted in service unit s of hospital $i = \{1, 2\}$

Total Reward for each service network s :

$$\text{Total Reward} = R1 \cdot \left(\sum_{i=1}^{2} \lambda_{si} p_i^s \right) + R2 \cdot \left(\sum_{j=1}^{6} \lambda_{sj} \left(\sum_{i=1}^{2} p_{ji}^s \right) \right)$$

where $p_{ji}^s = 0$ for $i = j$

In order to simulate the actual system accurately, some important simulation parameters are to be determined such as number of replications, replication length and warm up period. In order to ensure a 95 % confidence interval for performance indicators, 20 replications are launched. The minimal amount of replications n is calculated using the following formula $\bar{X}(n) \pm t_{n-1,1-\alpha/2} \sqrt{\frac{S^2(n)}{n}}$ where $S^2(n)$ is the estimated standard deviation and $\bar{X}(n)$ is the estimated average for the $X_1, X_2, \ldots, X_n$ values of an indicator for n replications. The value $t_{n-1,1-\alpha/2}$ is the critical point of the distribution t with $n-1$ degrees of liberty and a covering equal to $\frac{1-\alpha}{2}$ where $1 - \alpha = 0.95$.

In order to determine run length and warm up period, we have observed the bed occupancies in the system. Since our model is a pure loss system, there is no queue, therefore entities in the system do not grow exponentially in time and simulation model can reach its steady state condition quickly. For each control policy scenario, SIM model is run with 20 replications, with a run length of 10000 days and with a warm-up period of 100 days. We were able to verify the simulation model by counting on different methods. Firstly, we monitored the flow of entities

and confirmed the conformed behavior of the model to the reality. Secondly, we were able to use Markov Chain modeling for a specific scenario in order to verify the SIM model. The system under *no control* policy (all arrivals are accepted unless hospitals are full) has fewer variables to deal with, therefore we were able to model this system as a Markov Chain. For that specific case, the exact results of Markov model were very close to the results obtained from SIM model, set with the parameters above.

5.5.5.1 Computational Results

In this section, we present the SIM model results and comparison with the MDP optimal policies obtained for obstetric ($s = 1$) and basic neonatal ($s = 2$) services. NICUs are not considered in the simulation runs. In the network only the target hospitals ($H1$ and $H2$) contain NICU units; no external overflow streams (from small hospitals) occurs, so that no assumption of poisson is needed to be considered. Therefore, policy proposed by MDP model is optimal.

The optimal policies obtained from MDP model for obstetrics and neonatal services are used as base scenarios for SIM model. MDP model gives an optimal policy of $x_1^* = 58$ and $x_2^* = 48$ (base scenario-OB) and $x_1^* = 29$ and $x_2^* = 19$ (base scenario-Neo) respectively for Obstetrics and Neonatal Units in $H1$ and $H2$. Several control policy scenarios are generated around these base scenarios to be tested in SIM model. Scenarios are determined by modifying the control parameters x_1^* and x_2^* of base scenario-OB and base scenario-Neo.

Resulting admission probabilities of each scenario and their 95% confidence intervals are given in Table 5.8 and Table 5.9 respectively for Obstetric and Neonatal units. (All of the policies tested could not be presented here due to page restrictions). As x_1^* and x_2^* get smaller, it is expected that admission probabilities of *class-1* patients p_i^s increase while admission probabilities of *class-2* patients p_{ji}^s decrease since more beds are reserved for *class-1* patients.

Figure 5.7 presents Total Rewards calculated for each control scenario simulated for Obstetrics service network. The maximum total reward is 870 and it is achieved at the control points $x_1^* = 55$ and $x_2^* = 43$. SIM optimal policy provides 7% increase in the total reward compared to *no control* policy case. The optimal control points proposed by SIM model is different than the ones obtained from MDP model as expected due to Markovian assumption. However, it is important to note that % difference between total rewards obtained from both policies is not significant such that total best reward of SIM is 1.5% higher than the one of MDP. Futhermore, it can be clearly observed from Figure 5.7 that reserving two beds in each hospital (as proposed by MDP) brings the biggest rise in total reward. Reserving more beds until reaching the SIM optimal points helps to increase total reward to its max level but with a declined slope. Additionally, reserving more beds after reaching the optimal points proposed by SIM does not have any significant effect on total reward.

Table 5.8: Admission Probabilities computed for various scenarios for Obstetrics Units

x_1^*	x_2^*	p_1	p_2	p_{12}	p_{21}	p_{31}	p_{32}	p_{41}	p_{42}	p_{51}	p_{52}	p_{61}	p_{62}	Total Rew.
60	50	0.674 ± 0.02	0.61 ± 0.018	0.153 ± 0.02	0.104 ± 0.013	0.099 ± 0.002	0.098 ± 0.0017	0.135 ± 0.003	0.133 ± 0.002	0.109 ± 0.0024	0.106 ± 0.0011	0.094 ± 0.002	0.092 ± 0.0014	807.1
59	49	0.766 ± 0.008	0.688 ± 0.008	0.082 ± 0.01	0.049 ± 0.006	0.086 ± 0.0003	0.092 ± 0.0008	0.11 ± 0.0006	0.12 ± 0.0008	0.093 ± 0.0005	0.1 ± 0.0008	0.081 ± 0.0003	0.087 ± 0.0007	845.6
58	48	0.796 ± 0.004	0.715 ± 0.0044	0.05 ± 0.006	0.027 ± 0.003	0.083 ± 0.0	0.09 ± 0.0006	0.11 ± 0.0002	0.122 ± 0.0008	0.091 ± 0.0001	0.098 ± 0.0004	0.078 ± 0.0001	0.084 ± 0.0004	856.8
56	47	0.819 ± 0.0016	0.734 ± 0.0021	0.02 ± 0.003	0.015 ± 0.002	0.082 ± 0.0	0.087 ± 0.0003	0.112 ± 0.00	0.119 ± 0.0002	0.09 ± 0.0	0.095 ± 0.0005	0.077 ± 0.0002	0.082 ± 0.0002	865.4
56	44	0.821 ± 0.0015	0.743 ± 0.0008	0.02 ± 0.0016	0.003 ± 0.0004	0.081 ± 0.0002	0.087 ± 0.0004	0.111 ± 0.0	0.12 ± 0.0003	0.089 ± 0.0	0.096 ± 0.0004	0.077 ± 0.0002	0.082 ± 0.0	867.1
55	43	0.826 ± 0.0009	0.748 ± 0.0006	0.01 ± 0.0017	0.002 ± 0.0002	0.082 ± 0.0001	0.086 ± 0.0002	0.112 ± 0.0001	0.119 ± 0.0001	0.09 ± 0.0	0.095 ± 0.0004	0.077 ± 0.0002	0.081 ± 0.0001	870
55	42	0.825 ± 0.0009	0.748 ± 0.0004	0.013 ± 0.0017	0.001 ± 0.0001	0.081 ± 0.0002	0.087 ± 0.0003	0.112 ± 0.0	0.119 ± 0.0	0.09 ± 0.0	0.095 ± 0.0004	0.077 ± 0.0002	0.082 ± 0.0001	869
55	41	0.825 ± 0.0009	0.746 ± 0.0006	0.014 ± 0.0017	0.001 ± 0.0	0.082 ± 0.0	0.087 ± 0.0002	0.112 ± 0.0	0.118 ± 0.0	0.09 ± 0.0	0.095 ± 0.0004	0.077 ± 0.0003	0.082 ± 0.0001	868.4
0	0	0.831 ± 0.0002	0.751 ± 0.0001	0	0	0.082 ± 0.0001	0.085 ± 0.0003	0.112 ± 0.0	0.117 ± 0.0	0.09 ± 0.0	0.094 ± 0.0002	0.077 ± 0.0003	0.081 ± 0.0001	869.7

Table 5.9: Admission Probabilities computed for various scenarios for Neonatal Units

x_1^*	x_2^*	p_1	p_2	p_{12}	p_{21}	p_{31}	p_{32}	p_{41}	p_{42}	Total Reward
30	20	0.862 ± 0.0006	0.845 ± 0.0005	0.091 ± 0.0003	0.074 ± 0.0002	0.135 ± 0.0001	0.141 ± 0.0003	0.183 ± 0.0003	0.193 ± 0.0003	**367.38**
29	19	0.885 ± 0.002	0.88 ± 0.004	0.06 ± 0.0024	0.053 ± 0.004	0.127 ± 0.0011	0.14 ± 0.0003	0.172 ± 0.0012	0.192 ± 0.0001	**371.28**
28	18	0.894 ± 0.003	0.894 ± 0.005	0.044 ± 0.003	0.042 ± 0.006	0.123 ± 0.0016	0.139 ± 0.0002	0.168 ± 0.002	0.19 ± 0.0002	**372.09**
27	18	0.897 ± 0.003	0.896 ± 0.005	0.036 ± 0.004	0.041 ± 0.007	0.122 ± 0.0016	0.139 ± 0.0002	0.166 ± 0.0021	0.19 ± 0.0003	**372.24**
25	18	0.901 ± 0.004	0.899 ± 0.005	0.025 ± 0.004	0.04 ± 0.008	0.121 ± 0.0018	0.137 ± 0.0003	0.166 ± 0.0022	0.188 ± 0.0005	**372.45**

Figure 5.8 presents Total Rewards calculated for each control scenario simulated for basic neonatal service network. The maximum total reward is 372.5 and it is achieved at the control points $x_1^* = 25$ and $x_2^* = 18$, on the other hand total rewards obtained from the control scenarios lie in the blue region in Figure 5.8 are quite close to the attained maximum reward, therefore those policies can all be

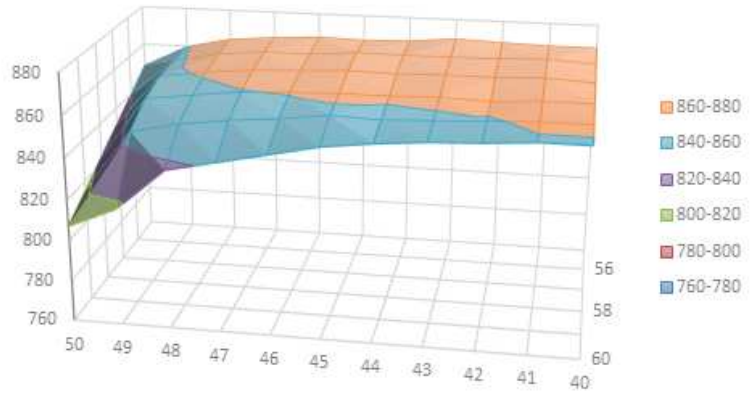

Figure 5.7: Total Rewards obtained for each scenario in Obstetric Units

considered as alternative optimum policies. Again as expected, the optimal control policies proposed by SIM model are different than the MDP optimal control policy due to the Markovian assumption. However, % difference between total rewards is not significant such that the max total reward obtained from SIM is 0.3% higher than the total reward of MDP optimal policy which is computed as 371.28.

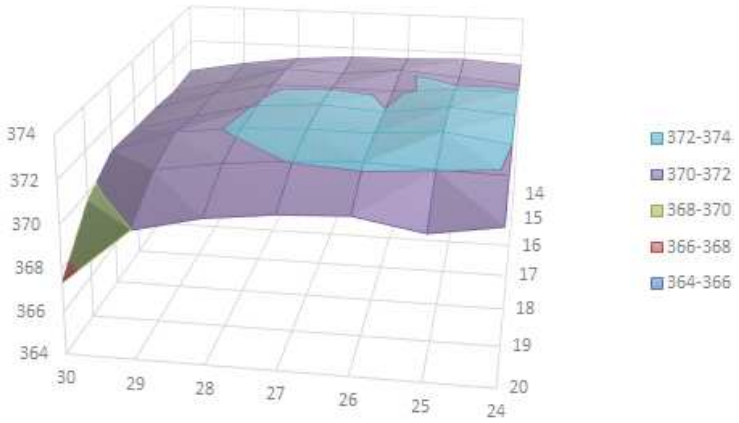

Figure 5.8: Total Rewards obtained for each scenario in Neonatal Units

In general it can be concluded for neonatal units that reserving beds (posing an admission control policy) do not make a big impact on admission probabilities (given in Table 5.9), consequently we do not observe a significant % improvement in total rewards. If we compare the total rewards obtained from the optimal policy of SIM and *no control* policy, we can state that % difference does not exceed 1.5%.

5.5.6 Conclusion

In this section, we considered the problem of finding an optimal admission control policy for two target hospitals in a pure-loss perinatal network. In order to model the system analytically, a Markovian assumption was necessary to make. Under this assumption, the system is modeled as an MDP process and an MDP optimal policy is computed. Afterwards, simulation model is used to strengthen the decision making process by incorporating the complexity in the model which can not be captured by MDP and measure the impact of Markovian assumptions on total reward. For that aim, MDP optimal policy is considered as a base scenario and various control scenarios are generated around the base scenarios and evaluated with the simulation model. It has been observed that Simulation model proposes to reserve more beds in both of the hospitals in its optimal scenarios compared to MDP optimal policies. Final total reward obtained from SIM optimal policies is higher than the total reward obtained from MDP optimal policies. However, the percentage differences are not always significant changing with the considered instance.

In this study, overflowed patients are assumed to arrive to two target hospitals and admission control policies are investigated only for those hospitals. Even though this assumption is realistic for our case study held in a specific perinatal network, in general in pure-loss networks such as "network of emergency units" overflow arrivals are not restricted with two target hospitals. For a future work, we believe it is scientifically interesting to consider a pure-loss healthcare network where overflows are less restricted as in reality and deal with the problem of finding optimal admission control policies for all hospitals in the network.

5.6 Admission Control Policy for I-Hospital Loss Network

5.6.1 System Description

In this section, we extend the 2-hospital loss system to multi-hospital (I>2) loss network. Consider a healthcare service network where there is a set $\mathcal{I} = 1, 2, ..., I$ of $i, j, k, t... \in \mathcal{I}$ hospitals offering service to patients in the region.

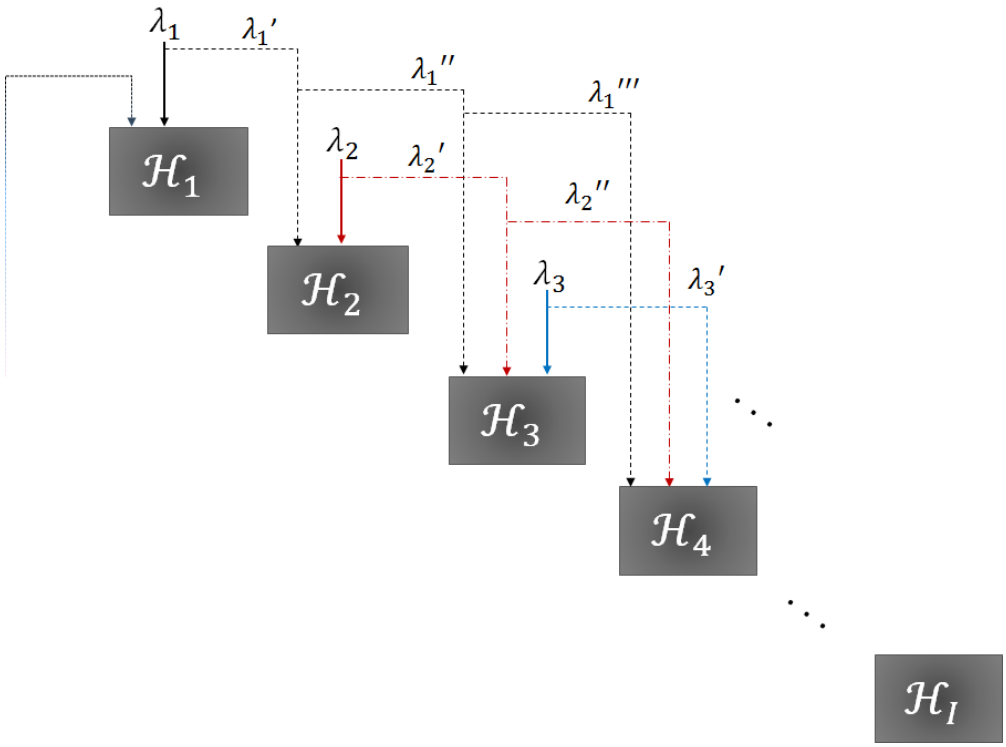

Figure 5.9: Representation of a I-Hospital Overflow Network

Assumption 1: In the network, patients have the flexibility to choose at which "facility" to be served according to their preference issues such as proximity, trustworthiness, etc. "The facility" which is chosen by a set of patients in the first place is called "preferred hospital" for that set of patients. And those patients create the major arrival stream of that hospital. In this work, it is assumed that patient set i arrive to their most preferred hospital i with an independent Poisson process of rate λ_i.

Assumption 2: The service time of a patient in a hospital is a random variable of exponential distribution and it is same for all patients. No classification among patients has been considered in terms of service time.

Remark 1: For our application field, perinatal network, Markovian assumptions are valid. Poisson arrivals are often observed in systems without appointments but with arrivals controlled by nature such as birth delivery (Green, 2004). Furthermore, in heavily stochastic systems, the coefficient of variation (CV) of length of stay (LOS), which is defined as the ratio of the standard deviation to the mean, is typically close to one, satisfying the usage of negative exponential distribution as a service time assumption.

Assumption 3: Each hospital i has N_i number of identical servers (beds). Each server serves one patient at a time.

Assumption 4: Healthcare network considered in this study is an overflow loss network as presented in Figure 5.9. If all N_i servers of a hospital i is occupied, incoming patients to hospital i are deferred to other hospitals in the network in accord with their preferences (will be explained later) or out of network.

Assumption 5: Each patient i has a preference list of facilities $Pref_i = \{i, j, k, ...t\}$ which is given in a descending order of preference. According to $Pref_i = \{i, j, k, ..., t\}$, a patient i arriving at hospitals $i, j, k, ..., t$ are called *class-1, class-2, class-3,..., class-I* arrival, respectively. As inherent, the first entry in the list is the most preferred hospital for patient i which is hospital i.

Assumption 6: Acceptance of each patient brings a reward to the system. The highest reward is associated with the admission of patients to their most preferred hospitals (*class-1* arrivals). As the number of overflow increases (*class-w arrivals w>1*), the reward obtained from its admission decreases respectively. No reward is associated with the patients who are rejected to out of network.

Remark 2: In our application field, perinatal networks, deferring a patient to another hospital may cause long travel distances which consequently might cause fatal occurrences. Therefore, admission rate of each patient to their 1^{st} preferred hospital is an important and prioritized performance measure for perinatal networks.

Assumption 7: Any incoming patient arriving to any hospital can be accepted or rejected at that hospital. The admission decision is given by a controller (bed manager) in order to maximize the total discounted reward over an infinite horizon.

Assumption 8: If a patient i is not accepted to any of the hospitals listed in $Pref_i$, the overflowed patient is rejected to out of network with other words the patient is lost.

5.6.2 MDP Formulation

In this section, a Markov Decision Process (MDP) formulation is presented for $I > 2$ loss network. The system can be modeled as a continuous time Markov Chain model with state $x = (x_1, x_2, ..., x_I)$ defined as the number of occupied beds (number of patients) in each hospital where $\Omega = \{(x_1, x_2, ..., x_I) : 0 \leqslant x_i \leqslant N_i \, \forall i \in I = \{1, 2, ..., I\}\}$ is the state space. Control action space is defined by a IxI dimensional mapping matrix with $a_{ji}(X) = 1(0)$ indicates that the control action is to admit(1) or reject(0) the patient $j \in \mathcal{I}$ to hospital $i \in \mathcal{I}$ where a_{ii} denotes the control of *class-1* arrivals.

In some regions of the state space, there is no admission control such that decisions are already defined inherently:

- If none of the hospitals are full ($x_i < N_i \, \forall i \in \mathcal{I}$), all patients are accepted to their most preferred hospitals ($a_{ii} = 1 \, \forall i \in \mathcal{I}$).
- If all hospitals are full ($x = \mathbf{N}$ where $x_i = N_i \, \forall i \in \mathcal{I}$), all patients are rejected ($a_{ji} = 0 \, \forall i, j \in \mathcal{I}$).

In the formulations we make use of the following notation and conventions:

- $Pref_i = \{i, j, k, ..., t\}$ denotes the list of hospitals in a descending order of preference for patient i to visit
- r_{ij} : reward associated with acceptance of a patient i to hospital $j \in Pref_i = \{i, j, k, ..., t\}$ where $r_{ii} \geqslant r_{ij} \geqslant r_{ik} \geqslant ... \geqslant r_{it} \geqslant 0$
- e_i is the i^{th} unity vector.
- Uniformization rate is taken as $\gamma = \sum_{i=1}^{I} (\lambda_i + N_i \mu_i)$ and parameters used in uniformized model are $p_i = \lambda_i / \gamma$, $q_i = (N_i \mu) / \gamma$

Our objective is to find an optimal admission control policy π^* that maximizes the expected total α-discounted reward over an infinite horizon. Optimality Equation of the event-based dynamic programming can be written as:

$$Tf(x) = \alpha \sum_{i=1}^{I} (p_i T_{A_i} f(x) + q_i T_{D_i} f(x))$$

Arrival operator T_{A_i} models admission control of a patient i.

$$T_{A_i} f(x) = \max\{r_{ij} + f(x + e_j), f(x)\} \qquad \forall j : j \in Pref_i = \{i, j, k, ..., t\} \wedge 0 \leqslant x_j < N_j$$

The Departure operator T_{D_i} models departures of a patient from hospital $i \in \mathcal{I}$

$$T_{D_i} f(x) = \frac{x_i}{N_i} f(x - e_i) + \frac{N_i - x_i}{N_i} f(x) \qquad \forall x \in \Omega$$

Due to the involvement of numerous hospital states ($x = (x_1, x_2, ..., x_I)$) in each admission decision, proving the structure of the optimal policy is a challenging task. Furthermore, the optimal control pattern of the optimal policy is expected to highly depend on the structure of the overflow routing between hospitals (preference lists). On the other side, large-scale of the loss network makes the problem quite complex such that even finding a numerical solution via value iteration algorithm gets computationally demanding very fast as the number of hospitals increases.

Therefore, in the next section we propose a near-optimal heuristic policy to approximate the optimal policy in large-scale overflow loss networks. The proposed heuristic is based on a local threshold admission control policy. In order to assess the performance of the proposed heuristic policy, in subsection 5.6.4 we construct several LP models that provide upper-bounds to the problem.

5.6.3 Local Admission Control Policy

Local admission policy is built upon the idea of considering each hospital and its all possible arrival streams locally. Our objective is to obtain optimal control points for each hospital which indicate how many beds to reserve for each arriving stream (with other words when to admit/deny each arriving stream). Therefore, an admission control policy will be determined for each arriving stream in each hospital.

At first, we give the following notation that will be used repeatedly in this section to explain the proposed policy:

- $Pref_j = \{j, ..i, ..\}$: List of hospitals in descending order of preference of patient j. From this list, we can extract the ranking (preference degree) of hospital i for a patient j.
- w_{ji} : the ranking (preference degree) of hospital i for a patient j where $w_{ii} = 1$ and $w_{ji} > 1$
- $P_{w_{ji}}$: Overflow probability from hospital j to hospital i.
- r_{ji} : reward associated with acceptance of a patient j to hospital $i \in Pref_j$. Reward gets smaller as the ranking of hospital i increases in $Pref_j \quad , \forall j \in \mathcal{I}$.

$$r_{ii} \geqslant r_{ji} \geqslant r_{ki} \geqslant ... \geqslant r_{si} \geqslant \text{given that } w_{ii} = 1 < w_{ji} \leqslant w_{ki} \leqslant ... \leqslant w_{si} = I$$

Given assumption 5 in the previous section, each hospital $i \in \mathcal{I}$ may have various types of arrival streams:

- A hospital i is always fed by *class-1* arrivals (λ_i)
- A patient $j \in \mathcal{I} - i$ may overflow to hospital i only if $i \in Pref_j$

- A patient $j \in \mathfrak{I} - i$ may overflow to hospital $i \in Pref_j$ only if patient j is rejected from all other hospitals with smaller rankings (higher preference degrees) than hospital i such that $w_{js} < w_{ji}, \forall s \in Pref_j - i$.

Therefore, a hospital i may have a *Class-*w_{ji} arrival overflowed from a hospital $j \in \mathfrak{I}$ where $i \in Pref_j$. Incoming overflow rates $\lambda_j * P_{w_{ji}}$ depends on the ranking (w_{ji}) of hospital i in $Pref_j$ and gets smaller as w_{ji} increases. The calculation of overflow rates is given in the next section.

All possible incoming arrival streams to hospital i with arrival rates $(\lambda_j * P_{w_{ji}})$ and associated rewards (r_{ji}) are presented in Figure 5.10.

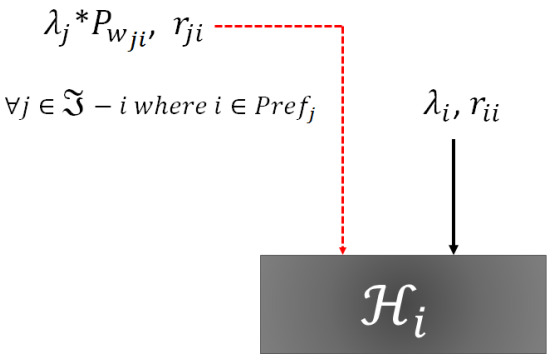

Figure 5.10: A representation of the possible incoming streams to a hospital i

With the assumption of Poisson arrivals for overflow streams, each hospital can be approximated as a 1-hospital multi-arrival stream system addressed early in this section 5.3. For each considered hospital $i \in \mathfrak{I}$, state space is determined as $x_i : \{0 \leqslant x_i \leqslant N_i\}$ number of beds occupied in the hospital. Let,

AR_{si}= set of patients (of hospital) $s \in \mathfrak{I}$ where hospital $i \in Pref_s$

The event-based dynamic-programming optimality equation is given below. By solving this equation to optimality, we can obtain the local control points for each arriving stream of hospital $i \in \mathfrak{I}$:

$$Tf(x_i) = \alpha \left(\sum_{j \in AR_{si}} \frac{\lambda_j * P_{w_{ji}}}{\gamma_i} T_{A_i} f(x_i) + \frac{N_i \mu_i}{\gamma_i} T_{D_i} f(x_i) \right)$$

where $\gamma_i = \sum_{j \in AR_{si}} \lambda_j * P_{w_{ji}} + N_i \mu_i$ is uniformization rate for each hospital i and α is discount factor where $0 < \alpha < 1$.

The Arrival operator T_{A_i} models admission control of any patient to hospital i

$$T_{A_i} f(x_i) = \min \begin{cases} \max\{r_{ji} + f(x_i + 1), f(x_i)\} & \text{for } x_i < N_i \\ f(x_i) & \text{for } x_i = N_i \end{cases}$$

The Departure operator T_{D_i} models departures from hospital i

$$T_{D_i} f(x_i) = \frac{x_i}{N_i} f(x_i - 1) + \frac{N_i - x_i}{N_i} f(x_i) \qquad \text{for all } x_i$$

An optimal control policy should be chosen in each state x_i to maximize the future expected discounted revenue as given by the optimality equation. We use the well-known computational algorithm "Value Iteration" to obtain the optimal policy in each hospital i.

Calculation of Overflow Rates

Overflowed patients create non-Poisson arrival streams at hospital i. However, we assume a Markovian system in order to be able to use MDPs. The overflow rates are calculated based on Erlang Loss Blocking Probability $B(\lambda/\mu, N)$. In this study, service times are identical for all patients and normalized $\mu = 1$. Thus, from here, Erlang loss blocking probability is represented as $B(\lambda, N)$.

Step 1. Calculation of Initial overflow rates

In the 1st step, we assign initial/tentative values to overflow probabilities ($P_{w_{ji}}$). For determining these tentative values we assumed that a patient gets rejected and overflows only if the visited hospital is full (no consideration of threshold points) and hospitals are assumed to be independent of each other without overflow among hospitals. Therefore, a patient j overflows to (visits) hospital i only if her first preferred hospital (hospital j) and all other hospitals that patient j needs to visit before arriving to hospital i are full. By assuming independency among hospitals, we used Erlang loss probability to calculate the blocking of a patient j in a hospital.

Since in the proposed local control policy we focus on one hospital at a time, below (Table 5.10) we present the computation of overflow probabilities (of all class of streams) to hospital i. A patient j is used as reference arrival. At every line we assume that patient j needs to visit one more hospital before arriving to hospital i. Therefore, all possible arrival classes to hospital i are considered.

If patient j is the *class-1* arrival, she visits hospital i as the first hospital (most preferred hospital), thus no overflow probability incurs. As a *class-2* arrival, patient j arrives to hospital i after she is rejected from its most preferred hospital j, which is computed by using $B(\lambda_j, N_j)$. Similarly, as a *class-3* arrival, patient j can arrive to hospital i, after she is rejected from her most preferred hospital (j) and second preferred hospital (k).

Table 5.10: Calculation of Initial Overflow Rates

Possible Class of Arrivals to hospital i	**Overflow Rates from hospital** j **to hospital** i $\lambda_j * P_{w_{ji}}$
Class-1 arrivals ($w_{ji} = 1$)	λ_j
Class-2 arrivals ($w_{ji} = 2$)	$\lambda_j * B(\lambda_j, N_j)$
Class-3 arrivals ($w_{ji} = 3$)	$\lambda_j * B(\lambda_j, N_j) * B(\lambda_k, N_k)$ where $w_{jk} = 2$
Class-4 arrivals ($w_{ji} = 4$)	$\lambda_j * B(\lambda_j, N_j) * B(\lambda_k, N_k) * B(\lambda_t, N_t)$ where $w_{jk} = 2, w_{jt} = 3$
$\vdots$	$\vdots$
Class-I arrivals ($w_{ji} = I$)	$\lambda_j * \prod_{\forall k \in Pref_j - i} * B(\lambda_k, N_k)$ where $w_{jk} = \{2, 3, \cdots, I-1\}$

Step 2. Solving Structured Value Iteration Algorithm

Each hospital $i \in \mathcal{I} - i$ is focused independently. A VI algorithm (which is developed in section 5.3 for a single hospital with m-arrivals) is solved for each hospital i with its several class of arrivals. For each class of arrival, the arrival traffic is calculated by using the initial overflow rates presented in step 1. Solutions of VI algorithms give us the local switching (control) points for each stream arriving at each hospital i.

Step 3. Modification of overflow rates

Using merely "being full probability of a hospital" to calculate overflow rates may result in underestimated rates. An overflowed patient arriving to a hospital k is not rejected only when the hospital k is full, but also when hospital k is occupied more than a certain threshold (different for each class of patient). Thus, in step 3, we incorporate the pre-defined control points in computation of overflow rates. We iteratively use the previously computed control points for each hospital in network to update the overflow rates as presented in Table 5.11

We continue to assume that hospitals are independent of each other. Similar to step 1, a patient j is used as a reference arrival to hospital i. If patient j is a *class-1* arrival, she visits hospital i as the first hospital (most preferred hospital), thus no overflow probability incurs. As a *class-2* arrival, patient j arrive to hospital i after she is rejected from its most preferred hospital j. No control policy is employed to patient j in hospital j since she is a *class-1* arrival; she gets rejected only if hospital j is full. Control policies get involve to the computations starting from *class-3* arrivals. As a *class-3* arrival, patient j overflows to hospital i, if hospital j is full and hospital k (the second preferred hospital) is occupied more than a certain threshold. Patient j arrives to hospital k as a *class-2* arrival, thus patient j is rejected from hospital k only if $x_k \geqslant C_k^2$. We approximate the probability of $(x_k \geqslant C_k^2)$ as a conditional probability of "hospital k being occupied more than C_k^2 given that hospital j is full". In the computation of this conditional probability, we

assumed that there are two class of arriving streams to hospital k; its own *class-1* patients (λ_k) and *class-2* arrivals (patients of hospital j (λ_j). Therefore, the computation of blocking probability that hospital k employs for any class of arrival w_{jk} from hospital j can be defined as:

$$P\left(C_k^{w_{jk}} \leqslant x_k \leqslant N_k | x_j = N_j\right) = \frac{\sum_{C_k^{w_{jk}}}^{N_k} (\lambda_k + \lambda_j)^n / n!}{\sum_{n=0}^{N_k} (\lambda_k + \lambda_j)^n / n!}$$

where $C_k^{w_{jk}}$ = control point in hospital k for class-w_{jk} arrivals.

Remark: Blocking experienced by patient j in hospital k is computed by isolating two hospitals from the rest and assuming that the arrival load of hospital k is $\lambda_k + \lambda_j$. In this approximation, we do not consider the other possible arrival streams neither to hospital j nor to hospital k from other hospitals (being full possibility of other hospitals).

Table 5.11: Calculation of Updated Overflow Rates

Type of Arrivals to hospital i	**Overflow Rates from hospital** j **to hospital** i: $\lambda_j * P_{w_{ji}}$
Class-1 arrivals ($w_{ji} = 1$)	λ_j
Class-2 arrivals ($w_{ji} = 2$)	$\lambda_j * B(\lambda_j, N_j)$
Class-3 arrivals ($w_{ji} = 3$)	$\lambda_j * B(\lambda_j, N_j) * P\left(C_k^{w_{jk}} \leqslant x_k \leqslant N_k \mid x_j = N_j\right)$ where $w_{jk} = 2$
Class-4 arrivals ($w_{ji} = 4$)	$\lambda_j * B(\lambda_j, N_j) * P\left(C_k^{w_{jk}} \leqslant x_k \leqslant N_k \mid x_j = N_j\right) * P\left(C_t^{w_{jt}} \leqslant x_t \leqslant N_t \mid x_j = N_j\right)$ where $w_{jk} = 2, w_{jt} = 3$
$\vdots$	$\vdots$
Class-I arrivals ($w_{ji} = I$)	$\lambda_j * \prod_{\forall k \in Pref_j - i} * P\left(C_k^{w_{jk}} \leqslant x_k \leqslant N_k \mid x_j = N_j\right)$ where $w_{jk} = \{2, 3, \cdots, I-1\}$

Step 4. Iteration

A fix point iteration process is used to improve overflow rates and accordingly the switching points in an iterative way. In each iteration, the overflow rates are modified with control points obtained from the previous iteration, the value iteration algorithm in step 2 is solved with modified overflow rates and new control points are obtained. Finally, with the convergence of the algorithm, we obtain the control points for each arriving stream of each hospital i in the network suggesting the bed manager how many beds to reserve for each arriving stream.

Algorithm 4 Local Admission Policy

1. Initialization:
$t := 0$
for all $i \in \mathcal{I}$ do
1.1 Calculate initial $P_{(w_{ji})}$ for all $j \in \mathcal{I}$ where $i \in Pref_j$
1.2 Solve VI algorithm to obtain initial control points $[C_i^{class-w}]^0$ where $w = 1, 2, ..., I$
2. $t := t + 1$
for all $i \in \mathcal{I}$ **do**
2.1 Update $P_{(w_{ji})}$ for all $j \in \mathcal{I}$ where $i \in Pref_j$ using $[C_i^{class-w}]^{(t-1)}$
2.2 Solve VI algorithm to obtain $[C_i^{class-w}]^t$ where $w = 1, 2, ..., I$
3. Compute:
if $[C_i^{class-w}]^t - [C_i^{class-w}]^{(t-1)} \leqslant \varepsilon$ **then**
Stop
else
Go to step 2.

The proposed local control policy is required to be tested computationally and its performance should be compared against the global optimal policy. For small-scale network ($I < 4$), it is possible to solve the global admission control policy to optimality. However, a big-scale hospital network where each hospital has multiple servers (more than 30 beds in perinatal network) and different class of arrival streams becomes a complex problem and easily exceeds the limits of the programs used. Therefore, in the next section we propose several upper bound formulations in order to be able to assess the optimality gap of the local admission control policy in big-scale networks.

5.6.4 Upper-Bounds of the Optimal Rewards

Several LP models are developed to provide upper-bounds to the problem of admission control in large-scale loss networks. The best LP model which computes the tightest upper-bound is presented below and is used to assess the performance of the proposed heuristic policy. For the sake of the construction of the idea, we present other developed LP models, including the ones which do not provide good bounds in Appendix F.

All models presented in Appendix F are tested on some instances and we observe that in order to obtain a tighter upper bound, we need to better bound the maximum *class-1* load that can be accepted in a hospital i. Below we introduce the best UB formulation with updated constraints 5.5 and 5.6.

Notation:

G	set of the hospitals where $i \in G = \{0, 1, ..., I\}$
a_i	arrival load of hospital i, where $a_i = \lambda_i/\mu = \lambda_i$ since $\mu = 1$ and $a(G) = \sum_i a_i$
N_i	number of beds in hospital i
$P(G)$	global overflow probability if all hospital beds are grouped in a single hospital where $P(G) = B(a_1 + ...a_I, N_1 + ...N_I)$
P_i	isolated overflow probability if each hospital only accepts its own patients, $P_i = B(a_i, N_i)$
A_i	Sum of arrival rates of hospital i and hospitals $j \in \mathcal{J}$ who may overflow to hospital i where $i \in Pref_j$ $A_i = \sum_{j \in \mathcal{J} \text{ if } i \in Pref_j} a_j$
C_i	Possible switching points in hospital i which can take values between $(0, N_i)$
r_{ij}	reward of a patient of hospital i admitted in hospital j

Decision Variable:

x_{ij}	load of patients of hospital i admitted in hospital j .

Our objective is to maximize the total reward obtained from the accepted load to each hospital:

$$UB = \max \sum_{i \in G} \sum_{j \in G} r_{ij} x_{ij}$$

subject to

$$\sum_j x_{ij} \leqslant a_i \tag{5.4}$$

$$\sum_j x_{ji} \leqslant (1 - B(A_i, N_i))\, A_i \tag{5.5}$$

$$x_{ii} \leqslant a_i \left[1 - BB\left(a_i, A_i, C\left(\sum_j x_{ij}\right), N_i\right)\right] \tag{5.6}$$

$$\sum_{i,j} x_{ij} \leqslant (1 - P(G))\, a(G) \tag{5.7}$$

$$x_{ij} \geqslant 0 \tag{5.8}$$

Constraint 5.4 ensures that the total accepted load of patients i to all hospital cannot bigger than its major load. Constraint 5.5 limits the maximum load can be accepted at a hospital i. Unlike previous formulation where all system arrivals ($a(G)$) are considered, in constraint 5.5 we consider simply sum of the arrivals of hospitals who may visit hospital i since hospital i is in their preferred hospital list. Thus, we feed the N_i servers of hospital i with the maximum possible arrival rate that hospital can receive (A_i), and calculate the admission probability of A_i. The resulting right-hand-side admission rate gives us the **maximum total-accepted-load** to hospital i (including all class of arrivals). From here, aL_i denotes the total-accepted-load to hospital i such that $aL_i = \sum_j x_{ji}$.

Constraint 5.6 puts an upper bound to *class-1* accepted load at each hospital assuming that each hospital receives two arrival streams (*class-1* and *class-2*) and the hospital imposes a control policy for the *class-2* arrivals. In constraint 5.6, we find an upper bound for the admission of *class-1* patients to hospital i by imposing control points. Here, we define a pure loss system for each hospital i (a_i, A_i, C_i, N_i) of a single group of N_i servers serving two arrival streams a_i and $(A_i - a_i)$ with switching (control) point $C_i \in \{0, 1, ..., N_i\}$. The arrivals of the stream $(A_i - a_i)$ is the overflow stream and lost for all states $X_i \geqslant C_i$. The arrivals of stream a_i (*class-1* arrivals) is assumed to be lost at state $X_i = N_i$. It is a birth-death process with birth rate equal to A_i or a_i depending on state X_i and death rate is equal to X_i. $BB(a_i, A_i, C_i, N_i)$ is the loss probability of stream a_i, i.e. stationary probability of being in state $X_i = N_i$; π_N.

A value is selected for C_i by the model such that the total-accepted-load of hospital i with admission control (a_i, A_i, C_i, N_i) will be exactly equal to aL_i which is the total-accepted-load defined in constraint 5.5. Furthermore, we claim that the control policy C_i chosen by the optimization model actually minimizes the blocking rate of a_i, therefore maximizes the accepted load of a_i to hospital $i(x_{ii})$ among all policies achieving exactly the same total-accepted-load (aL_i). This argument is proved at the end of this section.

As in some previous models, constraint 5.7 ensures an upper bound on the total-accepted-load in the network by considering that all hospital beds are grouped in a single hospital ($N(G)$) receiving the total arrival rate in the network ($a(G)$).

The non-linear constraints of UB can be linearized with additional binary variables defining the select value of C.

y_{ic}: auxiliary binary variable for choosing the appropriate control point $c = C$

A mix-integer linear programming model (UB') is formulated as follows:

$$UB' = \sum_{i \in G} \sum_{j \in G} r_{ij} x_{ij}$$

s.t.

$$\sum_{j} x_{ij} \leqslant a_i \tag{5.9}$$

$$\sum_{i} x_{ji} \leqslant (1 - B(A_i, N_i))\, A_i \tag{5.10}$$

$$x_{ii} \leqslant \sum_{c=0}^{N_i} a_i \,(1 - BB(a_i, A_i, c, N_i))\, y_{ic} \tag{5.11}$$

$$\sum_{j} x_{ji} \leqslant \sum_{c=0}^{N_i} BL(a_i, A_i, c+1, N_i) y_{ic} \tag{5.12}$$

$$\sum_{c=0}^{N_i} y_{ic} = 1 \tag{5.13}$$

$$\sum_{i,j} x_{ij} \leqslant (1 - P(G))\, a(G) \tag{5.14}$$

$$x_{ij} \geqslant 0, y_{ic} \in 0, 1 \tag{5.15}$$

where, $BL(a_i, A_i, c, N_i)$ is the accepted-load in hospital i with a threshold c and by convention, $BL(a_i, A_i, N_i + 1, N_i) = BL(a_i, A_i, N_i, N_i)$
$y_{ic} = 1$ if the accepted-load $BL(a_i, A_i, c, N_i) \leqslant \sum_j x_{ji} < BL(a_i, A_i, c+1, N_i)$.

Compared to the previous upper bound formulations, UB performs much better in providing a tight upper bound. However, in the construction of UB we rely on some facts that need to be proved in order to verify that UB provides an upper bound. The facts and proofs of properties are presented in the following.

Fact : There exists a control policy C minimizes the rejection rate of *class-1* arrivals (a_i) among all other control policies achieving the same total accepted load of $aL(C)$.

Let,
$aL1(C)$ = accepted load of stream $a1$ with control point C
$aL2(C)$ = accepted load of stream $a2$ with control point C
$aL(C) = aL1(C) + aL2(C)$

Property 1: Consider a single loss queue with multiple servers fed by two different class of arrival streams. Admission of a *class-2* patient leads to a reward 1, whereas admission of a *class-1* patient leads to a reward r>1. The optimal admission control of such a system is a threshold policy $C(r)$ for controlling *class-2* arrivals.
Proof 1: Proof of existence of a threshold optimal policy for a single loss queue with m-class of arrivals is presented in section 5.3. Proof is given in Appendix E.1.

Property 2: As a result of Property 1, $C(r)$ is the control point that maximizes the total reward, i.e. $TR^* = r.aL1(C(r)) + aL2(C(r))$.

Property 3:
(a) As r changes from 1 to infinity, optimal control point changes continuously from N to 0. For all $r \leqslant r'$, $C(r) \geqslant C(r')$
(b) For $r = 1$, no control policy exists such that $C(1) = N$
(c) For some big enough r, optimal control point tends to zero; $C(r) = 0$
Proof. Easy consequences of Property 2.

Property 4: $aL1(C)$ decreases in C and $aL2(C)$ increases in C. $aL1(r)$ increases in r and $aL2(r)$ decreases in r.

Property 5: $C(r)$-policy maximizes $aL1$ or equivalently minimizes rejection rate of stream a_1 among all policy π such that $aL2^\pi \geqslant aL2(C(r))$ where $aL^\pi = aL(C(r))$.

Proof by contradiction. Assume that $aL1^\pi > aL1(C(r))$. Let $TR^\pi(r)$ and $TR(C(r))$ be the total reward of policy π and the optimal reward, respectively.

$TR^\pi(r) = r.aL1^\pi + aL2^\pi > r.aL1(C(r)) + aL2(C(r)) = TR(C(r))$

which contradicts the optimality of $C(r)$-policy and concludes the proof. Q.E.D.

5.6.5 Numerical Results

In this section, we analyzed the performance of the proposed local optimal policy on different numerical examples where we consider different size networks and different overflow scenarios. For big size networks, where a global optimal policy is not computable, the model UB is used to obtain a good upper bound to the problem which gives an idea about the optimality gap of the local optimal policy.

In our examples, we avoid to use extremely heavy and light arrival loads regarding the results of the previous numerical examples for 2-hospital case in section 5.4. As explained in detail, for light and heavy loads, the optimal control policy tends to resemble no control policy or full control policy, respectively. Furthermore, medium load represents better the realistic system compared to light and heavy load in a perinatal network.

We generate three instances: 3-hospital overflow network, 6-hospital overflow network with two different overflow routing. Non-identical hospitals with different number of servers are considered in order to represent the real system better. In this section we are interested in observing the performance of the two proposed methodologies: the local admission policy to compute a near-optimal solution and UB model to find a tight upper bound. Additionally we analyzed the effect of overflow structure on the performance of those methodologies. The results show that proposed methodologies perform well for the considered instances. These are not proved facts but they are expected to hold in most cases.

CASE 1: 3-Hospital Loss Network

In this first computational study, we choose to work with a small size network (a network of three multi-server loss queue) since the dimensionality of the problem allow us to compute the global optimal policy, and consequently allows us to analyze the performance of the proposed local control policy. Our objective is to compare the resulting optimal control points and evaluate the optimality gap and its significance between the global optimal and local policy. The parameters of the test instance are presented in Table 5.12.

Table 5.12: Arrival rates, Bed Capacity and Preference List of Hospitals

Hospital i	λ_i	N_i	**Preference list of** i
1	14	15	$\{1, 2, 3\}$
2	18	20	$\{2, 3, 1\}$
3	27	30	$\{3, 1, 2\}$

In order to compute the optimal control policies for both MDP models, two different value iteration algorithms are coded in C++ and solved respectively for global and local policies. Rewards considered in the study are r1=10, r2=5, r3=2 respectively for the admission of *class-1*, *class-2* and *class-3* patients. The discount

factor is taken very close to 1, in order to better approach reality (α=0.999).

The optimal admission switching points obtained from global and local policy are presented in the figures below. Figures 5.11, 5.12, 5.13 present the joint optimal switching points for patients rejected from hospital 1, 2, 3, respectively, under global (left) and local (right) admission control policy.

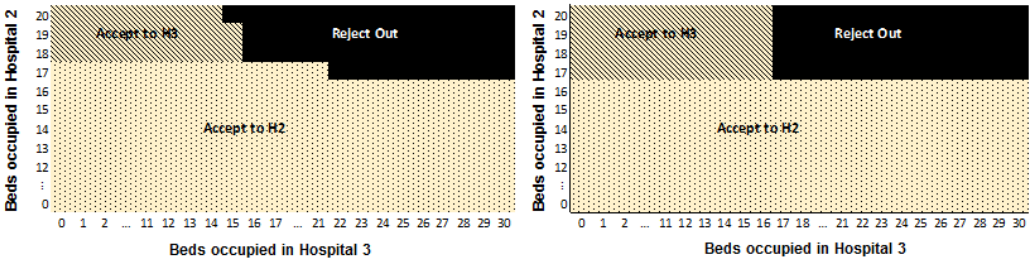

Figure 5.11: Optimal switching points for overflowed patients of Hospital 1 under Global and Local Admission Control Policy

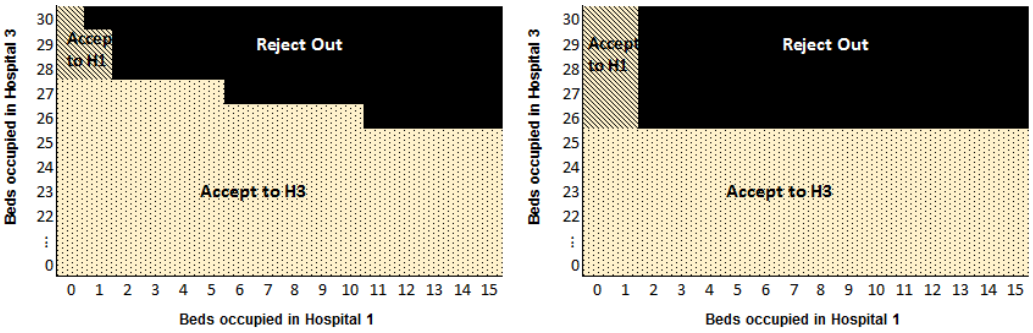

Figure 5.12: Optimal switching points for overflowed patients of Hospital 2 under Global and Local Admission Control Policy

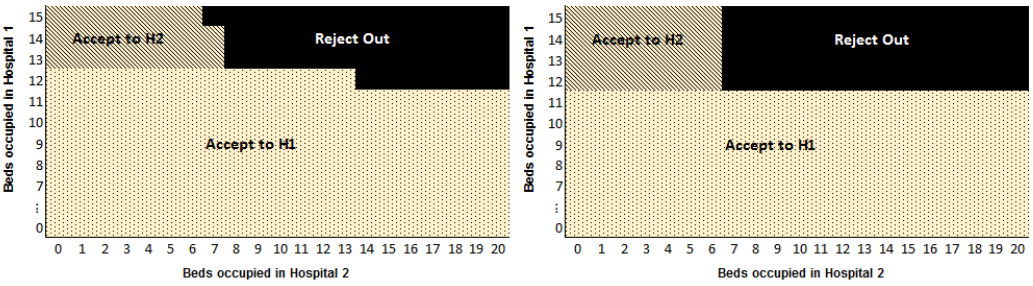

Figure 5.13: Optimal switching points for overflowed patients of Hospital 2 under Global and Local Admission Control Policy

Remark: As a common and expected result of both policies, *class-1* arrivals to a hospital i are always admitted in any state of the system (x_1, x_2, x_3) unless hospital i is full. Therefore, the figures above presents the optimal control points for overflowed patients of hospital $i = \{1, 2, 3\}$ when hospital i was full.

The resulting admission probabilities for each arrival stream to each hospital are calculated under the optimal switching points proposed by both policies. We compared those results also with the results of no-control-policy (Under **no-control policy**, any arrival stream to any hospital will be accepted unless the corresponding hospital is full.) Total rewards obtained under **no-control policy, local control policy and global optimal policy** are presented at Table 5.13. As expected, the admission probabilities for class-1 patients increase while the ones for class-2 and class-3 decrease under both global and local policy compared to the no-control policy. Even though the global optimal policy gives the highest total reward, the difference between global and local policy is not significant. The optimality gap is quite small (0.04%). This result indicates the possibility of using a local control policy in a big scale network (more than 3 hospitals) where it is not possible to use global control policy due to curse of dimensionality.

For this specific instance, the improvement(% relative difference) obtained with a local control policy is 4.3% compared to no-control policy. It is important to mention that the results and rewards are highly dependent on the chosen data set and the gain is expected to be higher as the number of hospitals, number of beds increase.

Table 5.13: Performance Comparison of Different Control Policies

	Admission probabilities	**No-control Policy**	**Local Policy**	**Global Policy**
Class-1	p11in	0.730	0.836	0.839
Admission	p22in	0.775	0.873	0.873
Probabilities	p33in	0.838	0.907	0.907
Class-2	p12in	0.169	0.058	0.062
Admission	p23in	0.155	0.048	0.049
Probabilities	p31in	0.081	0.020	0.022
Class-2	p13in	0.055	0.001	0.000
Admission	p21in	0.024	0.000	0.000
Probabilities	p32in	0.036	0.000	0.000
TOTAL REWARD		**508.9**	**530.8**	**531.0**
% relative difference			4.30%	4.34%

The proposed local admission control policy gives near-optimal results for 3 hospital case. Biases (possible to occur due to the assumptions and approximations we have made) seem to be very small for 3 hospital case whereas for a bigger network, they can easily get bigger and deteriorate the results. Therefore, local policy should

be tested on a bigger network before we can conclude about the performance and robustness of the policy.

CASE 2: 6-Hospital Loss Network

In case 2, we address a 6-hospital loss network where we consider a hierarchical overflow rule between hospitals. The bed capacities, major arrivals to each hospital, and preference (overflow) list of hospitals is presented in Table 5.14.

Table 5.14: Arrival rates, Bed Capacity and Preference List of Hospitals

Hospital i	λ_i	N_i	**Preference list of Hospital** i
1	21	20	$\{1,2,3,4,5,6\}$
2	24	25	$\{2,3,4,5,6,1\}$
3	22	20	$\{3,4,5,6,1,2\}$
4	14	15	$\{4,5,6,1,2,3\}$
5	20	25	$\{5,6,1,2,3,4\}$
6	28	25	$\{6,1,2,3,4,5\}$

Rewards considered in the study are r1=10, r2=5, r3=3, r4=2, r5=1, r6=0.5 for the admission of *class-1*, *class-2*, *class-3*, *class-4*, *class-5* and *class-6* patients, respectively. The discount factor is taken very close to 1, in order to better approach reality ($\alpha = 0.999$). Determining a global optimal policy for such a big-scale hospital network is a complex and computationally demanding problem. Thus, the local control policy is employed and the resulting local control points for each arrival stream in each hospital are presented in Table 5.15. For assessing the performance of local policy, an upper-bound is calculated by using UB proposed in section 5.6.4.

$C_i^w = (C_i^1, C_i^2, C_i^3, C_i^4, C_i^5, C_i^6)$ are the control points in hospital i defined for *class-w* patients where $w = \{1, 2, ...6\}$. A *class-w* patient is rejected from hospital i if $x_i \geqslant C_i^w$. Class-1 patients are always accepted unless the hospital is full such that $C_i^1 = N_i$. It can be also observed in Table 5.15 that the control points get tighter (close to 0) as the number of overflows (class of arrivals) increase.

Table 5.15: Optimal Local Control Points

	C^1	C^2	C^3	C^4	C^5	C^6
H1	20	16	11	5	0	0
H2	25	22	17	11	0	0
H3	20	16	11	5	0	0
H4	15	13	8	3	0	0
H5	25	23	20	16	7	0
H6	25	21	16	11	0	0

In order to search for the possible improvements, we employed an experimental

Table 5.16: Performance Comparison of Different Control Policies

	Admission Probabilities	**No-control Policy**	**Local Control Policy**	**Upper Bound**
Class-1 Admission Probabilities	p11in	0.57	0.81	0.79
	p22in	0.63	0.87	0.83
	p33in	0.55	0.79	0.76
	p44in	0.52	0.82	0.77
	p55in	0.68	0.93	0.92
	p66in	0.62	0.79	0.78
Class-2 Admission Probabilities	p12in	0.21	0.03	0.14
	p23in	0.15	0.01	0.06
	p34in	0.17	0.04	0.14
	p45in	0.27	0.09	0.23
	p56in	0.15	0	0.06
	p61in	0.16	0.02	0.06
Class-3 Admission Probabilities	p13in	0.07	0	0
	p24in	0.07	0	0
	p35in	0.14	0.02	0.02
	p46in	0.09	0	0
	p51in	0.06	0	0
	p62in	0.09	0	0
Class-4 Admission Probabilities	p14in	0.04	0	0
	p25in	0.07	0	0
	p36in	0.05	0	0
	p41in	0.04	0	0
	p52in	0.04	0	0
	p63in	0.04	0	0
TOTAL REWARD		937.95	1093.01	1110.29
%relative difference			16.5%	18.3%

setting around the local control points. The control points for a hospital (C_i^2, C_i^3, C_i^4, C_i^5) are considered jointly and as one factor (C_i) counting on the fact that they would most probably increase/decrease together (inherent correlation among them). Each factor is tested with three levels ($C_i - 1, C_i, C_i + 1$). We use factorial design setting to generate test scenarios and conduct simulation optimization. The maximum reward is attained once again at local control points proposed by local control policy. However, it is important to note that the relative difference between rewards of different scenarios is quite small.

The admission probabilities and total rewards are computed by using a simulation model for **no-control policy** and **local control policy**. The 6-hospital overflow network is modeled in Arena and the simulation model is run with 30 replications each for 10000 days which was sufficient to reach 95% CI. The results obtained under no-control policy, local control policy as well as an upper-bound are presented in Table 5.16. The local control policy lead to 16.5% improvement in total

reward compared to no-control policy. That much increase in reward is significant for performance of any kind of network, given that there is no cost involved in taking those control actions.

Furthermore, the relative gap between local policy and the upper-bound is calculated as 1.6%, which is quite small. Therefore, it is possible to conclude that the proposed UB calculates a very tight upper-bound for the optimal policy and performance of the local control policy is very close to the optimal control policy.

CASE 3: 6-Hospital Loss Network with Centralized Overflows

We consider a 6-hospital loss network where we consider a centralized overflow rule between hospitals. Each patient overflows to the bigger and centralized hospitals as similar to example considered in section 5.5. The bed capacities, major arrivals to each hospital, and preference (overflow) list of hospitals are presented in Table 5.17.

Table 5.17: Arrival rates, Bed Capacity and Preference List of Hospitals

Hospital i	λ_i	N_i	**Preference list of Hospital** i
1	6	5	$\{1,2,3,4\}$
2	9	10	$\{2,3,4\}$
3	19	20	$\{3,4\}$
4	18	20	$\{4,3\}$
5	8	10	$\{5,4,3\}$
6	5	5	$\{6,5,4,3\}$

Rewards considered in the study are r1=10, r2=5, r3=2, r4=1 for the admission of *class-1*, *class-2*, *class-3* and *class-4* patients, respectively. The discount factor is taken very close to 1, in order to better approach reality ($\alpha = 0.999$).

The resulting local optimal control points for each arrival stream in each hospital are presented in Table 5.18. Hospital 1 and 6 receive only their *class-1* patients and do not receive any overflow patients, thus no admission policy is employed. Hospital 2 and 5 receive *class-1* and *class-2* patients (from hospital 1 and hospital 6 respectively), thus they have one control point C_i^2 to control *class-2* arrivals. Hospital 3 receives *class-1* patients (its own patients), *class-2* patients (overflowed from hospital 2 and 4), *class-3* patients (overflowed from hospital 1 and 5), and *class-4* patients (from hospital 6). Similarly, hospital 4 receives *class-1* patients (its own patients), *class-2* patients (overflowed from hospital 3 and 5), *class-3* patients (overflowed from hospital 2 and 6), and *class-4* patients (from hospital 1). Thus, they have C_i^2, C_i^3, C_i^4 three control points for controlling *class-2*, *class-3* and *class-4* arrivals, respectively.

The admission probabilities and total rewards under no-control policy and local optimal control policy are computed via simulation. The results are presented in

Table 5.18: Optimal Local control points in each hospital

	C^1	C^2	C^3	C^4
H1	5	-	-	-
H2	10	7	-	-
H3	20	16	0	0
H4	20	13	0	0
H5	10	8	0	0
H6	5	-	-	-

Table 5.19: Performance Comparison of Different Control Policies

	Admission Probabilities	**No-control Policy**	**Local Control Policy**	**Upper Bound**
Class-1 Admission Probabilities	p11in	0.64	0.64	0.64
	p22in	0.74	0.80	0.80
	p33in	0.69	0.86	0.83
	p44in	0.71	0.87	0.83
	p55in	0.79	0.84	0.80
	p66in	0.64	0.64	0.64
Class-2 Admission Probabilities	p12in	0.24	0.11	0.19
	p23in	0.16	0.01	0.20
	p34in	0.16	0.03	0.09
	p43in	0.15	0.00	0.01
	p54in	0.14	0.04	0.20
	p65in	0.26	0.13	0.34
Class-3 Admission Probabilities	p13in	0.07	0	0
	p24in	0.05	0	0
	p53in	0.04	0	0
	p64in	0.06	0	0
Class-4 Admission Pr.	p14in	0.02	0	0
	p63in	0.02	0	0
TOTAL REWARD		524.38	550.633	563
%relative difference			5.01%	7.36%

Table 5.19. It is observed that the local optimal control policy lead to 5% improvement in total reward compared to no-control policy. One of the key observation is that the relative gain obtained from an optimal admission control policy compare to no-control policy increases as the network gets bigger or the interaction among hospitals increases. The performance of the local optimal control policy is evaluated against an upper bound value computed via UB. The relative gap is computed as 2.2 % which can be considered as a fair gap to claim that local policy performs well in a network with centralized overflows.

5.7 Conclusion & Future Work

In this chapter we studied the admission control problem in a multi-server loss network of hospitals where rejected patients may overflow to other hospitals. We adapted a research strategy evolving from simple networks to big scale networks. For the systems of "one-hospital with m class of arrivals" and "two-hospital overflow network", we showed that the optimal control policy is of threshold form. Afterwards, numerous computational studies are conducted to study sensitivity of system parameters and the optimal policy behavior rigorously. The effects of number of servers, identical and distinguishable stations, different arrival intensity and unit reward variations on the resulting optimal policy are comprehensively analyzed in an overflow network of two hospitals.

For the big scale overflow loss systems, we developed a near-optimal heuristic policy built upon the idea of treating each hospital separately, yet considering the possible overflows from other hospitals. In order to assess the performance of the proposed heuristic, several LP models are proposed to calculate a good upper bound on total reward. The last model (UB) evolved through the preceding LPs and prove itself to be a good performing upper bound model. Several case studies are generated to measure the effect of different size networks and different overflow rules on the performance of local policy and UB. As a result of those numerical studies we were able to highlight a series of key observations. As the system gets bigger (number of hospitals and the number of servers increase), the improvement obtained compared to no-control policy increases, besides % relative gap between local optimal and upper bound increases too, yet the gap still can be considered small (2.2%). The source of the gap is hard to detect since both are approximation methodologies. However, it is important to note that in a big size network even though the optimality gap increases as the network gets bigger, the gain obtained from employing an admission control policy gets much higher compared to small size networks. This result is indeed intuitive as in bigger size systems the model has more room to improve.

On the other side, it is shown that as the intensity of load gets extremities (very light or very heavy) the optimal policy becomes closer to either no control or full control policy. The highest performance is achieved from an optimal admission policy is when the system is not under or over utilized so that there is an opportunity to take a control decision. Even though those are not proved facts, they are expected to hold in most of the cases. In order to better strengthen the final conclusions about the good performance of proposed models, enlarging the scope of the numerical analysis where more diverse and bigger systems are analyzed should be considered as a necessary prospective study.

5.8 Résumé du chapitre

Dans ce chapitre, nous abordons un problème de pilotage d'admissions dans un réseau de soins sans attente multi-serveurs où chaque hôpital est alimenté par un flux de patients (loi d'arrivée de Poisson) et un flux de patients provenants d'autres hôpitaux (loi d'arrivée différente). Un décideur (gestionnaire de lits) qui a une connaissance complète du nombre de lits occupés dans chaque hôpital du réseau décide d'accepter ou de rejeter chaque flux d'arrivée selon un certain critère. Une "récompense" est obtenue lorsque le patient est admis dans un hôpital en fonction de sa provenance. Dans cette étude, les principales arrivées suivant une loi de Poisson sont prioritaires sur celles provenant d'autres hôpitaux. Notre objectif est de trouver la politique optimale qui maximise les récompenses totales remisés. Dans ce travail, afin de simplifier, nous menons notre étude sur un type de patients (femmes enceintes) et ne considérons que les ressources en lits comme serveurs dans chaque unité. Comme nous le supposons également dans le reste de la thèse, chaque patient est servi par un serveur avec une durée de service exponentielle, identique pour chaque catégorie de patient et pour chaque hôpital.

Notre stratégie de recherche est présentée de manière progressive, depuis les réseaux simples jusqu'aux plus complexes. Sur les réseaux de petite taille comme un hôpital avec m classes d'arrivées et deux hôpitaux dans un réseau sans attente avec rejet, nous avons montré que la politique de pilotage optimale est de la forme d'un seuil. Par la suite, de nombreuses études sont effectuées pour étudier la sensibilité des paramètres du système et le comportement optimal de politiques de rigueur. L'effet du nombre de serveurs, de stations identiques et distinctes, l'intensité d'arrivées différentes et les variations de fidélité unitaires sur la politique optimale obtenue est analysée en détail dans un réseau de deux hôpitaux.

Pour les systèmes plus grands, nous avons développé une politique heuristique quasi-optimale construite sur l'idée de traiter séparément chaque hôpital en tenant compte des éventuels transferts depuis d'autres hôpitaux. Afin d'évaluer la performance de l'heuristique proposée, plusieurs modèles de programmation linéaire sont proposés pour calculer une bonne borne supérieure sur la récompense totale. Le dernier modèle (UB) a évolué à travers les modèles précédents et se révéle être un bon modèle de la limite supérieure. Plusieurs études de cas sont générés pour mesurer l'effet des différents réseaux. À la suite de ces études numériques, nous avons pu mettre en évidence une série d'observations clés. Quand le système devient plus grand (nombre d'hôpitaux et nombre de serveurs importants), l'écart relatif entre les augmentations liées optimales et supérieures locales, mais encore peut être considéré comme faible (2.2%). La source de l'écart est difficile à suivre puisque les deux sont des modèles d'approximation. Cependant, il est important de noter que dans un réseau de grande taille, même si l'écart d'optimalité augmente à mesure que le réseau s'agrandit, le gain obtenu d'employer une politique de contrôle d'admission devient beaucoup plus élevé par rapport aux réseaux de petite taille.

Ce résultat n'est en effet intuitif que dans les systèmes de plus grande taille, où le modèle laisse une plus grande place à l'amélioration.

D'autre part, nous avons montré que lorsque l'intensité de la charge devient extrême (très faible ou très lourde), la politique optimale se rapproche de deux pas de pilotage ou de la politique de pilotage total. La meilleure performance obtenue à partir d'une politique d'admission est optimale lorsque le système n'est pas sous ou surutilisé afin qu'il y ait la possibilité de prendre une décision.

Chapter 6

Performance Evaluation of Perinatal Network via Simulation

Contents

6.1 Introduction

A simulation model can represent complex systems in a more realistic way compared to analytical models where usually we have to make several assumptions for the sake of construction of model and the development of interesting solution methodologies.

In this chapter, we develop a joint "agent-based" "discrete-event-system" simulation model of a hierarchical overflow loss healthcare network (perinatal network of Nord Hauts-de-Seine). In agent-based part, the health-state evolution of women in antenatal period (∼9 months phase of pregnancy before childbirth) is modeled by

quantifying the possible passages between health states via Markov Chain modeling. The women with resulting health state information generates the arrival of entities in discrete-event-simulation where the flow of women through obstetrics and consecutively the flow of newborns through neonatal service units are modeled. With this model, we represent the process of mother and babies together which are inherently highly dependent processes nevertheless commonly modeled separately in literature. Furthermore, with the addition of modeling health evolution of pregnant women, the whole process of pregnancy is addressed. To the best of our knowledge, this is the first effort to combine the health state evolution of patients with the operational decisions in network such that health state evolution of pregnant women is monitored to determine different types of demand in each commune and each hospital.

Main objective of this chapter is to evaluate the strength of the optimal results of our analytical models and the validity of underlying assumptions in comparison with the results obtained from the simulation model which represents the system in a more accurate way. In this particular study, we focus on operational level decisions. The analytical model proposed in chapter 3 is simplified into a single period capacity planning model (CP) without considering facility location decisions. By this way, we obtain two models that both focuses on operational decisions and are compatible to be compared:
- Analytical capacity planning model (CP): considers a pure loss system (no overflows), include Markovian and demand distribution assumptions.
- Simulation model: considers overflows between units and maintain a more accurate system representation.
After CP is solved to optimality, we search for possible improvements around optimal solutions of CP and propose more fit-to-reality solutions through simulation-optimization approach.

The remainder of this chapter is organized as follows. Section 6.2 reviews the related literature on simulation studies and defines our contribution. Section 6.3 presents the pregnancy process and patient flows as they are represented in our simulation model (along with necessary assumptions). We use a big scale hospital data in order to introduce a close-to-reality experimental setting for perinatal network Nord Hauts-de-Seine. Section 6.4 presents the input data analysis and resulting experimental setting of the case study. In Section 6.5, the simulation model is implemented and validated. In Section 6.6, CP model is solved under this experimental setting and optimal capacity decisions are obtained. Afterwards, several capacity scenarios are generated around the optimal solution and simulation runs are realized under both optimal CP solution and the proposed scenarios. Finally, conclusion and future research directions are given in Section 6.7.

6.2 Literature Review

In service industry, dicrete-event simulation is considered to be one of the best tools to analyze real world systems ([Kelton 2000]). Particularly in healthcare systems, due to their complex and stochastic nature, simulation had found widespread application where it is used as a tool to analyze critical parts of healthcare systems such as facility design (emergency departments, operating rooms, etc.), staff planning and scheduling and bed capacity management ([Jun 1999], [Augusto 2008]).

Over the last decade, there have been many efforts in developing simulation models to show decision-makers a realistic reproduction of the healthcare system at work. [Syam 2010], [Ferreira 1999], [Swisher 2001], [Blasak 2003], [Sinreich 2005], and [Charfeddine 2010] use simulation models to mimic the behavior of a healthcare system in order to evaluate its performance and analyze the outcome of different scenarios. In literature, there have been also some efforts in simulating obstetrics units as in [Mahachek 1984] and recently [Griffin 2012].

There have also been extensive amount of studies focused on integrated usage of simulation and optimization models applied to the healthcare sector in recent years. De Angelis et al. [De Angelis 2003] present a methodology that interactively uses system simulation and optimization to calculate and validate the optimal configuration of servers in a healthcare facility. In their study simulation is used to estimate the relationship between input parameters and service performance such that a nonlinear function is estimated from simulated data. The estimated function is used as objective function in optimization model to calculate and validate the optimal configuration of servers. Oddoye et al. ([Oddoye 2009]) presents a methodology that combines simulation and goal programming for healthcare planning in a medical assessment unit (MAU). The simulation model enables different scenarios to be tested to eliminate bottlenecks in order to achieve optimal clinical workflow. The results of simulation are analyzed with a goal programming model which employs a multiobjective perspective. Ahmed et al. ([Ahmed 2009]) integrated simulation with optimization to design a decision support tool for the operation of an emergency department unit at a governmental hospital in Kuwait, specifically to determine the optimal staffing levels required to maximize patient throughput and to reduce patient time in the system subject to budget restrictions. Instead of dealing with an approximated mathematical model, they use simulation to evaluate stochastic objective function and the stochastic set of constraints of the optimization model which have no analytical form. With combined usage of optimization and simulation, they evaluate the impact of various staffing levels on service efficiency. Instead of dealing with an approximated mathematical model of the system.

In this study we present a realistic reproduction of the perinatal network with a comprehensive joint agent-based discrete-event simulation model. Key features include patient classification, path-based modeling, blocking and overflow, statis-

tical fitting of LOS, realistic patient distribution (arrivals) to hospitals. With all those details incorporated in simulation, we could better capture complexity of a healthcare system compared to an analytical model. Our aim is to bring together the relative strengths of simulation (capture reality and complexity) and analytical model (quick solution over a range of criteria). Finally, simulation is used to support and improve the decision making process initialized by an analytical model.

6.3 Pregnancy Process and Patient Flows

In this section, we describe the pregnancy process and perinatal network modeled along with our assumptions. Pregnancy process and patient flow in a perinatal network can be described in the following four phases:

Phase 1: Prenatal Period

This phase corresponds to the period before delivery (child-birth) which is known as the prenatal (antenatal) period and considered to proceed three trimesters (9 months) for a healthy pregnancy. In this period each pregnant woman has their routine medical and nursing care in a hospital or at a private clinic with a gynecologist/midwife they choose in their network, most likely close to their domicile. In this period, most of the arrivals to the hospitals/clinics are scheduled arrivals. Each pregnant woman is supposed to attain those routine controls once or twice in a month dependent on the evolution of their pregnancy. Mainly, the resources in the network are sufficient for this phase. Even though, there might exist some specific bottleneck resources (i.e. MRI) needed to be tightly scheduled. This phase is quite important in order to monitor the health evolution of each pregnant woman, already existing health problems such as pre-diabetes as well as the possible disease developments such as diabetes gestational, pregnancy-induced hypertension, etc. Monitoring allows us to attach the inherent variability in the progress of pregnancy to possible health syndromes and eventually better estimate the resource requirements of each woman throughout pregnancy and be able to direct them to the appropriate hospitals in the end of their pregnancy. From global perspective, monitoring women in that period let us to better estimate the capacity requirements in each commune as well as in each hospital in the network.

The assumptions considered throughout the modeling of Phase 1 are given as in the following:

Assumption 1: Pregnant women may have different criticalities. Furthermore, there might be changes in their health states throughout the process of pregnancy. For this simulation study we assume two types of pregnant woman; normal and critical. In the end of each trimester, classification of woman may change from normal

to critical with predefined transition probabilities.(see Figure 6.1)

Assumption 2: A woman who has once passed to a critical state is assumed to stay critical until the end of pregnancy, therefore no transition is foreseen from critical to normal state.

Assumption 3: Each trimester (except 3rd trimester of critical women) is assumed to proceed 90 days (∼14 weeks). Due to the criticality of patient, early delivery (premature birth) is quite common in critical pregnancy, therefore 3rd (last) trimester of critical women is assumed to proceed 75 days.

Assumption 4: In the beginning of Phase 1, each woman is defined with four layer information of *"domicile, health state, 1st preferred hospital, 2nd preferred hospital"*. The 1st and 2nd preferred hospitals of a woman are determined according to the joint information of domicile and health state of woman. The relations are extracted by analyzing the historical data, and precised later in detail.

Assumption 5: No critical woman can leave Phase 1 as being assigned to a type 1 hospital where no neonatal services exist. Throughout phase 1, if a normal woman has passed to a critical state, in the end of phase 1 she will be reassigned to the next preferred appropriate hospital so that her place of birth is changed accordingly to her needs.

Phase 2: Patient Flow in Obstetrics

Phase 2 corresponds to the process of child-birth and restoration of women after-birth in obstetrics unit of a maternity facility.

Assumption 6: Two types of deliveries are considered; caesarean (C-Section) and vaginal birth (VB) whose service times differ significantly.

Assumption 7: As explained in detail in the section 2.2 in chapter 2 there are several units through which woman flows in OB such as Triage, Antepartum, Delivery, PACU, Postpartum. For keeping the model simple due to the big scale of work, we combined those units and behave as one process block for each type of delivery: C-section and VB block.

Assumption 8: After the end of Phase 1, each woman arrives to her 1st preferred hospital for delivery. If a woman is rejected from her 1st preferred hospital, she overflows to her 2nd preferred hospital pre-defined in the beginning of Phase 1. If the woman is also rejected from 2nd preferred hospital, she is rejected out of network (lost) such that we assume one overflow for each patient in this study. Since pregnant women are urgent, it is a reasonable assumption not letting women overflow more than once.

Assumption 9: Multiple babies such as twins, triples constitute a small percentage of total births (∼1%). For the purpose of simplification, in this study each woman is assumed to give birth to one baby.

Assumption 10: A normal woman is assumed to give birth to either a healthy baby or a low/medium risk baby requiring basic neonatal care. No normal woman is supposed to give birth to a high-risk baby requiring an NICU.

Assumption 11: A critical woman may give birth to a healthy baby, a low/medium-risk baby requiring basic neonatal care or a high-risk baby requiring intensive neonatal care as well as she may have a still-birth.

Phase 3: Patient Flow in Basic Neonatal

Newborn arriving to neonatal unit stays here until s/he completely recovers. After recovery, newborn either is transferred to postpartum (if the mother still recovers in the hospital) or leave the unit.

Assumption 12: After the delivery, if the required neonatal services cannot be provided by the current hospital due to various reasons (required service is full or required service unit does not exist in the current hospital) baby is transferred (over-

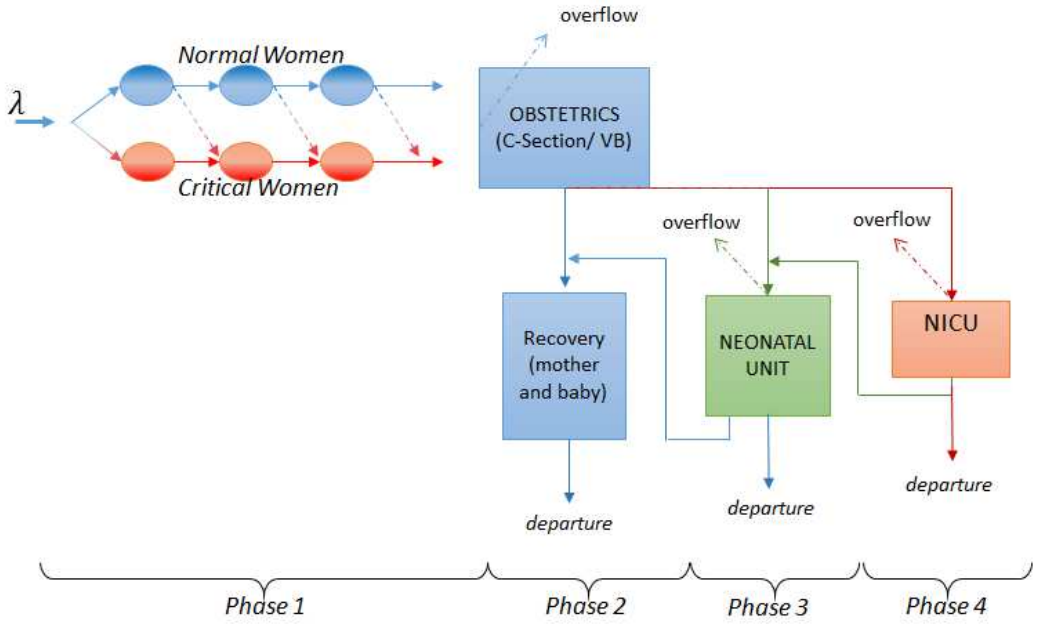

Figure 6.1: Phases of simulation model

flows to) to the closest available hospital where neonatal services exist.

Assumption 13: Babies can be transferred to another hospital after birth independently of the mother.

Phase 4: Patient Flow in Neonatal Intensive Care Unit

Newborns arriving to NICU stay here until their criticality is over. Commonly after babies quit NICU, they are deferred to the neonatal unit for a complete recovery. Once a newborn arrives to a fully occupied NICU, it is transferred (overflows) to the closest NICU in the region. Considering that in our application area, there exist only one NICU, being rejected from this NICU also means being rejected from the network (no overflow in network at NICU level).

6.4 Input Data Analysis

In each hospital in the network, the arrivals of pregnant women and their personal flow in the hospital is recorded with an adequate detail level. ARS provided us the big-scale hospital data of each maternity facility in Hauts-de-Seine perinatal network of the year 2012. In the raw data, it is possible to reach specific information about each woman such as:

- domicile (address)
- hospital treated
- arrival and departure dates
- woman age
- parental history of pregnancy
- checklist of numerous diseases might be critical for the pregnancy, i.e. pre-diabetes, diabetes gestational, HTA, rupture utinerine, hemotome, preeclampsy, eclampsy serious, HELLP, placenta complications, cordon complications, etc.
- type of delivery (C-section, VB)
- newborn LOS in neonatal unit/NICU
- newborn gestational age

6.4.1 Health State Evolution of Pregnant Women

In this study, pregnant women can attain two health states: normal or critical. In Phase 1, we intend to monitor the health condition of each woman in each trimester until delivery in order to be able to better forecast the allocation of pregnant women to the hospitals in network where the women and the newborns can get the right service. Therefore, it is required to set the health state of each woman in the beginning of pregnancy, define transition rates between health states (normal to critical) in the end of each trimester during pregnancy.

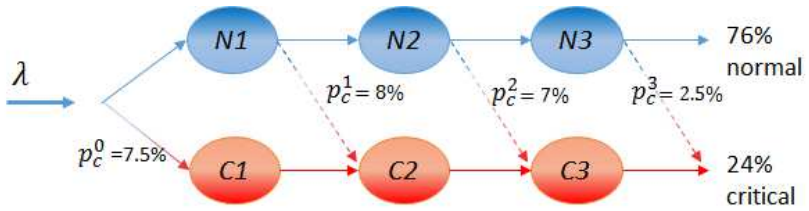

Figure 6.2: Health Evolution of Pregnant Women in Antenatal Period

Figure 6.2 presents the three trimesters for normal and critical pregnant women until child-birth along with the possible transitions between health states.

Women enter the process with a Poisson rate of λ. From the first day of pregnancy, it is possible to classify a pregnant woman as critical or normal. Mostly women with chronic diseases such as chronic hypertension (HTA), pre-diabetes or women with family history (ATCD), assisted conception, obesity, low pre-pregnancy body mass index, extreme maternal age (<15, >38) are considered in the risk group and needed to be closely monitored from the beginning of pregnancy, therefore critical. From hospital data, we extract the proportion of women suffer from such conditions and compute the transition probability leading to C1 as $p_c^0 = 7.5\%$. Complementary probability $p_n^0 = 92.5\%$ gives the proportion of women who have low-risk pregnancy start.

The health state of a woman may change during the pregnancy due to many changes which pregnancy causes in a woman's body. In the end of each trimester, a transition from normal to a critical health state is possible. Multiple pregnancies, placental abnormalities, feeling stressed (including number of stressful or traumatic events), high blood pressure can be considered as risk factors which may develop during the early stages of gestational period (1^{st} trimester) and may give rise to deterioration of health state or a miscarriage. By classifying normal women who shows those symptoms in hospital data, transition probability from N1 to C2 is roughly estimated as $p_c^1 = 8\%$.

During the 2^{nd} trimester, pregnancy itself can trigger some severe complications like "pregnancy-induced hypertension" which is commonly diagnosed after 20 weeks of gestation. There exist several hypertensive states of pregnancy which are known as preeclampsia, eclampsia, HELLP syndrome. By classifying women who had gestational hypertension in hospital data, transition probability from N2 to C3 is roughly estimated as $p_c^2 = 7\%$.

In the last trimester, the most commonly observed risk factor is the development of gestational diabetes. Women who developed this type of diabetes in the hospital data give rise to the transition probability of 2.5%.

In the end of phase 1, the percentage of a woman being critical or normal is calculated as 76% normal, 24% critical.

Remark 1: In this chapter, our ultimate goal is to better set the capacity requirements and better allocate the capacity in maternity facilities throughout the network. Knowing the percentage of critical patients to normal patients (have different capacity requirements) and their spacio distribution let us to model a more realistic system and also allow us to direct the patients to the appropriate facilities and accordingly newborns. On the other side, modeling the health-evolution of

woman "during the pregnancy" and knowing the possible rates of critical and normal woman in all trimesters/months would be extremely useful to estimate the seasonal capacity requirements of antenatal period (before delivery), especially for planification of scarce resources such as MRI. However, as we focus on capacity planning on perinatal phase in this study, this is not the scope of this thesis.

Remark 2: The analysis of health-state evaluation employed in this work is a premature study built-upon the historical data and related literature analysis. This study is mainly presented here for pointing out the promising idea and the possibility of prompting prospective research directions. Such a multidisciplinary study requires a comprehensive analysis involving medical expertise and several professional specializations. For capturing the real evolution, a broad classification of women is necessary that will cover different criticalities. Furthermore, transition probabilities should be determined via guidance of a team of health professionals and through a rigorous analysis of related literature.

6.4.2 Existing Capacity

The current capacity exists in each service unit in each hospital in the network is presented in Table 6.1. The information about the type of maternity facilities can be extracted from the existing capacities. In Nord Hauts-de-Seine Network, there are 2 *type-1* (H1 and H2), 3 *type-2* (H3, H4 and H5) and 1 *type-3* hospital (H6); represented in Figure 6.3.

Table 6.1: Existing number of staffed-beds in network

	H1	H2	H3	H4	H5	H6
Obstetric Unit	25	40	30	35	50	60
Neonatal Unit	0	0	15	15	20	20
NICU Unit	0	0	0	0	0	13

6.4.3 Arrival Rates of Women and Newborns

Arrivals of Pregnant Women:

Women who are cared in perinatal network of Nord Hauts-de-Seine originate from both the communes of the network itself and also the neighboring departments. In Figure 6.3, all population centers (the communes of Nord Hauts-de-Seine and departments in the neighborhood) are represented along with the existing maternity facilities in the network.

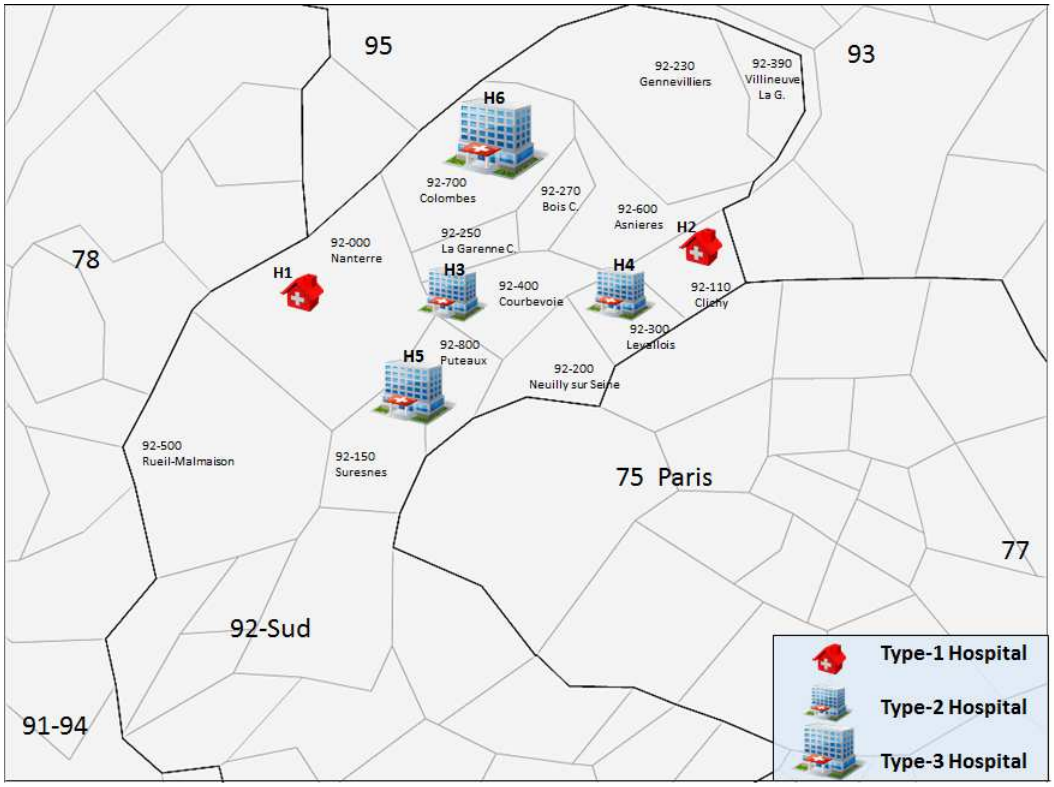

Figure 6.3: Representation of Perinatal Network Nord Hauts-de-Seine

Hospital data provides us the arrival information of each woman to each hospital in the network. By organizing the data around the information of commune and hospital treated, the amount of demand generated from each population center and woman arrival rate to each hospital from each population center are computed. All rates are presented in Table 6.2.

There are 12665 pregnant women treated in the network in year 2012. No seasonality is encountered in the arrivals to hospitals. Therefore, without loss of generality daily arrival rates are computed by using a simple division: (total yearly arrival rates)/365 which is 34.7 for the whole network.

Percentage of arrivals from each population center to each hospital let us to extract the preference order of hospitals for each population center. It is important to note that normal and critical woman from the same population center may have different hospital preferences. While for a normal woman proximity is prioritized, a critical women may prefer a fully equipped hospital since their newborns may need NICU services.

Table 6.2: Distribution of Arrival Rates from Each Population Center to Each Hospital and Overflow Preferences of Women

				Women Arrival Rates Distributed To Each Hospital						2nd Preferred (Overflow) Hospitals											
				H1	H2	H3	H4	H5	H6	Normal Women						Critical Women					
Abb.	Code of Population center	Name of Population center	Arrival Rates generated from each Population Center	Hospital Nanterre	Hospital Beaujon	Hospital Neuilly	Hospital Franco Brit.	Hospital Foch	Hospital Louis M.	From H0 To	From H1 To	From H2 To	From H3 To	From H4 To	From H5 To	From H0 To	From H1 To	From H2 To	From H3 To	From H4 To	From H5 To
0	75-77	PARIS and east	3.09	0.03	0.46	0.50	1.40	0.54	0.16	3	3	3	4	3	3	3	3	3	4	3	3
1	78	west	1.41	0.55	0.13	0.16	0.06	0.30	0.22	4	4	4	4	0	4	4	4	4	4	5	4
2	93	nord east	1.37	0.05	0.92	0.09	0.13	0.06	0.12	1	3	1	1	1	1	3	3	3	5	3	3
3	95	nord west	1.10	0.26	0.13	0.14	0.09	0.11	0.38	5	5	5	5	5	0	5	5	5	5	5	4
4	92390	VILLENEUVE LA GARENNE	1.08	0.04	0.24	0.02	0.02	0.05	0.70	5	5	5	5	5	1	5	5	5	5	5	4
5	92230	GENNEVILLIERS	1.91	0.04	0.41	0.05	0.06	0.05	1.31	5	5	5	5	5	1	5	5	5	5	5	3
6	92600	ASNIERES SUR SEINE	3.46	0.09	0.59	0.53	0.68	0.37	1.20	5	5	5	5	5	3	5	5	5	5	5	3
7	92110	CLICHY	2.12	0.02	1.34	0.12	0.57	0.03	0.06	1	3	1	1	1	1	3	3	3	2	3	3
8	92300	LEVALLOIS PERRET	1.89	0.00	0.32	0.28	1.20	0.08	0.02	3	3	3	2	3	3	3	3	3	2	3	3
9	92700	COLOMBES	3.36	0.29	0.04	0.11	0.08	0.11	2.74	5	5	5	5	5	0	5	5	5	5	5	3
10	92270	BOIS COLOMBES	0.92	0.03	0.01	0.09	0.10	0.09	0.60	5	5	5	5	5	3	5	5	5	5	5	3
11	92250	LA GARENNE COLOMBES	0.65	0.06	0.00	0.08	0.06	0.11	0.33	5	5	5	5	5	4	5	5	5	5	5	4
12	92400	COURBEVOIE	2.82	0.10	0.02	1.89	0.36	0.39	0.06	2	2	4	2	2	2	2	2	4	2	2	2
13	92200	NEUILLY SUR SEINE	0.90	0.01	0.00	0.65	0.18	0.06	0.00	2	2	3	2	2	2	2	2	4	2	2	2
14	92800	PUTEAUX	1.46	0.06	0.00	0.36	0.06	0.96	0.01	4	4	4	4	2	4	4	4	4	4	2	4
15	92000	NANTERRE	2.89	2.00	0.02	0.10	0.04	0.57	0.17	4	0	0	0	0	0	4	4	4	4	5	4
16	92150	SURESNES	1.46	0.07	0.00	0.03	0.02	1.34	0.01	4	4	4	4	3	4	4	4	4	4	3	4
17	92500	RUEIL MALMAISON	1.45	0.18	0.01	0.06	0.03	1.12	0.04	4	4	4	4	0	4	4	4	4	4	2	4
18	91-94-92sud	Sud	1.36	0.04	0.09	0.14	0.14	0.85	0.10	4	4	4	4	3	4	4	4	4	4	2	4
		TOTAL ARRIVAL RATE	34.70	3.91	4.71	5.41	5.29	7.16	8.22												

Arrivals of Newborns:

A newborn baby is generated after the delivery process of woman is terminated. Due to assumption 9 (one woman give birth to one baby), after child-birth the entity for woman continues as the entity for baby. Therefore, arrival rates of newborn babies are dependent on the arrival rates and service times of woman. Furthermore, the health state of a newborn can be anticipated through health-state of the mother. By using the woman-baby match in the hospital data, the health state breakdown percentages are calculated for newborns and presented in Figure 6.4.

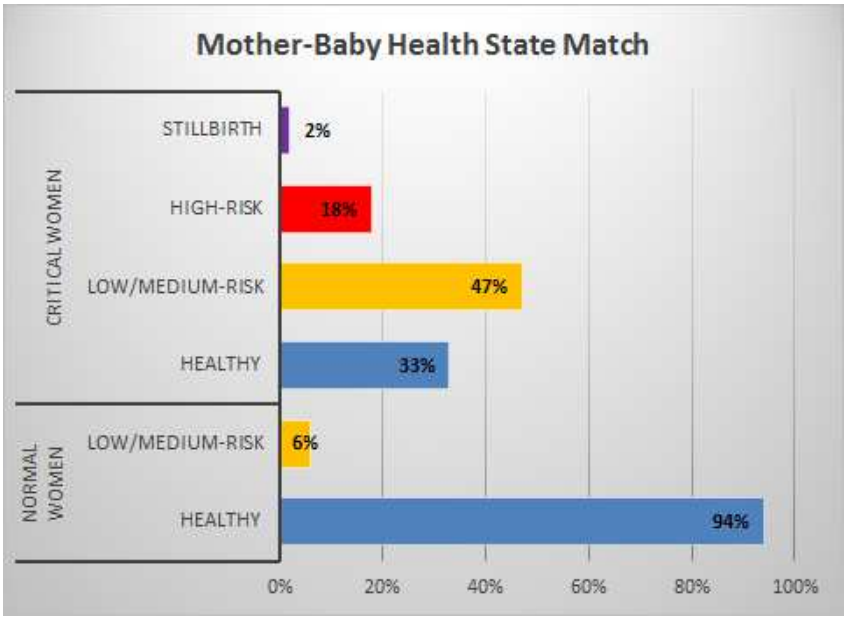

Figure 6.4: Health State Breakdown of Newborns

Considering that there exist 76% normal and 24% critical woman in the system, the health state ratio of babies in the whole network is presented in Figure 6.5.

A newborn requiring neonatal care yet not admitted to its current hospital, overflows to the closest type 2 or type 3 hospital in the network. A newborn requiring intensive neonatal care yet not accepted to only one NICU in the network, overflows to the closest type 3 hospital located out of network. There exists only one type 3 hospital in the network which is H5.

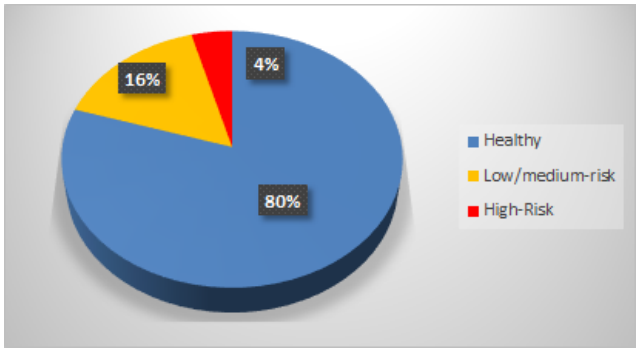

Figure 6.5: Health-state ratio of newborns in the network

6.4.4 Length of Stay (LOS)

In maternity facilities childbirth is achieved through either a caesarean section or a vaginal birth. LOS of woman received C-section is expected to be longer compared to the one received vaginal birth ($1/\mu_c > 1/\mu_v$). Among all pregnant women, C-section is employed with ∼21%. As computed from hospital data, 18% of normal women received C-section while the percentage is higher (32%) for critical women. On the other side, it is observed that service time distributions (for both C-section and VB) do not significantly differ with respect to the health-state of pregnant woman.

Patient length of stays are available in each hospital data for both women and babies. The service time distributions used in simulation study are presented in Table 6.3.

Table 6.3: LOS Distributions used in simulation study

Service Activity	**Fit Distribution**	**Mean**	**Min**	**Max**
Childbirth (C-Section)	GAMMA(5,1.13,0.5)	6.15	0.5	48
Childbirth (Vaginal Birth)	GAMMA(9.15,0.496,0)	4.52	0	19
Neonatal Activity	Min{63,LOGN(1.66,0.958,0)}	8.1	0	63
NICU Activity	Min{100,WEIBULL(10.2,0.752)}	11.94	0	100

where GAMMA $(\alpha, \beta, \min)$, WEIBULL (α, β), lognormal $(\mu, \sigma, \min)$

Remark 3: Weibull and lognormal distributions are truncated with the maximum observed service time (outlier) in order not to feed the model with extreme, unrealistic values.

6.5 Agent-based & Discrete Event Simulation Model Implementation

In this section, we present the implementation, calibration and validation of "agent-based & DES" simulation model. Phase 1 of the pregnancy process is modeled as an agent-based simulation where each women entering the system is an agent and each agent leaves Phase 1 by carrying the joint information of *"health state", "population center", "1st preferred hospital", "2nd preferred (Overflow) hospital"*. The flow of agents in Phase 2, 3 and 4 are modeled as a discrete-event-simulation. Phase 2 (flow of pregnant woman) is a prior process for Phase 3 (flow of low/medium-risk newborn) and Phase 4 (flow of high-risk newborn) while Phase 3 and Phase 4 are parallel processes. The simulation model is constructed by using a commercial software called AnyLogic.

6.5.1 Performance Measures and Objective

Performance indicators used in this study are the rejection probabilities (α_{is}) from each unit s (OB, Neonatal, NICU) of each hospital i. Our objective is to find the minimum total cost solution suggesting a capacity combination which guarantees a rejection probability $\alpha \leqslant 0.05$ for each unit in each hospital.

In Step I, a single period capacity planning model (CP) is derived from multi-period dynamic location & capacity planning presented in chapter 3. In this study, we focus on only capacity decisions in a single period. Since we do not address long-term decisions in CP, we didn't assume any demographic changes in demand throughout the period, relatively no facility location decisions are considered. Basically, we adopt a simplified, more operational version of the mathematical model proposed in 3 with an objective to minimize the total cost (sum of purchase, decrement, holding, transfer and assignment cost) as given in the following:

$$\min \sum_{(i,s)\in IS} (PC_s n_{is} + DC_s z_{is} + HC_s c_{is}) + \sum_{i,j\in I|(i,s)\in IS\&(j,s)\in IS} TC_{ijs} y_{ijs} + \sum_{k\in K}\sum_{i\in I} e_{ki} x_{ki}$$

Constraints

$$\sum_{i\in I_k} x_{ki} = 1, \qquad \forall k \in K$$

$$a_{is} = \sum_{k\in K_s \cup K_{s+1} \ldots \cup K_{\overline{S}}} b_{ks} x_{ki}, \qquad \forall (i,s) \in IS$$

$$c_{is} = c_{is0} + n_{is} - z_{is} - \sum_j y_{ijs} + \sum_j y_{jis}, \qquad \forall (i,s) \in IS$$

$$z_{is} + \sum_j y_{ijs} \leqslant c_{is0}, \qquad \forall (i,s) \in IS$$

$$c_{is0} + n_{is} + \sum_j y_{jis} \leqslant UB_{is}, \qquad \forall (i,s) \in IS$$

$$LB_{is} \leqslant c_{is} \leqslant UB_{is}, \qquad \forall (i,s) \in IS$$

$$B(c_{is}, a_{is}) \leqslant \alpha_{is}, \qquad \forall (i,s) \in IS$$

$$n_{is}, z_{is}, y_{ijs}, c_{is} \in \mathbb{N}, x_{ki} \in \{0,1\}$$

CP model is solved to optimality and the optimal capacity decisions for each unit are obtained.

In Step II, we investigate the possibility to improve over the optimal CP solution by using the proposed simulation model which can mimic the real system much better than an analytical model does. We generate several feasible and coherent scenarios around the optimal solution obtained in Step I, and simulation model is executed for those scenarios. Among the solutions which satisfy the service level constraint ($\alpha \leqslant 0.05$), the one computing the minimum total cost is selected as the best solution.

Remark 4: The cost structure used in chapter 3 is not suitable for one period capacity planning model. The transfer cost is modified as in the following Table 6.4, by taking into account the fundamental relations set between the cost parameters in Appendix A.

Table 6.4: Cost Data

Costs	OB Unit (s=1)	Neonatal (s=2)	NICU (s=3)
Purchase Cost (PC)	1	3	7
Holding Cost (HC)	7	10	20
Transfer Cost (TR)	5	8	15

6.5.2 Calibration of data

We had to make some strong assumptions in CP model in terms of demand structure and assignment of demand in order to facilitate the modeling. However, in simulation model, demand assignments could be handled with less strict and closer to reality assumptions. Therefore, there exist crucial differences the way we model the demands in both models which do not let us use the same input data. The distinctive assumptions of CP and SIM model are presented in Table 6.5.

Table 6.5: Comparison of Assumptions of CP and Simulation Model

Capacity Planning Model	**vs.**	**Simulation Model**
Demands of pregnant woman, neonatal baby, NICU baby is defined in the beginning of the process and it is permanent		Demand of pregnant woman is generated as agents in the beginning of the model. Baby demands are a function of women and generated from women after delivery.
Demand in a *demand-zone* is integer(discretized) and is assigned to one hospital. To increase precision, various demand-zones are created in one population center.		Demand of women is generated in each population center and is assigned to different hospitals with some probability of preference.
No overflows, but it is possible to assign a woman to a less preferred hospital if it is compromised with capacity costs		Direct assignment to the most preferred hospital but possible to overflow
Exponential service time distribution		Distribution Fit to Data
No consideration of different health states		Normal/Critical Woman

Distinctive demand structures prevent us to feed the input data of simulation data directly to MP model. In order to obtain comparable results from simulation and MP model, it is necessary to convert the input data of simulation in a way to achieve the necessary precision to acquire comparable results. Therefore, a comprehensive study is realized to feed the two models correspondingly correct data structures such that both model will generate similar percentage of arrivals for each hospital in the network.

6.5.3 Validation of Simulation Model

In order to simulate the actual system accurately, some important simulation parameters are to be determined such as number of replications, replication length and warm up period. In order to determine run length and warm up period, we have observed the bed occupancies in the system. Since our model is a loss system, there is no queue, therefore entities in the system do not grow exponentially in time and simulation model can reach its steady state condition quickly. For each test scenario, SIM model is run with 50 replications in order to ensure a 95% confidence interval for performance indicators, with a run length of 25000 days and with a warm-up period of 500 days.

The simulation model is verified and validated in several ways. By using animation screen, we monitored the flow of agents and entities along with the relevant dynamic statistics and graphs. Furthermore, by using statistical approaches, we verified the accurateness of some important simulation outputs (arrival rates, per-

centages of newborns, LOS distributions) or performance indicators (rejection rate of the each type of patient to out of network). The verification is realized by using the hospital data and the related literature, since it was not possible to observe the system and collect data regarding the size of the network.

Percentages of Newborn Arrivals to Hospitals

In simulation model, as in reality, process of newborns starts after child-birth. Newborn arrivals to each hospital are modeled as a function of women arrival and departure information (arrival rates/ratios to each hospital, LOS distributions, overflowed hospital, etc.) by incorporating the mother-baby match ratios (given in 6.4.3). Due to this inherent dependency, we don't expect simulation model to make 100% accurate assignment of babies to hospitals compared to hospital data, but at least the difference should not be so big. Therefore, in order to claim that simulation model is assigning close percentage of babies to each hospital, the output of simulation model is compared with the percentages calculated from historical data.

Table 6.6 presents the arrival percentages of women and newborns to each hospital in comparison with the hospital data (real) values. For women, the distributions of women to obstetrics units of hospitals are given directly to simulation model. Therefore the resulting simulated percentages are very close to computed values. However this is not the case for newborns as expected. For hospitals H4 and H6 the ratios are very close, while there exists some difference in the ratios of H3 and H5. The fact that we let the women overflow to some other hospitals when the intended hospital is full, changes the birth place of newborn which eventually cause the arrival information to deviate from the expected one.

Table 6.6: Arrival Percentages of Women and Newborns to Each Hospital

	H1	H2	H3	H4	H5	H6
Sim OB Ratios	11.29%	13.56%	15.60%	15.18%	20.63%	23.75%
Real OB Ratios	11.26%	13.57%	15.58%	15.25%	20.65%	23.70%
Sim Neonatal Ratios	-	-	14.16%	18.86%	22.92%	44.06%
Real Neonatal Ratios	-	-	16.49%	18.32%	21.99%	43.19%

Newborn Arrival rates

The total number of newborn baby arrivals should also be validated with respect to women arrivals. The ratio of total neonatal and NICU baby arrivals to the total women arrivals were computed as 19.8% and 4.2% respectively from the network historical data. The ratio intervals for total neonatal and NICU baby arrivals obtained from 50 simulation runs are presented in Table 6.7. It is observed that the expected ratios computed from hospital data stay in the limits of 95% confidence

intervals constructed around the mean ratios obtained from simulation runs.

Table 6.7: Ratio of Newborn Arrivals to Total Women Arrivals

	Hospital Data Results	**Simulation Range of 95% CI**
NICU Arrivals	4.20%	(4.16%, 4.21%)
NICU+Neonatal Arrivals	19.80%	(19.76%, 19.83%)

LOS Distributions

The selected LOS distributions are also verified in terms of if they conform to the expected mean computed from hospital historical data. We conduct 50 replications of simulation model and in the end of runs, by using the realizations of LOS distributions we construct a 95% confidence interval for the mean of each service type in the network. From Table 6.8, it can be observed that the mean obtained from hospital data stays in the limits of confidence intervals set for each service type.

Table 6.8: Verification of Input LOS Distributions

Service Type	**LOS Distribution**	**Mean**	**Std**	**95% CI**
C-Section	GAMMA(5,1.13,0.5)	6.15	2.53	(6.143, 6.155)
VB	GAMMA(9.15, 0.496, 0)	4.52	1.5	(4.513, 4.537)
Neonatal	Min{63, LOGN(1.66, 0.958, 0)}	8.1	8.98	(8.093, 8.16)
NICU	Min{100, WEIBULL(10.2, 0.752)}	11.94	15.6	(11.93, 12.09)

System Rejection Rates

Data in hand do not include the information of rejection/transfer of patients. Only the women served in the hospital find place in hospital data. Therefore, we consult external trustable information sources and literature for determining the expected probability of patients not being served in the network and eventually rejected to out of network, in France. In literature [Asaduzzaman 2010] and [Beck 2010] estimate neonatal rejection rates in developed countries as; ∼5-7% for neonatal units and ∼25-30% for an NICU. The rejection probabilities obtained from the simulation model are ∼0% for pregnant women, ∼5% for neonatal babies, and ∼34% for NICU babies. It can be interpreted that only NICU rejection rate is computed higher than the one presented in literature. Nevertheless, it is important to mention that such an augmented rate is reasonable considering the geopolitical and demographic structure of the department Hauts-de-Seine. This department is known as one of the most crowded department in Ile-de-France, furthermore it has a lot of attraction from neighboring departments. Even though the quality of service is high and the premature birth rates are relatively low 6.3% ([EFC 2010]), the rejection rates can become quite high since there exist only one NICU in the network

which receives an augmented demand for its service. Therefore we believe that the simulation model accurately represents the real situation in the perinatal network Nord Hauts-de-Seine.

6.6 Numerical Results

6.6.1 Optimum Capacity Planning of CP Model

The input data setting for case study is presented in detail in Section 6.4. The capacity planning model (CP) is solved with CPLEX. Optimal capacity decisions obtained from the model and the total cost associated with the optimal solution are presented in Table 6.9. No change in capacity of obstetric units is proposed by the model which can be interpreted as that the number of existing resources of OBs is sufficient for the network. For neonatal units, the model proposes "to transfer a staffed-bed from H3 to H5" and "to invest on new 5 basic neonatal staffed-beds" and "9 NICU staffed-beds at H6" in order to ensure the desired service level in each unit ($\alpha \leqslant 0.05$) in the network.

After the optimum capacity planning is obtained from CP model, simulation model is executed under this optimal CP capacity setting. The system rejection probabilities obtained under both model is presented in Table 6.9. It is important to note that simulation model computes lower system rejection rates (for obstetric and neonatal service units) compared to CP model under the same optimal CP capacity setting. The major reason of this result is the consideration of overflows in simulation model. While, in CP model, a rejected patient is considered as lost, in simulation model s/he is allowed to overflow to another hospital in the network where she may be accepted, which indeed drops rejection rates significantly.

Table 6.9: Comparison of SIM and CP results

				System Rejection Probabilities under Optimal Capacity	
	Capacity Abb.	Existing Capacity	Optimal CP Capacity	CP model	SIM model
Obstetric Units	x11	25	25	2.48%	0.01%
	x21	30	30	2.05%	0.02%
	x31	35	35	0.75%	0.15%
	x41	35	35	2.41%	0.20%
	x51	50	50	1.03%	0.08%
	x61	60	60	0.01%	0.01%
Neonatal Units	x32	15	14	3.38%	0.86%
	x42	15	15	3.65%	1.41%
	x52	20	21	4.68%	1.57%
	x62	20	25	5.00%	3.62%
NICU	x63	13	22	4.78%	4.98%
			TOTAL COST	326	387

On the other side, different than OB or basic neonatal, rejection rate of H6 NICU (4.98%) is higher than the one computed by CP model (4.78%). This is because there exists only one NICU in network, which obviously cannot profit from overflow phenomena. The observed difference mostly stems from the distinctive service distributions used in CP and simulation models; which are defined as exponential and weibull respectively.

6.6.2 Simulation-Optimization

Under the optimal CP capacity, simulation model computes significantly smaller rejection probabilities than targeted maximum level ($\alpha = 0.05$). Since simulation model is a better representation of the real system, this result encourages us to claim that the targeted service level is possible to be guaranteed with fewer numbers of staffed-beds through less costly capacity decisions.

Motivated from the latter results, the optimal CP solution is taken as the base-scenario and several coherent scenarios are generated around the base scenario. The *control factors* are the number of staffed-beds in each basic neonatal and NICU service unit. The values tested are determined from the direction of possible improvement which can be deducted from the Table 6.9. Obstetric units are not considered in the test scenarios since it is already self-sufficient. A full factorial experimental design setting is constructed with the values presented in Table 6.10.

Table 6.10: Experimental Design Setting

Control Factors		**Level 1**	**Level 2**	**Level 3**	**Responses**
Neonatal Units	x32	12	13	14	rej32
	x42	13	14	15	rej42
	x52	19	20	21	rej52
	x62	23	24	25	rej62
NICU	x63	-	22	23	rej63

50 Simulation runs are generated for each of the 162 test scenarios and corresponding rejection probabilities obtained. The scenarios, which satisfy the service level constraint (5%) in each service unit, are selected and presented in Tables 6.11 and 6.12 along with their associated costs.

Remark : Costs associated with location of facilities such as Opening Cost, Closing Cost and Fix Cost are discarded since no location (opening/closing) decision is taken in the optimal CP results. Therefore, Total Costs are computed considering **Purchase Cost, Holding Cost, Transfer Cost** and **Assignment Cost** (corresponds to **Overflow Cost** in simulation model).

Table 6.11: Test Scenarios (computing $\leqslant 0.05$ rejection probabilities) and associated Capacity Costs

	Control Variables					Responses					number of beds			Cost				
Scenarios (rej<0.05)	x32	x42	x52	x62	x63	rej32	rej42	rej52	rej62	rej63	# Inc. Neo	# Inc. NICU	#of Trans.	Purchase Cost	Holding Cost	Transfer Cost	OV Cost	TOTAL COST
1	13	13	21	24	22	1.88%	3.20%	2.49%	4.81%	4.76%	1	9	4	66	190	32	82.1	370.1
2	13	14	21	24	22	1.61%	2.31%	2.21%	4.58%	4.83%	2	9	3	69	200	24	80.6	373.6
3	14	14	21	24	22	1.01%	2.24%	2.17%	4.47%	4.99%	3	9	2	72	210	16	77.8	375.8
4	13	15	21	24	22	1.52%	1.68%	2.18%	4.38%	4.78%	3	9	2	72	210	16	75.5	373.5
5	**13**	**14**	**20**	**24**	**22**	**2.02%**	**2.36%**	**2.89%**	**4.92%**	**4.86%**	**1**	**9**	**3**	**66**	**190**	**24**	**80.1**	**360.1**
6	13	12	20	25	22	1.96%	4.23%	3.08%	4.73%	4.86%	0	9	5	63	180	40	84.3	367.3
7	13	13	20	25	22	2.10%	3.35%	2.88%	4.57%	4.82%	1	9	4	66	190	32	80.3	368.3
8	14	13	20	25	22	1.24%	3.07%	2.49%	4.39%	4.90%	2	9	3	69	200	24	77.8	370.8
9	12	15	20	25	22	2.45%	1.86%	2.80%	4.55%	4.93%	2	9	3	69	200	24	76.2	369.2
10	13	15	20	25	22	1.47%	1.45%	2.21%	4.13%	4.90%	3	9	2	72	210	16	73.0	371.0
11	**14**	**15**	**21**	**25**	**22**	**0.86%**	**1.41%**	**1.57%**	**3.62%**	**4.98%**	**5**	**9**	**1**	**78**	**230**	**8**	**71**	**387.4**
12	14	14	20	24	23	1.29%	2.27%	2.65%	4.91%	3.93%	2	10	2	76	220	16	79.7	391.7
13	14	15	20	24	23	1.29%	1.63%	2.69%	4.94%	3.91%	3	10	1	79	230	8	78.2	395.2
14	14	12	21	24	23	1.43%	4.12%	2.41%	4.89%	3.85%	1	10	4	73	210	32	86.5	401.5
15	14	13	21	24	23	1.26%	3.19%	2.34%	4.84%	3.90%	2	10	3	76	220	24	82.9	402.9
16	12	14	21	24	23	2.45%	2.52%	2.80%	4.91%	4.09%	1	10	4	73	210	32	81.8	396.8
17	13	14	21	24	23	1.77%	2.43%	2.21%	4.49%	4.14%	2	10	3	76	220	24	80.7	400.7
18	14	14	21	24	23	1.07%	1.97%	2.04%	4.36%	3.53%	3	10	2	79	230	16	77.6	402.6
19	12	15	21	24	23	2.17%	1.71%	2.45%	4.82%	3.54%	2	10	3	76	220	24	80.8	400.8
20	13	15	21	24	23	1.53%	1.67%	2.22%	4.75%	3.90%	3	10	2	79	230	16	79.6	404.6
21	14	15	21	24	23	1.05%	1.60%	2.04%	4.61%	4.07%	4	10	1	82	240	8	77.4	407.4
22	13	12	20	25	23	2.02%	4.34%	2.83%	4.59%	3.97%	0	10	5	70	200	40	82.3	392.3

Table 6.12: Test Scenarios (computing $\leqslant 0.05$ rejection probabilities) and associated Capacity Costs (Continues...)

	Control Variables					Responses					number of beds			Cost				
Scenarios (rej<0.05)	x32	x42	x52	x62	x63	rej32	rej42	rej52	rej62	rej63	# Inc. Neo	# Inc. NICU	#of Trans.	Purchase Cost	Holding Cost	Transfer Cost	OV Cost	TOTAL COST
23	14	12	20	25	23	1.46%	3.86%	2.52%	4.48%	3.76%	1	10	4	73	210	32	82.4	397.4
24	12	13	20	25	23	2.49%	3.40%	3.16%	4.83%	3.44%	0	10	5	70	200	40	82.6	392.6
25	13	13	20	25	23	1.98%	3.17%	2.79%	4.51%	3.42%	1	10	4	73	210	32	81.2	396.2
26	14	13	20	25	23	1.40%	3.03%	2.73%	4.61%	4.03%	2	10	3	76	220	24	79.8	399.8
27	12	14	20	25	23	2.25%	2.39%	2.70%	4.55%	3.33%	1	10	4	73	210	32	77.3	392.3
28	13	14	20	25	23	1.68%	2.36%	2.45%	4.46%	3.40%	2	10	3	76	220	24	77.3	397.3
29	14	14	20	25	23	0.94%	2.16%	2.02%	4.14%	3.91%	3	10	2	79	230	16	74.1	399.1
30	12	15	20	25	23	2.41%	1.67%	2.72%	4.55%	4.22%	2	10	3	76	220	24	76.3	396.3
31	13	15	20	25	23	1.46%	1.81%	2.53%	4.30%	4.08%	3	10	2	79	230	16	75.4	400.4
32	14	15	20	25	23	1.13%	1.54%	2.14%	4.09%	4.18%	4	10	1	82	240	8	71.4	401.4
33	12	12	21	25	23	2.62%	4.35%	2.74%	4.68%	3.43%	0	10	6	70	200	48	84.3	402.3
34	13	12	21	25	23	2.20%	4.33%	2.69%	4.56%	3.61%	1	10	5	73	210	40	84.2	407.2
35	14	12	21	25	23	1.34%	4.08%	2.22%	4.16%	3.75%	2	10	4	76	220	32	82.1	410.1
36	12	13	21	25	23	2.49%	3.45%	2.44%	4.32%	3.95%	1	10	5	73	210	40	82.4	405.4
37	13	13	21	25	23	1.59%	2.94%	2.04%	4.08%	3.48%	2	10	4	76	220	32	77.3	405.3
38	14	13	21	25	23	1.29%	2.84%	1.86%	4.20%	3.51%	3	10	3	79	230	24	79.0	412.0
39	12	14	21	25	23	2.16%	2.37%	2.28%	4.14%	3.51%	2	10	4	76	220	32	77.2	405.2
40	13	14	21	25	23	1.45%	2.19%	2.03%	3.96%	3.53%	3	10	3	79	230	24	77.0	410.0
41	14	14	21	25	23	1.02%	2.06%	1.88%	3.80%	3.80%	4	10	2	82	240	16	74.2	412.2
42	12	15	21	25	23	1.99%	1.78%	2.12%	4.00%	3.94%	3	10	3	79	230	24	75.5	408.5
43	13	15	21	25	23	1.58%	1.43%	1.86%	3.85%	3.82%	4	10	2	82	240	16	72.9	410.9
44	14	15	21	25	23	0.86%	1.39%	1.76%	3.75%	3.23%	5	10	1	85	250	8	72.4	415.4

The 11^{th} scenario in Table 6.11 is the simulation output of optimal capacity setting proposed by CP whose total cost is computed as 387.4. Among all scenarios, the minimum total cost is achieved at the 5^{th} scenario which computes 360.1. As we compare 5^{th} and 11^{th} scenario it can be observed clearly that simulation optimum basically proposes to transfer two more neonatal staffed-beds instead of purchasing more neonatal beds as in optimal CP solution. Under this setting the network is still able to perform a rejection rate below 5% ($\alpha \leqslant 0.05$). On the other side, it is observed that the service level desired in NICUs can be obtained with at least 23 NICU staffed-beds.

Another interesting result extracted from simulation study is that it is possible to validate the expected behavior of CP model defined through its assumptions by using simulation output, and reversibly simulation model itself. Since CP model does not consider overflows, it decides on capacity planning simultaneously with allocation of demand (such that CP might decide to increase the capacity of a preferred unit in order not to assign patients to a less preferred unit which is penalized with a higher cost). Then, it is expected from CP optimal to compute the lowest Overflow cost(Assignment cost) among other scenarios. From Table 6.11, the 11^{th} scenario (the simulation output of the optimal CP planning) computes the smallest OV cost (71) which confirms both the validity of CP and simulation model. Optimal CP scenario proposes to purchase the highest number of basic neonatal beds which directly increases the model ability of assigning patients to their most preferred hospitals. Consequently, the lowest assignment cost is achieved which is theoretically expected from a model where no overflows are considered.

In this study, the capacity of obstetric units is not considered as a decision variable since obstetric rejection rate is already very low in each hospital.

6.7 Conclusion & Futurework

In this chapter, we constructed a joint “agent-based” “discrete-event” simulation model for representing both health-state evaluation of pregnant women during antenatal period and the flow of women, consequtively flow of babies in a hierarchical overflow loss healthcare network (perinatal network of Nord Hauts-de-Seine).

From the provided hospital historical data, we conducted a comprehensive input analysis to determine the realistic parameter setting to feed the model. The same data is used (with support of related literature) to validate the constructed simulation model regarding several important outputs.

Analytical models are built upon many assumptions. In this chapter we tested the validity of one of our analytical models (single period capacity planning model), its ability to represent a perinatal network via the proposed simulation model. For this purpose, several coherent scenarios are generated around optimum capacity decisions obtained from CP model and run under the same experimental setting. It is observed that CP model slightly overestimates the required capacity of each neonatal unit compared to simulation model; which likely stems from the assumptions of "no consideration of overflows" and "exponential service times". However, from the analysis of scenarios, it can be claimed that the analytical model behaves as intended, its validity is assured.

In this very chapter, we used the simulation model to reproduce the behavior of the perinatal network with an adequate detail level in order to evaluate its performance by analyzing outcomes of different scenarios. As a future work, analytical modeling and simulation modeling can be used interactively where simulation model can easily capture the complexity of the system and its outcomes recursively can provide the necessary input data for analytical model. In our analytical models, we attempted to explicitly represent the functioning of the system. As a result of this, either model got too complicated (i.e., nonlinear functions for rejection probabilities) or we had to make several simplifying assumptions to build the model. It would be an interesting prospective study to adopt a simpler analytical model which is solved recursively with the rejection probabilities estimated from simulated data until the convergence is achieved. This type of modeling would yield quite realistic results and the solution approach can be adopted to a flexible tool that the managers can use to evaluate the system performance metrics and making several operational decisions.

A short term perspective consists in adding antenatal patient pathway flows, such as visits to the gynecologist, imaging exams, emergency visits in maternity facilities, etc. As a matter of fact, the same model can be used to predict the optimal capacity for all resources used in the antenatal phase, using the benefit of the health state model during pregnancy. Finally, another future work consists in

focusing on a subset of patients of interest (such as pregnant women with a high risk of premature delivery) to better size the network in terms of rare resources. The provided simulation model is flexible enough to achieve such objectives.

6.8 Résumé du chapitre

Dans ce chapitre, nous avons construit un modèle de simulation unifié "multi-agents" et "à événements discrets" pour représenter conjointement évaluation de l'état de santé des femmes enceintes pendant la période prénatale et la modélisation du flux des femmes dans les différentes structures de soins, ainsi que les flux de bébés (application au réseau périnatal Hauts-de-Seine Nord).

D'après les données historiques hospitalières fournies, nous procédons à une analyse des données d'entrée complètes pour déterminer la valeur du paramètre réaliste pour alimenter le modèle. Les mêmes données sont utilisées (avec l'appui de la littérature connexe) pour valider le modèle de simulation construit sur grâce à plusieurs indicateurs importants.

Les modèles analytiques sont construits sur de nombreuses hypothèses. Dans ce chapitre, nous avons testé la validité de l'un de nos modèles analytiques (modèle de planification de la capacité sur une seule période), sa capacité à représenter un réseau périnatal via le modèle de simulation proposé. A cet effet, plusieurs scénarios cohérents sont générés autour des décisions de capacité optimale obtenues à partir du modèle CP et fonctionnent sous le même cadre expérimental. On constate que le modèle surestime légèrement la capacité requise de chaque unité néonatale par rapport aux modèles de simulation, ce qui découle probablement des hypothèses selon lesquelles nous ne tenions pas compte des transferts entre hôpitaux et nous considérions des durées de service exponentielles. Cependant, à partir de l'analyse de scénarios, on peut affirmer que le modèle analytique se comporte comme prévu, sa validité est assurée.

Dans ce chapitre, nous avons utilisé le modèle de simulation pour reproduire le comportement du réseau périnatal avec un niveau de détail suffisant pour évaluer sa performance par l'analyse des résultats des différents scénarios. La modélisation analytique et la simulation peut être utilisé de manière interactive où le modèle de simulation peut facilement capturer la complexité du système et de ses résultats de manière récursive peut fournir les données d'entrée nécessaires pour le modèle analytique. Dans nos modèles d'analyse, nous avons tenté de représenter explicitement le fonctionnement du système. À la suite de cela, les deux modèles devenu trop compliqués (par exemple, les fonctions non linéaires pour la probabilité de rejet) ou nous avons dû faire plusieurs hypothèses simplificatrices pour construire le modèle. Ce serait une étude prospective intéressant d'adopter un modèle analytique simple qui est résolu de façon récursive avec les probabilités de rejet estimées à partir des données simulées jusqu'à ce que la convergence soit atteinte. Ce type de modélisation donnerait des résultats très réalistes et l'approche de la solution peut être adoptée à un outil flexible que les gestionnaires peuvent utiliser pour évaluer les indicateurs de performance du système et faisant plusieurs décisions opérationnelles.

Chapter 7

General Conclusion

7.1 Conclusion

In this thesis, we addressed design, evaluation and flow control of stochastic overflow healthcare networks where patients get rejected if the required service capacity is not available at arrival and they may overflow to another hospital or to out of network. By motivated from perinatal networks, we studied several challenging issues from both strategic and operational perspectives. The proposed methodologies for each considered problem and possible extensions are summarized as in the following.

One of the most challenging issues of perinatal and particularly neonatal networks is the inadequacy of resources and as its natural extension: high rejection rates. In chapter 3, we proposed a dynamic location & capacity planning model where we consider simultaneously location (opening/closing) and service capacity decisions (increase, decrease, transfer) in order to ensure a minimum desired customer acceptance rate for each service unit in each facility. The resulting model was nonlinear due to the utilization of Erlang Loss function to represent stochasticity in the network. We were able to linearize the nonlinear model by various linearization models which allow us to solve big size problems to optimality in a reasonable time. Structural properties of these linearization models are proved.

Due to its strategic context, possibility of overflows between hospitals was kept out of scope of chapter 3. However, in operational level, overflow of patients in a hierarchical network is an important feature and creates an interesting and rich research area, hence constitutes mainly the remaining part of our study.

In Chapter 4, we studied the performance evaluation of overflow loss networks. Several approximation methodologies are developed for estimating the blocking probabilities for each arrival stream of each hospital in different overflow routing structures. We presented a new method called BinomIPP which is based on IPP characterized arrivals and binomial moment transformation. The proposed method is shown to outperform existing approaches in an overflow network with forward routing. The success of the method lies in its ability to overcome Poisson error as well as correlation error; two type of errors usually manifest themselves in overflow loss networks. For evaluating the performance of an overflow network with feedback routing, we presented several efforts, some of which give promising results under special cases yet development of a generic approximation methodology which per-

forms robustly well in such type of networks remains as a promising prospective research idea.

In Chapter 5, we focused on the control of overflows in various size hierarchical networks with different types of routing structure. We studied optimal admission control policies via Markov Decision Processes on overflow loss networks with increasing complexities which have not been studied in the existing literature. For smaller size networks, we provided the structural properties of optimal admission control policies. For big scale networks, we proposed a near-optimal heuristic policy (local control policy) whose performance is evaluated by an LP model, developed to compute a tight upper-bound on the problem. In the end of the section, the properties of LP formulation are proved so that we could conclude that the proposed LP computes an upper-bound for the problem. In the light of several numerical examples, we analyzed the performance of both local control policy and the upper bound formulation and it is observed that both models perform quite well with respect to the small % relative gaps computed.

In Chapter 6, we presented the implementation and validation process of a joint "agent-based" "discrete-event" simulation model of a stochastic hierarchical overflow-loss perinatal network (Nord Hauts-de-Seine). Our main objective was to reproduce the behavior of the perinatal network with an adequate detail level in order to evaluate its performance by analyzing outcomes of different scenarios and more importantly evaluate the strength of the optimal results of our analytical models built upon numerous underlying assumptions. In this particular study, we focused on the analytical model (CP) which was a simplified version of the location and capacity model proposed in chapter 3. After fairly identical inputs are fed into both CP and simulation model, the results are compared. It is revealed that CP model slightly overestimated the planning of each unit compared to simulation model optimum, as expected. We concluded the chapter by stating that underlying assumptions (no overflows, exponential service times) of CP have certainly an impact on the results, yet in small magnitude and in the expected direction.

The development of COVER platform was conducted in parallel with the research project realized in this thesis. All scientific tools developed are included in a web-platform with a mapping tool based on Google Maps and a user-friendly interface which also includes collaboration tools such as private document sharing to improve the collaborative performance of the network. Real data is provided by ARS and the perinatal network from Hauts-de-Seine which is already fed to the system.

In this thesis, our main motivation source and application area manifested as perinatal networks. However it is important to note that all scientific tools developed in this thesis are kept robust enough to be applied in any stochastic hierarchical service networks, i.e., emergency units network. Thus, by using the same information

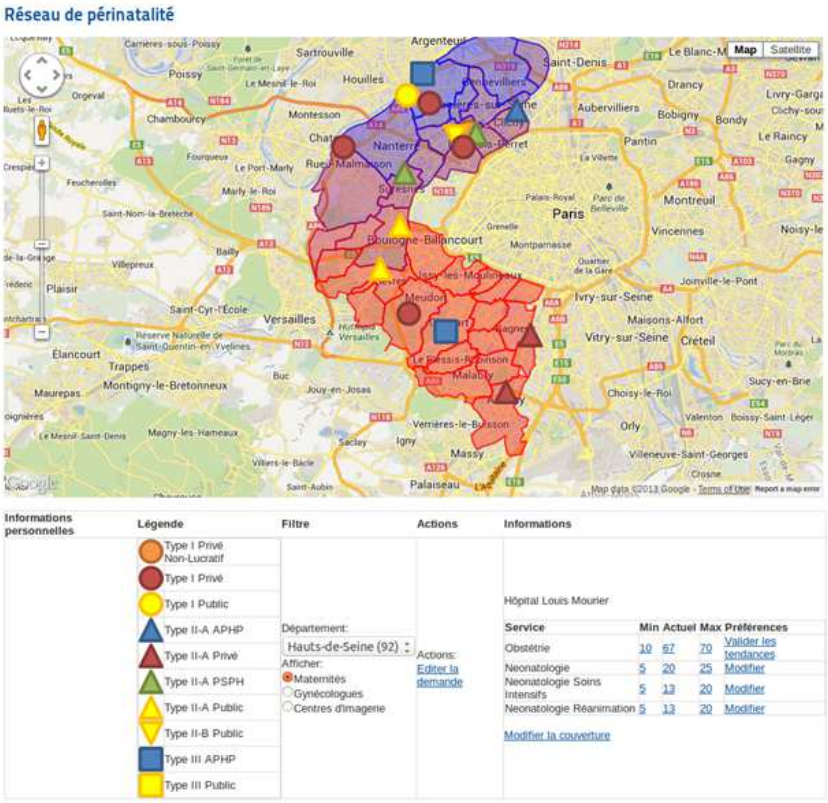

Figure 7.1: Screenshot of the COVER webplatform user interface

system the proposed models and programs can be applied directly to any department of France or any other type of geographical region with different features. We argue that COVER platform can serve as a decision support tool that would aid decision makers, healthcare managers, health professionals to better assess the system performance and assist them in giving important decisions more efficiently in a continuously changing and complex environment.

7.2 Perspectives

Several research directions can be determined in the frame of this thesis:

- In this study, the resources in hospitals are generalized as staffed-beds. As a prospective study, all related scarce hospital resources (gynecologists, nurses, specific equipment) can be taken into account in capacity planning problem in order to determine the necessary number individual resources instead of staffed-beds.

- For "Feedback (cyclic) overflow loss network", the aggregated approach performed well in 3-hospital case. However, the method is based on a two state Markov

Chain computation in the aggregated state and the extension of this method to big size networks is blocked by a scalability problem. Thus, investigating a more sophisticated probabilistic selection rule for developing a better approximation methodology to evaluate "Feedback (cyclic) overflow loss networks" manifests as an interesting research direction.

- In performance evaluation via simulation, we focused on only capacity planning problem as an analytical model. The same strategy can be used to optimize the control policies used in the various hospitals of the network. By constructing a design of experiment, it is possible to test various scenarios to select the best control policy for all services taking into account the results from Chapter 5.

7.3 Résumé de la thèse

Dans cette thèse, nous avons le problème de conception et de pilotage de flux d'un réseau de soins stochastique sans attente, où les patients sont rejetés vers d'autres hôpitaux si la capacité de service requis n'est pas disponible à l'arrivée. Une application à la périnatalité est proposée. Nous avons étudié plusieurs questions difficiles du point de vue à la fois stratégique et opérationnel. Les méthodes proposées pour chaque problème considéré et leurs extensions possibles sont résumés comme suit.

L'une des questions les plus difficiles en lien avec les réseaux périnatals et néonatals en particulier est l'insuffisance des ressources et par conséquent : le taux de rejet élevé. Dans le chapitre 3, nous avons proposé un modèle de planification des capacités dans lequel nous considérons simultanément la localisation (ouverture/fermeture de structures) et les décisions de la capacité de service (augmentation, diminution, transfert) afin d'assurer l'acceptation d'un nombre de patients minimum en tenant compte du coût global du réseau. Le modèle qui en résulte est non linéaire en raison de l'utilisation de fonctions de Erlang pour modéliser le caractère stochastique du réseau. Nous avons réussi à linéariser le modèle non linéaire, nous permettent de résoudre des problèmes de grande taille à l'optimalité dans un délai raisonnable. Les propriétés structurelles de ces modèles de linéarisation sont prouvées.

Grâce à son cadre stratégique, la possibilité de transferts entre les hôpitaux n'a pas été abordée dans le chapitre 3. Cependant, au niveau opérationnel, le transfert des patients dans un réseau hiérarchique est une caractéristique importante, constituant un verrou scientifique intéressant et riche, traité dans la suite de notre étude.

Dans le chapitre 4, nous avons étudié l'évaluation de la performance des réseaux sans attente avec rejet. Plusieurs méthodes d'approximation sont développés pour estimer les probabilités de blocage pour chaque flux d'arrivée de chaque hôpital dans les différentes structures de routage des rejets. Nous avons présenté une nouvelle méthode appelée BinomIPP, basée sur les arrivées caractérisées IPP et la transformation de moment binomial. La méthode proposée s'est révélée plus performante que les approches existantes dans un réseau sans attente avec rejet et avec routage vers l'avant. Le succès de la méthode réside dans sa capacité à éliminer l'erreur liée à la loi de Poisson, ainsi que l'erreur de corrélation, deux types d'erreurs qui se manifestent généralement dans ce type de réseaux. Pour évaluer la performance d'un réseau sans attente avec rejet et avec routage de feedback, nous avons présenté plusieurs initiatives, dont certaines donnent des résultats prometteurs dans des cas particuliers. Le développement d'une méthodologie de générique et robuste constitue une perspective de recherche prometteuse.

Dans le chapitre 5, nous nous sommes concentrés sur le pilotage des rejets dans les différents réseaux avec différents types de structure de routage. Nous avons étudié les politiques de contrôle d'admission optimales en utilisant des processus de déci-

sion markoviens. Pour les réseaux de petite taille, nous fournissons les propriétés structurelles de la politique de pilotage d'admission optimale. Pour les réseaux d'envergure, nous avons proposé une politique heuristique quasi-optimale dont la performance est évaluée par un modèle d'optimisation linéaire, développé pour calculer une borne supérieure de qualité sur le problème.

Dans le chapitre 6, nous avons présenté un modèle de simulation mixte multi-agents à événements discrets d'un réseau périnatal (Hauts-de-Seine Nord), sa mise en oeuvre et sa validation. Notre objectif principal consistait à reproduire le comportement du réseau périnatal avec un niveau de détail suffisant pour évaluer sa performance par l'analyse des résultats des différents scénarios et surtout d'évaluer la solidité des résultats optimaux de nos modèles analytiques construits sur de nombreuses hypothèses sous-jacentes. Le modèle se révèle flexible, et permet de modérer les résultats obtenus précédemment, et permet également de prodiguer des recommandations à destination des gestionnaires de ce type de réseaux de santé.

Appendix A

Design of Experiments (DOE) for Cost Parameter Setting

DOE is constructed for the three most difficult-to-set cost parameters (FC, DC, TC) in order to observe their impact on model decisions. Cost parameters are tested at different values (levels), which are set relative to the predefined relationships with HC (Section 3.6.2). Tested values for each cost parameter are presented in Table A.1.

Table A.1: DOE Setting

	Level 1	**Level 2**	**Level 3**	**Level 4**
	$s = 1, 2, 3$	$s = 1, 2, 3$	$s = 1, 2, 3$	$s = 1, 2, 3$
	10,10,10	50,50,50	100,100,100	
	7,10,20	40,60,100	100,150,200	
	7,10,20	15,20,40	40,60,100	100,150,200

Main effect plots allow us to point out the most significant cost parameter(s) for each capacity and facility location decision considered by presenting the proportion between the change in the value of parameter and resulting change in the corresponding decision. The two cost parameters which have the biggest impact on facility location decisions are FC and DC as presented in Fig.A.1. On capacity planning decisions, Fig.A.2 points out that there is a trade-off between DC and TC while FC has an ignorable effect on downsizing decisions. Furthermore, Fig.A.3 indicates that TC is the most important factor for determining the amount of transportation in the network whereas FC and DC work in an opposing direction.

Interaction plots allow us to determine a possible correlation among cost parameters for a corresponding decision. The plots for facility relocation decisions are presented in Fig. A.4. It can be observed that TC does not have a significant effect on the closure of services. However there is a compromise between FC and DC. If FC is chosen too high, (since it is a periodic cost it incurs for each existing service) it easily dominates the cost structure such that DC or TC do not have any significant impact on the decisions. However if FC is set too small, then it starts to have almost no effect on facility relocation decisions.

On the capacity planning side we have two interaction plots presenting the compromise between DC and TC in terms of two important decisions in the network:

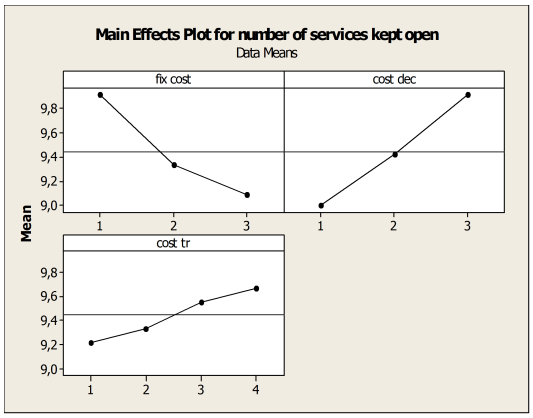

Figure A.1: Main effect plots of cost factors on number of services kept open

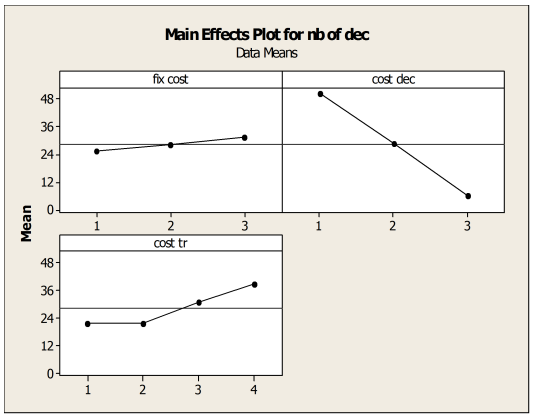

Figure A.2: Main effect plots of cost factors on number of staffed-beds downsized

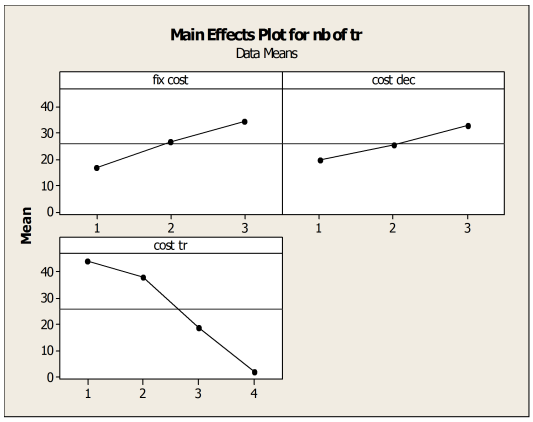

Figure A.3: Main effect plots of cost factors on number of staffed-beds downsized

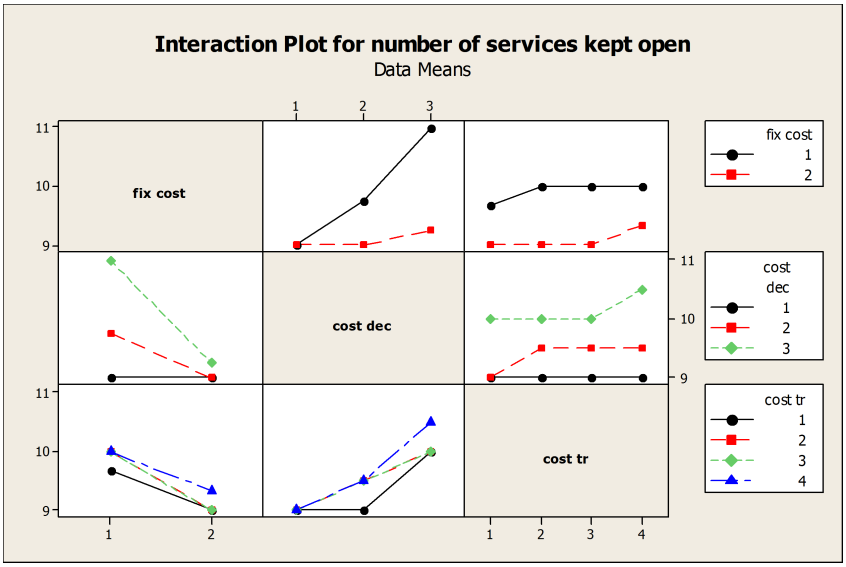

Figure A.4: Interaction plots of cost factors on number of services kept open (on facility relocation decisions)

downsizing and transferring resources presented in Fig. A.5 and A.6 respectively. Intuitively when DC is low, the highest amount of downsize occurs. When DC is at its highest value (level 3), there is no decrement. Generally TC has a clear negative correlation with DC in terms of downsized beds such that as TC gets smaller, less number of beds gets downsized. However, small values of TC (lower than level 2) do not have any effect on number of staffed-beds downsized.

On the other side in terms of transfer, relation between TC and DC depends on the tradeoff between holding cost of not-downsized resources and purchase+holding costs of new resources. Fig.14 presents that for all values of TC smaller level 4, even if DC is low the model tends to transfer beds rather than downsize them even though it means paying holding cost for each resource not downsized. However, when DC is at its highest level, downsizing gets too costly such that more services are kept open which in turn results in finding less possibility to transfer even if TC is set its lowest level.

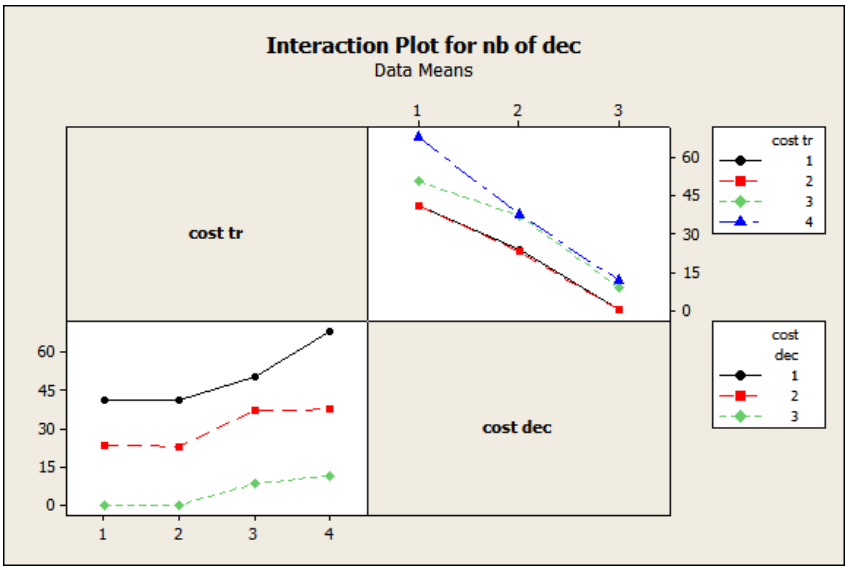

Figure A.5: Interaction plots of cost factors on number of staffed-beds downsized

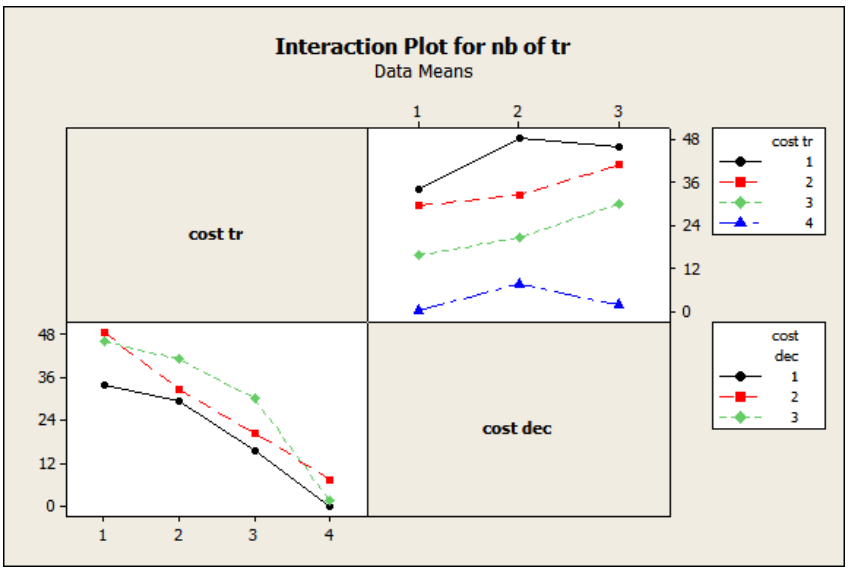

Figure A.6: Interaction plots of cost factors on number of staffed-beds transferred

Appendix B

Capacity planning plots for Scenario 2

The capacity planning plots for Scenario 2 (where no cooperation/transfer is assumed) is given in Fig. B.1, B.2 and B.3 for three service units. It is not any more profitable to close an existing facility since its cost cannot be compensated by other decisions such as downsizing or transferring. The model has to make new investments in order to ensure required service level, thus the number of staffed-beds purchased is 2 times more than the ones in the base scenario, which results in a higher holding cost. Furthermore, in a network where cooperation is not possible, the model chooses to assign patients to their "not primarily preferred" hospitals due to their idle capacity rather than increase the capacity (by making new investments) on their preferred hospitals.

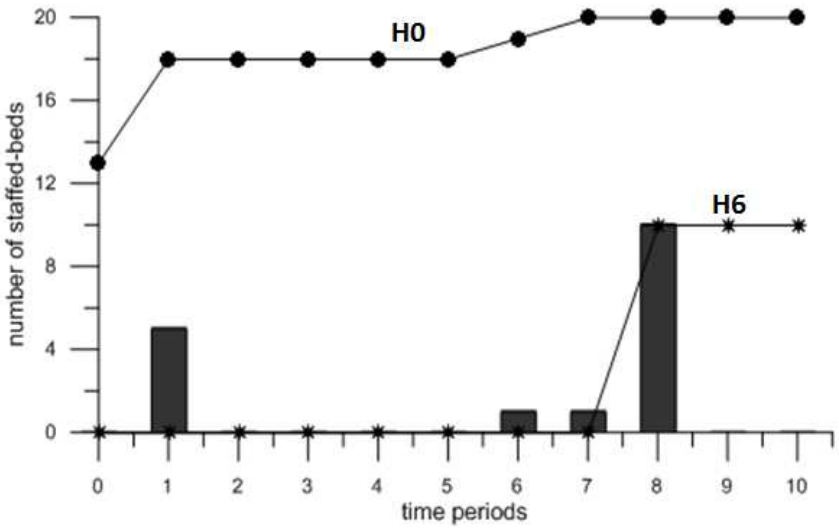

Figure B.1: Optimum planning in NICUs (no cooperation)

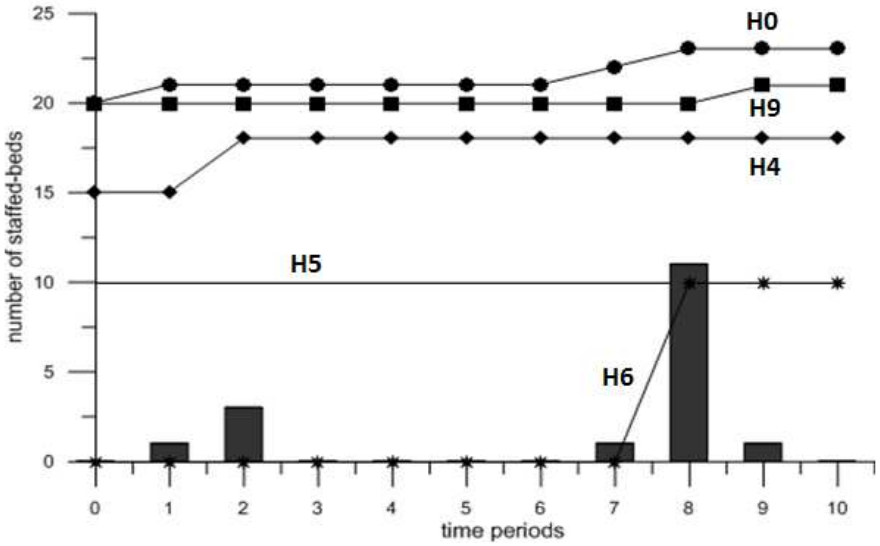

Figure B.2: Optimum planning in Basic Neonatals (no cooperation)

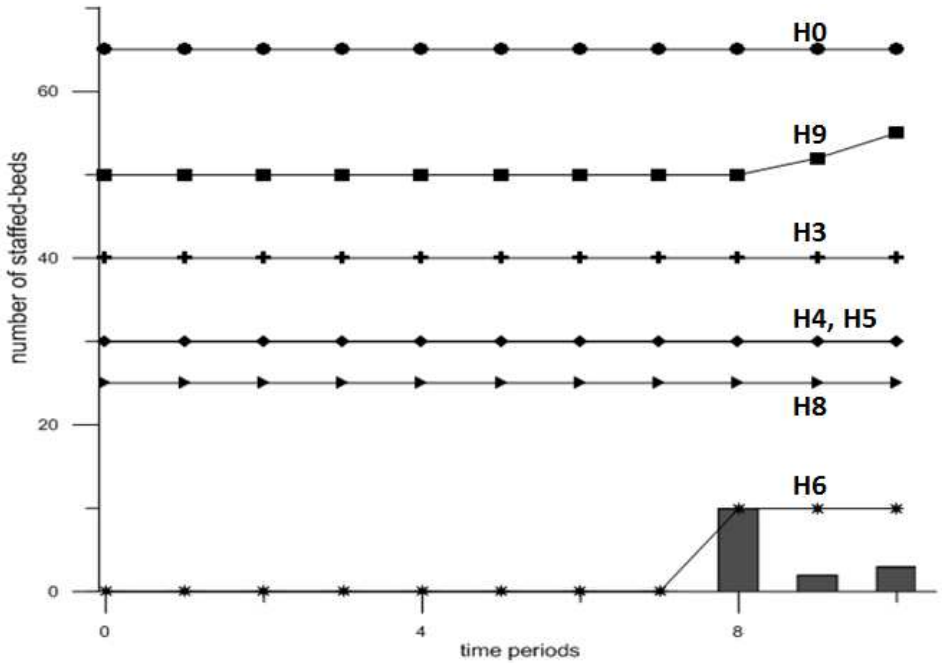

Figure B.3: Optimum planning in OBs (no cooperation)

Appendix C

Capacity planning plots for Scenario 3

The capacity planning plots for Scenario 3, where service level requirement is 99% for NICUs and neonatals, is given in Fig. C.1, C.2 and C.3.

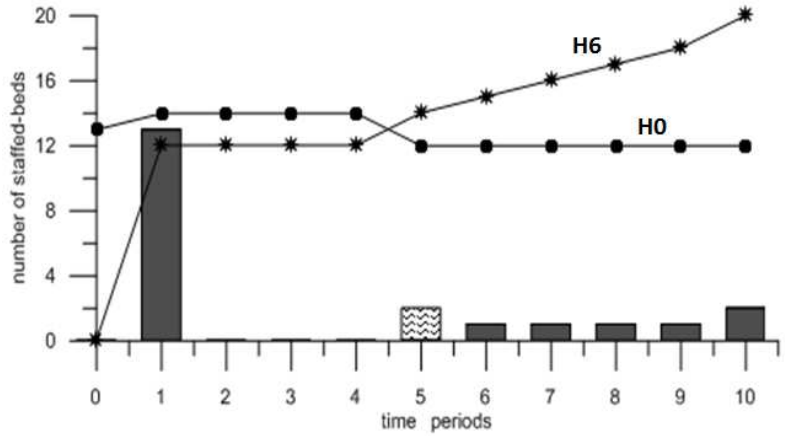

Figure C.1: Optimum planning in NICUs (higher service level)

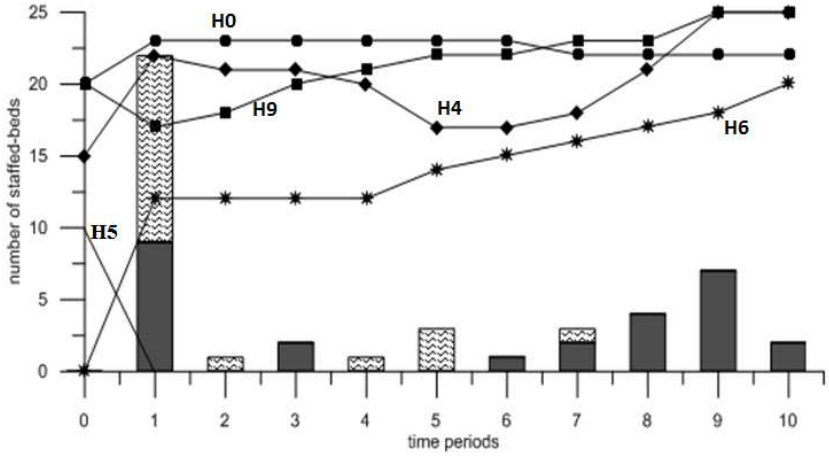

Figure C.2: Optimum planning in Basic Neonatals (higher service level)

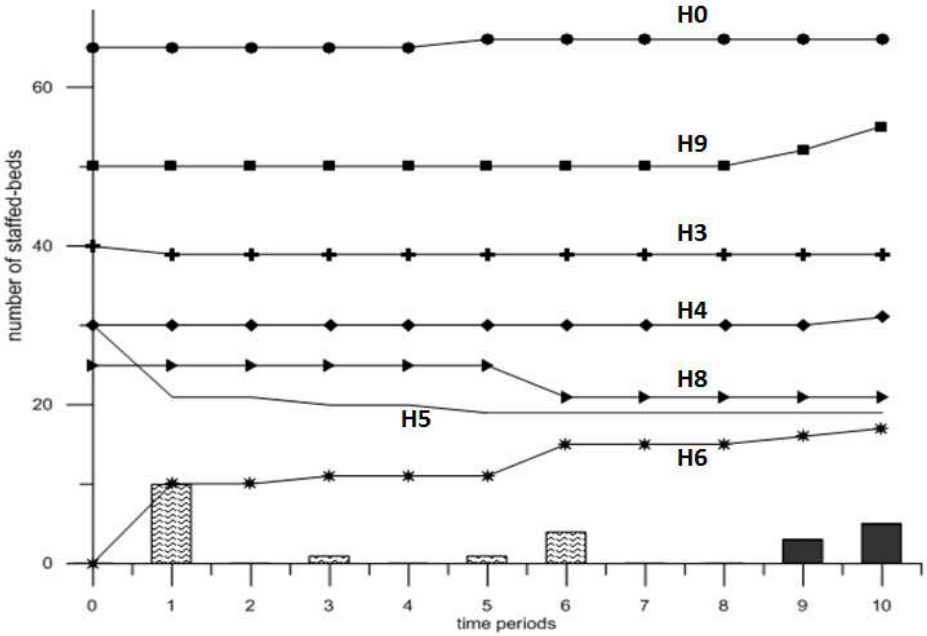

Figure C.3: Optimum planning in OBs (higher service level)

Appendix D

Binomial Moment Transformation

In this appendix, we present the derivation of finite system equations from infinite system balance equations of multidimensional MC through binomial moment transformation.

We multiply both sides of the MC infinite balance equations (given in subsection 4.3.2) by $(1+x)^k$ and summing on k from 0 to infinity:

For $0 \leqslant n < N_i$

$$\begin{aligned}
&\sum_{k=0}^{\infty} P(1,1,n,k)(a_1+\gamma_1+a_2+\gamma_2+\lambda+n+k)(1+x)^k \\
&\qquad = \sum_{k=0}^{\infty} P(0,1,n,k)w_1(1+x)^k + \sum_{k=0}^{\infty} P(1,0,n,k)w_2(1+x)^k \\
&\qquad + \sum_{k=0}^{\infty} P(1,1,n-1,k)(a_1+a_2+\lambda)(1+x)^k \\
&\qquad + \sum_{k=0}^{\infty} P(1,1,n+1,k)(n+1)(1+x)^k + \sum_{k=0}^{\infty} P(1,1,n,k+1)(k+1)(1+x)^k
\end{aligned} \tag{D.1}$$

Rewriting equation D.1 as:

$$\begin{aligned}
&(a_1+\gamma_1+a_2+\gamma_2+\lambda+n+k)\sum_{k=l}^{\infty}\binom{k}{l}P(1,1,n,k) \\
&\qquad = w_1\sum_{k=l}^{\infty}\binom{k}{l}P(0,1,n,k) + w_2\sum_{k=l}^{\infty}\binom{k}{l}P(1,0,n,k) \\
&\qquad + (a_1+a_2+\lambda)\sum_{k=l}^{\infty}\binom{k}{l}P(1,1,n-1,k) \\
&\qquad + (n+1)\sum_{k=l}^{\infty}\binom{k}{l}P(1,1,n+1,k) + \sum_{k=l}^{\infty}(k+1)\binom{k}{l}P(1,1,n,k+1)
\end{aligned}$$

Rewriting equation D.1 by using the relation $\sum_{k=l}^{\infty}\binom{k}{l}P(1,1,n,k) = B_l(1,1,n)$:

$$(a_1+\gamma_1+a_2+\gamma_2+\lambda+n)B_l(1,1,n)+\sum_{k=l}^{\infty}\binom{k}{l}kP(1,1,n,k)$$
$$=w_1B_l(0,1,n)+w_2B_l(1,0,n)+(a_1+a_2+\lambda)B_l(1,1,n-1)$$
$$+(n+1)B_l(1,1,n+1)+\sum_{k=l}^{\infty}(k+1)\binom{k}{l}P(1,1,n,k+1)$$

While the transformation is straightforward for finite variables, for the r.v. l which can take infinite values, we proceed as in the following:

Note that $(l+1)\sum_{k=l}^{\infty}\binom{k}{l}P(1,1,n,k+1)$ can be written as $\sum_{k'=l+1}^{\infty}\binom{k'-1}{l}P(1,1,n,k')$

Therefore,

$$\sum_{k=l}^{\infty}k\binom{k}{l}P(1,1,n,k)-\sum_{k'=l+1}^{\infty}\binom{k'-1}{l}P(1,1,n,k')$$
$$=lP(1,1,n,l)+\sum_{k=l+1}^{\infty}k\binom{k}{l}P(1,1,n,k)-\sum_{k=l+1}^{\infty}k\binom{k-1}{l}P(1,1,n,k)$$
$$=lP(1,1,n,l)+\sum_{k=l+1}^{\infty}k\left[\binom{k}{l}-\binom{k-1}{l}\right]P(1,1,n,k)$$

From $\binom{k}{l}=\binom{k-1}{l}+\binom{k-1}{l-1}$

$$=lP(1,1,n,l)+\sum_{k=l+1}^{\infty}l\binom{k}{l}P(1,1,n,k)$$
$$=lP(1,1,n,l)+\left(\sum_{k=l}^{\infty}l\binom{k}{l}P(1,1,n,k)-lP(1,1,n,l)\right)$$
$$=l\sum_{k=l}^{\infty}\binom{k}{l}P(1,1,n,k)$$
$$=lB_l(1,1,n)$$

Then equation D.1 can be written as:

$$(a_1+\gamma_1+a_2+\gamma_2+\lambda+n+\boldsymbol{l})B_l(1,1,n)$$
$$=w_1B_l(0,1,n)+w_2B_l(1,0,n)+(a_1+a_2+\lambda)B_l(1,1,n-1)$$
$$+(n+1)B_l(1,1,n+1)$$

With similar transformations the remaining equations for the case $n < N$ can also be written as:

$$\begin{aligned}(a_1+\gamma_1+w_2+\lambda+n+\boldsymbol{l})B_l(1,0,n) = {} & w_1B_l(1,1,n)+\gamma_2B_l(0,0,n)\\ & +(a_1+\lambda)B_l(1,0,n-1)\\ & +(n+1)B_l(1,0,n+1)\end{aligned} \tag{D.2}$$

$$\begin{aligned}(w_1+a_2+\gamma_2+\lambda+n+\boldsymbol{l})B_l(0,1,n) = {} & \gamma_1B_l(1,1,n)+w_2B_l(0,0,n)\\ & +(a_2+\lambda)B_l(0,1,n-1)\\ & +(n+1)B_l(0,1,n+1)\end{aligned} \tag{D.3}$$

$$\begin{aligned}(w_1+w_2+\lambda+n+\boldsymbol{l})B_l(0,0,n) = {} & \gamma_1B_l(1,0,n)+\gamma_2B_l(0,1,n)\\ & +(\lambda)B_l(0,0,n-1)+(n+1)B_l(0,0,n+1)\end{aligned} \tag{D.4}$$

On the boundary space, similar transformations are done:

For $n = N$ the first equation:

$$\begin{aligned}&(a_1+\gamma_1+a_2+\gamma_2+\lambda+N+l)\sum_{k=l}^{\infty}\binom{k}{l}P(1,1,N,k)\\ &\quad= w_1\sum_{k=l}^{\infty}\binom{k}{l}P(0,1,N,k)+w_2\sum_{k=l}^{\infty}\binom{k}{l}P(1,0,N,k)\\ &\quad+(a_1+a_2+\lambda)\sum_{k=l}^{\infty}\binom{k}{l}P(1,1,N-1,k)\\ &\quad+(a_1+a_2+\lambda)\sum_{k=l}^{\infty}\binom{k}{l}P(1,1,N,k-1)+\sum_{k=l}^{\infty}(k+1)\binom{k}{l}P(1,1,N,k+1)\end{aligned} \tag{D.5}$$

Rewriting equation D.5 by using the same relation $\sum_{k=l}^{\infty}\binom{k}{l}P(1,1,n,k) = B_l(1,1,n)$ and the binomial transformation presented in the case before:

$$\begin{aligned}&(a_1+a_2+\lambda)\sum_{k=l}^{\infty}\binom{k}{l}P(1,1,N,k)+(\gamma_1+\gamma_2+N+\boldsymbol{l})B_l(1,1,N)\\ &\quad= w_1B_l(0,1,N)+w_2B_l(1,0,N)+(a_1+a_2+\lambda)B_l(1,1,N-1)\\ &\quad+(a_1+a_2+\lambda)\sum_{k=l}^{\infty}\binom{k}{l}P(1,1,N,k-1)\end{aligned}$$

On the boundary space equation, when n=N, the arrivals are accepted to the infinite server station that changes its state space from k-1 to k. Thus, there is one extra term that exists in the boundary space equations which is needed to be transformed into Binomial moments. If we focus on this term:

$$
\begin{aligned}
&(a_1+a_2+\lambda)\left(\sum_{k=l}^{\infty}\binom{k}{l}P(1,1,N,k-1)-\sum_{k=l}^{\infty}\binom{k}{l}P(1,1,N,k)\right)\\
&=(a_1+a_2+\lambda)\left(\sum_{k'=l-1}^{\infty}\binom{k'+1}{l}P(1,1,N,k')-\sum_{k=l}^{\infty}\binom{k}{l}P(1,1,N,k)\right)\\
&=(a_1+a_2+\lambda)\\
&\quad\left((P(1,1,N,l-1)+\sum_{k'=l}^{\infty}\binom{k'+1}{l}P(1,1,N,k')-\sum_{k=l}^{\infty}\binom{k}{l}P(1,1,N,k)\right)\\
&=(a_1+a_2+\lambda)\left(P(1,1,N,l-1)+\sum_{k=l}^{\infty}\left[\binom{k+1}{l}-\binom{k}{l}\right]P(1,1,N,k)\right)\\
&=(a_1+a_2+\lambda)\left(P(1,1,N,l-1)+\sum_{k=l}^{\infty}\binom{k}{l-1}P(1,1,N,k)\right)\\
&=(a_1+a_2+\lambda)\sum_{k=l-1}^{\infty}\binom{k}{l-1}P(1,1,N,k)\\
&=(a_1+a_2+\lambda)B_{l-1}(1,1,N)
\end{aligned}
$$

Then equation D.5 can be written as:

$$
\begin{aligned}
(\gamma_1+\gamma_2+N+\boldsymbol{l})B_l(1,1,N)&=w_1B_l(0,1,N)+w_2B_l(1,0,N)\\
&\quad+(a_1+a_2+\lambda)B_l(1,1,N-1)\\
&\quad+(a_1+a_2+\lambda)B_{l-1}(1,1,N)
\end{aligned}
$$

With similar transformations, the remaining equations for the case $n=N$ can also be written as:

$$
\begin{aligned}
(\gamma_1+w_2+N+l)B_l(1,0,N)&=w_1B_l(0,0,N)+\gamma_2B_l(1,1,N)\\
&\quad+(a_1+\lambda)B_l(1,0,N-1)+(a_1+\lambda)B_{l-1}(1,0,N)
\end{aligned}
\tag{D.6}
$$

$$
\begin{aligned}
(w_1+\gamma_2+N+l)B_l(0,1,N)&=\gamma_1B_l(1,1,N)+w_2B_l(0,0,N)\\
&\quad+(a_2+\lambda)B_l(0,1,N-1)+(a_2+\lambda)B_{l-1}(0,1,N)
\end{aligned}
\tag{D.7}
$$

$$(w_1 + w_2 + N + l)B_l(0, 0, N) = \gamma_1 B_l(1, 0, N) + \gamma_2 B_l(0, 1, N) + (\lambda)B_l(0, 0, N - 1) + (\lambda)B_{l-1}(0, 0, N) \tag{D.8}$$

Appendix E

Establishing Structural Properties of Optimal Value Functions

Contents

E.1 For Single Hospital

Proofs are established by using induction on value function operators.

Lemma 2: If B and D hold for f(x), then D holds for $T_D f(x)$.

Proof. The lemma is proved by establishing the inequality

$$T_D f(x+1) - T_D f(x) \geqslant -r_1, \forall x \in \Omega|^{\cdot\cdot}0 \leqslant x \leqslant N-1$$

which holds as

$$\begin{aligned} T_D f(x+1) - T_D f(x) &= \frac{x+1}{N} f(x) + \frac{N-x-1}{N} f(x+1) - \frac{x}{N} f(x-1) - \frac{N-x}{N} f(x) \\ &= \frac{x}{N}\Delta f(x) + \frac{N-x}{N}\Delta f(x+1) + \frac{1}{N}\left(f(x) - f(x+1)\right) \\ &\geq \frac{x}{N}\Delta f(x) + \frac{N-x}{N}\Delta f(x+1) \qquad \text{(by B)} \\ &\geq -r_1 \qquad \text{(by D)} \end{aligned}$$

Q.E.D

Lemma 3: If E and D hold for f(x), then D holds for $T_{A_i} f(x)$.
Proof: We need to show

$$T_{A_i} f(x+1) - T_{A_i} f(x) \geqslant -r_1, \qquad \forall x \in \Omega|^{\cdot\cdot}0 \leqslant x \leqslant N-1$$

which holds for two cases (I) and (II):
For Case (I) where $x < N-1$:

$$\begin{aligned} T_{A_i}f(x+1) - T_{A_i}f(x) &= f(x+1) + \max\{r_i + f(x+2) - f(x+1), 0\} \\ &\quad - f(x) - \max\{r_i + f(x+1) - f(x), 0\} \\ &(\text{by } \max\{a,0\} - \max\{b,0\} \geqslant -\max\{b-a,0\}) \\ &\geq f(x+1) - f(x) \\ &\quad - \max\{2f(x+1) - f(x) - f(x+2), 0\} \qquad \text{(by E)} \\ &= f(x+1) - f(x) - 2f(x+1) + f(x) + f(x+2) \\ &= f(x+2) - f(x+1) \\ &\geq -r_1 \end{aligned}$$

For Case (II) where $x = N-1$:

$$T_{A_i}f(N) - T_{A_i}f(N-1) = f(N) - \max\{r_i + f(N), f(N-1)\}$$

if $r_i + f(N) \geq f(N-1)$:

$$\begin{aligned} &= f(N) - r_i - f(N) \\ &= -r_i \\ &\geq -r_1 \end{aligned}$$

if $r_i + f(N) \leq f(N-1)$:

$$\begin{aligned} &= f(N) - f(N-1) \\ &\geq -r_1 \qquad \text{(by D)} \end{aligned}$$

Q.E.D

Lemma 4: If B holds for f(x), then B holds for $T_D f(x)$.
Proof: We need to show

$$T_D f(x+1) - T_D f(x) \leqslant 0 \qquad \forall x \in \Omega | \ddot{} 0 \leqslant x \leqslant N-1$$

which hold as:

$$\begin{aligned} T_D f(x+1) - T_D f(x) &= \frac{x+1}{N} f(x) + \frac{N-x-1}{N} f(x+1) - \frac{x}{N} f(x-1) - \frac{N-x}{N} f(x) \\ &= \frac{x}{N}\left(f(x) - f(x-1)\right) + \frac{N-x-1}{N}\left(f(x+1) - f(x)\right) \\ &\leq 0 \qquad \text{(by B)} \end{aligned}$$

Q.E.D

Lemma 5: If B holds for f(x), then B holds for $T_{A_i} f(x)$.
Proof: We need to show

$$T_{A_i} f(x+1) - T_{A_i} f(x) \leqslant 0 \qquad \forall x \in \Omega | \ddot{} 0 \leqslant x \leqslant N-1$$

which holds for two cases (I) and (II):

For Case (I) where $x < N-1$:

$$\begin{aligned}
T_{A_i}f(x+1) - T_{A_i}f(x) &= \max\{r_i + f(x+2), f(x+1)\} - \max\{r_i + f(x+1), f(x)\} \\
&= f(x+1) - f(x) + \max\{r_i + f(x+2) - f(x+1), 0\} \\
&\quad - \max\{r_i + f(x+1) - f(x), 0\} \\
&\leq \max\{r_i + f(x+2) - f(x+1), 0\} \\
&\quad - \max\{r_i + f(x+1) - f(x), 0\} \qquad \text{(by B)} \\
&\text{(from } \max\{a,0\} - \max\{b,0\} \leqslant \max\{a-b,0\}\text{)} \\
&\leq \max\{f(x+2) - 2f(x+1) + f(x), 0\} \\
&= \max\{\Delta f(x+1) - \Delta f(x), 0\} \qquad \text{(by E)} \\
&\leq 0
\end{aligned}$$

For Case (II) where $x = N-1$:

$$T_{A_i}f(N) - T_{A_i}f(N-1) = f(N) - \max\{r_i + f(N), f(N-1)\}$$

if $r_i + f(N) \geq f(N-1)$:

$$\begin{aligned}
&= f(N) - r_i - f(N) \\
&= -r_i \\
&\leq 0
\end{aligned}$$

if $r_i + f(N) \leq f(N-1)$:

$$\begin{aligned}
&= f(N) - f(N-1) \\
&\leq 0 \qquad \text{(by B)}
\end{aligned}$$

Q.E.D

Lemma 6: If E holds for f(x), then E holds for $T_D f(x)$.
Proof. It is enough to show

$$\Delta T_D f(x) - \Delta T_D f(x+1) \geqslant 0 \qquad \forall x \in \Omega | \ddot{} 0 \leqslant x \leqslant N-1,$$

which hold as:

$$\begin{aligned}
&\Delta T_D f(x) - \Delta T_D f(x+1) \\
&= 2\left(\frac{N-x}{N}f(x) + \frac{x}{N}f(x-1)\right) - \left(\frac{N-x+1}{N}f(x-1) + \frac{x-1}{N}f(x-2)\right) \\
&\quad - \left(\frac{N-x-1}{N}f(x+1) + \frac{x+1}{N}f(x)\right) \\
&= \frac{N-x}{N}((2f(x) - f(x-1) - f(x+1)) \\
&\quad + \frac{x}{N}((2f(x-1) - f(x) - f(x-2) + \frac{1}{N}(-f(x-1) + f(x-2) + f(x+1) - f(x)) \\
&= \frac{N-x-1}{N}(2f(x) - f(x-1) - f(x+1)) + \frac{x-1}{N}(2f(x-1) - f(x) - f(x-2)) \\
&\geq 0 \qquad \text{(by E)}
\end{aligned}$$

Q.E.D

Lemma 7: If E holds for f(x), then E holds for T_{A_i}.
Proof. It is enough to show:

$$\Delta T_{A_i} f(x) - \Delta T_{A_i} f(x+1) \geqslant 0,\ \forall x \in \Omega |^{\cdot\cdot} 0 \leqslant x \leqslant N-1$$

which holds for two cases (I) and (II):

For Case (I) where $x < N-1$:

$$\begin{aligned}
\Delta T_{A_i} f(x) - \Delta T_{A_i} f(x+1) &= T_{A_i} f(x) - T_{A_i} f(x-1) + T_{A_i} f(x) - T_{A_i} f(x+1) \\
&= \max\{r_i + f(x+1), f(x)\} - \max\{r_i + f(x), f(x-1)\} \\
&\quad + \max\{r_i + f(x+1), f(x)\} \\
&\quad - \max\{r_i + f(x+2), f(x+1)\} \\
&= f(x) - f(x-1) + \max\{r_i + f(x+1) - f(x), 0\} \\
&\quad - \max\{r_i + f(x) - f(x-1), 0\} + f(x) \\
&\quad - f(x+1) + \max\{r_i + f(x+1) - f(x), 0\} \\
&\quad - \max\{r_i + f(x+2) - f(x+1), 0\} \\
&= 2f(x) - f(x-1) - f(x+1) \\
&\quad + \max\{r_i + f(x+1) - f(x), 0\} \\
&\quad - \max\{r_i + f(x) - f(x-1), 0\} \\
&\quad + \max\{r_i + f(x+1) - f(x), 0\} \\
&\quad - \max\{r_i + f(x+2) - f(x+1), 0\} \\
&\geq 2f(x) - f(x-1) - f(x+1) \\
&\quad - \max\{2f(x) - f(x-1) - f(x+1), 0\} \\
&\quad - \max\{f(x+2) - 2f(x+1) + f(x), 0\} \\
&= 0 \qquad \text{(by E)}
\end{aligned}$$

For Case (II) where $x = N-1$:

$$\begin{aligned}
\Delta T_{A_i} f(N-1) - \Delta T_{A_i} f(N) &= T_{A_i} f(N-1) - T_{A_i} f(N-2) + T_{A_i} f(N-1) - T_{A_i} f(N) \\
&= \max\{r_i + f(N), f(N-1)\} \\
&\quad - \max\{r_i + f(N-1), f(N-2)\} \\
&\quad + \max\{r_i + f(N), f(N-1)\} - f(N) \\
&= f(N-1) - f(N-2) + \max\{r_i + f(N) - f(N-1), 0\} \\
&\quad - \max\{r_i + f(N-1) - f(N-2), 0\} + f(N-1) \\
&\quad + \max\{r_i + f(N) - f(N-1), 0\} - f(N) \\
&= 2f(N-1) - f(N-2) - f(N) \\
&\quad + \max\{r_i + f(N) - f(N-1), 0\} \\
&\quad - \max\{r_i + f(N-1) - f(N-2), 0\} \\
&\quad + \max\{r_i + f(N) - f(N-1), 0\} \\
&\geq 2f(N-1) - f(N-2) - f(N) \\
&\quad - \max\{2f(N-1) - f(N-2) - f(N), 0\} \\
&\quad + \max\{r_i + f(N) - f(N-1), 0\} \\
&\geq 0 + \max\{r_i + f(N) - f(N-1), 0\} \\
&\geq 0 \qquad \text{(by E)}
\end{aligned}$$

Q.E.D

E.2 For 2-Hospital

Proofs are established by using induction on value function operators.

Lemma 2: If B and D hold for f(x), then D holds for $T_{D_i}f(x)$.
Proof. The lemma is proved by establishing two inequalities

$$T_{D_i}f(x+e_i) - T_{D_i}f(x) \geqslant -r_{ii},\ \forall x \in \Omega|^{\cdot\cdot}0 \leqslant x_i \leqslant N_i - 1$$
$$T_{D_j}f(x+e_i) - T_{D_j}f(x) \geqslant -r_{ii},\ \forall x \in \Omega|^{\cdot\cdot}0 \leqslant x_i \leqslant N_i - 1$$

which holds as

$$\begin{aligned} T_{D_i}f(x+e_i) - T_{D_i}f(x) &= \frac{x_i}{N_i}\Delta^i f(x) + \frac{N_i - x_i}{N_i}\Delta^i f(x+e_i) + \frac{1}{N_i}\left(f(x) - f(x+e_i)\right) \\ &\geq \frac{x_i}{N_i}\Delta^i f(x) + \frac{N_i - x_i}{N_i}\Delta^i f(x+e_i) \qquad \text{(by B)} \\ &\geq -r_{ii} \qquad \text{(by D)} \end{aligned}$$

and

$$\begin{aligned} T_{D_j}f(x+e_i) - T_{D_j}f(x) &= \frac{x_j}{N_j}\Delta f(x+e_i-e_j) + \frac{N_j - x_j}{N_j}\Delta f(x+e_i) \\ &\geq -r_{ii} \qquad \text{(by D)} \end{aligned}$$

Q.E.D

Lemma 3: If C and D hold for f(x), then D holds for $T_{A_i}f(x)$.
Proof: We need to show

$$T_{A_i}f(x+e_j) - T_{A_i}f(x) \geqslant -r_{ii}(= -r_{jj}), \qquad \forall x \in \Omega|^{\cdot\cdot}0 \leqslant x_j \leqslant N_j - 1$$

Case $x_i \leqslant N_i - 1$, $e_i \neq e_j$

$$\begin{aligned} T_{A_i}f(x+e_j) - T_{A_i}f(x) &= [r_{ii} + f(x+e_j+e_i)] - [r_{ii} + f(x+e_i)] \text{ (by C,D)} \\ &\geq -r_{jj} \qquad \text{(by D)} \end{aligned}$$

Case $x_i \leqslant N_i - 1$, $e_i = e_j$

$$\begin{aligned} T_{A_i}f(x+e_j) - T_{A_i}f(x) &\geq [f(x+e_j)] - [r_{ii} + f(x+e_i)] \text{ (by definition, C, D)} \\ &\geq -r_{ii} \qquad (\text{by } e_i{=}e_j) \end{aligned}$$

Case $x_i = N_i, x_j \leqslant N_j - 1$

$$\begin{aligned} T_{A_i}f(x+e_j) - T_{A_i}f(x) &\geq [f(x+e_j)] - \max\{r_{ij} + f(x+e_j), f(x)\} \text{ (by definition)} \\ &\geq \min\{-r_{ij}, f(x+e_j) - f(x)\} \\ &\geq -r_{ii} \qquad \text{(by D)} \end{aligned}$$

Q.E.D

Lemma 4: If C holds for $f(x)$, then C holds for $T_D f(x)$.
Proof: We need to show

$$r_{ii} + T_D f(x + e_i) \geqslant r_{ij} + T_D f(x + e_j), \forall x \in \Omega | \ddot{} 0 \leqslant x_i \leqslant N_i - 1, 0 \leqslant x_j \leqslant N_j - 1$$

which hold as:

$$\begin{aligned} T_D f(x + e_i) - T_D f(x + e_j) = &\left(\frac{N - x_i - x_j - 1}{N} f(x + e_i) + \frac{x_i + 1}{N} f(x) \right. \\ &\left. + \frac{x_j}{N} f(x + e_i + e_j) \right) - \left(\frac{N - x_i - x_j - 1}{N} f(x + e_j) \right. \\ &\left. + \frac{x_i}{N} f(x + e_j - e_i) + \frac{x_j + 1}{N} f(x) \right) \\ = &\frac{N - x_i - x_j - 1}{N} \left(f(x + e_i) - f(x + e_j) \right) + \frac{x_i}{N} \left(f(x) \right. \\ &\left. - f(x - e_i + e_j) \right) + \frac{x_j}{N} \left(f(x + e_i - e_j) - f(x) \right) \\ \geq &\frac{N - 1}{N} (r_{ij} - r_{ii}) \qquad \text{(by C)} \\ \geq &\, r_{ij} - r_{ii} \end{aligned}$$

Q.E.D

Remark: this property does not hold for $T_{D_i} f(x)$.

Lemma 5: If C and D hold for $f(x)$, then C holds for $T_{A_i} f(x)$.
Proof: We need to show

$$r_{ii} + T_{A_i} f(x + e_i) \geqslant r_{ij} + T_{A_i} f(x + e_j), \, \forall x \in \Omega | \ddot{} 0 \leqslant x_i \leqslant N_i - 1, 0 \leqslant x_j \leqslant N_j - 1$$
$$r_{jj} + T_{A_j} f(x + e_i) \geqslant r_{ji} + T_{A_j} f(x + e_j), \, \forall x \in \Omega | \ddot{} 0 \leqslant x_i \leqslant N_i - 1, 0 \leqslant x_j \leqslant N_j - 1$$

which hold as:

$$\begin{aligned} T_{A_i} f(x + e_i) - T_{A_i} f(x + e_j) = &\, T_{A_i} f(x + e_i) - (r_{ii} + f(x + e_j + e_i)) \qquad \text{(by C, D)} \\ \geq &\, (r_{ij} + f(x + e_j + e_i)) - (r_{ii} + f(x + e_j + e_i)) \\ &\text{(by definition)} \\ \geq &\, r_{ij} - r_{ii} \end{aligned}$$

and

$$\begin{aligned} T_{A_j} f(x + e_i) - T_{A_j} f(x + e_j) = &\, (r_{jj} + f(x + e_j + e_i)) - T_{A_j} f(x + e_j) \, \text{(by C, D)} \\ \geq &\, (r_{jj} + f(x + e_j + e_i)) - f(x + e_j) \, \text{(by definition)} \\ \geq &\, 0 \, \text{(by D)} \\ \geq &\, r_{ji} - r_{jj} \end{aligned}$$

Q.E.D

Lemma 6: If B holds for $f(x)$, then B holds for $T_D f(x)$.
Proof. It is enough to show

$$T_D f(x+e_i) - T_D f(x) \leqslant 0, \qquad \forall x \in \Omega | \ddot{} 0 \leqslant x_i \leqslant N_i - 1$$

which hold as

$$\begin{aligned} T_D f(x+e_i) - T_D f(x) &= \left(\frac{N - x_i - x_j - 1}{N} f(x+e_i) + \frac{x_i + 1}{N} f(x) + \frac{x_j}{N} f(x+e_i - \right. \\ &\quad \left. e_j) \right) - \left(\frac{N - x_i - x_j}{N} f(x) + \frac{x_i}{N} f(x-e_i) + \frac{x_j}{N} f(x-e_j) \right) \\ &= \frac{N - x_i - x_j - 1}{N} \left(f(x+e_i) - f(x) \right) \\ &\quad + \frac{x_i}{N} \left(f(x) - f(x-e_i) \right) \frac{x_j}{N} \left(f(x+e_i+e_j) - f(x-e_j) \right) \\ &\leq 0 \qquad \text{(by B)} \end{aligned}$$

Q.E.D

Lemma 7: If B, C and D hold for $f(x)$, then B holds for $T_{A_i} f(x)$.
Proof. It is enough to show

$$T_{A_i} f(x+e_i) - T_{A_i} f(x) \leqslant 0, \, \forall x \in \Omega | \ddot{} 0 \leqslant x_i \leqslant N_i - 1 \tag{E.1}$$

$$T_{A_j} f(x+e_i) - T_{A_j} f(x) \leqslant 0, \, \forall x \in \Omega | \ddot{} 0 \leqslant x_i \leqslant N_i - 1 \tag{E.2}$$

(I) holds as:

$$\begin{aligned} T_{A_i} f(x+e_i) - T_{A_i} f(x) &= T_{A_i} f(x+e_i) - (r_{ii} + f(x+e_i)) \, \text{(by C, D)} \\ &= \max\{\Delta^i f(x+2e_i), r_{ij} - r_{ii} + \Delta^j f(x+e_i+e_j), -r_{ii}\} \\ &\quad \text{(without term 1 if } x_i = N_i - 1 \text{ and term 2 if } x_j = N_j) \\ &\leq 0 \qquad \text{(by B)} \end{aligned}$$

Two cases for (II).

Case $x_j < N_j$,

$$\begin{aligned} T_{A_j} f(x+e_i) - T_{A_j} f(x) &= (r_{jj} + f(x+e_i+e_j)) - (r_{jj} + f(x+e_j)) \text{ (by C, D)} \\ &\leq 0 \qquad \text{(by B)} \end{aligned}$$

Case $x_j = N_j$,

$$\begin{aligned} T_{A_j} f(x+e_i) - T_{A_j} f(x) &\leq T_{A_j} f(x+e_i) - (r_{ji} + f(x+e_i)) \\ &= \max\{r_{ji} + f(x+2e_i), f(x+e_i)\} - (r_{ji} + f(x+e_i)) \\ &\quad \text{(without term 1 in max if } x_i = N_i - 1 \text{)} \\ &= \max\{f(x+2e_i) - f(x+e_i), -r_{ji}\} \\ &\leq 0 \qquad \text{(by B)} \end{aligned}$$

Q.E.D

Lemma 8: If E hold for $f(x)$, then E holds for $T_D f(x)$.
Proof: For $x \in \Omega | \ddot{} 1 \leqslant x_i \leqslant N_i - 1$,

$$\begin{aligned}
\Delta^i T_D f(x) - \Delta^i T_D f(x+e_i) &= 2\left(\frac{N-x_i-x_j}{N} f(x) + \frac{x_i}{N} f(x-e_i) + \frac{x_j}{N} f(x-e_j)\right) \\
&\quad - \left(\frac{N-x_i-x_j+1}{N} f(x-e_i) + \frac{x_i-1}{N} f(x-2e_i) + \right. \\
&\quad \left. \frac{x_j}{N} f(x-e_i-e_j)\right) - \left(\frac{N-x_i-x_j-1}{N} f(x+e_i) \right. \\
&\quad \left. + \frac{x_i+1}{N} f(x) + \frac{x_j}{N} f(x+e_i-e_j)\right) \\
&= \frac{N-x_i-x_j-1}{N}(2f(x) - f(x-e_i) - f(x+e_i)) \\
&\quad + \frac{x_i-1}{N}(2f(x-e_i) - f(x-2e_i) - f(x)) \\
&\quad + \frac{x_j}{N}(2f(x-e_j) - f(x-e_i-e_j) - f(x+e_i-e_j)) \\
&\geq 0 \qquad \text{(by E)}
\end{aligned}$$

Q.E.D

Lemma 9: If B, C, D, E hold for $f(x)$, then E holds for T_{A_i}.
Proof. It is enough to show:

$$\Delta^i T_{A_i} f(x) - \Delta^i T_{A_i} f(x+e_i) \geqslant 0,\ \forall x \in \Omega | \ddot{} 1 \leqslant x_i \leqslant N_i - 1$$
$$\Delta^i T_{A_j} f(x) - \Delta^i T_{A_j} f(x+e_i) \geqslant 0,\ \forall x \in \Omega | \ddot{} 1 \leqslant x_i \leqslant N_i - 1$$

Consider (I)

$$\begin{aligned}
\Delta^i T_{A_i} f(x) - \Delta^i T_{A_i} f(x+e_i) &= 2T_{A_i} f(x) - T_{A_i} f(x-e_i) - T_{A_i} f(x+e_i) \\
&= 2(r_{ii} + f(x+e_i)) - (r_{ii} + f(x)) - T_{A_i} f(x+e_i) \\
&\quad \text{(by C, D)} \\
&= r_{ii} + 2f(x+e_i) - f(x) - T_{A_i} f(x+e_i)
\end{aligned}$$

Case $x_i < N_i - 1$

$$\begin{aligned}
\Delta^i T_{A_i} f(x) - \Delta^i T_{A_i} f(x+e_i) &= r_{ii} + 2f(x+e_i) - f(x) - (r_{ii} + f(x+2e_i)) \text{ (by C, D)} \\
&= 2f(x+e_i) - f(x) - f(x+2e_i) \geqslant 0 \qquad \text{(by E)}
\end{aligned}$$

Case $x_i = N_i - 1$

$$\begin{aligned}
\Delta^i T_{A_i} f(x) - \Delta^i T_{A_i} f(x+e_i) &= r_{ii} + 2f(x+e_i) - f(x) \\
&\quad - \max\{r_{ij} + f(x+e_i+e_j), f(x+e_i)\}
\end{aligned}$$

$$= \min \begin{cases} \{r_{ii} + 2f(x+e_i) - (f(x) + r_{ij} + f(x+e_i+e_j)) \equiv X \\ r_{ii} + f(x+e_i) - f(x) \equiv Y \qquad (\text{without term } X \text{ if } x_j = N_j) \end{cases}$$

by D, $Y \geqslant 0$. by C,

$$\begin{aligned} X &\geq r_{ii} + 2f(x+e_i) - (f(x+e_i-e_j) + r_{ii} + f(x+e_i+e_j)) \\ &= 2f(x+e_i) - f(x+e_i-e_j) - f(x+e_i+e_j) \\ &\geq 0 \qquad (\text{by E}) \end{aligned}$$

which implies $\Delta^i T_{A_i} f(x) - \Delta^i T_{A_i} f(x+e_i) \geqslant 0$.

Consider (II)

It holds if $x_j < N_j$ as $\Delta^i T_{A_j} f(x) - \Delta^i T_{A_j} f(x+e_i) = \Delta^i f(x+e_j) - \Delta^i f(x+e_j+e_i)$. Assume $x_j = N_j$.

$$\begin{aligned} \Delta^i T_{A_j} f(x) &= T_{A_j} f(x) - T_{A_j} f(x-e_i) \\ &= \max\{r_{ji} + f(x+e_i), f(x)\} - \max\{r_{ji} + f(x), f(x-e_i)\} \\ &= f(x) - f(x-e_i) + \max\{0, r_{ji} + f(x+e_i) - f(x)\} \\ &\quad - \max\{0, r_{ji} + f(x) - f(x+e_i)\} \\ &\geq f(x) - f(x-e_i) - \max\{0, 2f(x) - f(x+e_i) - f(x-e_i)\} \\ &\quad (\text{by } \max\{0,a\} - \max\{0,b\} \geqslant -\max\{0, b-a\}) \\ &= f(x) - f(x-e_i) - (2f(x) - f(x+e_i) - f(x-e_i)) \qquad (\text{by E}) \\ &= -f(x) + f(x+e_i) \end{aligned}$$

Case $x_i < N_i - 1$

$$\begin{aligned} \Delta^i T_{A_j} f(x+e_i) &= T_{A_j} f(x+e_i) - T_{A_j} f(x) \\ &= \max\{r_{ji} + f(x+2e_i), f(x+e_i)\} - \max\{r_{ji} + f(x+e_i), f(x)\} \\ &= f(x+e_i) - f(x) + \max\{0, r_{ji} + f(x+2e_i) - f(x+e_i)\} \\ &\quad - \max\{0, r_{ji} + f(x+e_i) - f(x)\} \\ &\leq f(x+e_i) - f(x) + \max\{0, f(x+2e_i) + f(x) \\ &\qquad - 2f(x+e_i)\} \qquad (\text{by } \max\{0,a\} - \max\{0,b\} \leqslant \max\{0, a-b\}) \\ &= f(x+e_i) - f(x) \qquad (\text{by E}) \end{aligned}$$

Case $x_i = N_i - 1$

$$\begin{aligned} \Delta^i T_{A_j} f(x+e_i) &= \Delta^i T_{A_j} f(x+e_i) - \Delta^i T_{A_j} f(x) \\ &= f(x+e_i) - \max\{r_{ji} + f(x+e_i), f(x)\} \\ &\leq f(x+e_i) - f(x) \end{aligned}$$

As a result, $\Delta^i T_{A_i} f(x) - \Delta^i T_{A_i} f(x+e_i) \geqslant 0$.

Q.E.D

Appendix F

Upper bounds

Several upper bound formulations are presented in the following.

A very straightforward upper-bound can be obtained by grouping all hospitals into a single hospital and merge their individual arrival streams into one arrival stream with an arrival rate of $a(G)$. Then, we can use the aggregated system blocking probability in order to set an upper bound to total accepted load in the total system. The necessary notation and the LP model of UB1 are defined as in the following.

Notation:

G	set of the hospitals where $i \in G = \{0, 1, ..., I\}$
a_i	arrival load of hospital i, where $a_i = \lambda_i/\mu = \lambda_i$ since $\mu = 1$ and $a(G) = \sum_i a_i$
N_i	number of beds in hospital i
$P(G)$	global overflow probability if all hospital beds are grouped in a single hospital where $P(G) = B(a_1 + ...a_I, N_1 + ...N_I)$
P_i	isolated overflow probability if each hospital only accepts its own patients, $P_i = B(a_i, N_i)$
r_{ij}	reward of a patient of hospital i admitted in hospital j

Decision Variable:

x_{ij}	load of patients of hospital i admitted in hospital j .

The first linear programming model (UB1) which achieves an upper bound is formulated as follows:

$$UB1 = \max \sum_{i \in G} \sum_{j \in G} r_{ij} x_{ij}$$

subject to

$$\sum_{j} x_{ij} \leqslant a_i \tag{F.1}$$

$$\sum_{i} x_{ij} \leqslant N_j \tag{F.2}$$

$$x_{ii} \leqslant a_i(1 - P_i) \tag{F.3}$$

$$\sum_{i,j} x_{ij} \leqslant (1 - P(G))a(G) \tag{F.4}$$

$$x_{ij} \geqslant 0 \tag{F.5}$$

Our objective is to maximize the total reward obtained from the accepted load to each hospital. Constraint F.1 ensures that the total accepted load of patients i cannot bigger than its total arriving load. Constraint F.2 ensures that total load accepted to hospital j cannot be bigger than the total number of servers in that hospital regarding 100% utilization of hospital beds. Assuming that patients can be accepted only to their preferred hospitals, constraint F.3 defines the maximum amount of *class-1* load i that can accepted to a hospital i regarding the isolated blocking probability. Constraints F.4 sets an upper bound on total number of accepted load in the whole system by assuming the system is one big aggregated hospital with one big arrival stream. Constraint F.5 is the non-negativity constraint.

UB1 formulation is updated by introducing the phenomena of subsets of hospitals and a correction factor for achieve more realistic utilization of beds. The improved formulation UB2 is given as follows:

$$UB2 = \max \sum_{i \in G} \sum_{j \in G} r_{ij} x_{ij}$$

subject to

$$\sum_{j} x_{ij} \leqslant a_i \tag{F.6}$$

$$\sum_{i} x_{ij} \leqslant \frac{1}{(1 + 1/a(G))} N_j \tag{F.7}$$

$$\sum_{i,j \in S} x_{ij} \leqslant (1 - P(S))a(S) \quad \forall S \subseteq G \tag{F.8}$$

$$x_{ij} \geqslant 0 \tag{F.9}$$

where $P(S)$ is the global overflow probability if all beds of hospitals in S are grouped together and only patients of these hospitals are considered, i.e., $P(S) = B(a(S), N(S)), a(S) = \sum_{i \in S} a_i, N(S) = \sum_{i \in S} N_i$. In UB1, we let 100% usage of hospital beds which is not possible in practice. As there is no queue in the system, whenever a bed becomes available, the bed remains idle at least till the arrival of

a patient. The average waiting time is then at least $1/a(G)$, the waiting time for arrival of a patient even he is not served by the bed. Note that the average service time of a patient is normalized to 1. As a result, the constraints F.7 $\subseteq$ F.2.

Constraint F.8 guarantees an upper bound on the total accepted load in S where all possible subsets (S) of all hospitals (G) are taken into account. For example, for three hospital case, i.e. $G = \{1, 2, 3\}$, there are 7 constraints related to each nonempty subset S, i.e. $\{1\}$, $\{2\}$, $\{3\}$, $\{1, 2\}$, $\{1, 3\}$, $\{2, 3\}$, $\{1, 2, 3\}$. Therefore, for $S = \{i\}$, constraints F.3 $\subseteq$ F.8 whereas for $S = \{G\}$, constraints F.4 $\subseteq$ F.8, and additionally constraints F.8 includes constraints for all other subsets S of G. Therefore, UB2 computes a tighter upper-bound where UB2 $\leqslant$ UB1.

The two following models further improve the previous ones by bounding total accepted load of each hospital j with constraint F.11. We observe that the load accepted in each hospital j is no more than the accepted load of j with all patients first arrived at j, i.e. the accepted load of the pure loss system with N_j servers and $a(G)$ offered load.

$$UB3 = \max \sum_{i \in G} \sum_{j \in G} r_{ij} x_{ij}$$

subject to

$$\sum_{j} x_{ij} \leqslant a_i \tag{F.10}$$

$$\sum_{i} x_{ij} \leqslant a(G)(1 - B(a(G), N_j)) \tag{F.11}$$

$$x_{ii} \leqslant a_i(1 - P_i) \tag{F.12}$$

$$\sum_{i,j \in S} x_{ij} \leqslant (1 - P(G))a(G) \tag{F.13}$$

$$x_{ij} \geqslant 0 \tag{F.14}$$

The following model UB4 is the updated version of UB3 by introducing the previously mentioned phenomena of subsets where constraints F.12 and F.13 are replaced with the constraint F.17

$$UB4 = \max \sum_{i \in G} \sum_{j \in G} r_{ij} x_{ij}$$

subject to

$$\sum_j x_{ij} \leqslant a_i \tag{F.15}$$

$$\sum_i x_{ij} \leqslant a(G)(1 - B(a(G), N_j)) \tag{F.16}$$

$$\sum_{i,j \in S} x_{ij} \leqslant (1 - P(S))a(S) \quad \forall S \subseteq G \tag{F.17}$$

$$x_{ij} \geqslant 0 \tag{F.18}$$

The following models UB5 and UB6 are two natural extensions to UB4.

$$UB5 = \max \sum_{i \in G} \sum_{j \in G} r_{ij} x_{ij}$$

subject to

$$\sum_j x_{ij} \leqslant a_i$$

$$\sum_{i \in G} \sum_{j \in S} x_{ij} \leqslant a(G)(1 - B(a(G), N(S)) \quad \forall S \subseteq G$$

$$\sum_{i \in S} \sum_{j \in S} x_{ij} \leqslant a(S)(1 - B(a(S), N(S)) \quad \forall S \subseteq G$$

$$x_{ij} \geqslant 0$$

$$UB6 = \max \sum_{i \in G} \sum_{j \in G} r_{ij} x_{ij}$$

subject to

$$\sum_{i \in S'} \sum_{j \in S} x_{ij} \leqslant a(S')(1 - B(a(S'), N(S)) \quad \forall S' \subseteq G, \forall S \subseteq G$$

$$x_{ij} \geqslant 0$$

All models are tested on some instances and we observe that UB3 improves UB2 (UB3 $\leqslant$ UB2) whereas no significant difference is observed among others such that UB6=UB5=UB4=UB3. In all of the formulations above, we improve the constraints in a way to better bound the maximum total load that can be accepted in a hospital i.

Appendix G

Liste de Publications

Publications dans des revues internationals

C.Pehlivan, V.Augusto, X.Xie, Dynamic capacity planning and location of hierarchical service networks under service level constraints, *IEEE Transactions on Automation Science and Engineering (TASE)* to appear in 2014.

Conférences et communications internationales

Avec actes

Canan Pehlivan, Vincent Augusto, Xiaolan Xie and Catherine Crenn-Hebert. Multi-period capacity planning for maternity facilities in a perinatal network: A queuing and optimization approach. *IEEE International Conference on Automation Science and Engineering (CASE)*, pages 137-142. IEEE, 2012.

Pehlivan, Canan, Vincent Augusto, and Xiaolan Xie. Admission control in a pure loss healthcare network: MDP and DES approach. *WSC 2013 (Winter Simulation Conference)*. 2013.

Sans actes

C.Pehlivan, V.Augusto, X.Xie, (2011). Resource Capacity Planning in Perinatal Neonatal Network,*Conference on Operational Research Applied to Health Services (ORAHS 2011)*. Cardiff, Wales, 2011.

Bibliography

[Ahmed 2009] Mohamed A Ahmed and Talal M Alkhamis. *Simulation optimization for an emergency department healthcare unit in Kuwait.* European Journal of Operational Research, vol. 198, no. 3, pages 936–942, 2009. (Cited on pages 23 and 159.)

[Aiken 2002] Linda H Aiken, Sean P Clarke, Douglas M Sloane, Julie Sochalski and Jeffrey H Silber. *Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction.* JAMA: the journal of the American Medical Association, vol. 288, no. 16, pages 1987–1993, 2002. (Cited on page 23.)

[Altman 2002] Eitan Altman. *Applications of Markov decision processes in communication networks.* In Handbook of Markov decision processes, pages 489–536. Springer, 2002. (Cited on page 104.)

[Asaduzzaman 2010] Md Asaduzzaman, Thierry J Chaussalet and Nicola J Robertson. *A loss network model with overflow for capacity planning of a neonatal unit.* Annals of Operations Research, vol. 178, no. 1, pages 67–76, 2010. (Cited on pages 19, 36, 37, 41, 56 and 175.)

[Attar 2006] MA Attar, K Hanrahan, SW Lang, MR Gates and SL Bratton. *Pregnant mothers out of the perinatal regionalization's reach.* Journal of perinatology, vol. 26, no. 4, pages 210–214, 2006. (Cited on page 18.)

[Augusto 2008] Vincent Augusto. *Modélisation, analyse et pilotage de flux en milieu hospitalier à l'aide d'UML et des réseaux de Petri.* PhD thesis, Ecole Nationale Supérieure des Mines de Saint-Etienne, 2008. (Cited on page 159.)

[Baray 2012] Jérôme Baray and Gérard Cliquet. *Optimizing locations through a maximum covering/p-median hierarchical model: Maternity hospitals in France.* Journal of Business Research, 2012. (Cited on pages vii, 12 and 13.)

[Beck 2010] Stacy Beck, Daniel Wojdyla, Lale Say, Ana Pilar Betran, Mario Merialdi, Jennifer Harris Requejo, Craig Rubens, Ramkumar Menon and Paul FA Van Look. *The worldwide incidence of preterm birth: a systematic review of maternal mortality and morbidity.* Bulletin of the World Health Organization, vol. 88, no. 1, pages 31–38, 2010. (Cited on page 175.)

[Bhargava 1999] Hemant K Bhargava, Suresh Sridhar and Craig Herrick. *Beyond spreadsheets: tools for building decision support systems.* Computer, vol. 32, no. 3, pages 31–39, 1999. (Cited on page 25.)

[Blasak 2003] Ruby E Blasak, Darrell W Starks, Wendy S Armel and Mary C Hayduk. *Healthcare process analysis: the use of simulation to evaluate hospital operations between the emergency department and a medical telemetry unit.*

In Proceedings of the 35th conference on Winter simulation: driving innovation, pages 1887–1893. Winter Simulation Conference, 2003. (Cited on page 159.)

[Blondel 2011] Béatrice Blondel, Nicolas Drewniak, Hugo Pilkington and Jennifer Zeitlin. *Out-of-hospital births and the supply of maternity units in France.* Health & place, vol. 17, no. 5, pages 1170–1173, 2011. (Cited on pages 31 and 34.)

[Brailsford 2007] Sally C Brailsford. *Advances and challenges in healthcare simulation modeling: tutorial.* In S.B. Henderson, B. Biller, M.H. Hsieh, J. Shortle, J.D. Tew and R.R. Barton, editeurs, Proceedings of the 2007 Winter Simulation Conference, pages 1436–1448, Piscataway, New Jersey, 2007. Institute of Electrical and Electronics Engineers (IEEE). (Cited on page 122.)

[Brandeau 2003] Margaret L Brandeau, Gregory S Zaric and Anke Richter. *Resource allocation for control of infectious diseases in multiple independent populations: beyond cost-effectiveness analysis.* Journal of health economics, vol. 22, no. 4, pages 575–598, 2003. (Cited on page 23.)

[Brandeau 2004] Margaret L Brandeau, François Sainfort and William P Pierskalla. Operations research and health care: a handbook of methods and applications, volume 70. Springer, 2004. (Cited on page 40.)

[Bravi 2012] F Bravi, D Gibertoni, A Marcon, C Sicotte, E Minvielle, P Rucci, A Angelastro, T Carradori and MP Fantini. *Hospital network performance: A survey of hospital stakeholders' perspectives.* Health policy, 2012. (Cited on page 1.)

[Bréart 2003] G Bréart, F Puech and JC Rozé. *Mission périnatalité: conclusions: 20 propositions pour une politique périnatale.* Paris: Ministére de la Santé, 2003. (Cited on page 16.)

[Bretthauer 1998] Kurt M Bretthauer and Murray J Cŏté. *A model for planning resource requirements in health care organizations.* Decision Sciences, vol. 29, no. 1, pages 243–270, 1998. (Cited on page 37.)

[Budge 2009] Susan Budge, Armann Ingolfsson and Erhan Erkut. *Technical Note-Approximating Vehicle Dispatch Probabilities for Emergency Service Systems with Location-Specific Service Times and Multiple Units per Location.* Operations Research, vol. 57, no. 1, pages 251–255, 2009. (Cited on page 71.)

[Carrizosa 1998] Emilio Carrizosa, E Conde and M Munoz-Marquez. *Admission policies in loss queueing models with heterogeneous arrivals.* Management Science, vol. 44, no. 3, pages 311–320, 1998. (Cited on page 105.)

[Chang 2003] K-H Chang and W-F Chen. *Admission control policies for two-stage tandem queues with no waiting spaces.* Computers & Operations Research, vol. 30, no. 4, pages 589–601, 2003. (Cited on page 105.)

[Charfeddine 2010] Moez Charfeddine and Benoit Montreuil. *Integrated agent-oriented modeling and simulation of population and healthcare delivery network: application to COPD chronic disease in a Canadian region.* In B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan and E. Yucesan, editeurs, Proceedings of the 2010 Winter Simulation Conference, pages 2327–2339, Piscataway, New Jersey, 2010. Institute of Electrical and Electronics Engineers (IEEE). (Cited on pages 27, 122 and 159.)

[Chevalier 2003] Philippe Chevalier and Nathalie Tabordon. *Overflow analysis and cross-trained servers.* International Journal of Production Economics, vol. 85, no. 1, pages 47–60, 2003. (Cited on pages 69, 70 and 71.)

[Cochran 2006] Jeffery K Cochran and Aseem Bharti. *Stochastic bed balancing of an obstetrics hospital.* Health care management science, vol. 9, no. 1, pages 31–45, 2006. (Cited on pages 36, 37 and 41.)

[Cunich 2011] Michelle Cunich, Glenn Salkeld, Jack Dowie, Joan Henderson, Clare Bayram, Helena Britt and Kirsten Howard. *Integrating Evidence and Individual Preferences Using a Web-Based Multi-Criteria Decision Analytic Tool.* The Patient: Patient-Centered Outcomes Research, vol. 4, no. 3, pages 153–162, 2011. (Cited on page 28.)

[Daskin 2005] Mark S Daskin and Latoya K Dean. *Location of health care facilities.* In Operations research and health care, pages 43–76. Springer, 2005. (Cited on pages 21 and 36.)

[Davis 1977] E Davis. *Optimal control of arrivals to a two-server queueing system with separate queues.* PhD thesis, PhD dissertation, Program in Operations Research, North Carolina State University, Raleigh, NC, 1977. (Cited on page 104.)

[De Angelis 2003] Vanda De Angelis, Giovanni Felici and Paolo Impelluso. *Integrating simulation and optimisation in health care centre management.* European Journal of Operational Research, vol. 150, no. 1, pages 101–114, 2003. (Cited on pages 24 and 159.)

[De Bruin 2010] AM De Bruin, René Bekker, Lillian Van Zanten and GM Koole. *Dimensioning hospital wards using the Erlang loss model.* Annals of Operations Research, vol. 178, no. 1, pages 23–43, 2010. (Cited on page 41.)

[Delasay 2012] Mohammad Delasay, Bora Kolfal and Armann Ingolfsson. *Maximizing throughput in finite-source parallel queue systems.* European Journal of Operational Research, vol. 217, no. 3, pages 554–559, 2012. (Cited on pages 105 and 123.)

[Dexter 2001] Franklin Dexter and Alex Macario. *Optimal number of beds and occupancy to minimize staffing costs in an obstetrical unit?* Canadian Journal of Anesthesia, vol. 48, pages 295–301, 2001. (Cited on page 39.)

[Earnshaw 2002] Stephanie R Earnshaw, Anke Richter, Stephen W Sorensen, Thomas J Hoerger, Katherine A Hicks, Michael Engelgau, Ted Thompson, KM Venkat Narayan, David F Williamson, Edward Gregg*et al.* *Optimal allocation of resources across four interventions for type 2 diabetes.* Medical Decision Making, vol. 22, no. suppl 1, pages s80–s91, 2002. (Cited on page 23.)

[EFC 2010] *European benchmarking report.* EFCNI EU European Foundation for the care of newborn infants, 2009–2010. (Cited on page 175.)

[Essoussi 2009] IE Essoussi and P Ladet. *Towards resource pooling in cooperative health care networks: Case of medical supply centralization.* In Computers & Industrial Engineering, 2009. CIE 2009. International Conference on, pages 600–605. IEEE, 2009. (Cited on page 22.)

[Ferreira 1999] Jorge A Ferreira and JP Estima de Oliveira. *Modelling hybrid systems using statecharts and Modelica.* In Emerging Technologies and Factory Automation, 1999. Proceedings. ETFA'99. 1999 7th IEEE International Conference on, volume 2, pages 1063–1069. IEEE, 1999. (Cited on page 159.)

[Flessa 2000] Steffen Flessa. *Where efficiency saves lives: A linear programme for the optimal allocation of health care resources in developing countries.* Health Care Management Science, vol. 3, no. 3, pages 249–267, 2000. (Cited on pages 23 and 36.)

[Franx 2006] Geert Jan Franx, Ger Koole and Auke Pot. *Approximating multi-skill blocking systems by hyperexponential decomposition.* Performance Evaluation, vol. 63, no. 8, pages 799–824, 2006. (Cited on pages 70 and 71.)

[Fredericks 1980] AA Fredericks. *Congestion in blocking systems-a simple approximation technique.* BSTJ, vol. 59, no. 6, pages 805–827, 1980. (Cited on pages 69 and 71.)

[Galvão 2006] Roberto D Galvão, Luis Gonzalo Acosta Espejo and Brian Boffey. *Practical aspects associated with location planning for maternal and perinatal assistance in Brazil.* Annals of Operations Research, vol. 143, no. 1, pages 31–44, 2006. (Cited on page 36.)

[Ghoneim 1985] Hussein A Ghoneim and Shaler Stidham Jr. *Control of arrivals to two queues in series.* European Journal of Operational Research, vol. 21, no. 3, pages 399–409, 1985. (Cited on page 104.)

[Gorunescu 2002] Florin Gorunescu, Sally I McClean and Peter H Millard. *A queueing model for bed-occupancy management and planning of hospitals.* Journal

of the Operational Research Society, pages 19–24, 2002. (Cited on pages 36, 37 and 56.)

[Govind 2008] Rahul Govind, Rabikar Chatterjee and Vikas Mittal. *Timely access to health care: Customer-focused resource allocation in a hospital network*. International Journal of Research in Marketing, vol. 25, no. 4, pages 294–300, 2008. (Cited on pages 23 and 36.)

[Green 2002] Linda V Green. *How many hospital beds?* Journal Information, vol. 39, no. 4, 2002. (Cited on page 23.)

[Grenier 2004] A Grenier. *Exercice groupé et réseaux de santé. La pratique de la médecine libérale sort de l isolement*. Actualité et Dossier en Santé Publique, vol. 176, page 49, 2004. (Cited on page 9.)

[Griffin 2008] Paul M Griffin, Christina R Scherrer and Julie L Swann. *Optimization of community health center locations and service offerings with statistical need estimation*. IIE Transactions, vol. 40, no. 9, pages 880–892, 2008. (Cited on pages 21 and 36.)

[Griffin 2012] Jacqueline Griffin, Shuangjun Xia, Siyang Peng and Pinar Keskinocak. *Improving patient flow in an obstetric unit*. Health care management science, vol. 15, no. 1, pages 1–14, 2012. (Cited on page 159.)

[Gulliford 2002] Martin Gulliford, Jose Figueroa-Munoz, Myfanwy Morgan, David Hughes, Barry Gibson, Roger Beech and Meryl Hudson. *What does access to health care mean?* Journal of health services research & policy, vol. 7, no. 3, pages 186–188, 2002. (Cited on page 18.)

[Günes 2009] ED Günes and Hande Yaman. *Health network mergers and hospital re-planning*. Journal of the Operational Research Society, vol. 61, no. 2, pages 275–283, 2009. (Cited on page 36.)

[Hajek 1984] Bruce Hajek. *Optimal control of two interacting service stations*. Automatic Control, IEEE Transactions on, vol. 29, no. 6, pages 491–499, 1984. (Cited on page 104.)

[Hakimi 1964] S Louis Hakimi. *Optimum locations of switching centers and the absolute centers and medians of a graph*. Operations Research, vol. 12, no. 3, pages 450–459, 1964. (Cited on page 21.)

[Hariharan 1990] Rema Hariharan, VG Kulkarni and S Stidham Jr. *Optimal control of two parallel infinite-server queues*. In Decision and Control, 1990., Proceedings of the 29th IEEE Conference on, pages 1329–1335. IEEE, 1990. (Cited on page 104.)

[Harper 2002] PR Harper and AK Shahani. *Modelling for the planning and management of bed capacities in hospitals*. Journal of the Operational Research Society, pages 11–18, 2002. (Cited on pages 34, 36 and 37.)

[Harrison 1975] J Michael Harrison. *Dynamic scheduling of a multiclass queue: Discount optimality*. Operations Research, vol. 23, no. 2, pages 270–282, 1975. (Cited on page 104.)

[Hordijk 1992] Arie Hordijk and Ger Koole. *On the shortest queue policy for the tandem parallel queue*. Probability in the Engineering and Informational Sciences, vol. 6, no. 01, pages 63–79, 1992. (Cited on page 105.)

[Horev 2004] Tuvia Horev, Irena Pesis-Katz and Dana B Mukamel. *Trends in geographic disparities in allocation of health care resources in the US*. Health policy, vol. 68, no. 2, pages 223–232, 2004. (Cited on page 23.)

[Huang 2008] Qian Huang, King-Tim Ko and Villy Bcek Iversen. *Approximation of loss calculation for hierarchical networks with multiservice overflows*. Communications, IEEE Transactions on, vol. 56, no. 3, pages 466–473, 2008. (Cited on pages 70 and 71.)

[IGAS 2013] IGAS. *Li¿½Hôpital : rapport 2012*, 2013. ISBN : 978-2-11-009064-5. (Cited on pages 31 and 33.)

[Jagerman 1974] David L Jagerman. *Some properties of the Erlang loss function*. Bell System Tech. J, vol. 53, no. 3, pages 525–551, 1974. (Cited on pages 51, 52 and 53.)

[Jagers 1986] AA Jagers and Erik A Van Doorn. *On the continued Erlang loss function*. Operations Research Letters, vol. 5, no. 1, pages 43–46, 1986. (Cited on pages 52 and 53.)

[Jarvis 1985] JP Jarvis. *Approximating the equilibrium behavior of multi-server loss systems*. Management Science, vol. 31, no. 2, pages 235–239, 1985. (Cited on page 71.)

[Javitt 1994] Jonathan C Javitt, Lloyd Paul Aiello, Yenpin Chiang, Frederick L Ferris, Joseph K Canner and Sheldon Greenfield. *Preventive eye care in people with diabetes is cost-saving to the federal government: implications for health-care reform*. Diabetes care, vol. 17, no. 8, pages 909–917, 1994. (Cited on page 28.)

[Jun 1999] JB Jun, SH Jacobson, JR Swisher*et al.* *Application of discrete-event simulation in health care clinics: a survey*. Journal of the operational research society, vol. 50, no. 2, pages 109–123, 1999. (Cited on page 159.)

[Kao 1981] Edward PC Kao and Grace G Tung. *Bed allocation in a public health care delivery system*. Management Science, vol. 27, no. 5, pages 507–520, 1981. (Cited on page 37.)

[Kelton 2000] W David Kelton and Averill M Law. Simulation modeling and analysis. McGraw Hill Boston, MA, 2000. (Cited on page 159.)

[Kim 1999] Seung-Chul Kim, Ira Horowitz, Karl K Young and Thomas A Buckley. *Analysis of capacity management of the intensive care unit in a hospital*. European Journal of Operational Research, vol. 115, no. 1, pages 36–46, 1999. (Cited on pages 36 and 37.)

[Koole 2000] Ger Koole. *Stochastic scheduling with event-based dynamic programming*. Mathematical Methods of Operations Research, vol. 51, no. 2, pages 249–261, 2000. (Cited on page 108.)

[Ku 1997] Cheng-Yuan Ku and Scott Jordan. *Access control to two multiserver loss queues in series*. Automatic Control, IEEE Transactions on, vol. 42, no. 7, pages 1017–1023, 1997. (Cited on page 105.)

[Ku 2002] Cheng-Yuan Ku and Scott Jordan. *Access control of parallel multiserver loss queues*. Performance Evaluation, vol. 50, no. 4, pages 219–231, 2002. (Cited on pages 105 and 123.)

[Ku 2006] Cheng-Yuan Ku, David C Yen, I Chang, Shi-Ming Huang, Scott Jordan*et al.* *Near-optimal control policy for loss networks*. Omega, vol. 34, no. 4, pages 406–416, 2006. (Cited on page 105.)

[Kuczura 1973] Anatol Kuczura. *The interrupted Poisson process as an overflow process*. Bell System Technical Journal, 1973. (Cited on pages 74 and 75.)

[Kwak 1997] NK Kwak and Changwon Lee. *A linear goal programming model for human resource allocation in a health-care organization*. Journal of Medical Systems, vol. 21, no. 3, pages 129–140, 1997. (Cited on page 23.)

[Larson 1974] Richard C Larson. *A hypercube queuing model for facility location and redistricting in urban emergency services*. Computers & Operations Research, vol. 1, no. 1, pages 67–95, 1974. (Cited on page 70.)

[Lavieri 2009] Mariel S Lavieri and Martin L Puterman. *Optimizing nursing human resource planning in British Columbia*. Health care management science, vol. 12, no. 2, pages 119–128, 2009. (Cited on page 23.)

[Lee 2013] Eva K Lee, Ferdinand Pietz, Bernard Benecke, Jacquelyn Mason and Greg Burel. *Advancing Public Health and Medical Preparedness with Operations Research*. Interfaces, vol. 43, no. 1, pages 79–98, 2013. (Cited on page 26.)

[Lerzan Örmeci 2001] E Lerzan Örmeci, Apostolos Burnetas and Jan van der Wal. *Admission policies for a two class loss system*. 2001. (Cited on pages 102 and 105.)

[Li 2008] Xinmei Li, Patrick Beullens, Dylan Jones and Mehrdad Tamiz. *An integrated queuing and multi-objective bed allocation model with application to a hospital in China*. Journal of the Operational Research Society, vol. 60, no. 3, pages 330–338, 2008. (Cited on pages 36, 37 and 56.)

[Lippman 1971] Steven A Lippman and Sheldon M Ross. *The streetwalker's dilemma: A job shop model.* SIAM Journal on Applied Mathematics, vol. 20, no. 3, pages 336–342, 1971. (Cited on page 104.)

[Lippman 1975] Steven A Lippman. *Applying a new device in the optimization of exponential queuing systems.* Operations Research, vol. 23, no. 4, pages 687–710, 1975. (Cited on page 108.)

[Mahachek 1984] Arnold R Mahachek and Terry L Knabe. *Computer simulation of patient flow in obstetrical/gynecology clinics.* Simulation, vol. 43, no. 2, pages 95–101, 1984. (Cited on page 159.)

[Mahar 2011] Stephen Mahar, Kurt M Bretthauer and Peter A Salzarulo. *Locating specialized service capacity in a multi-hospital network.* European Journal of Operational Research, vol. 212, no. 3, pages 596–605, 2011. (Cited on pages 24 and 36.)

[Miller 1969] Bruce L Miller. *A queueing reward system with several customer classes.* Management Science, vol. 16, no. 3, pages 234–245, 1969. (Cited on page 104.)

[Miller 2009] Martin Miller, David Ferrin and Niloo Shahi. *Estimating patient surge impact on boarding time in several regional emergency departments.* In M.D. Rossetti, R.R. Hill, B. Johansson, A. Dunkin and R.G. Ingalls, editeurs, Proceedings of the 2009 Winter Simulation Conference, pages 1906–1915, Piscataway, New Jersey, 2009. Institute of Electrical and Electronics Engineers (IEEE). (Cited on page 122.)

[Montreuil 2011] Benoit Montreuil, Moez Charfeddine, Vincent Augusto, Caroline Cloutier and Christelle Montreuil. *Live, Open, Voluntary and Collaborative Mapping of Healthcare Delivery Networks.* In INFORMS Healthcare conference, MontrÃ©al, QuÃ©bec, Canada, 2011. (Cited on page 27.)

[Ndiaye 2008] Malick Ndiaye and Hesham Alfares. *Modeling health care facility location for moving population groups.* Computers & Operations Research, vol. 35, no. 7, pages 2154–2161, 2008. (Cited on page 22.)

[Neal 1971] Scotty Neal. *Combining correlated streams of nonrandom traffic.* Bell System Technical Journal, vol. 50, no. 6, pages 2015–2037, 1971. (Cited on pages 70, 72 and 77.)

[Nesbitt 1997] Thomas S Nesbitt, Eric H Larson, Roger A Rosenblatt and L Gary Hart. *Access to maternity care in rural Washington: its effect on neonatal outcomes and resource use.* American Journal of Public Health, vol. 87, no. 1, pages 85–90, 1997. (Cited on pages 31 and 34.)

[Nguyen 2005] JM Nguyen, Patrick Six, D Antonioli, Pascal Glemain, Gilles Potel, Pierre Lombrail and Pierre Le Beux. *A simple method to optimize hospital*

beds capacity. International journal of medical informatics, vol. 74, no. 1, pages 39–49, 2005. (Cited on page 34.)

[Oddoye 2009] John Paul Oddoye, Dylan F Jones, Mehrdad Tamiz and P Schmidt. *Combining simulation and goal programming for healthcare planning in a medical assessment unit.* European Journal of Operational Research, vol. 193, no. 1, pages 250–261, 2009. (Cited on page 159.)

[Örmeci 2006] E Lerzan Örmeci and Jan van der Wal. *Admission policies for a two class loss system with general interarrival times.* Stochastic models, vol. 22, no. 1, pages 37–53, 2006. (Cited on page 105.)

[Pehlivan 2012] Canan Pehlivan, Vincent Augusto, Xiaolan Xie and Catherine Crenn-Hebert. *Multi-period capacity planning for maternity facilities in a perinatal network: A queuing and optimization approach.* In Automation Science and Engineering (CASE), 2012 IEEE International Conference on, pages 137–142. IEEE, 2012. (Cited on page 34.)

[Pierskalla 1994] William P Pierskalla and David J Brailer. *Applications of operations research in health care delivery.* Handbooks in operations research and management science, vol. 6, pages 469–505, 1994. (Cited on page 21.)

[Pilkington 2008] Hugo Pilkington, Béatrice Blondel, Marion Carayol, Gérard Breart and Jennifer Zeitlin. *Impact of maternity unit closures on access to obstetrical care: the French experience between 1998 and 2003.* Social science & medicine, vol. 67, no. 10, pages 1521–1529, 2008. (Cited on pages vii, 16, 18, 31 and 33.)

[ReVelle 1986] Charles ReVelle. *The Maximum Capture or "Sphere of Influence" Location Problem: Hotelling Revisited on a Network.* Journal of Regional Science, vol. 26, no. 2, pages 343–358, 1986. (Cited on page 21.)

[Ridge 1998] JC Ridge, SK Jones, MS Nielsen and AK Shahani. *Capacity planning for intensive care units.* European journal of operational research, vol. 105, no. 2, pages 346–355, 1998. (Cited on pages 36 and 37.)

[Roberts 2011] Stephen D Roberts. *Tutorial on the simulation of healthcare systems.* In S. Jain, R. R. Creasey, J. Himmelspach, K. P. White and M. Fu, editeurs, Proceedings of the 2011 Winter Simulation Conference, pages 1408–1419, Piscataway, New Jersey, 2011. Institute of Electrical and Electronics Engineers (IEEE). (Cited on page 122.)

[Rodriguez-Verjan 2012] Carlos Rodriguez-Verjan, Vincent Augusto, Thierry Garaix, Xiaolan Xie and Valerie Buthion. *Healthcare at home facility location-allocation problem.* In Automation Science and Engineering (CASE), 2012 IEEE International Conference on, pages 150–155. IEEE, 2012. (Cited on page 22.)

[Roth 2009] Loren H Roth, Kathleen Criss, Xavier Stewart and Kevin McCann. *PrepLink: a novel web-based tool for healthcare emergency planning and response*. Biosecurity and Bioterrorism, vol. 7, no. 1, pages 85–92, 2009. (Cited on page 26.)

[Royston 2009] Geoff Royston. *One hundred years of Operational Research in Health*. Journal of the Operational Research Society, pages S169–S179, 2009. (Cited on page 17.)

[Şahin 2007] Güvenç Şahin and Haldun Süral. *A review of hierarchical facility location models*. Computers & Operations Research, vol. 34, no. 8, pages 2310–2331, 2007. (Cited on page 21.)

[Santibáñez 2009] Pablo Santibáñez, Georgia Bekiou and Kenneth Yip. *Fraser Health uses mathematical programming to plan its inpatient hospital network*. Interfaces, vol. 39, no. 3, pages 196–208, 2009. (Cited on pages 23 and 36.)

[Schilling 1993] David A Schilling, Vaidyanathan Jayaraman and Reza Barkhi. *A REVIEW OF COVERING PROBLEMS IN FACILITY LOCATION*. Computers & Operations Research, 1993. (Cited on page 21.)

[Schuurman 2008] Nadine Schuurman, Margo Leight and Myriam Berube. *International Journal of Health Geographics*. International journal of health geographics, vol. 7, page 49, 2008. (Cited on page 27.)

[Sheu 2010] Ru-Shuo Sheu and Ilze Ziedins. *Asymptotically optimal control of parallel tandem queues with loss*. Queueing Systems, vol. 65, no. 3, pages 211–227, 2010. (Cited on page 106.)

[Sinreich 2005] David Sinreich and Yariv Marmor. *Emergency department operations: the basis for developing a simulation tool*. IIE Transactions, vol. 37, no. 3, pages 233–245, 2005. (Cited on page 159.)

[Stein 2006] Mike Stein. *The Map of Medicine®-an Innovative Knowledge Management Tool*. In AMIA Annual Symposium Proceedings, volume 2006, page 1196. American Medical Informatics Association, 2006. (Cited on page 28.)

[Stein 2012] Mart Stein, James Rudge, Richard Coker, Charlie van der Weijden, Ralf Krumkamp, Piya Hanvoravongchai, Irwin Chavez, Weerasak Putthasri, Bounlay Phommasack, Wiku Adisasmito*et al.* *Development of a resource modelling tool to support decision makers in pandemic influenza preparedness: The AsiaFluCap Simulator*. BMC public health, vol. 12, no. 1, page 870, 2012. (Cited on page 26.)

[Stidham Jr 1993] Shaler Stidham Jr and Richard Weber. *A survey of Markov decision models for control of networks of queues*. Queueing Systems, vol. 13, no. 1-3, pages 291–314, 1993. (Cited on page 105.)

[Stummer 2004] Christian Stummer, Karl Doerner, Axel Focke and Kurt Heidenberger. *Determining location and size of medical departments in a hospital network: A multiobjective decision support approach.* Health Care Management Science, vol. 7, no. 1, pages 63–71, 2004. (Cited on page 36.)

[Swisher 2001] James R Swisher, Sheldon H Jacobson, J Brian Jun and Osman Balci. *Modeling and analyzing a physician clinic environment using discrete-event (visual) simulation.* Computers & operations research, vol. 28, no. 2, pages 105–125, 2001. (Cited on page 159.)

[Syam 2010] Siddhartha S Syam and Murray J Côté. *A location-allocation model for service providers with application to not-for-profit health care organizations.* Omega, vol. 38, no. 3, pages 157–166, 2010. (Cited on pages 21, 36 and 159.)

[Tanonkou 2008] G-A Tanonkou, Lyès Benyoucef and Xiaolan Xie. *Design of stochastic distribution networks using Lagrangian relaxation.* Automation Science and Engineering, IEEE Transactions on, vol. 5, no. 4, pages 597–608, 2008. (Cited on page 36.)

[Vallancien 2006] Guy Vallancien. *L'évaluation de la sécurité, de la qualité et de la continuité des soins chirurgicaux dans les petits hôpitaux publics en France.* Commissioned report of the French Ministry of Health, 2006. (Cited on pages 31 and 33.)

[Verter 2002] Vedat Verter and Sophie D Lapierre. *Location of preventive health care facilities.* Annals of Operations Research, vol. 110, no. 1-4, pages 123–132, 2002. (Cited on page 36.)

[Wilkinson 1956] Roger I Wilkinson. *Theories for Toll Traffic Engineering in the USA.* Bell System Technical Journal, vol. 35, no. 2, pages 421–514, 1956. (Cited on page 69.)

[Wong 2007] Eric WM Wong, Andrew Zalesky, Zvi Rosberg and Moshe Zukerman. *A new method for approximating blocking probability in overflow loss networks.* Computer Networks, vol. 51, no. 11, pages 2958–2975, 2007. (Cited on page 69.)

[www.ars.fr] www.ars.fr. (Cited on page 10.)

[www.sante.gouv.fr] www.sante.gouv.fr. (Cited on page 9.)

[Zeitlin 2004] Jennifer Zeitlin, Emile Papiernik and Gérard Bréart. *Regionalization of perinatal care in Europe.* In Seminars in Neonatology, volume 9, pages 99–110. Elsevier, 2004. (Cited on page 18.)

[Zhang 2010] Yue Zhang, Oded Berman, Patrice Marcotte and Vedat Verter. *A bilevel model for preventive healthcare facility network design with congestion.* IIE Transactions, vol. 42, no. 12, pages 865–880, 2010. (Cited on page 22.)

[Zhang 2013] Bo Zhang and Hayriye Ayhan. *Optimal admission control for tandem queues with loss.* Automatic Control, IEEE Transactions on, vol. 58, no. 1, pages 163–167, 2013. (Cited on page 106.)

École Nationale Supérieure des Mines
de Saint-Étienne

NNT : 2014 EMSE 0731

Canan PEHLIVAN

DESIGN AND FLOW CONTROL OF STOCHASTIC HEALTHCARE NETWORKS WITHOUT WAITING ROOMS: A PERINATAL APPLICATION

Speciality : Industrial Engineering

Keywords : Stochastic Healthcare Network, Overflow, Rejection, Perinatal Network, Performance Evaluation, Admission Control Policies, Location and Capacity Planning

Abstract :

In this thesis, by being motivated from the challenges in perinatal networks, we address design, evaluation and flow control of a stochastic healthcare network where there exist multiple levels of hospitals and different types of patients. Patients are supposed urgent; thus they can be rejected and overflow to another facility in the same network if no service capacity is available at their arrival. Rejection of patients due to the lack of service capacity is the common phenomenon in overflow networks. We approach the problem from both strategic and operational perspectives. In strategic part, we address a location & capacity planning problem for adjusting the network to better meet demographic changes. In operational part, we study the optimal patient admission control policies to increase flexibility in allocation of resources and improve the control of patient flow in the network. Finally, in order to evaluate the performance of the network, we develop new approximation methodologies that estimate the rejection probabilities in each hospital for each arriving patient group, thus the overflow probabilities among hospitals. Furthermore, an agent-based discrete-event simulation model is constructed to adequately represent our main application area: Nord Hauts-de-Seine Perinatal Network. The simulation model is used to evaluate the performance of the complex network and more importantly evaluate the strength of the optimal results of our analytical models. The developed methodologies in this thesis are combined in a decision support tool, foreseen under the project “COVER”, which aims to assist health system managers to effectively plan strategic and operational decisions of a healthcare network and evaluate the performance of their decisions.

École Nationale Supérieure des Mines
de Saint-Étienne

NNT : 2014 EMSE 0731

Canan PEHLIVAN

CONCEPTION ET PILOTAGE DE FLUX D'UN RESEAU DE SOINS STOCHASTIQUE SANS ATTENTE: APPLICATION A LA PERINATALITE

Spécialité: Génie Industriel

Mots clefs : réseau de soins stochastique, probabilité de transfert, probabilité de rejet, réseaux de périnatalité, planification de la capacité et localisation, l'évaluation de performance, politiques de pilotage d'admission optimales, simulation multi-agents et événements discrets

Résumé :

Cette thèse porte sur l'étude d'un réseau de soins hiérarchique stochastique avec rejet où les patients sont transférés lorsque la capacité de l'hôpital d'accueil n'est pas suffisante. Les patients sont alors redirigés vers un autre hôpital, ou hors du réseau. Une application concrète sur les réseaux de périnatalité est proposée, et nous avons identifié plusieurs verrous scientifiques fondamentaux d'un point de vue stratégique et opérationnel. Dans la partie stratégique, nous nous sommes intéressés à un problème de planification de capacité dans le réseau. Nous avons développé un modèle de localisation et de dimensionnement non-linéaire qui tient compte de la nature stochastique du système. La linéarisation du modèle permet de résoudre des problèmes de taille réelle en temps raisonnable. Nous avons développé dans un second temps de nouvelles méthodologies d'approximation permettant d'évaluer la performance du réseau en termes de probabilité de rejet et de transfert entre hôpitaux. Dans la partie opérationnelle, nous avons étudié des politiques de pilotage d'admission optimales pour différentes tailles de réseaux de manière utiliser au mieux les ressources hospitalières. Finalement, nous avons construit un modèle de simulation couplant multi-agents et événements discrets permettant la validation des résultats précédents et l'évaluation de performance du système de manière réaliste.