Bayes Business School

Doctoral Thesis

Reconfiguration of Inpatient Services to Reduce Bed Pressure in Hospitals

A thesis submitted in fulfilment of the requirements

for the degree of Doctor of Philosophy

in the

Operations Management

Department of Management

September 22, 2022

I, Bahar Arabzadeh, declare that this thesis titled, "Reconfiguration of Inpatient Services to Reduce Bed Pressure in Hospitals" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this Bayes Business School, City University of London.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Signed:*Bahar Arabzadeh*

Date: 16/02/2022

"*Yesterday I was clever, so I wanted to change the world. today I am wise, so I am changing myself.*"

Rumi

BAYES BUSINESS SCHOOL

Abstract

Operations Management

Department of Management

Doctor of Philosophy

Reconfiguration of Inpatient Services to Reduce Bed Pressure in Hospitals

by Bahar Arabzadeh

Healthcare systems around the world are facing an inpatient bed crisis. This crises has been highlighted, more than ever before, during the recent Covid-19 pandemic. Our aim in this doctoral dissertation is to propose a cost-effective solution to the ongoing bed-crisis with a focus on reconfiguration of inpatient services. The configuration of inpatient services, which identifies the set of specialties and bed numbers allocated to each ward, has a substantial impact on performance, which we measure by the cost of patients waiting for services or abandoning the services plus the cost of nursing teams. Reviewing the existing configurations proposed in the literature, we choose the clustered overflow configuration as the basis for our study due to its versatility. Given a set of specialties, a total number of beds, and a (potentially infinite) waiting time threshold for patients, we then propose a heuristic methodology for finding a good allocation of beds and specialties for this configuration. This methodology relies on a novel performance evaluation model for overflow delays systems, i.e., hierarchical queueing systems involving several dedicated pools and a single overflow pool. We illustrate the application of our methodology by applying it on a comprehensive inpatient dataset obtained from a UK hospital. A simulation study shows substantial savings can potentially be made by using the configurations proposed by our methodology as compared to the existing configuration of the hospital or other major configurations proposed in the literature.

Acknowledgements

Throughout the writing of this dissertation I have received a great deal of support and assistance. I would first like to thank my supervisor, Dr. Navid Izady, whose expertise was invaluable in the formulating of the research topic and methodology in particular. Dr Izady has been a great teacher, mentor and supervisor, I have learnt a great deal throughout my PhD from him, things that I will always cherish and carry in my personal and professional life with me.

I would like to extend my gratitude to the Director of Transformation of our partner hospital, Mr. Nick Sands and the Clinical Lead for Transformation, Dr. James Adams for their time, provision of access to the data and for their specialised input into the day to day operations of the hospital under study. I would also like to thank Dr. Ramin Ajami Head of Oncology and Medical Services in Guy's and St Thomas' hospital for providing expert consultations during different phases of the research project.

Countless people have supported me professionally and in my personal life through out these years. My PhD colleagues who have helped me find my footing when I started the PhD journey and majority of them are now Drs themselves, Dr. Matilde Lucheschi, Dr. Mislav Radic, Dr. Alessandro Tirapani, Dr. Alex Isakson, Dr. Parastoo Mousavi, Dr. Edwin Lee, Mrs. Saha Malaki, Mr. Adi Ghosh, Mrs. Fabienne Chedid and Mr. Salar Nikandish and many others which time and space limits me to name. I would also like to thank two of my very special friends, Shokouh and Sepideh who were of great support in deliberating over my problems, as well as providing happy distraction to rest my mind outside of my research.

I left the longest yet most important paragraph to dedicate to three individuals whom without them none of this would have been possible. My parents, Haleh and Farid and my partner Pedram. I start with Mum and Dad. My mother for teaching me how to be resilient, invincible and independent. My Father for teaching me to appreciate knowledge, music and history. Mum, your never ending sacrifices have got me where I am in life today and I am eternally thankful for that. Dad, your wisdom your calm, cool exterior when facing a crises has been the ultimate backbone for me, I am where I am for your compromises and sleepless nights. To my beloved Pedram, who has been nothing but a loving, encouraging, supportive partner that anyone could

ask for. I would like to thank him for challenging me intellectually and teaching me new things on a daily basis. Pedram's never ending support for me and for my work is unfathomable and I am endlessly thankful for that.

Contents

Declaration of Authorship ii

Abstract v

Acknowledgements vii

List of Figures xiii

List of Tables xvii

List of Abbreviations xix

1 Introduction 1

1.1 Motivation 1

1.2 Configurations of Inpatient Services 3

1.2.1 Fully Dedicated Configuration 4

1.2.2 Fully Flexible Configuration 5

1.2.3 Wing Formation Configuration 5

1.2.4 Earmarking Configuration 6

1.2.5 Clustered Overflow Configuration 7

1.3 Thesis Structure 9

2 Literature Review 11

2.1 Introduction 11

2.2 Inpatient Bed Allocation 11

2.3 Resource Flexibility 15

2.4 Performance Evaluation 16

2.5 Research Questions 18

3 Problem Formulation **21**

3.1 Introduction . . . 21

3.2 Mathematical Representation . . . 21

3.3 Optimization Formulation . . . 22

3.4 Conclusion . . . 26

4 The Performance Evaluation Models **27**

4.1 Introduction . . . 27

4.2 Exact Analysis . . . 28

4.3 Approximation Models . . . 33

4.3.1 Overflow Loss Systems . . . 35

4.3.2 Overflow Delay Systems . . . 37

4.3.3 Overflow Abandonment Systems . . . 39

4.4 Numerical Experiment . . . 41

4.5 Conclusion . . . 53

5 Solving Optimization Models **55**

5.1 Introduction . . . 55

5.2 The Intra-Cluster Bed Allocation Model . . . 55

5.3 The Partitioning and Inter-Cluster Bed Allocation Model . . . 57

5.4 Conclusions . . . 58

6 Case Study **61**

6.1 Introduction . . . 61

6.2 Current Status . . . 61

6.3 Applying Reconfiguration Models . . . 71

6.3.1 Parameter Estimation . . . 71

6.3.2 Best-Found Configurations with IWTT Assumption . . . 74

6.3.3 Best-Found Configurations with FWTT Assumption . . . 78

6.3.4 Sensitivity Analysis . . . 83

6.4 Conclusions . . . 86

7 Conclusions and Future Research **89**

7.1 Introduction . . . 89

7.2 Contributions 93
7.3 Future Research 94
7.3.1 Incorporating Size and Skill-Set of the Existing Nursing Teams . 95
7.3.2 Considering Different Waiting Time Thresholds 96
7.3.3 Considering Non-Exponential Inter-arrival Times and LOSs . . . 96

List of Figures

1.1 The spectrum of inpatient bed configurations 4
1.2 An example of the fully dedicated configuration 5
1.3 An example of the fully flexible configuration 5
1.4 An example of the wing formation configuration 6
1.5 An example of the earmarking configuration 7
1.6 An example of the clustered overflow configuration 7

2.1 Different categories of performance evaluation models 17

3.1 An example for the COF configuration 22

4.1 The schematic diagram of an overflow system 28
4.2 Delay probabilities obtained from the approximation vs simulation for: $\rho_1 = \rho_2 = 1$ (a , b); $\rho_1 = \rho_2 = 0.9$ (c , d); and $\rho_1 = \rho_2 = 0.8$ (e , f) in a logarithmic scale. 42
4.3 Mean waiting times obtained from the approximation vs simulation for: $\rho_1 = \rho_2 = 1$ (a , b); $\rho_1 = \rho_2 = 0.9$ (c , d); and $\rho_1 = \rho_2 = 0.8$ (e , f) in a logarithmic scale. 43
4.4 Mean dedicated busy severs obtained from the approximation vs simulation for: $\rho_1 = \rho_2 = 1$ (a , b); $\rho_1 = \rho_2 = 0.9$ (c , d); and $\rho_1 = \rho_2 = 0.8$ (e , f) in a logarithmic scale. 44
4.5 Mean overflow busy severs obtained from the approximation vs simulation for: $\rho_1 = \rho_2 = 1$ (a , b); $\rho_1 = \rho_2 = 0.9$ (c , d); and $\rho_1 = \rho_2 = 0.8$ (e , f) in a logarithmic scale. 45
4.6 Abandonment probabilities obtained from the approximation vs simulation for $\rho_1 = \rho_2 = 1$: $\gamma^{-1} = 5$ (a , b); $\gamma^{-1} = 10$ (c , d); and $\gamma^{-1} = 30$ (e , f) in a logarithmic scale. 47

4.7 Abandonment probabilities obtained from the approximation vs simulation for $\rho_1 = \rho_2 = 0.8$: $\gamma^{-1} = 5$ (a , b); $\gamma^{-1} = 10$ (c , d); and $\gamma^{-1} = 30$ (e , f) in a logarithmic scale. 48
4.8 Mean dedicated busy severs obtained from the approximation vs simulation for $\rho_1 = \rho_2 = 1$: $\gamma^{-1} = 5$ (a , b); $\gamma^{-1} = 10$ (c , d); and $\gamma^{-1} = 30$ (e , f) in a logarithmic scale. 49
4.9 Mean dedicated busy severs obtained from the approximation vs simulation for $\rho_1 = \rho_2 = 0.8$: $\gamma^{-1} = 5$ (a , b); $\gamma^{-1} = 10$ (c , d); and $\gamma^{-1} = 30$ (e , f) in a logarithmic scale. 50
4.10 Mean overflow busy severs obtained from the approximation vs simulation for $\rho_1 = \rho_2 = 1$: $\gamma^{-1} = 5$ (a , b); $\gamma^{-1} = 10$ (c , d); and $\gamma^{-1} = 30$ (e , f) in a logarithmic scale. 51
4.11 Mean overflow busy severs obtained from the approximation vs simulation for $\rho_1 = \rho_2 = 0.8$: $\gamma^{-1} = 5$ (a , b); $\gamma^{-1} = 10$ (c , d); and $\gamma^{-1} = 30$ (e , f) in a logarithmic scale. 52

5.1 Surface plots for the total mean daily cost of a 70-bed cluster with 2 specialties assuming $f_1 = 0.149$ and $f_2 = 0.149$. Panel (a) is for an IWTT system with $\lambda_1 = 5.2$, $\lambda_2 = 3.9$, $\nu_1 = 4.8$, $\nu_2 = 2.9$, and $c_w = 1030$, and panel (b) is for an FWTT system with $\lambda_1 = 6.2$, $\lambda_2 = 4.9$, $\nu_1 = 4.8$, $\nu_2 = 2.9$, $\gamma^{-1} = 5$ and $c_a = 1030$. 56

6.1 Daily occupancy levels for medical and surgical wards over the period 01/10/2017-01/10/2018 at RSCH. 63
6.2 Daily occupancy levels for EAU and escalation wards over the period 01/10/2017-01/10/2018 at RSCH. 64
6.3 The breakdown of a hospital spell 67
6.4 Percentage of outlying specialty spells for each specialty. 68
6.5 Contribution to workload by outlying patients for each ward 68
6.6 Average and maximum number of wards visited for different specialities at RSCH. 70
6.7 Mean LOS as a function of age band and number of ward changes for GAS specialty 70

6.8 The best-found configuration under IWTT assumption with $\Delta = 0.0$ . . 75
6.9 The best-found configuration under IWTT assumption with $\Delta = 0.1$ . . 76
6.10 The best-found configuration under FWTT with $\gamma^{-1} = 1$ day and $\Delta = 0.0$. 78
6.11 The best-found configuration under FWTT with $\gamma^{-1} = 1$ day and $\Delta =$ 0.1. 79
6.12 The best-found configuration under FWTT with $\gamma^{-1} = 15$ days and $\Delta = 0.0$. 80
6.13 The best-found configuration under FWTT with $\gamma^{-1} = 15$ days and $\Delta = 0.1$. 80
6.14 The best-found configuration under FWTT with $\gamma^{-1} = 30$ days and $\Delta = 0.0$. 82
6.15 The best-found configuration under FWTT with $\gamma^{-1} = 30$ days and $\Delta = 0.1$. 82
6.16 Sensitivity analysis on impact of bed numbers and traffic intensity on clustering of medical specialties under IWTT assumption for $\Delta = 0.0$ (a) and $\Delta = 0.1$ (b) 85

List of Tables

2.1 A summary of bed allocation literature in the first stream 13

4.1 Index conversion table for elements in $\mathcal{P}$. 32

4.2 Results of the numerical experiments with the exact performance evaluation model. 34

4.3 The parameters for our numerical experiments. For Series I, we set $\mu_1 = \mu_2 = 1$, and for Series II, we set $\mu_1 = 1.2$ and $\mu_2 = 0.75$. 41

4.4 The average relative accuracy of the approximation methodology for the overflow delay system. 46

4.5 The average relative accuracy of the approximation methodology for the overflow abandonment system. 53

5.1 The scenarios for which CDOS fails to find the optimal allocation. . . . 57

6.1 An example of episodes of care and corresponding bed visits within a hospital spell at RSCH. 62

6.2 Bed numbers and average occupancy levels of different inpatient wards at RSCH. 63

6.3 FCE-HRG codes and their corresponding specialties 65

6.4 Specialties at RSCH, and their corresponding divisions and acronyms. . 66

6.5 The primary wards of different specialties at RSCH. 67

6.6 Input parameters for our reconfiguration models. 72

6.7 Empirical and Theoretical CV under Poisson assumption for arrival counts 73

6.8 Empirical CV for LOSs. 74

6.9 Arrival rates and mean LOSs for inpatient wards at RSCH 77

6.10 The savings obtained from the best-found configuration (COF) as compared to other configurations under IWTT assumption. 78

6.11 The savings obtained from the best-found configuration (COF) as compared to other configurations under FWTT assumption with $\gamma^{-1} = 1$ day. 79
6.12 The savings obtained from the best-found configuration (COF) as compared to other configurations under FWTT assumption with $\gamma^{-1} = 15$ days. 81
6.13 The savings obtained from the best-found configurations as compared to other configurations under FWTT assumption with $\gamma^{-1} = 30$ days. For $\Delta = 0.0$ ($\Delta = 0.1$), WNG (COF) is the best-found configuration. . . 83
6.14 Sensitivity analysis on nurse skill cost under IWTT assumption. 83
6.15 Sensitivity analysis on nurse skill cost under FWTT assumption. 83
6.16 Sensitivity analysis on daily cost of waiting under IWTT assumption. . 84
6.17 Sensitivity analysis on cost of abandoning under FWTT assumption. . . 84

List of Abbreviations

NHS	National Health Service
LOS	Length of Stay
IWTT	Infinite Waiting Time Threshold
FWTT	Finite Waiting Time Threshold
ZWTT	Zero Waiting Time Threshold
RSCH	Royal Surrey County Hospital
COF	Clustered Overflow
WNG	Wing Formation
ERM	Earmarking
DED	Dedicated
FLX	Flexible
CTMC	Continuous Times Markov Chain
SERT	S (state space), E (set of event types), R (set of transition rates), and T (set of target states)
CDOS	Conjugate Direction with Orthogonal Shift
EAU	Emergency Assessment Unit
FCE-HRG	Finished Consultant Episode - Healthcare Resource Group
CRD	Cardiology
END	Endocrinology
ENT	Ear, Nose and Throat
GAS	Gastroenterology
GSR	General Surgery
GRT	Geriatric Medicine
GYN	Gynaecology & Obstetric
NRO	Neurology
ONC	Oncology & Haematology
OPL	Ophthalmology
ORM	Oral & Maxillo Facial
PAL	Palliative Medicine
RSP	Respiratory Medicine
RUM	Rheumatology
ORT	Trauma & Orthopaedics
URO	Urology
CV	Coefficient of Variation

Notation	Definition
B	total number of beds
n	total number of specialties
$\mathcal{S}$	$\{1, 2, \ldots, n\}$, i.e., the index set of specialties
$\boldsymbol{\mathcal{C}}$	$\{\mathcal{C}^1, \ldots, \mathcal{C}^m\}$, i.e., a partition of set $\mathcal{S}$ into m clusters
m	number of clusters for a given partition
$\mathbb{Z}$ ($\mathbb{Z}_+$)	set of non-negative (positive) integers
$\mathbf{d}$	$(d_1, \ldots, d_n)$, i.e., bed allocation vector for dedicated wards
$\mathbf{o}$	$(o^1, \ldots, o^m)$, i.e., bed allocation vector for the overflow wards
λ_i	arrival rate for specialty i patients
$\nu_i(d, \mathcal{A})$	mean LOS for patients of specialty $i \in \mathcal{A}$ cared for in a d-bed ward shared by a subset $\mathcal{A}$ of specialties
γ	abandonment rate of patients of all specialties
Z	objective function for the formulation given in (3.1)
$T(\mathcal{C}^j, (d_i; i \in \mathcal{C}^j), o^j)$	total average daily cost of cluster $\mathcal{C}^j$ with bed allocation $(d_i; i \in \mathcal{C}^j)$, o^j
$\mathbf{b}$	$(b^1, \ldots, b^m)$, i.e., the vector of bed numbers allocated to each cluster
Ψ	feasible region of problem (3.2)
$\phi(\mathcal{C}^j, b^j)$	minimum total average daily cost for cluster $\mathcal{C}^j$ given b^j beds as obtained by problem (3.3)
c_w	daily cost of a patient waiting in the queue
c_a	cost of a patient abandoning the queue
$Q(\mathcal{C}^j, (d_i; i \in \mathcal{C}^j), o^j)$	average daily number of patients waiting in the queue for cluster $\mathcal{C}^j$ with bed allocation $(d_i; i \in \mathcal{C}^j)$, o^j
$B(\mathcal{C}^j, (d_i; i \in \mathcal{C}^j), o^j)$	average daily number of patients abandoning for cluster $\mathcal{C}^j$ with bed allocation $(d_i; i \in \mathcal{C}^j)$, o^j
$R(\mathcal{C}^j, (d_i; i \in \mathcal{C}^j), o^j)$	average daily cost of nursing staff in cluster $\mathcal{C}_j$ with bed allocation $(d_i; i \in \mathcal{C}^j)$, o^j
$W_k(\mathcal{C}^j, (d_i; i \in \mathcal{C}^j), o^j)$	mean waiting time of specialty $k \in \mathcal{C}_j$ patients assuming bed allocation $(d_i; i \in \mathcal{C}^j)$, o^j for cluster $\mathcal{C}_j$
$A_k(\mathcal{C}^j, (d_i; i \in \mathcal{C}^j), o^j)$	probability of abandonment of specialty $k \in \mathcal{C}_j$ patients assuming bed allocation $(d_i; i \in \mathcal{C}^j)$, o^j for cluster $\mathcal{C}_j$
f_i	desired nurse-to-patient ratio for patients of specialty $i \in \mathcal{S}$
$r(\mathcal{A})$	daily cost of a nurse working in a ward admitting patients of specialties in $\mathcal{A} \subset \mathcal{S}$
$S_k^d((d_i; i \in \mathcal{C}^j), o^j)$	expected number of beds occupied by patients of specialty $k \in \mathcal{C}^j$ in their dedicated ward assuming bed allocation $(d_i; i \in \mathcal{C}^j)$, o^j for cluster $\mathcal{C}_j$
$S_k^o((d_i; i \in \mathcal{C}^j), o^j)$	expected number of beds occupied by patients of specialty $k \in \mathcal{C}^j$ in their overflow ward assuming bed allocation $(d_i; i \in \mathcal{C}^j)$, o^j for cluster $\mathcal{C}_j$
o	number of servers in the overflow beds of an overflow delay/abandonment system
ν_i	mean service time of class i customers in the corresponding dedicated pool of an overflow delay/abandonment system
ν_i'	mean service time of class i customers in the overflow pool of an overflow delay/abandonment system
μ_i	service rate of class i customers in the corresponding dedicated pool of an overflow delay/abandonment system
μ_i'	service rate of class i customers in the overflow pool of an overflow delay/abandonment system
$\mathcal{P}$	state space for the CTMC of a two-class single-server overflow delay system as defined in (4.1)
N	upper bound to make the state space $\mathcal{P}$ finite
$\mathcal{E}$	set of event types for the CTMC of a two-class single-server overflow delay system as defined in (4.2)

Notation	Definition
$e_1, \ldots, e_6$	different event types for the CTMC of a two-class single-server overflow delay system
$\mathcal{Q}$	set of transition rates for the CTMC of a two-class single-server overflow delay system as defined in (4.3)
$q_p^{e_i}$	rate of event e_i when the CTMC of a two-class single-server overflow delay system is in state $p \in \mathcal{P}$
r_p	transition rate out of any state $p \in \mathcal{P}$ for the CTMC of a two-class single-server overflow delay system
$\mathcal{T}$	set of target states for the CTMC of a two-class single-server overflow delay system as defined in (4.4)
$g_p^{e_i}$	target state when event e_i occurs in state $p \in \mathcal{P}$ for the CTMC of a two-class single-server overflow delay system
π_p	steady-state probability for the CTMC of a two-class single-server overflow delay system being in state $p \in \mathcal{P}$
L_i^q	queue length of class i customers in a two-class single-server overflow delay system
W_i^q	mean waiting time of class i customers in a two-class single-server overflow delay system
p_i^{loss}	probability of class i customers being turned away from a (truncated) two-class single-server overflow delay system
$\phi((i,j,k,l))$	indexing function mapping elements of $\mathcal{P}$ into a one-dimensional index as defined in (4.10)
v	the index obtained from the mapping function $\phi((i,j,k,l))$
$B(A(t), F(t), d)$	blocking probability in a loss system with arrival process $A(t)$, service time CDF $F(t)$ and d servers
$B_e(a,d)$	a continuous extension of the Erlang loss function for a loss system with offered load a and d servers
z	measure of peakedness, i.e., variance to mean ratio for steady state number of busy servers in an equivalent infinite-server system
a_i	$\lambda_i \nu_i$, i.e., the offered load in the dedicated pool of class i customers in an overflow loss system
ρ_i	$\nu_i / \nu_i^{'}$, i.e., ratio of mean service time in the dedicated pool to the overflow pool
α	offered load for the aggregate overflow stream in an overflow loss system
$\xi(a,d,\rho)$	peakedness of a Poisson stream overflowing a d-server dedicated pool with offered load a and mean service ratio ρ
β	peakedness of the aggregate overflow stream in an overflow loss system
L	blocking probability faced by the aggregated overflow stream in an overflow loss system
L_i	blocking probability faced by class i customers in an overflow loss system
P_i	probability of delay faced by class i customers in an overflow delay system
δ_i	offered load of class i customers in an overflow delay system
$\bar{\nu}_i$	weighted average of class i mean service times in dedicated and overflow pools of an overflow delay system
I_i^d	the idle service capacity in dedicated pool i of an overflow delay system
I_i^o	the portion of idle service capacity allocated to class i customers in the overflow pool of an overflow delay system
$\lambda_i^{'}$	the rate of class i customers overflow from dedicated pool in an overflow delay system
I^o	The overall idle capacity at the overflow pool of an overflow delay system
$\bar{\nu}$	the weighted average of mean service times in the overflow pool of an overflow delay system
s_i	number of servers allocated to class i customers in an overflow delay system
S_i^d	the mean number of class i customers in the dedicated pool of an overflow delay/abandonment system
S_i^o	the mean number of class i customers in the overflow pool of an overflow delay/abandonment system
$\Gamma(x)$	Gamma function
A_i	abandonment probability experienced by class i customers in an overflow abandonment system

Notation	Definition
c_i	a parameter required for evaluating abandonment probability defined as $s_i/\bar{\nu}_i\gamma$
η_i	a parameter required for evaluating abandonment probability defined as λ_i/γ
$\mathcal{N}$	a fixed sequence of specialties in $\mathcal{S}$
$Z_{\mathcal{N}}$	formulation of restricted objective function by sequence $\mathcal{N}$
Θ	the state space for the DP approach for solving the restricted problem $Z_{\mathcal{N}}$
$A(\theta)$	the set of actions associated with state $\theta \in \Theta$ for the DP approach for solving the restricted problem $Z_{\mathcal{N}}$
$V(\theta)$	the objective function given the current state $\theta \in \Theta$ for the DP approach for solving the restricted problem $Z_{\mathcal{N}}$
m_i	the existing meal LOS of specialty i patients at RSCH
τ_i	mean nominal LOS for specialty i patients at RSCH
Δ	parameter controlling the impact of focus on mean LOS
β and ϵ	parameters controlling the impact of workload on mean LOS
τ_i	mean nominal LOS for specialty i

To Farid, Haleh and Pedram ...

1 Introduction

1.1 Motivation

Now, more than ever before, healthcare systems around the world are facing an inpatient bed crisis. Inpatient beds are one of the most valuable resources of care delivery in hospitals (Green 2006), and their shortage may have dire consequences for both patients and staff. The (un)availability of inpatient beds is typically measured by the average bed occupancy level, defined as the total number of occupied bed days divided by the total number of available bed days in a given time period. As an example, the statistics published by the National Health Service (NHS) in the UK show that in 2019, the average bed occupancy across the country was above 90% for the third year in a row (King's Fund 2020a). The shortage of beds was highlighted even further during the recent COVID-19 pandemic. For example, Mateen et al. (2021) report that during the first wave of the pandemic in England, many hospitals operated above safe-occupancy thresholds for significant periods. Further, a surge in demand is expected once the pandemic is over as many non-urgent procedures which have been postponed to free up capacity for COVID-19 patients need to be rescheduled (Propper et al. 2020). The impact of the pandemic on bed pressure will therefore stretch out further and worsen in the future. Our aim in this dissertation is to propose a cost-effective solution to the ongoing bed-crisis, and investigate its impact using the data from a UK NHS hospital. We set the stage by elaborating on the causes and consequences of bed shortages below, followed by an outline of the solutions proposed in the literature. This then leads to the focus selected for this study.

One of the prominent causes of bed shortage is the reduction in the number of hospital beds over the years. As reported in King's Fund (2020a), for example, bed cutbacks have happened almost every year in the UK, some years more dramatically than others. The main logic behind bed cutbacks is the perception that demand for inpatient care has decreased overall as a result of expansion in community care, reduction in patients' mean length of stay (LOS), and replacement of some procedures requiring inpatient admission with day procedures. However, these trends seem to have been off-set by two main factors: (i) a substantial increase in the number of people who attend emergency departments and subsequently require admission to inpatient care (NHS Digital 2018); and (ii) the growth of the elderly population, who are

the most common users of beds, and the prevalence of long-term conditions and co-morbidity among them (BMA 2017). Ineffective discharge processes and prolonged turnover intervals, i.e., the interval between a bed becoming available and being occupied by a patient, have also added to the bed pressure in hospitals (Steventon et al. 2018).

Shortage of inpatient beds has many ramifications for both patients and staff. It prolongs the trolley wait, i.e., the time between a decision being made in the emergency department (ED) to admit a patient and admission to inpatient care. Trolley waits lower the quality of patient care and may result in patient fatalities (Campbell 2019). For example, a recent study of more than 5 million patients in England shows a linear increase in all-cause 30-day mortality from 5 hours after arrival to the ED up to 12 hours (Jones et al. 2022). Trolley waits also create large backlogs in emergency departments (King's Fund 2020b). Congestion in emergency departments is linked to higher morbidity rates and may also lead to *ambulance diversion*; see, e.g., Sprivulis et al. (2006) and Olshaker and Rathlev (2006). Further, overcrowded hospitals are exposed to a higher risk of hospital acquired infections (Kaier et al. 2012). Not only does this put patients' health at risk, but also prolongs LOS and may also result in bed and ward closures (Goldstein et al. 2017), exacerbating the bed shortage problem. Shortage of beds for post-operative care may result in cancellation of medical procedures. Patients may also be discharged pre-maturely, only to be re-admitted later with potentially worse conditions (Maguire 2015), or moved to lower acuity beds to make space for more critical patients (Green 2002). Readmitted patients are reported to cost the NHS £2.6 billion each year (Conroy and Dowsing 2012).

Patient outlying, i.e., admitting patients to clinically inappropriate wards, is a common phenomenon in hospitals which operate with high occupancy levels. Stowell et al. (2013) report a significant increase in the LOS and rate of mortality of outlied patients. Serafini et al. (2015) demonstrate that the quality of care and patient satisfaction are typically lower for outlied patients. The negative impact of outlying is not limited to outlied patients. In a recent study, Lim et al. (2021) show that the level and volatility of outlying will increase the LOS and the likelihood of readmission for non-outlied patients as well. Finally, the pressures that bed shortages create can have a damaging impact on staff morale and retention, which in turn impacts negatively on patient care (Roberston et al. 2017). Véricourt and Jennings (2011) report that overcrowded hospitals are associated with an increase in consumption of antidepressants among hospital staff.

Expanding the bed base of a hospital is the most straight-forward solution to bed shortages. However, it is only considered as a last resort as adding a hospital bed and staffing it is quite costly (Akcali et al. 2006). For example, Hall (2012) estimate that adding an average bed may cost over one million dollars. As such, the focus of hospitals is typically on more efficient use

of existing beds. This is achieved by improving bed management practices as well as discharge processes. Using a combination of flow management techniques and scheduling models, bed management aims to maximize the utilisation of existing inpatient beds (Boaden et al. 1999). Discharge management is, on the other hand, focused on identifying and eliminating the factors contributing to delays in the discharge process (Huang 1998). This is often achieved through ".... obtaining high-quality discharge information, coordination of care across health and social sector, and direct and timely communication between medical professionals ..." (Hesselink et al. 2014, p.1). In addition to bed and discharge management practices, there are some initiatives undertaken by hospitals to improve beds availability. According to the guideline published by the NHS in 2018, for example, patients with lengths of stay longer than 20 days must be reviewed weekly so that the issues delaying their treatment and discharge are resolved (NHS 2018). The same guideline advises hospitals to provide early senior assessment and intervention in the emergency department to reduce the need for hospitalization.

In this research, we focus on reconfiguration of inpatient services as a cost-effective way for dealing with bed shortages. Inpatient care in general hospitals is typically delivered through a number of clinical units or *wards*. Each ward has a specific number of beds and a dedicated nursing team, and provides care for a certain subset of specialties. The configuration of inpatient services identifies the number of beds as well as the subset of specialties allocated to each ward. Once a decision is made to admit a patient, the most appropriate ward(s) depending on the patient diagnosis is selected. We call this the primary ward(s) of the patient. If the primary ward is full, the patient would have to wait in a virtual queue until a bed becomes available. If the waiting time exceeds a certain threshold, the patient may be transferred to another hospital. The waiting time threshold depends on the urgency of patient condition, with emergency patients having a much shorter threshold than electives. It also varies in different healthcare systems. In Dutch hospitals, for example, the threshold is very short (Bekker et al. 2016), while in the UK it is very long with many patients waiting for several months before admission.

The performance and quality of inpatient care is greatly influenced by the configuration of inpatient services in the hospital. We identify five major configurations of inpatient services as illustrated in the spectrum in Figure 1.1. Each of theses configurations and their pros and cons are discussed in the following section.

1.2 Configurations of Inpatient Services

In this section, we explain each of the configurations illustrated in Figure 1.1 using an example with 12 beds and 4 specialities. We compare these configurations in terms of their level of

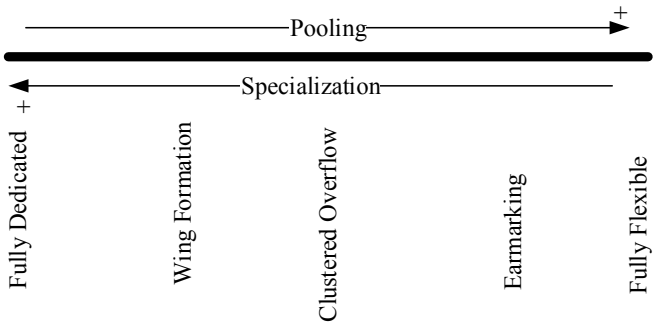


FIGURE 1.1: The spectrum of inpatient bed configurations

"focused care", the amount of "slack capacity" and "mix variability", and "cross-training" costs. Focus in a hospital setting means delivering care for a limited set of conditions. It lowers uncertainty, reduces complexity, and gives the clinical staff the opportunity to develop specialised expertise (Clark and Huckman 2012). Several studies suggest that focused care improves the quality of care, in particular, it reduces mean LOS (Best et al. 2015), rate of mortality (Clark and Huckman 2012), and risk of readmission (KC and Terwiesch 2011). Slack capacity refers to the number of available but unused beds in the hospital. It increases as the wards become more specialized. We define mix variability as the additional variability in LOS created by mixing specialties with different LOS distributions in a single ward. Queueing theory suggests that a higher variability in LOS typically leads to longer waiting times in the queue (Gross et al. 2011). Cross-training is about nurses being trained to deliver care for the specialties treated in their ward. Clearly, the cost of cross-training increases as more specialties are allocated to a ward.

1.2.1 Fully Dedicated Configuration

Starting from the far left of the spectrum in Figure 1.1, in a fully dedicated configuration each specialty is allocated to a single ward fed by a single queue (see Figure 1.2). This configuration benefits from the maximum level of focused care and the minimum amount of mix variability as a limited number of conditions are treated in each ward. Slack capacity will, however, be at its highest level as there will be situations wherein patients are waiting for admission to their primary ward whilst beds are available in other wards.

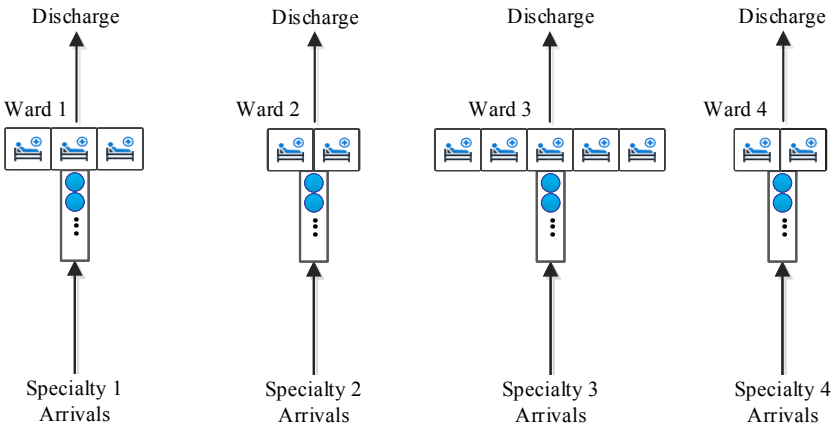


FIGURE 1.2: An example of the fully dedicated configuration

1.2.2 Fully Flexible Configuration

The fully flexible configuration is on the far right of the spectrum in Figure 1.1. It acts like a super ward, admitting patients of different specialties (who join a single queue) as long as a bed is available (see Figure 1.3). This configuration enjoys the benefits of pooling, i.e., the slack capacity will be minimized. However, mix variability will increase as a result of mixing patients with different LOS distributions. This, along with the potential increase in mean LOS due to losing focus, may off-set the advantages of pooling. The fully flexible configuration also requires full cross-training of nursing teams which is expensive.

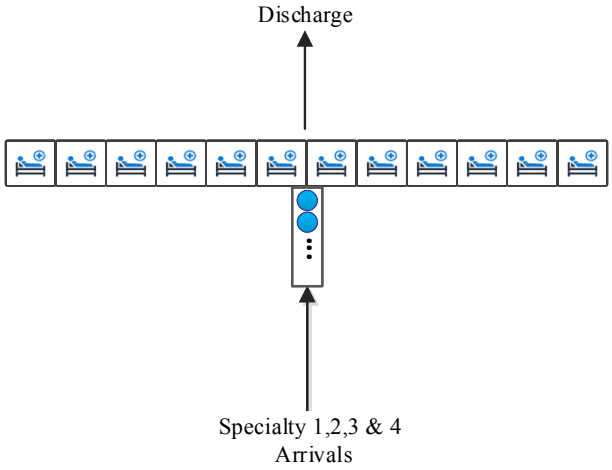


FIGURE 1.3: An example of the fully flexible configuration

1.2.3 Wing Formation Configuration

This configuration, proposed in Best et al. (2015), attempts to strike a balance between flexibility and specialization by partitioning the specialities into a number of clusters, and dedicating

each cluster to a ward (wing). Each ward is fed by a single queue including all the patients requiring care for specialties in the ward cluster (see Figure 1.4). This configuration captures the fully dedicated and flexible configurations as special cases. Its cross-training cost, the amount of slack capacity and mix variability, and the level of focus vary depending on the partitioning of specialties.

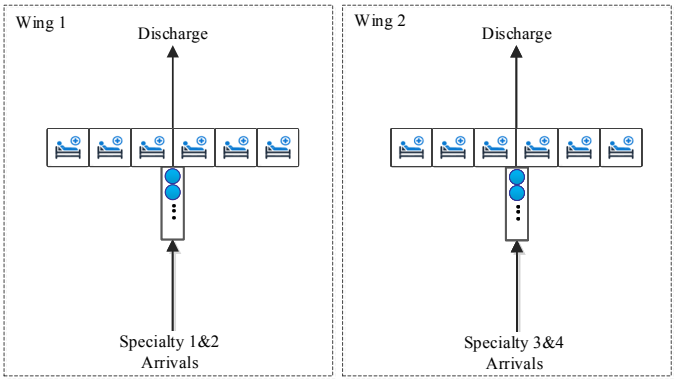


FIGURE 1.4: An example of the wing formation configuration

1.2.4 Earmarking Configuration

In this configuration, introduced in Bekker et al. (2016), each speciality has a ward dedicated to it which is linked to a single queue, similar to the fully dedicated configuration. But there also exists a shared overflow ward admitting patients whose dedicated wards are full (see Figure 1.5). The earmarking configuration benefits from focused care in its dedicated wards, and from pooling in its overflow ward. Since all specialities share one overflow ward, however, it requires costly full nurse cross-training.

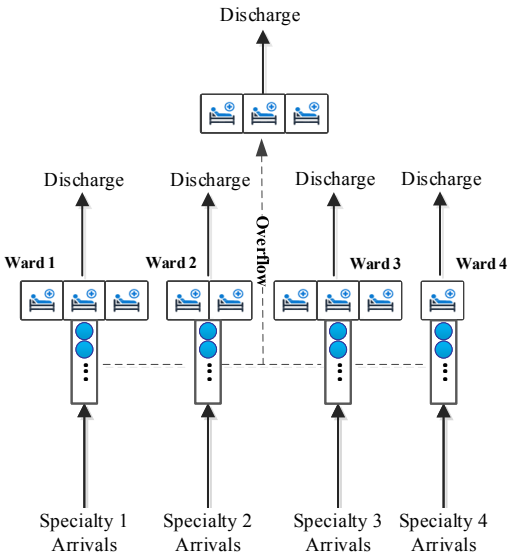


FIGURE 1.5: An example of the earmarking configuration

1.2.5 Clustered Overflow Configuration

In this configuration, proposed in Izady and Muhamed (2019), specialities are partitioned into a number of clusters similar to the wing formation configuration. Each cluster, however, includes a dedicated ward for each of the specialities in the cluster as well a single overflow ward shared among all the specialities of the cluster (see Figure 1.6). There is a single queue attached to each dedicated ward, feeding the patients first into the dedicated ward and next to the cluster overflow ward. Similar to earmarking, this configuration benefits from both pooling and focused care. The flexibility created by partitioning of specialties, however, helps reduce the cross-training cost as well as the amount of mix variability. The clustered overflow configuration (COF) captures all the configurations mentioned above as special cases; we elaborate on this in Chapter 3.

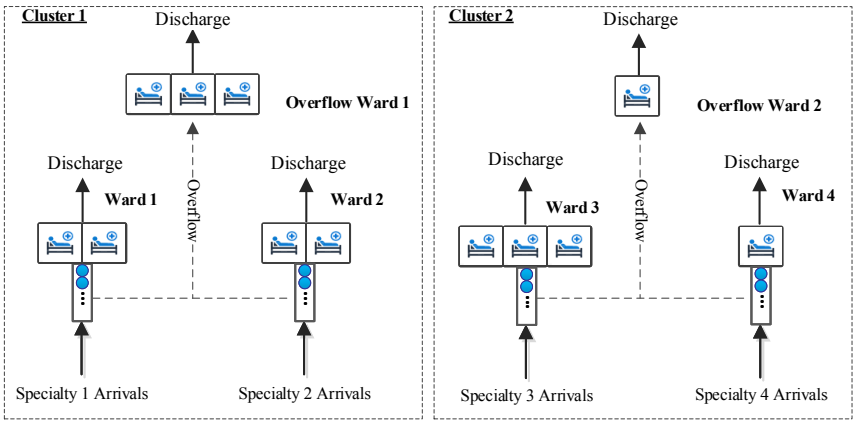


FIGURE 1.6: An example of the clustered overflow configuration

One of the five major configurations described above is typically adopted by the hospitals. However, there always exists a degree of patient outlying in practice. While this reduces the admission wait for the patients involved, it may negatively influence the care quality of outlied and non-outlied patients as explained earlier, and may also worsen the performance of the system overall. Further, a large percentage of patient outlying is an indicator that inpatient services are not organized properly. Our aim is to develop a methodology for finding a configuration of inpatient services which uses the existing capacity efficiently without the need to outlying patients.

1.3 Thesis Structure

The rest of the thesis is organised as follows. In Chapter 2, we review the relevant literature. This includes papers discussing inpatient bed planning and allocation as well as those that propose models for evaluating performance of a given configuration. This leads to identifying the research gaps and formulating our research questions.

In Chapter 3, we first provide a mathematical representation for the COF configuration, and explain why it is used as the basis for developing our methodology. We then formulate the optimization problem for finding the optimal partitioning and bed allocation for the COF configuration. We conclude this chapter by specifying the key performance metrics that must be evaluated as part of the optimization model, laying the foundation for our performance evaluation models in Chapter 4.

In Chapter 4, we propose methodologies for approximating the performance of the COF configuration assuming patients' waiting time threshold (before transfer to another hospital) is either infinitely long, or finite and given. We demonstrate the accuracy and reliability of our approximations by comparing their results to the ones obtained from a discrete-event simulation model which we have developed for the purpose of this study.

Chapter 5 is dedicated to solving the optimization model proposed in Chapter 3. We discuss the solution methodology for finding the optimal partitioning and bed allocation for the COF configuration. As our solution methodology involves heuristic approaches, we test their accuracy using a number of test cases for which the optimal solutions are obtained by complete enumeration.

Chapter 6 is designed to illustrate the application of our proposed methodology in a real-world setting. In particular, we report the result of our collaboration with an NHS hospital. We start by introducing the hospital and give a holistic overview of the then current status of inpatient services. We then illustrate how our methodology can reduce the bed pressure on hospital without expanding the bed base and with a moderate increase in daily nursing costs. Finally, in Chapter 7, we provide the reader with concluding remarks of this doctoral dissertation. We also discuss some potential directions for future research.

2 Literature Review

2.1 Introduction

In this chapter, we provide an overview of relevant literature to our research. We start by reviewing the inpatient bed allocation literature in Section 2.2. We divide this literature into three main streams, and highlight methodologies, objectives, settings, and salient points of important papers in each stream. This is followed by an outline of important results from the resource flexibility literature in Section 2.3. We then explain the critical role played by performance evaluation models in inpatient bed allocation methodologies and explore the relevant literature in Section 2.4. This leads to our research questions in Section 2.5.

2.2 Inpatient Bed Allocation

The subject of inpatient bed allocation has been studied by many researchers throughout the years. Bed allocation means "...assigning beds to various patient categories according to medical specialty, accommodation type and logistical considerations; presumably, patient needs, research goals, and educational requirements are taken into account, along with cost..." (Dumas 1985, p. 44). We divide the inpatient bed allocation literature into three main streams as below.

The first stream includes articles that seek to identify the number of beds for a clinical unit so as to achieve a given objective. The objective is often a target average bed occupancy (e.g., Vassilacopoulos 1985), delay probability (e.g., Green and Nguyen 2001) or loss probability (e.g., de Bruin et al. 2009). A cost minimization objective is also followed by some authors (e.g., Gorunescu et al. 2002b). It must be noted that reliance on target average occupancy levels for bed allocation, which is still common in practice, is problematic because average occupancy measures do not capture seasonal, weekly, and daily variations in patient census. Further, queueing theory shows that a given occupancy level may result in varying levels of performance depending on the size of the unit as well as the level of variability in the patient arrival process and LOS distribution (Green and Nguyen 2001).

A summary of the literature in the first stream is presented in Table 2.1. As illustrated in this table, some of the papers use discrete-event simulation models for performance evaluation (e.g., Vassilacopoulos 1985), while others apply analytical queueing models (e.g., Asaduzzaman and Chaussalet 2008). The performance evaluation model is embedded within an optimization framework in some studies (e.g., Belciug and Gorunescu 2015), while others simply follow a trial and error approach, i.e., experiment with increasing numbers of beds until the smallest value satisfying the pre-defined objective(s) is found (e.g., Green and Nguyen 2001).

Some studies are applied to a stand-alone unit (e.g., Belciug and Gorunescu 2015), while others consider a tandem (e.g, de Bruin et al. 2007) or hierarchical (e.g., Asaduzzaman et al. 2010) network structure. The papers in the first stream also differ in their assumption on patients' waiting time threshold before transfer. Some papers work with a zero waiting time threshold (ZWTT) assumption (e.g., de Bruin et al. 2009), while others consider an infinite waiting time threshold (IWTT) assumption (e.g., Green and Nguyen 2001). We did not find any paper with a (non-zero) finite waiting time threshold (FWTT) assumption in this stream.

Study	Objective	Methodology	Ward(s)	Assumption
de Bruin et al. (2007)	Target loss probability	Queueing theory applied within a trial and error approach	Cardiac care pathway	ZWTT
Gorunescu et al. (2002b)	Min. expected daily cost	Queueing theory applied within a trial and error approach	Geriatric ward	ZWTT
Gorunescu et al. (2002a)	Min. expected daily cost	Queueing theory applied within a trial and error approach	Geriatric ward	IWTT
Green and Nguyen (2001)	Target delay probability	Queueing theory applied within a trial and error approach	Obstetric ward	IWTT
Green and Nguyen (2001)	Target mean waiting time	Queueing theory applied within a trial and error approach	Surgery ward	IWTT
Asaduzzaman et al. (2010)	Target loss probability	Queueing theory applied within a trial and error approach	Neonatal care pathway	ZWTT
Kokangul (2008)	Generic	A simulation-based approximation applied within a non-linear optimization	Paediatric ICU	ZWTT
Vassilacopoulos (1985)	Target occupancy level	Simulation applied within a trial and error approach	Various wards	IWTT
de Bruin et al. (2009)	Target loss probability	Queueing theory applied within a trial and error approach	Various wards	ZWTT
Belciug and Gorunescu (2015)	Target loss probability	Queueing theory applied within a Genetic Algorithm heuristic	Various wards	ZWTT
Wu et al. (2019)	Min. daily cost	Queueing theory applied within a multi-objective optimization	Neurology pathway	ZWTT
Harper and Shahani (2002)	Overall improvement	Simulation applied within a trial and error approach	ICU	IWTT

TABLE 2.1: A summary of bed allocation literature in the first stream

The second stream includes studies that seek to find the optimal allocation of a given number of beds to a set of specialties under a fully dedicated configuration. As one of the earliest papers in this stream, Goldman et al. (1968) compare the impact of three bed allocation policies among nine different specialties using a discrete-event simulation model. Kao and Tung (1981) propose a two-stage optimization model to distribute beds among different specialties. In the first stage of their model, a baseline bed number is allocated to each specialty so as to achieve a given delay probability. The remaining beds are then allocated in the second stage minimizing the number of patients who have to wait. Dumas (1985) investigate the allocation of beds to different specialties following a trial-and-error approach using a discrete-event simulation model of the hospital. Cochran and Bharti (2006) use a combination of queueing theory and optimization to find the bed allocation balancing the utilization among different wards. Li et al. (2009) use a queueing model embedded within a goal programming optimization framework to find the best allocation of beds among departments aiming to minimize the deviation from target profit and admission levels. Holm et al. (2013) use a combination of simulation and optimization to allocate beds among specialties in a large hospital. In particular, they use a simulation model of the hospital with unlimited bed capacity to generate a bed-ward utilization matrix. This matrix is then fed into an optimization model which identifies the bed allocation vector that maximizes the overall bed utilization. Mohamed et al. (2016) use queueing theory along with complete enumeration to find the optimal allocation of beds among different specialties of a paediatric intensive care unit. All the studies we have cited in the second stream work under an IWTT assumption except that of Mohamed et al. (2016) which considers a ZWTT.

The third stream is the closest to our research, and includes articles that consider clustering a given set of specialties and identifying the corresponding bed allocation simultaneously. The two prominent papers in this stream are those of Best et al. (2015) and Izady and Muhamed (2019). Best et al. (2015) propose the "wing formation" configuration, aiming to find the optimal level of bed pooling. Given a total number of beds, a set of specialties, and a finite waiting time threshold, they seek to identify the optimal allocation of specialties to different wings and the corresponding bed allocation so as to maximise the total utility to the hospital. Best et al. (2015) use a diagnoses-related-group relative weight published in the federal database of the Center for Medicare and Medicaid Services in United States as a measure of utility. This measure represents the cost of resources consumed for a given specialty and so acts as a proxy for care complexity and also profitability. One of the salient points of Best et al. (2015)'s work is that it captures the impact of focus on the utility generated by each patient as well as her length of stay endogenously. It also proposes a dynamic programming (DP) solution methodology for solving a restricted version of the clustering and bed allocation model. Best

et al. (2015) apply their model to an 18-specialty hospital to generate insight. Their numerical experiments with varying levels of waiting time threshold and workload imply that hospitals with a longer waiting time threshold or higher levels of demand should form more specialized wings to benefit from the advantages of focused care.

Izady and Muhamed (2019) introduce the COF configuration and propose a heuristic methodology for its clustering and bed allocation. Assuming a ZWTT assumption, they propose two different formulations, a total cost minimization formulation and a constrained blocking minimization formulation. The former formulation aims to minimize the total average daily cost, including the cost of turning patients away plus the cost of nursing teams. The latter formulation seeks to minimize the total number of patients turned away subject to nursing cost falling below a given threshold. An important feature of both formulations in Izady and Muhamed (2019) is the inclusion of nursing costs. Their solution methodology involves an intra-cluster bed allocation model and a partitioning and inter-cluster bed allocation model. Izady and Muhamed (2019) apply their methodology to a 7-specialty paediatric department, and report that the configurations obtained from their methodology compare very well with other major configurations as long as patients' waiting time threshold is relatively short.

There are other studies in the bed allocation literature that do not fall within the three streams reviewed above. For example, van Essen et al. (2015) consider a setting in which the total number of beds for each ward is given, and the objective is to first form clusters of specialities, and then allocate these clusters to wards in such a way that the blocking probability for each cluster does not exceed a certain threshold. They propose exact and heuristic solution methodologies for this problem. Another example is the work of Bekker et al. (2016) which introduces the earmarking configuration and proposes an exact methodology for its performance evaluation under some specific conditions. A heuristic approach is also proposed in their paper for allocating beds among the dedicated wards assuming the number of beds in the overflow ward is given. Both van Essen et al. (2015) and Bekker et al. (2016) work with a ZWTT assumption.

2.3 Resource Flexibility

At a more general level, the flexibility of resources in manufacturing and service settings is addressed in two different streams of literature. The optimal configuration of multiple facilities in the presence of multiple customer classes with uncertain demand is explored in these two streams. Assuming that resources are naturally flexible, the first stream investigates whether facilities should be dedicated or pooled (see, e.g., van Dijk and van der Sluis 2008). The main

finding is that a pooled configuration performs better as long as different customer classes have the same service time distribution. When service time distributions are different, however, the positive impact of pooling maybe offset by the negative impact of mix variability. The second stream assumes that flexibility can be achieved at a cost, e.g., by cross-training human servers or using flexible machinery (see, e.g., Jordan and Graves 1995, Bassamboo et al. 2010). The key finding is that a limited amount of flexibility can go a long way. In particular, a chained configuration that uses only bi-flexible servers, i.e., servers that can serve two customer classes, can achieve almost all benefits of full flexibility.

2.4 Performance Evaluation

A performance evaluation model lies at the heart of all clustering and bed allocation methodologies. This model evaluates the performance metrics of a given partition of specialties with a given allocation of beds. Our focus in this section is on analytical performance evaluation models as applying simulation models within optimization frameworks for the purpose of clustering and bed allocation is challenging due to large numbers of scenarios that must be evaluated. We categorize these models based on two dimensions: (i) the type of interaction between different wards, and (ii) the waiting time threshold of patients. The interaction dimension includes three different types: no interaction, a hierarchical interaction, and a cross-facility interaction. The waiting time threshold dimension consists of three levels: zero, finite, and infinite thresholds. This leads to nine different categories of performance evaluation models as outlined in Figure 2.1. We review each of these categories below.

Category I: An Erlang loss queueing model, denoted by $M/GI/s/0$ — with a Poisson arrival process (the M), a general service time distribution that is independent of the arrival process (the GI), s servers and 0 waiting space — is generally applied. Examples include the works of De Bruin et al. (2007) and de Bruin et al. (2009). Extension to a time-dependent arrival process is useful due to the time-of-day and day-of-week trends observed in admission data (see, e.g., Shi et al. 2015). This extension is often achieved through approximations such as the modified offered load approximation (Massey and Whitt 1994) or the fixed-point approximation (Izady and Worthington 2011). See Bekker and de Bruin (2010) for an application of these approximations in bed planning with time-varying demand.

Category II: An Erlang abandonment queueing model, denoted by $M/M/s/\infty + M$ — with an Exponential service time distribution (the second M), infinite waiting space (the ∞), and an Exponential waiting time threshold (the $+M$) — is typically applied. For an example in the bed planning literature, see the work of Best et al. (2015). Approximations are proposed

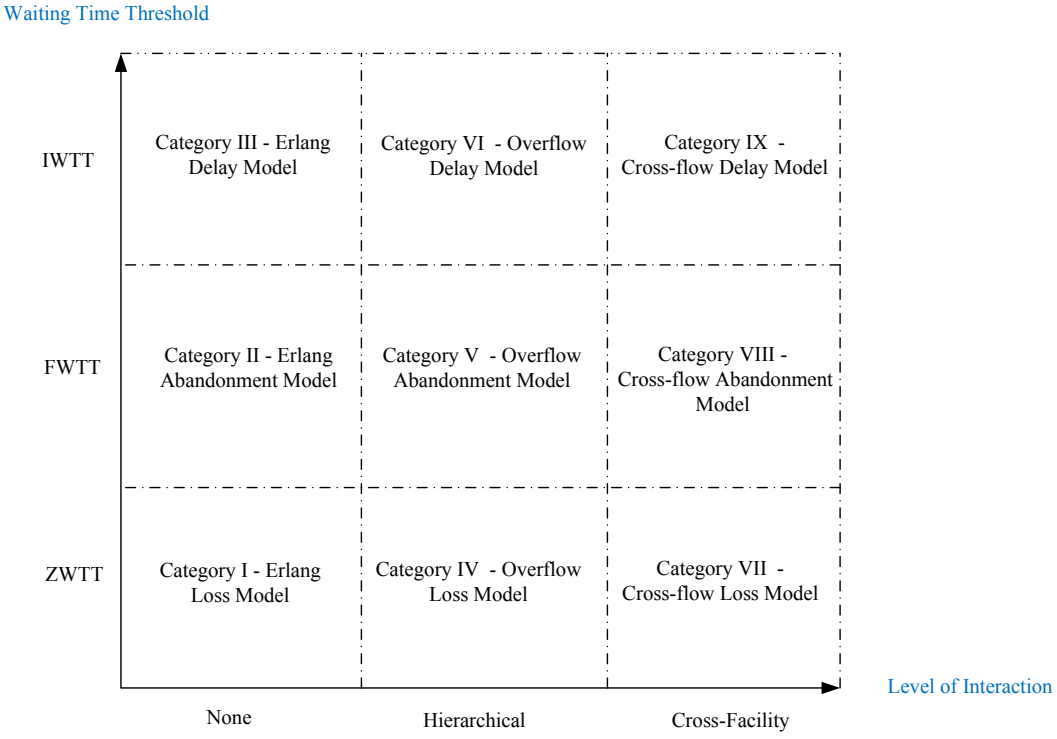


FIGURE 2.1: Different categories of performance evaluation models

in Mandelbaum and Zeltyn (2005) and Iravani and Balcioglu (2008) for $M/M/s/\infty + GI$ and $M/GI/s/\infty + GI$ models, respectively.

Category III: An Erlang delay model, denoted by $M/M/s/\infty$, is generally applied. See Green and Nguyen (2001) for an example.

Category IV: We refer to the relevant model in this category as the "overflow loss" model. It is applied for performance evaluation of hierarchical loss systems in which customers blocked from specialized primary facilities are served in a shared flexible overflow facility; see Chapter 4 for further detail. In an inpatient setting, the overflow loss model can be applied for performance evaluation of earmarking and clustered overflow configurations under a ZWTT assumption. An exact product-form performance evaluation methodology is proposed in Bekker et al. (2016) for the special case of overflow loss models wherein customers being served by the overflow facility move back to their primary facilities as soon as a primary server becomes available. For more general cases, approximations have been proposed in the literature. These include the equivalent random method (Cooper 1990, p.165), Haywards' approximation (Fredericks 1980), and the Hyper-Exponential decomposition (Franx et al. 2006). Izady and Muhamed (2019) extend the Hayward's model to situations in which mean service time of a customer class in the overflow facility is potentially different from the corresponding mean in the primary facility. This is useful for capturing the impact of focus on LOS in bed allocation models.

Category V: We refer to the relevant model as the "overflow abandonment" model. Such

model would be useful for performance evaluation of earmarking and clustered overflow configurations under an FWTT assumption. We have not found any analytical methodology for performance evaluation of overflow abandonment models in the literature.

Category VI: We refer to the relevant model as the "overflow delay" model. An approximation methodology has been proposed in Chevalier and Van den Schrieck (2009) for the special case where the mean LOS in dedicated and overflow wards are the same. Overflow delay models can be applied for performance evaluation of earmarking and clustered overflow configurations under an IWTT assumption.

Categories VII, VIII and, IX: We refer to the relevant models as "cross-flow loss", "cross-flow abandonment", and "cross-flow delay" models, respectively. These models are relevant for evaluating the performance of inpatient services with a substantial level of patient outlying. No analytical model is proposed in the literature for either of these categories. There are, however, detailed simulation models that attempt to capture the complexities of bed assignment to patients in settings with a high level of patient outlying. An early example is the simulation model of Dumas (1984). In their model, an ordered set of primary wards and an ordered set of non-primary wards are defined for each specialty. Once a patient makes a bed request, a primary bed is assigned to her if one is available, or becomes available before a pre-defined outlying threshold is reached. After this threshold, the patient can be assigned to the most preferable primary or non-primary bed, whichever becomes available first. A similar idea is applied in Shi et al. (2015), who add other complexities such as time-varying arrival rates, pre- and post-allocation delays, and time-dependent and correlated LOSs. Dong et al. (2019) take a data-driven approach in their simulation model by capturing the bed assignment decisions through fitting a discrete choice model.

2.5 Research Questions

Our literature review indicates that there is a gap for a methodology that finds the optimal configuration of inpatient services for an existing hospital. In particular, the papers that consider a hospital with a given bed number focus on a specific configuration. For example, Best et al. (2015) consider only the wing formation configuration or Bekker et al. (2016) focus only on the earmarking configuration. The only exception is the work of Izady and Muhamed (2019) which considers a COF configuration. This configuration captures other configurations as special cases, and so the methodology proposed in Izady and Muhamed (2019) would be able to find the best configuration in theory. However, due to the complexity of performance evaluation, Izady and Muhamed (2019) only consider the special case with a ZWTT assumption. This implies that patients are either admitted to their dedicated or

overflow wards upon arrival, or turned away immediately. This is not a realistic assumption in many care systems as patients do wait for admission. As such, we aim to consider both systems with a FWTT and those with an IWTT. We formulate our first research question as follows.

Q1. Given a total number of beds, a set of specialties, and a waiting time threshold, how can we find the optimal configuration of inpatient services and the corresponding bed allocation?

Similar to Izady and Muhamed (2019), we work with the COF configuration due to its versatility. This indicates that categories V and VI of performance evaluation models in Figure 2.1 are relevant to our research. As indicated above, there is no analytical methodology for category V models, and the only analytical methodology for category VI models does not capture the impact of focus on LOS. As such, we break down our first question into the following questions.

Q1.1. How can we analytically formulate the clustering and bed allocation optimisation problem for the COF configuration?

Q1.2. Can we analytically evaluate the performance of a COF configuration with given clustering and bed allocation, assuming an IWTT or an FWTT, accurately and efficiently while capturing the impact of focus?

Q1.3. Can we design a solution methodology for the formulation proposed in response to Q1.1?

Q1.1 is addressed in Chapter 3, while Q1.2 and Q1.3 are discussed in Chapters 4 and 5, respectively.

Our second research question focuses on the practical application of the methodology designed in response to Q1 and the potential savings that it can create. More specifically, the simulation experiments conducted in Izady and Muhamed (2019) suggest that COF is often the best configuration assuming patients have a short waiting time threshold. As such, it is not clear what configuration would perform better with longer thresholds that are common in many health systems such as the NHS. Also, the results in Izady and Muhamed (2019) are obtained based on a limited dataset from a small department with seven specialties. We would like to investigate if these results hold in more general settings with more specialties. As such we formulate the second and third questions as follows.

Q2. What are the potential cost savings of applying the methodology proposed in response to Q1 in a large-scale hospital?

Q3. Which configuration performs best under FWTT and IWTT assumptions?

Both Q2 and Q3 are addressed in Chapter 6.

3 Problem Formulation

3.1 Introduction

In this chapter, we first provide a mathematical representation for the COF configuration, and explain how it acts as a meta model capturing all the other configurations discussed in Chapter 1 as special cases. Outlining the underlying assumptions, we then formulate the optimization problem for finding the optimal partitioning and bed allocation for the COF configuration. We note that the mathematical representation and formulation provided in this chapter are based on the formulation proposed in Izady and Muhamed (2019). We make minor changes to simplify the notations, and revise the cost functions to capture waiting and abandonment costs. We conclude this chapter by specifying the key performance metrics that must be evaluated as part of the optimization model, laying the foundation for our performance evaluation models in Chapter 4.

3.2 Mathematical Representation

We assume there is a total of B inpatient beds providing care for a total of n specialties. As highlighted in Best et al. (2015), specialties should neither be too broad, as it would limit the benefits of clustering and optimization, nor too narrow to make it difficult to assign patients to relevant specialties upon admission. Let $\mathcal{S} = \{1, \ldots, n\}$ be the index set of specialties, and denote by $\boldsymbol{\mathcal{C}} = \{\mathcal{C}^1, \ldots, \mathcal{C}^m\}$ a partition of set $\mathcal{S}$ into $m \in \mathbb{Z}_+$ clusters. We use $\mathbb{Z}$ and $\mathbb{Z}_+$ to denote the set of non-negative and positive integers, respectively. For every cluster $\mathcal{C}^j \in \boldsymbol{\mathcal{C}}$ in the COF configuration, there exists a ward dedicated to patients of each speciality $i \in \mathcal{C}^j$, and an overflow ward j admitting overflowing patients of specialties in the cluster, for $j = 1, \ldots, m$. Let $\boldsymbol{d} = (d_1, \ldots, d_n)$ and $\boldsymbol{o} = (o^1, \ldots, o^m)$ be the dedicated and overflow bed allocation vectors, respectively, with $d_i \in \mathbb{Z}$ representing the number of beds in the ward dedicated to speciality i for $i \in \mathcal{S}$, and $o^j \in \mathbb{Z}$ the number of beds in the overflow ward of cluster j for $j = 1, \ldots, m$.

As an example, consider the configuration in Figure 3.1. Following the notation introduced above, we have $B = 12$, $\mathcal{S} = \{1, 2, 3, 4\}$, $\boldsymbol{\mathcal{C}} = \{\mathcal{C}^1, \mathcal{C}^2\}$ with $\mathcal{C}^1 = \{1, 2\}$ and $\mathcal{C}^2 = \{3, 4\}$,

$\boldsymbol{d} = (2, 2, 3, 1)$ and $\boldsymbol{o} = (3, 1)$.

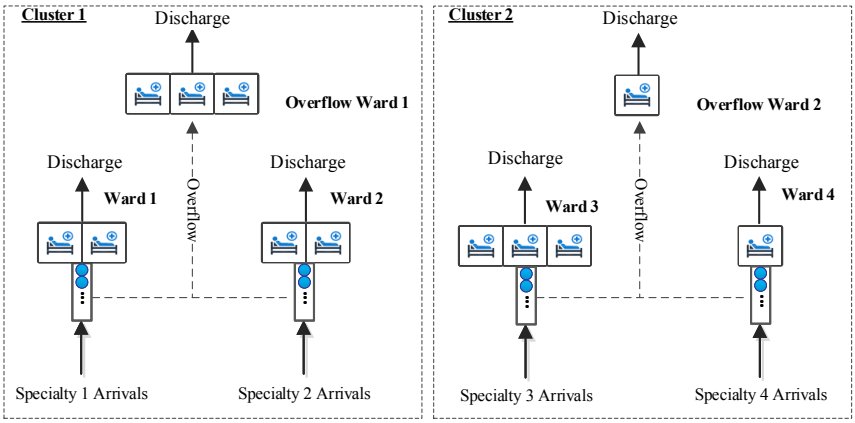


FIGURE 3.1: An example for the COF configuration

We now explain how COF represents the other four major configurations discussed in Chapter 1. A fully dedicated configuration is a COF configuration with each specialty assigned to a cluster, i.e., $\boldsymbol{\mathcal{C}} = \{\{1\}, \ldots, \{n\}\}$. With only one specialty in each cluster, there is no point in having both dedicated and overflow wards, hence we assume $\boldsymbol{d} > \boldsymbol{0}$ and $\boldsymbol{o} = \boldsymbol{0}$ for a fully dedicated configuration, where $\boldsymbol{0}$ is a vector of zeros with appropriate size. Earmarking is a COF configuration with all specialties allocated to one cluster, i.e., $\boldsymbol{\mathcal{C}} = \{\mathcal{S}\}$, $\boldsymbol{d} \geq \boldsymbol{0}$, and $\boldsymbol{o} = o^1 > 0$. A fully flexible configuration is a COF configuration with all specialties allocated to one cluster but with no dedicated wards, i.e. $\boldsymbol{\mathcal{C}} = \{\mathcal{S}\}$, $\boldsymbol{d} = \boldsymbol{0}$, and $\boldsymbol{o} = B$. The wing formation configuration is a COF configuration with wings acting as clusters and no dedicated wards in any of the clusters, i.e., $\boldsymbol{\mathcal{C}} = \{\mathcal{C}^1, \ldots, \mathcal{C}^m\}$, $\boldsymbol{d} = \boldsymbol{0}$, and $\boldsymbol{o} > \boldsymbol{0}$.

3.3 Optimization Formulation

We assume patients of each specialty request admission according to a stationary Poisson process, independently from other specialties, and their LOSs are independent and identically distributed (i. i. d.) as Exponential random variables. Both assumptions are followed in bulk of the bed allocation literature; see, for example, Best et al. (2015) and Bekker et al. (2016). We denote the rate of admission request for specialty i patients by λ_i. To capture the impact of focus and workload on LOS, we represent the mean LOS for specialty i patients admitted to a $d-$bed ward shared by a subset $\mathcal{A} \ni i$ of specialties by function $\nu_i(d, \mathcal{A})$. For systems with an IWTT, we assume patients wait in their queues until they are served. For systems with a FWTT, we assume waiting time thresholds are i. i. d. according to an Exponential distribution with rate γ for all specialties. This is also the assumption followed in Best et al. (2015). We assume arrival, service, and abandonment processes are mutually independent.

We highlight that some of the assumptions made above may not hold in reality. For example, admission requests may not be Poisson for some specialties, LOSs may not be Exponentially distributed, and patients of different specialties may have different waiting time thresholds. As we show in Chapter 4, however, these assumptions simplify the underlying queueing models, enabling us to develop analytical approximations for their performance evaluation.

Identifying the optimal configuration and the corresponding bed allocation for the COF configuration requires 4 sets of decision variables: i) the number of clusters m to create; ii) the set of specialties $\mathcal{C}^j$ to assign to cluster j for $j = 1, \ldots, m$; iii) the number of beds d_i to allocate to the ward dedicated to speciality i for $i \in \mathcal{S}$; and iv) the number of beds o^j to allocate to the overflow ward of cluster j for $j = 1, \ldots, m$. We define the optimal configuration as the one minimizing the mean total daily cost including the cost of patients abandoning (waiting in) the queue for FWTT (IWTT) assumption, and the cost of nursing teams. Cost minimization formulations are common in the bed allocation literature; see, for example, Izady and Muhamed (2019) and Wu et al. (2019) . The optimization problem is formulated as

$$Z = \min_{(m,\boldsymbol{\mathcal{C}},\mathbf{d},\mathbf{o})} \left\{ \sum_{j=1}^{m} T(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j) : \sum_{i=1}^{n} d_i + \sum_{j=1}^{m} o^j \leq B, \right.$$
$$\left. \boldsymbol{\mathcal{C}} \text{ is a feasible partition of } \mathcal{S}, m \in \mathbb{Z}_+, \mathbf{d} \in \mathbb{Z}^n \text{ and } \mathbf{o} \in \mathbb{Z}^m \right\}, \quad (3.1)$$

where $T(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j)$ gives the mean total daily cost of cluster $\mathcal{C}^j$ with $(d_i;\, i \in \mathcal{C}^j)$ dedicated beds and o^j overflow beds. In problem (3.1), we restrict clustering to feasible partitions of specialties. This is because, as we illustrate in Chapter 6, some partitions may not be feasible due to the limitations imposed by hospital layout or clinical requirements of different specialties. We propose

Proposition 1. *The optimization model in* (3.1) *is equivalent to*

$$Z = \min_{(m,\boldsymbol{b},\boldsymbol{\mathcal{C}})} \left\{ \sum_{j=1}^{m} \phi(\mathcal{C}^j, b^j) : (m, \boldsymbol{b}, \boldsymbol{\mathcal{C}}) \in \Psi \right\}, \quad (3.2)$$

where $\boldsymbol{b} = (b^1, \ldots, b^m)$,

$$\Psi = \left\{ (m, \boldsymbol{b}, \boldsymbol{\mathcal{C}}) : \sum_{j=1}^{m} b^j \leq B, \boldsymbol{\mathcal{C}} \textit{ is a feasible partition of } \mathcal{S}, m \in \mathbb{Z}_+, \textit{and } \boldsymbol{b} \in \mathbb{Z}^m \right\},$$

and

$$\phi(\mathcal{C}^j, b^j) = \min_{(d_i; i\in\mathcal{C}^j),\, o^j} \left\{ T(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j) : o^j + \sum_{i\in\mathcal{C}^j} d_i \le b^j, o^j \in \mathbb{Z}, \right.$$
$$\left. \text{and } d_i \in \mathbb{Z} \text{ for } i \in \mathcal{C}^j \right\}. \quad (3.3)$$

Proof. Inserting the model in (3.3) in the objective function in (3.2), we arrive at

$$Z = \min_{(m,\boldsymbol{b},\boldsymbol{\mathcal{C}})} \left\{ \sum_{j=1}^{m} \left(\min_{(d_i; i\in\mathcal{C}^j),\, o^j} \left\{ T(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j) : o^j + \sum_{i\in\mathcal{C}^j} d_i \le b^j \right\} \right) : \right.$$
$$\left. (m, \boldsymbol{b}, \boldsymbol{\mathcal{C}}) \in \Psi \right\},$$

where the integrality constraints on o^j and d_i's are suppressed for brevity. Since the decision variables for each of the m optimization models inside the first summation are different, the model above can be written as

$$Z = \min_{(m,\boldsymbol{b},\boldsymbol{\mathcal{C}})} \left\{ \min_{\boldsymbol{d},\boldsymbol{o}} \left\{ \sum_{j=1}^{m} T(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j) : o^j + \sum_{i\in\mathcal{C}^j} d_i \le b^j \text{ for } j = 1, \ldots, m \right\} : \right.$$
$$\left. (m, \boldsymbol{b}, \boldsymbol{\mathcal{C}}) \in \Psi \right\},$$

which is equivalent to

$$Z = \min_{(m,\boldsymbol{b},\boldsymbol{\mathcal{C}},\boldsymbol{d},\boldsymbol{o})} \left\{ \sum_{j=1}^{m} T(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j) : o^j + \sum_{i\in\mathcal{C}^j} d_i \le b^j \text{for } j = 1, \ldots, m, \right.$$
$$\left. \text{and } (m, \boldsymbol{b}, \boldsymbol{\mathcal{C}}) \in \Psi \right\}. \quad (3.4)$$

Since in the model above, the feasible region specified by the set of inequality constraints, $o^j + \sum_{i\in\mathcal{C}^j} d_i \le b^j$ for $j = 1, \ldots, m$, and $\sum_{j=1}^{m} b^j \le B$, is the same as the feasible region specified by the single constraint $\sum_{i=1}^{n} d_i + \sum_{j=1}^{m} o^j \le B$, we can replace the former set of constraints with the latter single constraint. We can also remove $\boldsymbol{b}$ from decision variables in the model given in (3.4) as it does not appear in the objective function. This yields the optimization model in (3.1). □

Note that in the COF configuration, clusters operate independently from each other as there is no sharing of patients or staff between them. The intuition behind Proposition 1 is

that due to this independence, the optimal allocation of a given number of beds among different wards of a cluster is not influenced by the allocation of beds among wards in other clusters.

Proposition 1 implies that the partitioning and bed allocation problem given in (3.1) can be decomposed into the partitioning and inter-cluster bed allocation problem given in (3.2), and the intra-cluster bed allocation problem given in (3.3). The former aims to partition the set of specialties into a number of clusters and identify the number of beds allocated to each cluster, and the latter aims to divide a given number of beds among dedicated and overflow wards of a cluster.

To evaluate the cost function T in Equation (3.3), we assume a cost of c_w is incurred each day a patient waits in the queue for systems with an IWTT assumption, and a cost of c_a is incurred for each patient abandoning the queue (as a result of her waiting time exceeding the threshold) for systems with an FWTT assumption. This gives

$$T(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j) = c_w Q(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j) + R(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j), \tag{3.5}$$

and

$$T(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j) = c_a B(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j) + R(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j), \tag{3.6}$$

for IWTT and FWTT systems, respectively, where R represents the mean daily cost of nursing, Q the mean number of patients waiting in the queue, and B the mean daily number of patients abandoning the system, all for cluster $\mathcal{C}^j$ with bed allocation $(d_i; i \in \mathcal{C}^j), o^j$. Note that for analytical simplicity, we did not include the waiting cost for FWTT systems. It is now easy to show that

$$Q(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j) = \sum_{k \in \mathcal{C}_j} \lambda_k W_k(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j), \tag{3.7}$$

$$B(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j) = \sum_{k \in \mathcal{C}_j} \lambda_k A_k(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j), \tag{3.8}$$

where W_k and A_k give the mean waiting time and probability of abandonment, respectively, of speciality $k \in \mathcal{C}^j$ patients. To evaluate the nursing cost R, we apply the minimum nurse-to-patient ratio approach following Izady and Muhamed (2019). As they explain, this approach makes analytical calculations easier, and is also the most common method for establishing nursing requirements in hospitals. Denote by f_i the desired nurse-to-patient ratio for specialty $i \in \mathcal{S}$ patients, and let $r(\mathcal{A})$ be the daily cost of a nurse working in a ward admitting patients

of a subset $\mathcal{A}$ of specialties. We then have

$$R(\mathcal{C}^j, (d_i;\, i \in \mathcal{C}^j),\, o^j) = \sum_{k \in \mathcal{C}_j} r(\{k\}) \left\lceil S_k^d(\mathcal{C}^j, (d_i; i \in \mathcal{C}^j), o^j) f_k \right\rceil + r(\mathcal{C}^j) \left\lceil \sum_{k \in \mathcal{C}^j} S_k^o(\mathcal{C}^j, (d_i; i \in \mathcal{C}^j), o^j) f_k \right\rceil, \quad (3.9)$$

where functions $S_k^d(\mathcal{C}^j, (d_i; i \in \mathcal{C}^j), o^j)$ and $S_k^o(\mathcal{C}^j, (d_i; i \in \mathcal{C}^j), o^j)$ give the mean number of beds occupied by specialty $k \in \mathcal{C}^j$ patients in their dedicated and overflow wards, respectively, and $\lceil x \rceil$ gives the smallest integer larger than or equal to x.

3.4 Conclusion

In this chapter, we provided a mathematical representation for the COF configuration, formulated a cost-minimization optimization model for finding the best partitioning and bed allocation for this configuration, and explained the underlying assumptions of our model. We reformulated the optimization model by breaking it into a partitioning and inter-cluster bed allocation model, and an intra-cluster bed allocation model.

The intra-cluster allocation model requires to evaluate a set of performance metrics, including the abandonment probability for FWTT systems, mean waiting time for IWTT systems, and mean busy servers in dedicated and overflow wards for both IWTT and FWTT systems, for a given cluster with a given bed allocation. Each cluster in the COF configuration can be modelled as an overflow queueing system as we shall explain the next chapter. In order to evaluate the desired performance metrics, we therefore need to analyse overflow queueing models. This is covered in the next chapter, wherein approximation methodologies are proposed for estimating key performance metrics.

4 The Performance Evaluation Models

4.1 Introduction

In this chapter, we develop methodologies for performance evaluation of overflow delay and abandonment systems, and investigate their speed and accuracy. We define an overflow system as a hierarchical multi-class queueing system with two types of server pools: i) dedicated pools, each specialized to serve a single class of customers; and ii) an overflow pool, cross-trained to serve all classes of customers. Upon arrival, customers will be served by their dedicated pool if it has an idle server available, and the overflow pool otherwise. If both pools are busy, customers wait in dedicated queues corresponding to their classes; see Figure 4.1. We assume that once a server in a dedicated pool becomes available, it serves the next customer in its dedicated queue according to a first-come first-served (FCFS) discipline. Once a server in the overflow pool becomes available, on the other hand, we assume that it serves the next customer from the longest queue following a FCFS discipline. The longest queue policy is found to outperform the other major policies in Sheikhzadeh et al. (1998) and Jordan et al. (2004). The FCFS discipline is a common assumption in many application areas, but it may not be realistic if customers joining a queue have different levels of urgency.

In overflow delay systems, customers are infinitely patient and wait in their queues until they are served. In overflow abandonment systems, customers are impatient and abandon the queue once their waiting time threshold is reached. A cluster in the COF configuration with an IWTT can therefore be represented as an overflow delay system, while a cluster with a FWTT can be represented as an overflow abandonment system.

Suppose there are n customer classes with $\mathcal{S} = \{1, 2, \ldots, n\}$ the corresponding index set. Let d_i be the number of servers in the pool dedicated to class i customers for $i \in \mathcal{S}$, and $\boldsymbol{d} = (d_1, \ldots, d_n)$ be the corresponding vector. Let o be the number of serves in the overflow pool. Suppose class i customers arrive to the system according to a Poisson process with rate λ_i, and their service times are i. i. d. as an Exponential distribution with mean ν_i if service is provided by the corresponding dedicated pool, and ν_i' if service is provided by the overflow

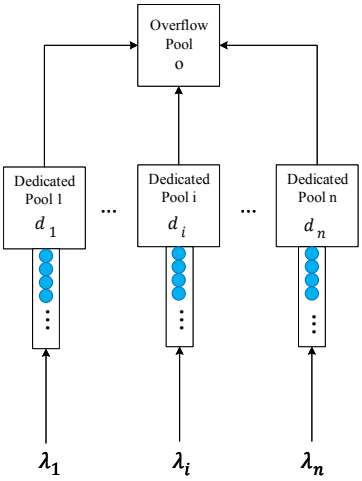


FIGURE 4.1: The schematic diagram of an overflow system

pool, for $i \in \mathcal{S}$. For the overflow abandonment system, in addition to the assumptions above, we assume times to abandon are i. i. d. following an exponential distribution with rate γ for all customer classes. Arrival, service and abandonment processes are assumed to be mutually independent. Following the formulation proposed in Chapter 3, we need to evaluate mean waiting time in the queue for the overflow delay system, and the probability of abandonment for the overflow abandonment system, for each class of customers. To evaluate nursing costs, we also need to evaluate mean busy servers in dedicated and overflow pools for both systems.

To demonstrate the complexities of performance evaluation in overflow systems, we provide exact analysis for a simplified overflow delay system in Section 4.2. In Section 4.3, we propose approximation methodologies for estimating performance metrics of overflow delays systems as well as overflow abandonment systems. The accuracy and speed of these approximations are investigated in Section 4.4, followed by our conclusions in Section 4.5.

4.2 Exact Analysis

We consider an overflow delay system with only two customer classes, i.e., $\mathcal{S} = \{1, 2\}$. We also assume that each server pool has only one server, i.e., $\boldsymbol{d} = (1, 1)$ and $o = 1$. We denote the service rate of class i customers in the their dedicated server by $\mu_i = 1/\nu_i$, and in the overflow server by $\mu_i' = 1/\nu_i'$, for $i \in \mathcal{S}$. We define the state of the system by a four dimensional vector (i, j, k, l), where $i(k)$ is the number of class one (two) customers being served by the first (second) dedicated server plus those waiting in the first (second) queue, and $j(l)$ is a binary variable equal to 0 if no type one (two) customer is in the overflow server and 1 otherwise. With Poisson arrivals and exponential service times, the system state evolves as a continuous time

Markov chain (CTMC) over time. The steady-state probabilities can therefore be obtained by solving the corresponding balance equations (see, e.g., Gross et al. 2011).

Given the complex nature of transitions of the proposed CTMC, we need a structured framework to derive the balance equations. We follow the SERT framework proposed in Gross and Miller (1984). It is called SERT as it involves four major sets: S (state space), E (set of event types), R (set of transition rates), and T (set of target states). The state space of the system is defined as

$$\mathcal{P} = \{(i,j,k,l) \in \mathbb{Z}^4 : i,k \leq N, j+l \leq 1, j+l \geq \min\{1, \max\{i-1, k-1\}\}\}, \tag{4.1}$$

where the first constraint sets an upper bound N on the second and fourth dimensions to make the state space finite, the second constraint ensures that there is no more than one customer in the overflow server, and the last constraint guarantees that the overflow server is not idle when there is at least one customer waiting in either first or second queue. Simple calculations show that size of $\mathcal{P}$ is $2(N+1)^2 + 4$. The set of event types is defined as

$$\mathcal{E} = \{e_1, e_2, e_3, e_4, e_5, e_6\}, \tag{4.2}$$

where e_1 (e_2) is the arrival of a class one (two) customer, e_3 (e_4) is the departure of a class one (two) customer from the first (second) dedicated server, and e_5 (e_6) is the departure of class one (two) customer from the overflow server. The set of transition rates is defined as

$$\mathcal{Q} = \{q_p^e : p \in \mathcal{P}, e \in \mathcal{E}\}, \tag{4.3}$$

where

$$q_p^{e_1} = \begin{cases} \lambda_1, & \text{for } p \in \{(i,j,k,l) \in \mathcal{P}, i \neq N\} \\ 0, & \text{otherwise,} \end{cases} \qquad q_p^{e_2} = \begin{cases} \lambda_2, & \text{for } p \in \{(i,j,k,l) \in \mathcal{P}, k \neq N\} \\ 0, & \text{otherwise,} \end{cases}$$

$$q_p^{e_3} = \begin{cases} \mu_1, & \text{for } p \in \{(i,j,k,l) \in \mathcal{P}, i \neq 0\} \\ 0, & \text{otherwise,} \end{cases} \qquad q_p^{e_4} = \begin{cases} \mu_2, & \text{for } p \in \{(i,j,k,l) \in \mathcal{P}, k \neq 0\} \\ 0, & \text{otherwise,} \end{cases}$$

$$q_p^{e_5} = \begin{cases} \mu_1', & \text{for } p \in \{(i,j,k,l) \in \mathcal{P}, j \neq 0\} \\ 0, & \text{otherwise,} \end{cases} \qquad q_p^{e_6} = \begin{cases} \mu_2', & \text{for } p \in \{(i,j,k,l) \in \mathcal{P}, l \neq 0\} \\ 0, & \text{otherwise} \end{cases}$$

The transition rate out of any state $p \in \mathcal{P}$ is therefore evaluated as

$$r_p = \sum_{e \in \mathcal{E}} q_p^e.$$

The set of target states is defined as

$$\mathcal{T} = \{g_p^e : p \in \mathcal{P}, e \in \mathcal{E}\}, \tag{4.4}$$

where

$$g_p^{e_1} = \begin{cases} (i+1,j,k,l), & p \in \{(i,j,k,l) \in \mathcal{P}, i < N, j+l = 1\} \cup \{(i,j,k,l) \in \mathcal{P}, i = j = l = 0\}, \\ (1,1,k,0), & p \in \{(i,j,k,l) \in \mathcal{P}, i = 1, j = l = 0\}, \end{cases}$$

$$g_p^{e_2} = \begin{cases} (i,j,k+1,l), & p \in \{(i,j,k,l) \in \mathcal{P}, k < N, , j+l = 1\} \cup \{(i,j,k,l) \in \mathcal{P}, j = l = k = 0\}, \\ (i,0,1,1), & p = \{(i,j,k,l) \in \mathcal{P}, j = l = 0, k = 1\}, \end{cases}$$

$$g_p^{e_3} = (i-1,j,k,l), \quad p \in \{(i,j,k,l) \in \mathcal{P}, i > 0\},$$

$$g_p^{e_4} = (i,j,k-1,l), \quad p \in \{(i,j,k,l) \in \mathcal{P}, k > 0\}, \tag{4.5}$$

$$g_p^{e_5} = \begin{cases} (i-1,1,k,0), & p \in \{(i,j,k,l) \in \mathcal{P}, i \geq 2, j = 1, k \leq i\}, \\ (i,0,k-1,1), & p \in \{(i,j,k,l) \in \mathcal{P}, i < k, j = 1, k \geq 2\}, \\ (i,0,k,0), & p \in \{(i,j,k,l) \in \mathcal{P}, i \leq 1, j = 1, k \leq 1\}, \end{cases}$$

$$g_p^{e_6} = \begin{cases} (i-1,1,k,0), & p \in \{(i,j,k,l) \in \mathcal{P}, i \geq 2, k \leq i, l = 1\}, \\ (i,0,k-1,1), & p \in \{(i,j,k,l) \in \mathcal{P}, i < k, k \geq 2, l = 1\}, \\ (i,0,k,0), & p \in \{(i,j,k,l) \in \mathcal{P}, i \leq 1, k \leq 1, l = 1\}. \end{cases}$$

The first case for $g_p^{e_1}$ ($g_p^{e_2}$) corresponds to situations where there is $N-2$ or fewer customers in the first (second) queue and the overflow server is busy, or the overflow server and the first (second) dedicated server are both idle. In these situations, a new arrival of class one (two) increases the first (third) dimension of the state vector by one. The second case for $g_p^{e_1}$ ($g_p^{e_2}$) corresponds to situations where the first (second) dedicated server is busy but the overflow server is idle so the new arrival starts service in the overflow server and thus the second (fourth) dimension increases to one. $g_p^{e_3}$ ($g_p^{e_4}$) reduce the first (third) dimension of the state vector by one as long as there is at least one customer being served by the first (second) server. The first case for $g_p^{e_5}$ ($g_p^{e_6}$) corresponds to situations where the overflow server finishes serving a class one (two) customer, there is at least one customer waiting in the first queue, and the first queue is larger than or equal to the second queue. In these situations, the first customer waiting in the first queue starts service in the overflow server and so the first dimension decreases by one whereas the second dimension remains one. The second case for $g_p^{e_5}$ ($g_p^{e_6}$) corresponds to situations where the overflow server finishes serving a class one (two) customer, there is at least one customer waiting in the second queue, and the second queue is

larger than the first queue. In these situations, the first customer waiting in the second queue starts service in the overflow server and so the third dimension decreases by one whereas the fourth dimension remains one. The last case for $g_p^{e_5}$ ($g_p^{e_6}$) corresponds to situations where the overflow server finishes serving a class one (two) customer and there is no customer waiting in the first or second queue. In these situations, the overflow server becomes idle once the departure occurs.

Given the sets $\mathcal{P}$, $\mathcal{E}$, $\mathcal{Q}$ and $\mathcal{T}$ defined in Equations (4.1), (4.2),(4.3) and (4.4), the balance equations can be written as

$$\pi_p r_p = \sum_{\substack{e\in\mathcal{E},p'\in\mathcal{P}\\ g_{p'}^e=p}} \pi_{p'} q_{p'}^e, \tag{4.6}$$

where π_p is the steady-state probability of the CTMC being in state $p \in \mathcal{P}$. Solving the balance equations along with the normalizing equation $\sum_{p\in\mathcal{P}} \pi_p = 1$ produces the steady-state probabilities. The length of the first and second queues are then obtained as

$$L_1^q = \sum_{p\in\{(i,j,k,l)\in\mathcal{P}:i>0\}} (i-1)\pi_p, \quad L_2^q = \sum_{p\in\{(i,j,k,l)\in\mathcal{P}:k>0\}} (k-1)\pi_p, \tag{4.7}$$

respectively. Applying Little's law (Gross et al. 2011) yields the mean waiting time of class i customers as

$$W_i^q = \frac{L_i^q}{\lambda_i(1-p_i^{loss})}, \tag{4.8}$$

for $i \in \mathcal{S}$, where

$$p_1^{loss} = \sum_{p\in\{(i,j,k,l)\in\mathcal{P}:i=N\}} \pi_p, \quad p_2^{loss} = \sum_{p\in\{(i,j,k,l)\in\mathcal{P}:k=N\}} \pi_p, \tag{4.9}$$

are the probabilities of a class one and class two customer being turned away, respectively.

To solve the steady-state equations numerically, we first map the multi-dimensional representation of the state space given in (4.1) into a single-dimension vector. This is because storing and retrieving elements from single-dimension vectors are more efficient than multi-dimensional arrays in computers (Izady 2010). To do this, we sort the elements of $\mathcal{P}$ in the order given in Table 4.1, and index each one starting from one.

Given the indexing in Table 4.1, the function

$$\iota((i,j,k,l)) = j(N+1)^2 + i(N+1) + k + 5, \tag{4.10}$$

index	State			
	i	j	k	l
1	0	0	0	0
2	0	0	1	0
3	1	0	0	0
4	1	0	1	0
5	0	0	0	1
6	0	0	1	1
⋮	⋮	⋮	⋮	⋮
$4+N+1$	0	0	N	1
$4+(N+1)+1$	1	0	0	1
$4+(N+1)+2$	1	0	1	1
⋮	⋮	⋮	⋮	⋮
$4+2(N+1)$	1	0	N	1
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
$4+(N+1)N+1$	N	0	0	1
$4+(N+1)N+2$	N	0	1	1
⋮	⋮	⋮	⋮	⋮
$4+(N+1)^2$	N	0	N	1
$4+(N+1)^2+1$	0	1	0	0
$4+(N+1)^2+2$	0	1	1	0
⋮	⋮	⋮	⋮	⋮
$4+(N+1)^2+(N+1)$	0	1	N	0
$4+(N+1)^2+(N+1)+1$	1	1	0	0
$4+(N+1)^2+(N+1)+2$	1	1	1	0
⋮	⋮	⋮	⋮	⋮
$4+(N+1)^2+2(N+1)$	1	1	N	0
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
$4+(N+1)^2+N(N+1)+1$	N	1	0	0
$4+(N+1)^2+N(N+1)+2$	N	1	1	0
⋮	⋮	⋮	⋮	⋮
$4+(N+1)^2+(N+1)^2$	N	1	N	0

TABLE 4.1: Index conversion table for elements in $\mathcal{P}$.

translates each element of the state space into its one-dimensional index. Conversely, each index $1 \leq v \leq 2(N+1)^2 + 4$ is converted into its four-dimensional representation through

$$\iota^{-1}(v) = \begin{cases} (0,0,0,0), & v = 1, \\ (0,0,1,0), & v = 2, \\ (1,0,0,0), & v = 3, \\ (1,0,1,0), & v = 4, \\ (i,j,k,l), & v \geq 5, \end{cases} \tag{4.11}$$

where $j = \left\lfloor \frac{v-4-1}{(N+1)^2} \right\rfloor$, $i = \left\lfloor \frac{v-4-1-j(N+1)^2}{(N+1)} \right\rfloor$, $k = v - 4 - j(N+1)^2 - i(N+1) - 1$, and $l = 0^j$ with 0^0 defined as 1.

We run a set of numerical experiments to evaluate the computational speed of the exact method discussed above. We consider 20 scenarios with λ_1 and λ_2 values set as in Table 4.2. The service rates of type 1 and 2 customers, respectively, are set to $\mu_1 = 0.85$ and $\mu_2 = 2$ for dedicated pools, and $\mu_1^{'} = 0.8$ and $\mu_2^{'} = 1.8$ for the overflow pool. In order to choose a sufficiently large value for N for each scenario, we start with $N = 16$, and increase it in steps of 2 until p_1^{loss} and p_2^{loss} are both less than 0.001. The N values, average waiting time for each customer class, and computation times for all the scenarios are given in Table 4.2.

The computation times provided in Table 4.2 show an exponential increase as the system size grows, with a system of size around 40 taking around 2 hours to produce the results. This result implies that an exact performance evaluation method would not be practical for finding the optimal bed allocation within a cluster, which may potentially include more than two specialties and require a larger system size. We therefore propose approximation methodologies for performance evaluation of hierarchical queues in the next section.

4.3 Approximation Models

In this section, we propose approximation methodologies for estimating performance metrics of overflow delay systems as well as overflow abandonment systems. Our approximations have two main steps. For both delay and abandonment systems, the first step involves estimating the blocking probabilities of different customer classes in an equivalent overflow loss system. An overflow loss system is an overflow abandonment system with a zero waiting time threshold. This implies that there is no waiting provision in this system so arrivals finding both dedicated and overflow pools busy are turned away immediately. For overflow delay systems, the second step of our approximation involves converting the estimated blocking probabilities in the overflow loss system to mean waiting times in the overflow delay system using the exact

Scenario	λ_1	λ_2	N	W_1	W_2	Time (s)
1	1.20	1.50	18	1.59	0.47	45
2	1.24		20	1.80	0.50	75
3	1.27		22	2.04	0.54	117
4	1.31		26	2.35	0.59	188
5	1.35		28	2.76	0.64	446
6	1.39		30	3.26	0.70	701
7	1.43		36	4.01	0.78	1988
8	1.48		44	5.13	0.87	7764
9	1.52		52	6.81	0.98	23154
10	1.57		62	9.47	1.11	45985
11	1.20	1.50	16	1.56	0.46	25
12		1.62	18	1.66	0.55	40
13		1.75	20	1.78	0.65	70
14		1.89	20	1.93	0.79	71
15		2.04	22	2.16	0.97	112
16		2.20	22	2.51	1.21	113
17		2.38	24	3.13	1.59	175
18		2.57	32	4.43	2.26	936
19		2.78	42	7.81	3.81	7654
20		3.00	226	78.95	32.61	345600

TABLE 4.2: Results of the numerical experiments with the exact performance evaluation model.

relation between loss probability in single-class $M/M/s/0$ loss queues and mean waiting time in single-class $M/M/s/\infty$ delay queues. For overflow abandonment systems, the second step involves converting the estimated blocking probabilities in the overflow loss system to abandonment probabilities in the overflow abandonment system using the exact relation between loss probability in single-class $M/M/s/0$ loss queues and abandonment probability in single-class $M/M/s/\infty + M$ abandonment queues.

Section 4.3.1 explains how loss probabilities are estimated in overflow loss systems. The approximations for evaluating performance in overflow delay and abandonment systems are elaborated in Sections 4.3.2 and 4.3.3, respectively.

4.3.1 Overflow Loss Systems

In this section, we estimate the blocking probabilities faced by different customer classes in an overflow loss system using the approximation proposed in Izady and Muhamed (2019). This approximation accounts for mean service times being different in dedicated and overflow pools, which is essential for capturing the impact of focus on LOS. We provide an outline of the approximation for completeness. Consider a single-class loss system with $d \in \mathbb{Z}_+$ servers, arrival process $A(t)$ with rate $\lambda > 0$, and service time cumulative distribution function (CDF) $F(t)$ with mean $\nu > 0$. By Hayward's approximation (see Fredericks 1980), the blocking probability in this system is estimated as

$$B(A(t), F(t), d) \approx B_H(a, z, d) = B_e(a/z, d/z), \tag{4.12}$$

where $a = \lambda\nu$ is the offered load, z is the "peakedness" measure, and $B_e(a, d)$ is a continuous extension of the Erlang loss function, such as the one proposed in Jagerman (1974) as follows

$$B_e(a, d) = \left(a \int_0^\infty \exp(-at)(1+t)^d dt \right)^{-1}. \tag{4.13}$$

The peakedness z is defined as the variance to mean ratio for the steady-state number of busy servers in an equivalent infinite-server system, i.e., an infinite-server system with arrival process $A(t)$ and service time CDF $F(t)$ (Fredericks 1980). It is a measure of variability in the arrival process as evaluated by its impact on queueing performance. When the arrival process is Poisson, it is easy to show that the distribution for the number of busy servers in an infinite-server system is Poisson regardless of the service time distribution (see Gross et al. 2011, Chap. 3), and so $z = 1$. This turns Equation (4.12) to the well-known Erlang loss formula, which is exact. When the arrival process is not Poisson, by Equation (4.12), it

suffices to characterize the arrival and service processes through the (a, z) pair to approximate the blocking probability.

Now consider an overflow loss system with the same arrival and service processes as the overflow delay or abandonment systems specified in Section 4.1. Since the arrivals are assumed to be Poisson, the blocking probability of class i customers in their dedicated pool is obtained as $B_e(a_i, d_i)$, where $a_i = \lambda_i \nu_i$ is the offered load, for $i \in \mathcal{S}$. To evaluate the blocking probability faced by class i customers at the overflow pool by Haywards's approximation, we need to evaluate the offered load and peakedness of the aggregate stream overflowing from dedicated pools to the overflow pool. The offered load of the stream overflowing dedicated pool i is $a_i B_e(a_i, d_i)/\rho_i$, where $\rho_i = \nu_i/\nu_i^{'}$ is the ratio of mean service time in the dedicated pool to the overflow pool. The offered load of the aggregate overflow stream is therefore

$$\alpha = \sum_{i \in \mathcal{S}} a_i B_e(a_i, d_i)/\rho_i. \tag{4.14}$$

Following Proposition 3 in Izady and Muhamed (2019), the peakedness of the stream overflowing from dedicated pool i is obtained as

$$\xi(a_i, d_i, \rho_i) = 1 - \frac{a_i B_e(a_i, d_i)}{\rho_i} + \frac{a_i(a_i+\rho_i)\,{}_3F_1(\rho_i, 1-d_i, a_i+\rho_i+1; a_i+\rho_i; -1/a_i)}{\rho_i(a_i+\rho_i+1)\,{}_3F_1(1-d_i, \rho_i+1, 2+a_i+\rho_i; a_i+\rho_i+1; -1/a_i)}, \tag{4.15}$$

where ${}_pF_q(a_1, ..., a_p; b_1, ...b_q; x)$ is a generalized hypergeometric function (Petkovsek et al. 1997). The equation above is exact and captures the impact of different mean service times in dedicated and overflow pools through parameter ρ_i. From the peakedness definition, we can now obtain the variance of the number of servers occupied by class i customers in the infinite-server system associated with the overflow pool as $a_i B_e(a_i, d_i)\xi(a_i, d_i, \rho_i)/\rho_i$. Since customers do not interfere with each other in the infinite server system, the variance of the number of busy servers in the infinite server system will be $\sum_{i \in \mathcal{S}} a_i B_e(a_i, d_i)\xi(a_i, d_i, \rho_i)/\rho_i$, hence the peakedness of the aggregate overflow stream is

$$\beta = \frac{1}{\alpha} \sum_{i \in \mathcal{S}} \frac{a_i}{\rho_i} B_e(a_i, d_i)\xi(a_i, d_i, \rho_i). \tag{4.16}$$

Having obtained the offered load and peakedness, the blocking probability faced by the aggregated overflow stream is estimated as

$$L \approx B_e(\alpha/\beta, o/\beta). \tag{4.17}$$

The estimate of overall blocking probability faced by class i customers is therefore obtained as

$$L_i = B_e(a_i, d_i)L \approx B_e(a_i, d_i)B_e(\alpha/\beta, o/\beta). \tag{4.18}$$

4.3.2 Overflow Delay Systems

In this section, we estimate the probability of delay and mean waiting time experienced by class i customers, denoted by P_i and W_i, respectively, in an overflow delay system using the estimate of L_i provided in Equation (4.18). To start, consider an $M/M/s/\infty$ delay queue with offered load $a < s$. The probability of delay P in this queue is related to the loss probability L in an equivalent $M/M/s/0$ loss queue through the following equation (see Cooper 1990, Chap. 10, Equation 5.31)

$$P = \frac{sL}{s - a(1 - L)}. \tag{4.19}$$

Motivated by the equation above, Chevalier and Van den Schrieck (2009) propose

$$P_i \approx \frac{s_i L_i}{s_i - \delta_i(1 - L_i)}, \tag{4.20}$$

for estimating the delay probability faced by class i customers in an overflow delay system. In Equation (4.20), L_i is estimated through Equation (4.18), s_i is a suitably defined number of servers allocated to class i customers as we shall explain below, and δ_i is the offered load of class i customers in the overflow delay system evaluated by

$$\delta_i = \lambda_i \bar{\nu}_i, \tag{4.21}$$

where $\bar{\nu}_i$ is the weighted average of class i mean service times in dedicated and overflow pools. To compute $\bar{\nu}_i$, we approximate the fraction of class i customers served by the overflow pool in the overflow delay system by the corresponding fraction in the associated overflow loss system. This yields

$$\bar{\nu}_i = (1 - B_e(a_i, d_i)(1 - L))\nu_i + B_e(a_i, d_i)(1 - L)\nu_i', \tag{4.22}$$

with $B_e(a_i, d_i)$ and L given in Equations (4.13) and (4.17), respectively.

To evaluate s_i, we first define I_i^d as the idle service capacity in dedicated pool i, and I_i^o as the portion of idle service capacity in the overflow pool allocated to class i customers. Both metrics have the same dimension as arrival and service rates, i.e., they are measured in customers per time period. To estimate I_i^d and I_i^o, we combine the fluid approximation proposed by Chevalier and Van den Schrieck (2009) with the results from the overflow loss

system. In particular, we estimate I_i^d as

$$I_i^d \approx (d_i/\nu_i - (\lambda_i - \lambda_i'))^+, \tag{4.23}$$

where $x^+ = \max\{x, 0\}$, and λ_i' is the rate of customer overflow from dedicated pool i. Using the overflow loss system results, we estimate λ_i' as

$$\lambda_i' \approx \lambda_i B_e(a_i, d_i)(1 - L). \tag{4.24}$$

The overall idle capacity at the overflow pool is therefore

$$I^o \approx (o/\bar{\nu} - \sum_{i \in \mathcal{S}} \lambda_i')^+, \tag{4.25}$$

where $\bar{\nu}$ is the weighted average of mean service times in the overflow pool given by

$$\bar{\nu} = \sum_{i \in \mathcal{S}} \lambda_i' \nu_i' / \sum_{i \in \mathcal{S}} \lambda_i'. \tag{4.26}$$

Assuming customer classes with higher loss probabilities in the equivalent loss overflow system will have a higher proportion of the overflow pool idle capacity, we have

$$I_i^o \approx \frac{\lambda_i L_i}{\sum_{i \in \mathcal{S}} \lambda_i L_i} I^o, \tag{4.27}$$

with L_i given in Equation (4.18). Note that our estimation of I_i^d and λ_i' given in Equations (4.23) and (4.24), receptively, are more accurate than those of Chevalier and Van den Schrieck (2009), i.e., $I_i^d \approx (d_i/\nu_i - \lambda_i)^+$ and $\lambda_i' \approx (\lambda_i - d_i/\nu_i)^+$, which are purely based on fluid approximation. We now propose

$$s_i = I_i^o \bar{\nu} + \lambda_i' \nu_i' + d_i. \tag{4.28}$$

Equation (4.28) implies that s_i equals the size of dedicated pool i plus the sum of idle and used capacity of the overflow pool allocated to class i customers. We use Equation (4.28) instead of the conservation equation $s_i = (\lambda_i + I_i^d + I_i^o)\nu_i$ proposed by Chevalier and Van den Schrieck (2009) as it does not account for different mean service times in dedicated and overflow pools. Substituting s_i and δ_i in Equation (4.20) with their corresponding values given in Equations (4.28) and (4.21), respectively, and simplifying, we obtain

$$P_i = \frac{\left(I_i^o \bar{\nu} + \lambda_i' \nu_i' + d_i\right) L_i}{I_i^o \bar{\nu} + \lambda_i' \nu_i' + d_i - \lambda_i \bar{\nu}_i (1 - L_i)}. \tag{4.29}$$

The next step is to convert the delay probability P_i to mean waiting time W_i. For this, we use the relation $W = P/(s/\nu - \lambda)$ between delay probability P and mean waiting time W in an $M/M/s/\infty$ queue with arrival rate λ and mean service time ν (Cooper 1990, Chap. 10, Equation 5.31). The denominator of this relation is in fact the idle service capacity. Motivated by this, we have the following approximation

$$W_i \approx \frac{P_i}{I_i^d + I^o}. \tag{4.30}$$

Note that if $I_i^d + I^o = 0$, the system is unstable for class i customers and so $P_i = 1$ and $W_i \to \infty$. Applying Little's law on server pools, we obtain the mean number of class i customers in the dedicated and overflow pools by

$$S_i^d \approx \lambda_i(1 - B_e(a_i, d_i)(1 - L))\nu_i, \tag{4.31}$$

and

$$S_i^o \approx \lambda_i B_e(a_i, d_i)(1 - L)\nu_i^{'}, \tag{4.32}$$

respectively, where the fraction of customers served by the overflow pool is estimated by the corresponding fraction in the overflow loss system.

4.3.3 Overflow Abandonment Systems

In this section, we propose an approximation for estimating the abandonment probability of each customer class in an overflow abandonment system. To start, consider an $M/M/s/\infty + M$ abandonment queue with arrival rate λ, mean service time ν, and abandonment rate γ. The abandonment probability A in this queue is related to the loss probability L in an equivalent $M/M/s/0$ loss queue through the following equation (see Equation 5.22 in Zhang 2010)

$$A = \frac{s(1 + f(c, \eta)(\eta/c - 1))L}{\lambda\nu\left(1 + (f(c, \eta) - 1)L\right)}, \tag{4.33}$$

where

$$f(c, \eta) = \sum_{i=0}^{\infty} \frac{\Gamma(c + 1)\eta^i}{\Gamma(c + i + 1)},$$

with $c = s/\nu\gamma$, $\eta = \lambda/\gamma$, and $\Gamma(x) = \int_0^\infty y^{x-1}e^{-y}dy$ is the Gamma function. Motivated by Equation (4.33), we propose the approximation

$$A_i \approx \frac{s_i(1 + f(c_i, \eta_i)(\eta_i/c_i - 1))L_i}{\lambda_i\bar{\nu}_i\left(1 + (f(c_i, \eta_i) - 1)L_i\right)}, \tag{4.34}$$

for the abandonment probability experienced by class i customers in the overflow abandonment system, where s_i is a suitably defined number of servers for class i customers, $c_i = s_i/\bar{\nu}_i\gamma$, $\eta_i = \lambda_i/\gamma$, $\bar{\nu}_i$ is the average mean service time of class i customers given in Equation (4.22), and L_i is estimated through Equation (4.18). For s_i, we use the same equation as for the overflow delay system, i.e.,

$$s_i = I_i^o \bar{\nu} + \lambda_i^{'} \nu_i^{\prime} + d_i, \tag{4.35}$$

with I_i^o, $\bar{\nu}$, and $\lambda_i^{'}$ given in Equations (4.27), (4.26), and (4.24), respectively. Once abandonment probabilities are obtained, we can use Equations (4.31) and (4.32) to estimate the mean number of class i customers in the dedicated and overflow pools, respectively, by replacing arrival rate λ_i with effective arrival rate $\lambda_i(1 - A_i)$.

4.4 Numerical Experiment

In order to investigate the accuracy of the approximation methodologies proposed in Sections 4.3.2 and 4.3.3, we compare the results obtained from these methodologies with those obtained from a discrete-event simulation model. We use the testing scenarios given in Chevalier and Van den Schrieck (2009) and expand them to include different values for γ and ρ. In particular, we consider an overflow system with $\mathcal{S} = \{1, 2\}$, and evaluate the relative accuracy for scenarios illustrated in Table 4.3 with $\rho \in \{1.0, 0.9, 0.8\}$ for the overflow delay system, and $\rho \in \{1.0, 0.8\}$ and $\gamma^{-1} \in \{5, 10, 30\}$ for the overflow abandonment system. This yields a total of 390 and 780 scenarios for the overflow delay and abandonment systems, respectively. Note that we only consider $\rho \leq 1$ in our scenarios as in our inpatient bed system, mean LOS may only increase as a result of losing focus in the overflow wards. The simulation model is replicated five times over 10^6 time periods for each scenario, producing estimates for P_i, W_i, S_i^d, and S_i^o for the overflow delay system, and A_i, S_i^d, and S_i^o for the overflow abandonment system, for $i \in \mathcal{S}$.

Series	λ_1	λ_2	d_1	d_2	o
I	2	2	2	2	$\{1, .., 5\}$
	2	3	2	2	$\{2, .., 6\}$
	2	5	3	3	$\{3, .., 7\}$
	3	10	5	5	$\{6, .., 10\}$
	4	5	3	3	$\{4, .., 8\}$
	5	5	4	4	$\{3, .., 7\}$
	2	2	$\{2, .., 6\}$	3	1
	2	3	$\{1, .., 5\}$	3	2
	2	5	$\{1, .., 5\}$	4	3
	3	10	$\{1, .., 5\}$	8	5
	4	5	$\{2, 4, 6, 8, 10\}$	4	5
	5	5	$\{2, 4, 6, 8, 10\}$	4	5
	2	2	2	$\{1, 3, 4, 5, 6\}$	2
	2	3	1	$\{1, .., 5\}$	5
	2	5	3	$\{4, .., 8\}$	2
	3	10	2	$\{10, .., 14\}$	2
	4	5	5	$\{4, .., 8\}$	3
	5	5	5	$\{4, .., 8\}$	3
II	2	2	2	2	$\{1, .., 5\}$
	3	4	2	2	$\{1, .., 5\}$
	3	5	2	5	$\{3, .., 7\}$
	5	9	5	5	$\{8, .., 12\}$
	4	5	3	3	$\{6, .., 10\}$
	5	5	4	4	$\{4, .., 8\}$
	2	3	$\{2, \ldots, 6\}$	3	2
	2	5	$\{2, \ldots, 6\}$	4	3

TABLE 4.3: The parameters for our numerical experiments. For Series I, we set $\mu_1 = \mu_2 = 1$, and for Series II, we set $\mu_1 = 1.2$ and $\mu_2 = 0.75$.

For the overflow delay system, the output of the approximation is plotted versus the output of the simulation in Figures 4.2, 4.3, 4.4, and 4.5 for delay probabilities, mean waiting times, mean dedicated busy severs, and mean overflow busy servers, respectively. These plots are created on a logarithmic scale in order to display a wide range of values in a compact way. They show a reliable level of accuracy for all the performance metrics evaluated.

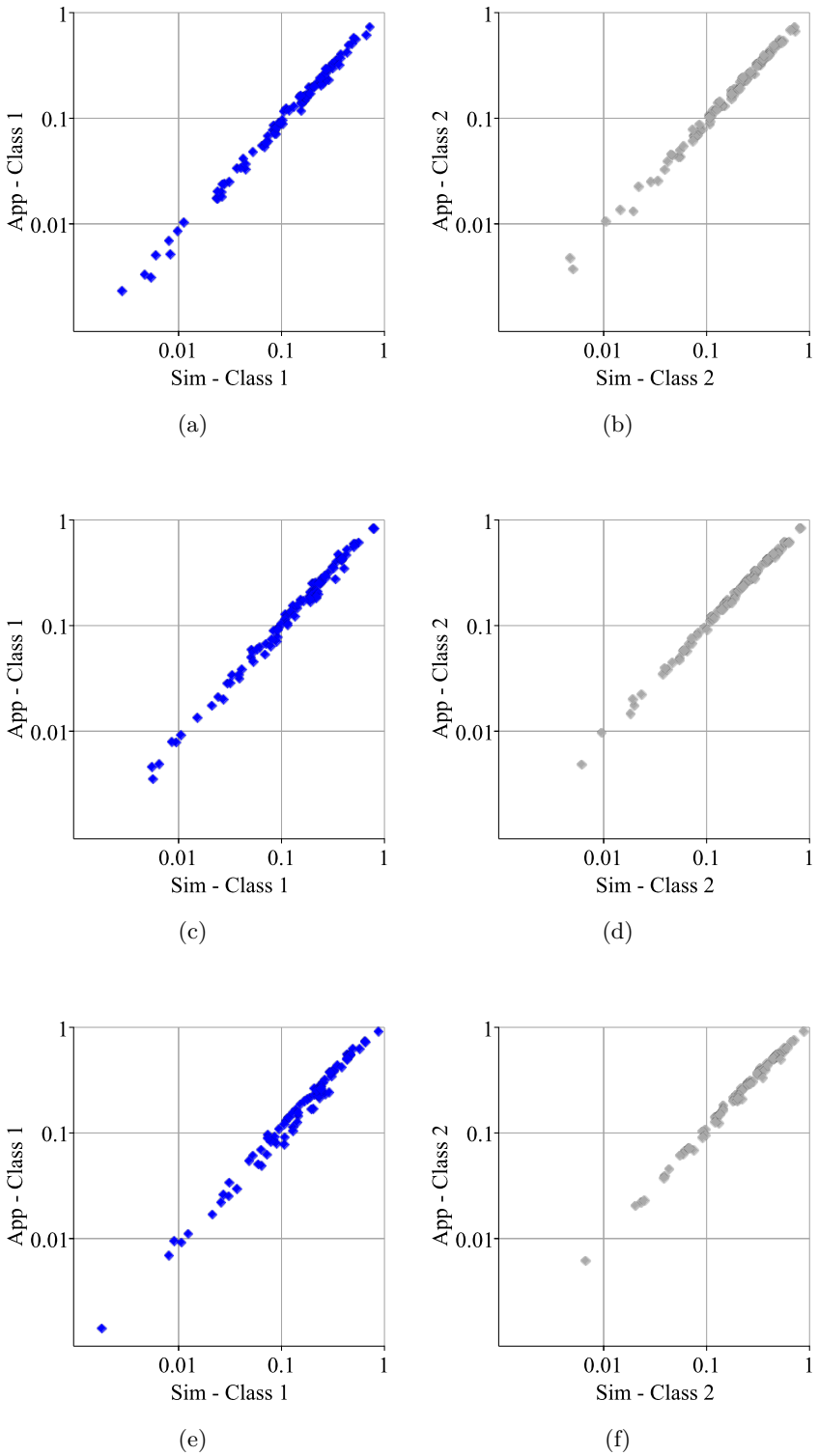


FIGURE 4.2: Delay probabilities obtained from the approximation vs simulation for: $\rho_1 = \rho_2 = 1$ (a , b); $\rho_1 = \rho_2 = 0.9$ (c , d); and $\rho_1 = \rho_2 = 0.8$ (e , f) in a logarithmic scale.

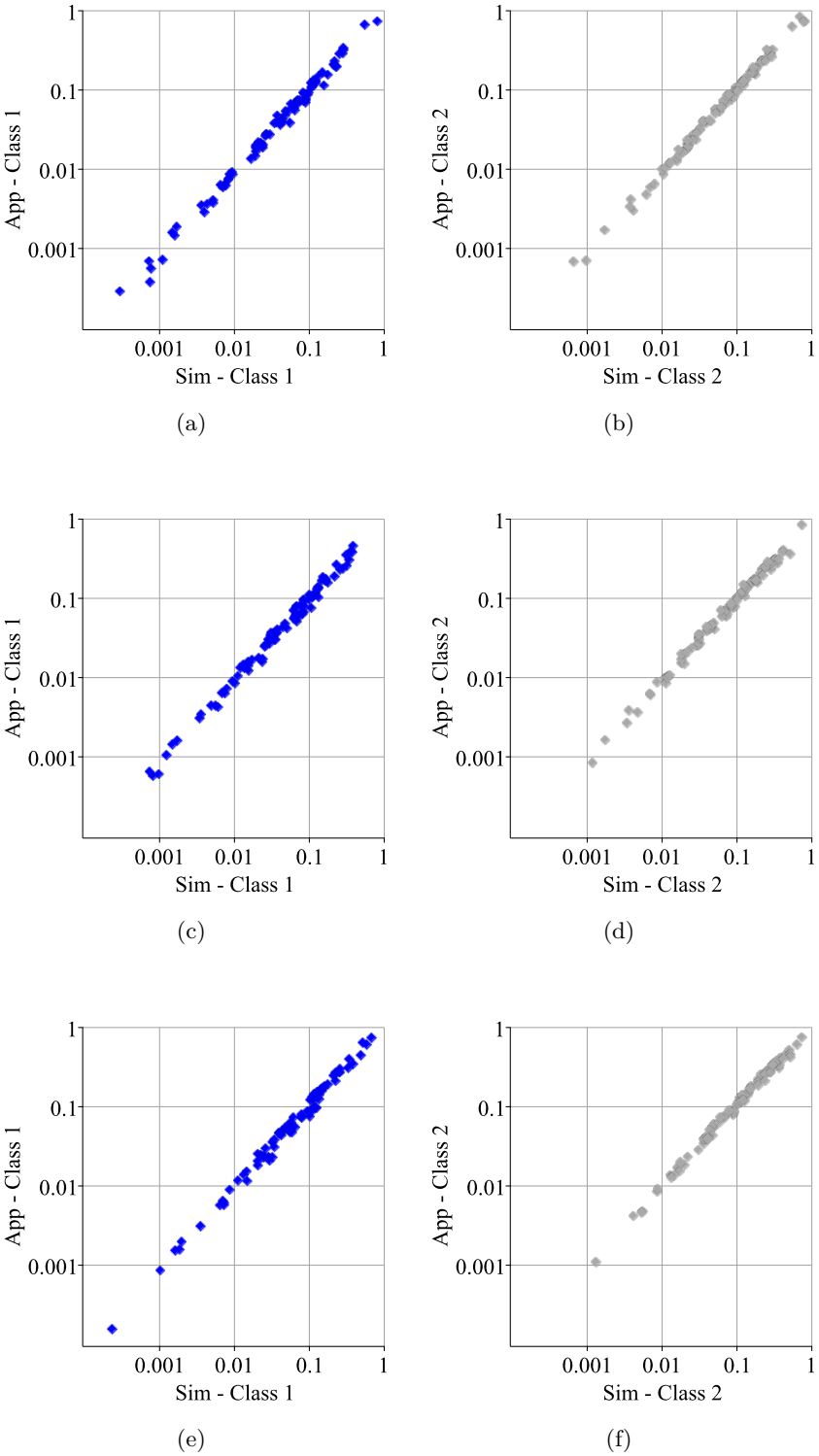


FIGURE 4.3: Mean waiting times obtained from the approximation vs simulation for: $\rho_1 = \rho_2 = 1$ (a , b); $\rho_1 = \rho_2 = 0.9$ (c , d); and $\rho_1 = \rho_2 = 0.8$ (e , f) in a logarithmic scale.

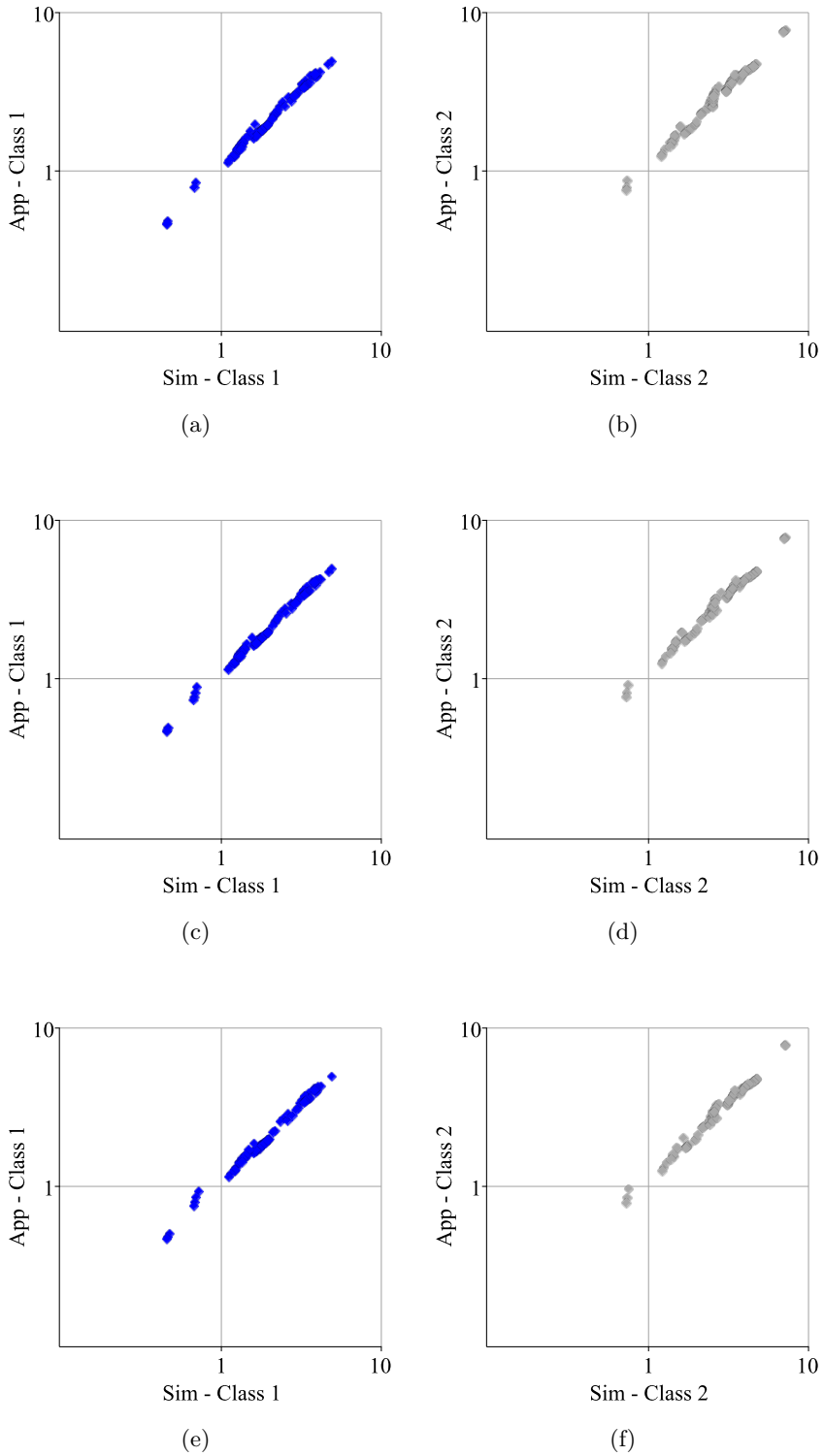


FIGURE 4.4: Mean dedicated busy severs obtained from the approximation vs simulation for: $\rho_1 = \rho_2 = 1$ (a , b); $\rho_1 = \rho_2 = 0.9$ (c , d); and $\rho_1 = \rho_2 = 0.8$ (e , f) in a logarithmic scale.

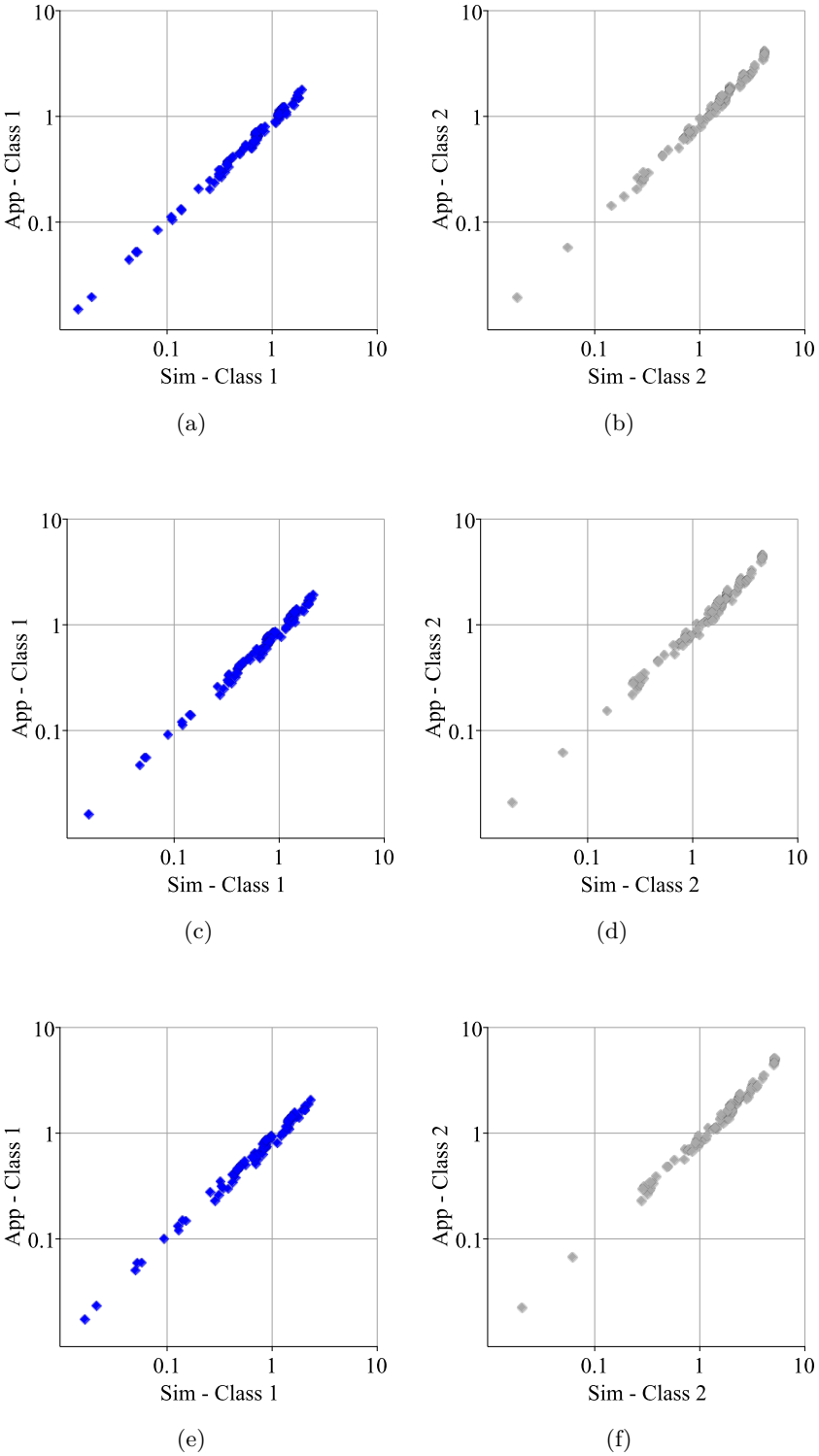


FIGURE 4.5: Mean overflow busy severs obtained from the approximation vs simulation for: $\rho_1 = \rho_2 = 1$ (a , b); $\rho_1 = \rho_2 = 0.9$ (c , d); and $\rho_1 = \rho_2 = 0.8$ (e , f) in a logarithmic scale.

To gain a better insight into the performance of the approximation methodology we have proposed, in Table 4.4 we illustrate the average relative accuracy across all the scenarios considered for the overflow delay system. The results illustrated in this table show a practical level of accuracy for all the measures included. They also show that mean busy server results are less accurate for the overflow pool than the dedicated pools, and that the accuracy reduces when $\rho \neq 1.0$. In Table 4.4, we also present in brackets the accuracy of the results obtained from the methodology in Chevalier and Van den Schrieck (2009). Since their model does not capture the impact of different mean service times in dedicated and overflow pools, these accuracies are only computed for $\rho = 1$. We observe a higher level of accuracy in Table 4.4 for the results from our methodology compared to those from the methodology proposed in Chevalier and Van den Schrieck (2009).

Scenario	P_1	P_2	W_1	W_2	S_1^d	S_2^d	S_1^o	S_2^o
$\rho = 1.0$	9% (14%)	7% (20%)	8% (16%)	7% (28%)	5% (35%)	9% (21%)	10% (74)%	13% (68%)
$\rho = 0.9$	12.21%	7.83%	11.35%	4.06%	5.58%	10.22%	13.64%	14.08%
$\rho = 0.8$	13.34%	9.30%	14.30%	8.24%	5.54%	10.10%	11.05%	14.03%

TABLE 4.4: The average relative accuracy of the approximation methodology for the overflow delay system.

For the overflow abandonment system, the output of the approximation is plotted versus the output of the simulation in Figures 4.6 and 4.7 for the abandonment probabilities, Figures 4.8 and 4.9 for mean dedicated busy servers, and Figures 4.10 and 4.11 for mean overflow busy servers. The average accuracy levels are also reported in Table 4.5. These results show a practical level of accuracy for all the measures included. They also show that the accuracy drops slightly for $\rho \neq 1.0$ as compared to $\rho = 1.0$, and that mean busy server results are less accurate than the abandonment probability results.

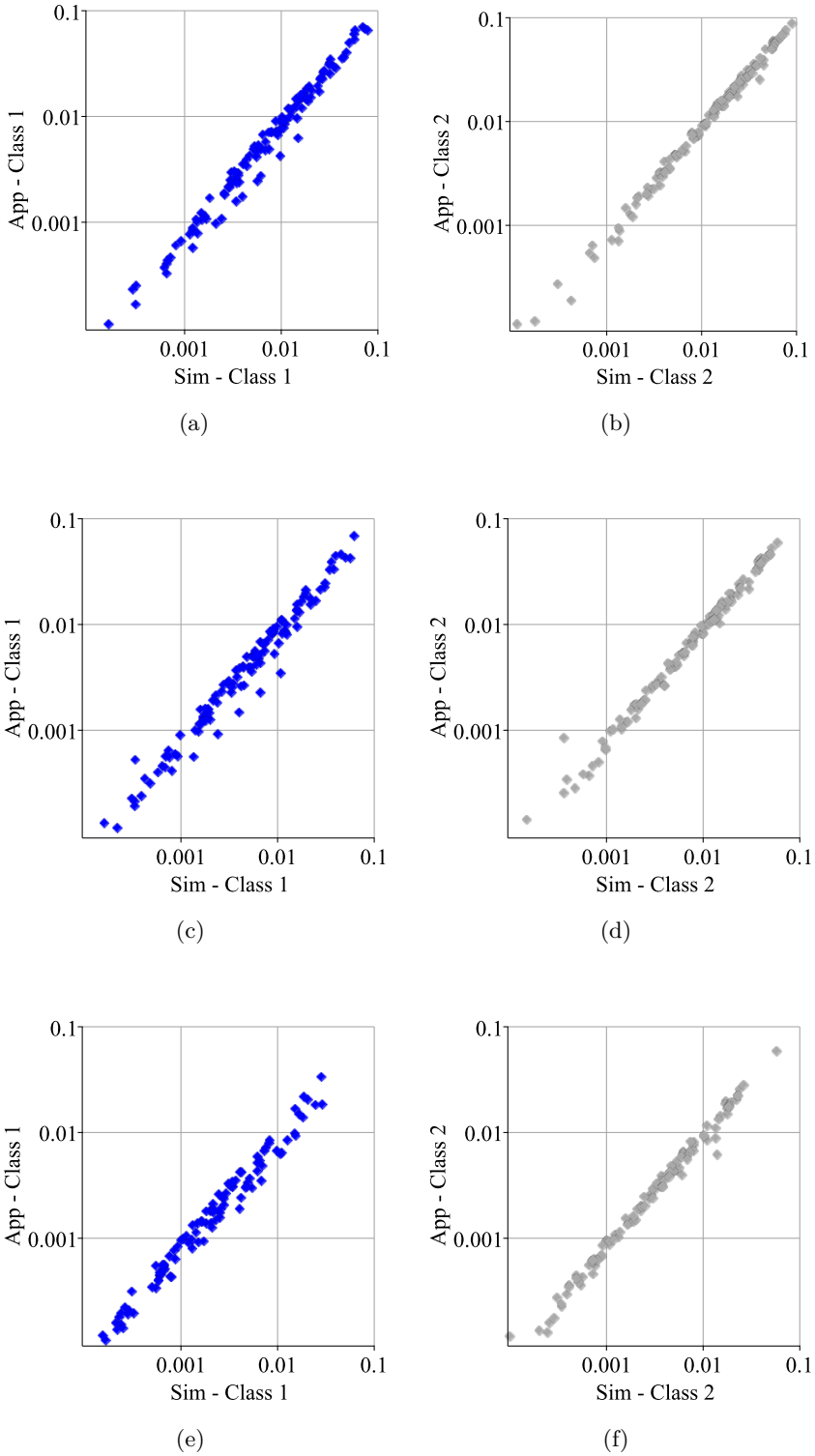


FIGURE 4.6: Abandonment probabilities obtained from the approximation vs simulation for $\rho_1 = \rho_2 = 1$: $\gamma^{-1} = 5$ (a , b); $\gamma^{-1} = 10$ (c , d); and $\gamma^{-1} = 30$ (e , f) in a logarithmic scale.

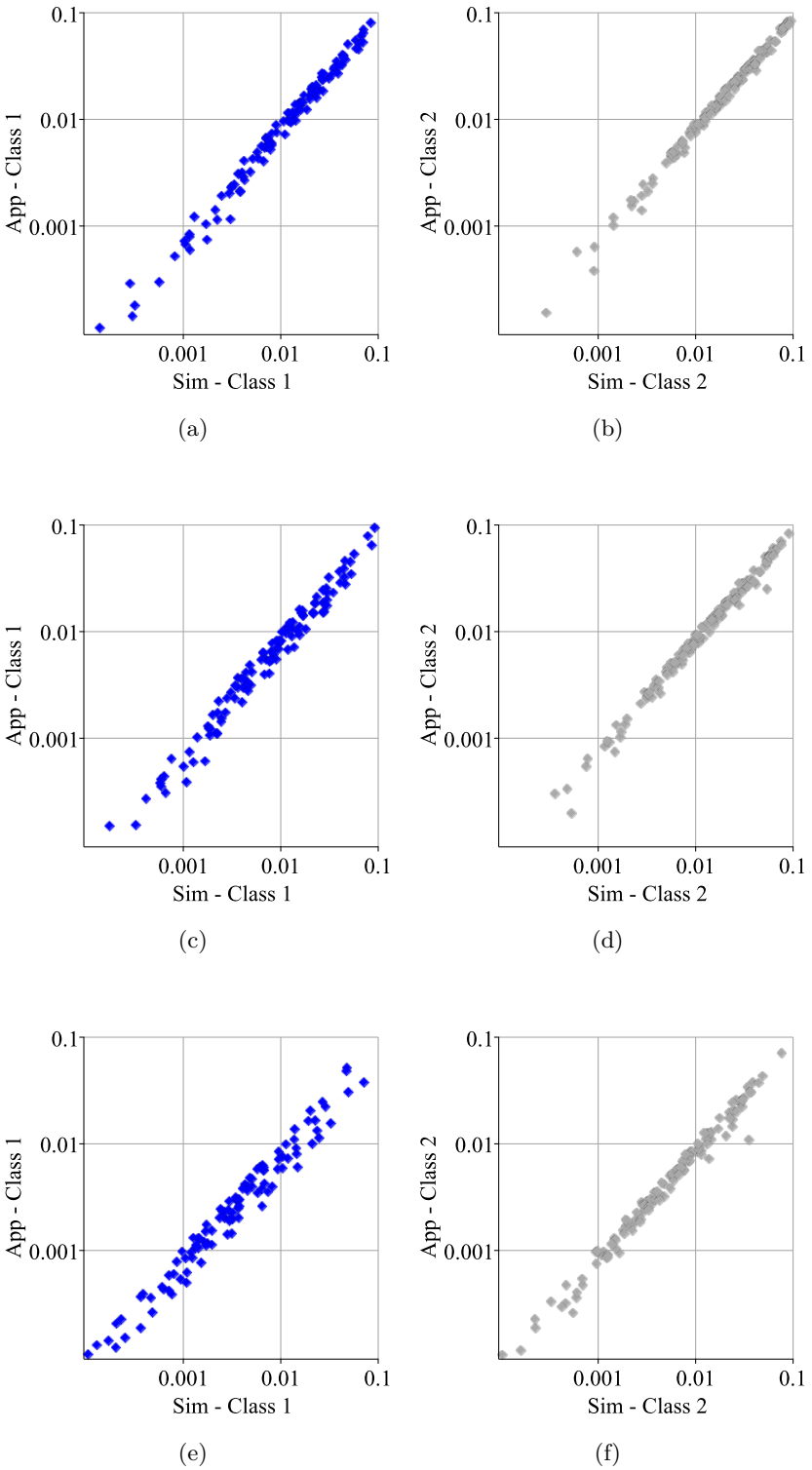


FIGURE 4.7: Abandonment probabilities obtained from the approximation vs simulation for $\rho_1 = \rho_2 = 0.8$: $\gamma^{-1} = 5$ (a , b); $\gamma^{-1} = 10$ (c , d); and $\gamma^{-1} = 30$ (e , f) in a logarithmic scale.

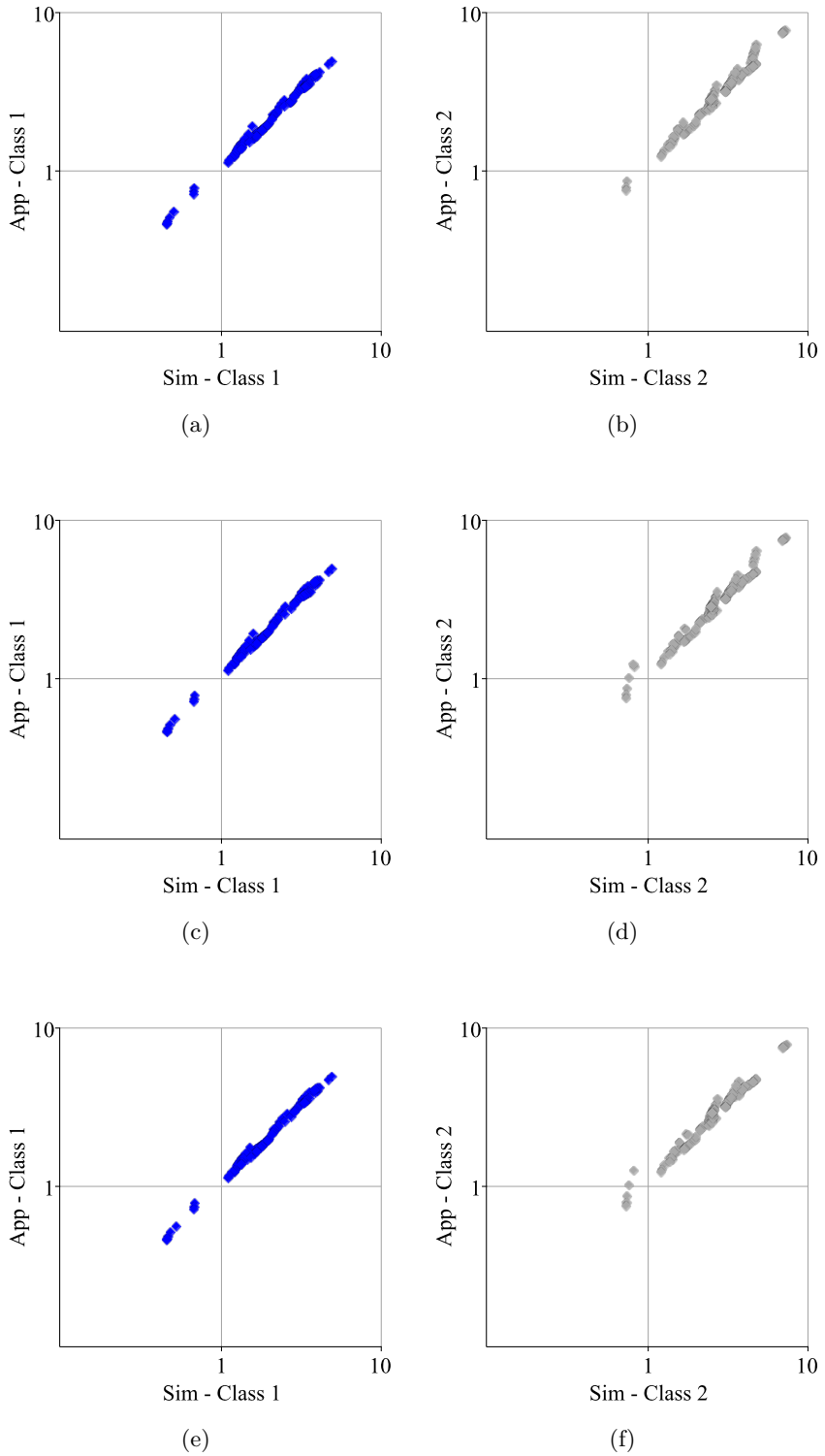


FIGURE 4.8: Mean dedicated busy severs obtained from the approximation vs simulation for $\rho_1 = \rho_2 = 1$: $\gamma^{-1} = 5$ (a , b); $\gamma^{-1} = 10$ (c , d); and $\gamma^{-1} = 30$ (e , f) in a logarithmic scale.

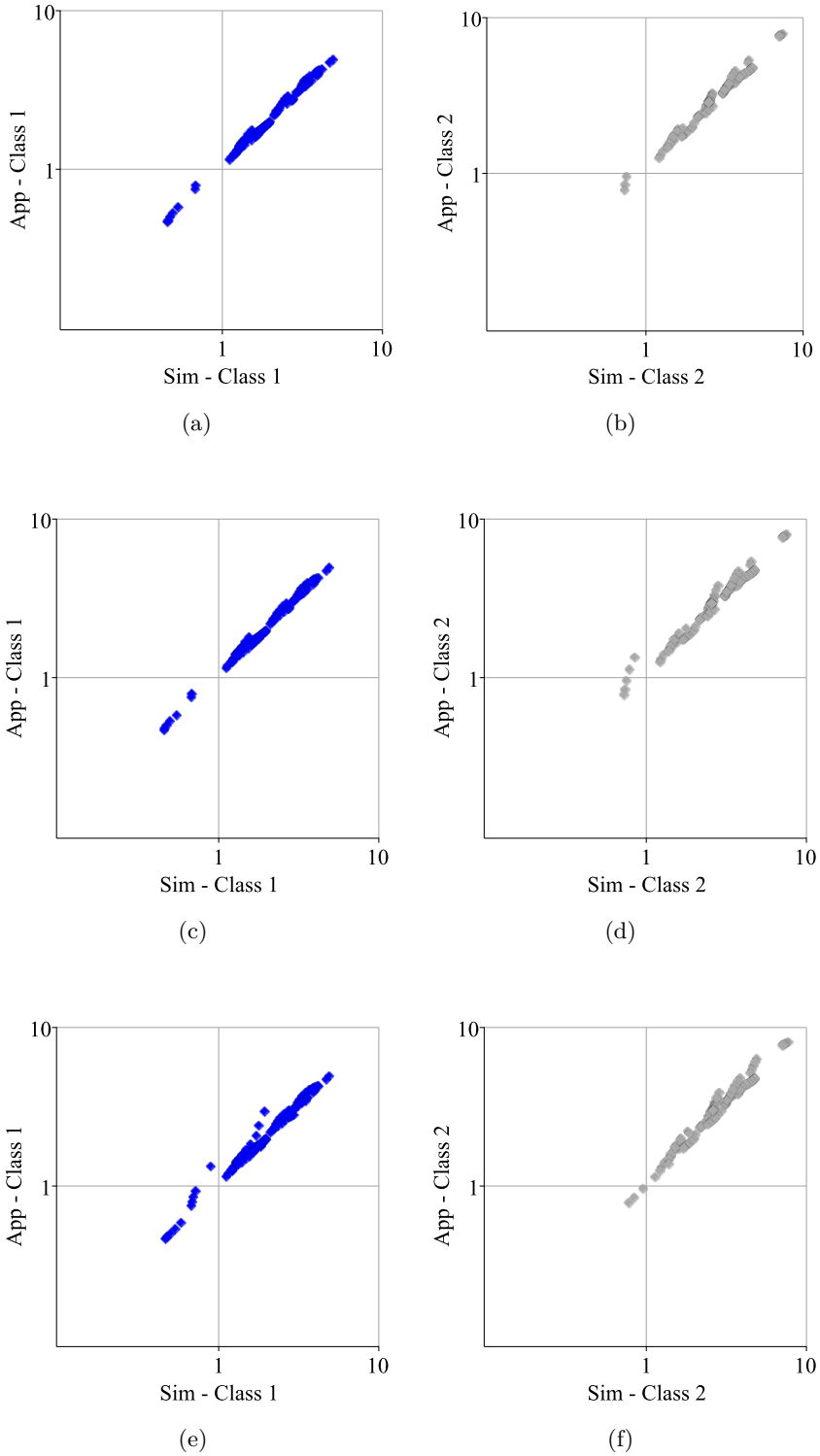


FIGURE 4.9: Mean dedicated busy severs obtained from the approximation vs simulation for $\rho_1 = \rho_2 = 0.8$: $\gamma^{-1} = 5$ (a , b); $\gamma^{-1} = 10$ (c , d); and $\gamma^{-1} = 30$ (e , f) in a logarithmic scale.

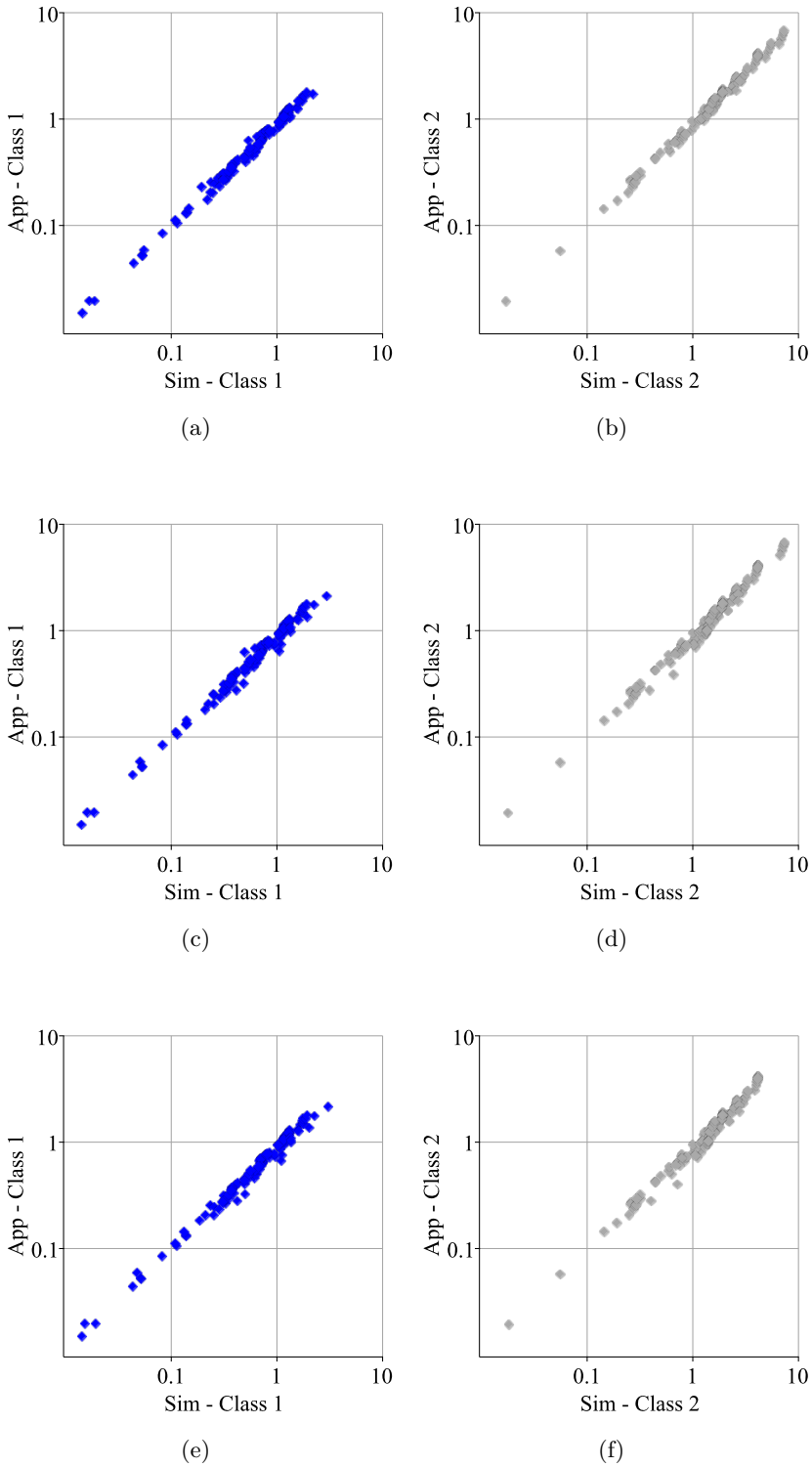


FIGURE 4.10: Mean overflow busy severs obtained from the approximation vs simulation for $\rho_1 = \rho_2 = 1$: $\gamma^{-1} = 5$ (a , b); $\gamma^{-1} = 10$ (c , d); and $\gamma^{-1} = 30$ (e , f) in a logarithmic scale.

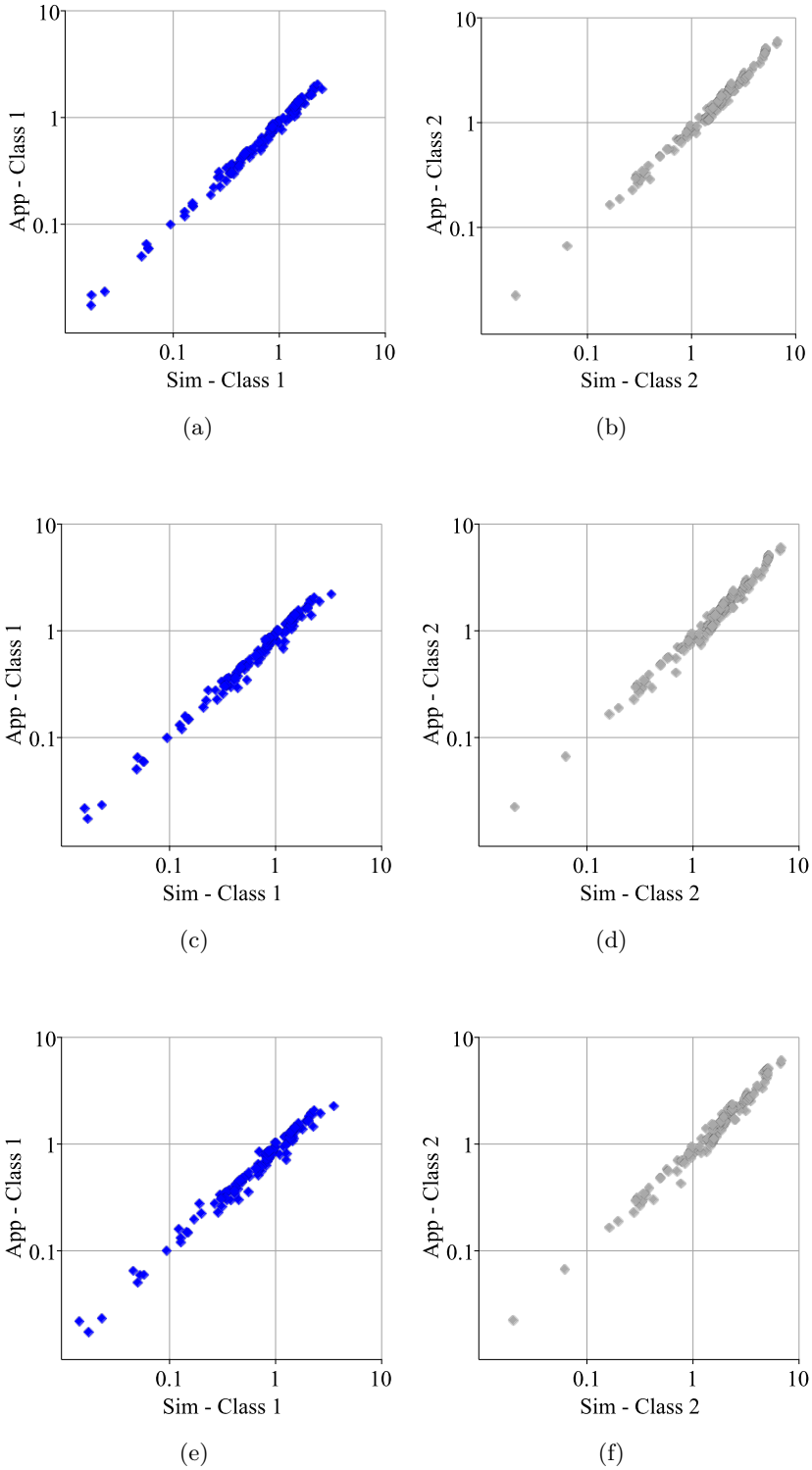


FIGURE 4.11: Mean overflow busy severs obtained from the approximation vs simulation for $\rho_1 = \rho_2 = 0.8$: $\gamma^{-1} = 5$ (a , b); $\gamma^{-1} = 10$ (c , d); and $\gamma^{-1} = 30$ (e , f) in a logarithmic scale.

Scenario	γ^{-1}	Mean Relative Error					
		A_1	A_2	S_1^d	S_2^d	S_1^o	S_2^o
$\rho = 1.0$	5	0.03%	0.05%	3.91%	8.57%	10.39%	12.75%
	10	5.66%	3.05%	4.19%	9.51%	11.23%	13.68%
	30	3.99%	1.25%	3.58%	8.46%	10.34%	12.95%
$\rho = 0.8$	5	10.58%	9.06%	4.13%	10.00%	12.82%	15.33%
	10	10.57%	7.62%	3.95%	9.69%	13.37%	15.34%
	30	7.89%	4.82%	5.12%	9.46%	15.36%	14.13%

TABLE 4.5: The average relative accuracy of the approximation methodology for the overflow abandonment system.

4.5 Conclusion

In this chapter, we first developed an exact model for performance evaluation of an overflow delay system with only two customer classes, assuming each server pool has only one server. We demonstrated that, even for this simplified system, the exact model works very slowly, requiring more than two hours to return performance metrics for a system with a maximum of 40 customers allowed to wait in each queue. This hinders the application of the exact model in configuration optimization frameworks wherein the performance of thousands of different bed allocations must be computed. We next proposed approximation methodologies for estimating the performance metrics of overflow delay and abandonment systems.

The approximations we have proposed build on the approximation proposed in Chevalier and Van den Schrieck (2009) for overflow delay system and extend it in the following ways. First, our approximations capture the impact of mean service times being different in dedicated and overflow pools. This enables us to reflect the impact of focus (or the lack of it) in overflow wards. This extension has been facilitated by using the analysis provided in Izady and Muhamed (2019) for overflow loss systems. Second, our approximation applies to overflow abandonment systems. This has been achieved by using a relation between abandonment and loss probabilities in single-class queues. Third, our approximation is more accurate than that of Chevalier and Van den Schrieck (2009). This is obtained by combining the fluid approximation deployed in Chevalier and Van den Schrieck (2009) with overflow loss system results.

We evaluated the accuracy of our approximations for a total of 390 scenarios for the overflow delay system and 780 scenarios for the overflow abandonment system by comparing the results with those obtained from a discrete-event simulation model. The results demonstrate a practical level of accuracy for mean waiting times, delay and abandonment probabilities, and mean dedicated and overflow busy servers. These measures enable us to evaluate the mean total daily cost of a cluster in the COF configuration. More specifically, for a cluster $\mathcal{C}$ with dedicated bed numbers given by $(d_i;\, i \in \mathcal{C})$ and o overflow beds, the mean daily nursing

cost, $R(\mathcal{C}, (d_i;\, i \in \mathcal{C}),\, o)$, is obtained by inserting the mean number of dedicated and overflow busy servers, denoted by $S_k^d(\mathcal{C}, (d_i;\, i \in \mathcal{C}),\, o)$ and $S_k^o(\mathcal{C}, (d_i;\, i \in \mathcal{C}),\, o)$ for $k \in \mathcal{C}$, respectively, obtained from our approximations in Equation (3.9). The waiting and abandonment costs, $c_w Q(\mathcal{C}, (d_i;\, i \in \mathcal{C}),\, o)$ and $c_a B(\mathcal{C}, (d_i;\, i \in \mathcal{C}),\, o)$, are then obtained by inserting mean waiting time and abandonment probability, denoted by $W_k(\mathcal{C}, (d_i;\, i \in \mathcal{C}),\, o)$ and $A_k(\mathcal{C}, (d_i;\, i \in \mathcal{C}),\, o)$ for $k \in \mathcal{C}$, obtained from our approximations in Equations (3.7) and (3.8), respectively. The total mean daily cost is finally evaluated using Equations (3.5) and (3.6) for IWTT and FWTT systems, respectively. The approximations we have proposed work very fast, returning the total cost of a cluster in a fraction of a second. This enables us to incorporate them within an optimization framework for finding the optimal bed configuration.

5 Solving Optimization Models

5.1 Introduction

In this chapter, we discuss the approach taken for solving the optimization models proposed in Chapter 3 for finding the optimal partitioning and bed allocation of the COF configuration. In Section 5.2, we apply the heuristic Conjugate Direction with Orthogonal Shift (CDOS) method for solving the intra-cluster bed allocation model. We also investigate the accuracy of this heuristic using a sample of problem instances. In Section 5.3, we explain how the dynamic programming (DP) approach proposed in Best et al. (2015) can be adapted for solving the partitioning and inter-cluster bed allocation model. The conclusions are provided in Section 5.4.

5.2 The Intra-Cluster Bed Allocation Model

We start with the intra-cluster bed allocation problem restated below

$$\phi(\mathcal{C}, b) = \min_{(d_i; i\in\mathcal{C}),\, o} \Bigg\{ T(\mathcal{C}, (d_i;\, i \in \mathcal{C}),\, o) : o + \sum_{i\in\mathcal{C}} d_i \leq b, o \in \mathbb{Z}, \text{ and } d_i \in \mathbb{Z} \text{ for } i \in \mathcal{C} \Bigg\},$$

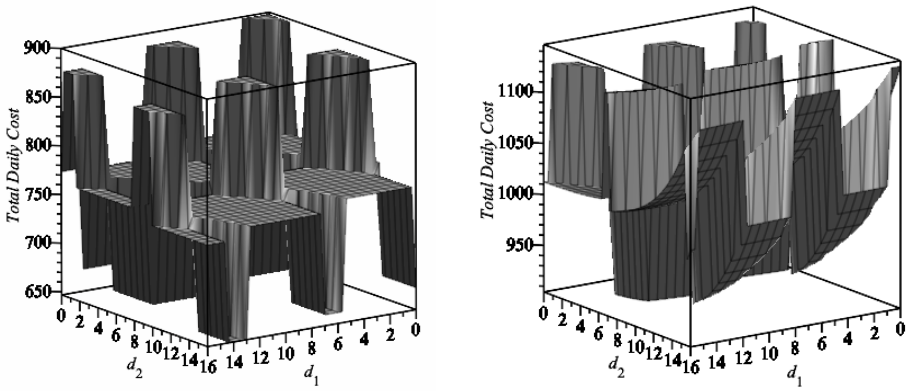
for a given cluster $\mathcal{C}$ and total bed numbers b. This problem seeks to find the allocation of b beds to dedicated and overflow wards of the cluster $\mathcal{C}$ minimizing the total mean daily cost given in Equations (3.5) and (3.6) for IWTT and FWTT systems, respectively. For a cluster $\mathcal{C}$ with bed allocation $(d_i;\, i \in \mathcal{C}), o$, the objective function of the intra-cluster bed allocation problem, i.e., $T(\mathcal{C}, (d_i;\, i \in \mathcal{C}),\, o)$, is evaluated using the approximations we proposed in Chapter 4.

Consider a hypothetical cluster $\mathcal{C} = \{1, 2\}$ with $b = 70$ beds. In Figure 5.1, we plot the objective function of the intra-cluster bed allocation model, i.e., $T(\mathcal{C}, (d_i;\, i \in \mathcal{C}),\, o)$, for this cluster with a given set of parameters as a function of d_1 and d_2 under IWTT as well as FWTT assumptions. The plots in this figure demonstrate that the objective function is neither convex nor differentiable for either assumption (the discontinuities in these plots are due to the presence of ceiling function in the nursing cost formula given in (3.9)). This implies that we

need to apply a gradient-free heuristic optimization method for finding a good solution. Izady and Muhamed (2019) propose the CDOS heuristic developed by Moiseev (2011) for finding a good bed allocation in a cluster under a ZWTT assumption, i.e., assuming the cluster works as an overflow loss system.

We briefly describe the CDOS algorithm; see Moiseev (2011) for a detailed description. CDOS uses conjugate line search directions. It has quadratic convergence for quadratic and near quadratic functions. More specifically, for $f(x_1, ..., x_n)$ an n-dimensional quadratic function, its minimum (maximum) is obtained by n moves in n various conjugate directions regardless of the starting point (Powell 1972). These conjugate directions are constructed using orthogonal shifts obtained by means of the Gram-Schmidt orthogonalization process or by QR decomposition Moiseev (2011). For non-quadratic functions, CDOS continues the search from the current minimum by constructing orthogonal shifts from the existing conjugate directions. This search is not greedy as it allows moving in the direction that does not contain the current minimum. The search stops when a set of convergence conditions are met.

CDOS is an attractive choice for the intra-cluster allocation problems in general since it works very well with inequality constraints, can be applied as a global optimizer as it uses a non-greedy search algorithm, and works well with high-dimensional optimization problems (Moiseev 2011). However, we need to test its accuracy for our bed allocation problem with IWTT and FWTT assumptions.



(a) (b)

FIGURE 5.1: Surface plots for the total mean daily cost of a 70-bed cluster with 2 specialties assuming $f_1 = 0.149$ and $f_2 = 0.149$. Panel (a) is for an IWTT system with $\lambda_1 = 5.2$, $\lambda_2 = 3.9$, $\nu_1 = 4.8$, $\nu_2 = 2.9$, and $c_w = 1030$, and panel (b) is for an FWTT system with $\lambda_1 = 6.2$, $\lambda_2 = 4.9$, $\nu_1 = 4.8$, $\nu_2 = 2.9$, $\gamma^{-1} = 5$ and $c_a = 1030$.

In order to test the accuracy of CDOS heuristic for intra-cluster bed allocation, we compare the allocations obtained from this heuristic with the optimal allocations obtained from complete enumeration for a hypothetical cluster with two specialties under both IWTT and FWTT

assumptions. For the system with IWTT assumption, we run our experiments with the following set of parameters: $\lambda_1 \in \{1, 2\}$, $\lambda_2 \in \{3, 4\}$, $\nu_1 \in \{1.0, 1.5\}$, $\nu_2 \in \{1.2, 2.0\}$, $c_w = 1030$, $b \in \{12, 16, 20, 24, 28, 32, 36\}$, and $f_1 = f_2 = 0.149$. For the system with the FWTT assumption, we run our experiments with the same set of parameters assuming $\gamma^{-1} = 10$ and $c_a = 1030$. The combination of these parameters results in 112 scenarios for each of the IWTT and FWTT systems.

In our numerical experiments, we observe that in 103 scenarios (92% of total scenarios) under the IWTT assumption and 111 scenarios (99% of total) under the FWTT assumption, the optimal bed allocation is produced by CDOS. The scenarios for which the optimal allocation is not obtained are illustrated in Table 5.1. As shown in this table, the optimality gap, evaluated as the percentage difference of the cost of the allocation obtained from CDOS as compared to the cost of optimal allocation, is always below 6 % (0.01%) for the IWTT (FWTT) system. These results indicates that CDOS provides a good level of accuracy for our intra-cluster bed allocation model.

Assumption	Complete Enumeration		CDOS		Optimality Gap
	Configuration	Cost	Configuration	Cost	
IWTT	(7,13,0)	210.94	(0,7,13)	219.36	3.99%
	(8,16,0)	208.43	(9,15,0)	209.08	0.31%
	(8,16,0)	208.17	(7,17,0)	208.39	0.10%
	(10,14,0)	209.08	(9,15,0)	211.58	1.19%
	(10,14,0)	208.83	(0,7,17)	219.36	5.01%
	(9,15,0)	208.20	(0,8,16)	219.36	5.36%
	(13,15,0)	208.28	(11,17,0)	209.08	0.38%
	(12,16,0)	208.10	(13,15,0)	208.13	0.01%
	(0,6,30)	219.36	(0,9,27)	219.40	0.01%
FWTT	(9,15,0)	208.07	(8,16,0)	208.10	0.01%

TABLE 5.1: The scenarios for which CDOS fails to find the optimal allocation.

5.3 The Partitioning and Inter-Cluster Bed Allocation Model

We now focus on the partitioning and inter-cluster allocation model restated below

$$Z = \min_{(m,\boldsymbol{b},\boldsymbol{\mathcal{C}})} \left\{ \sum_{j=1}^{m} \phi(\mathcal{C}^j, b^j) : (m, \boldsymbol{b}, \boldsymbol{\mathcal{C}}) \in \Psi \right\},$$

where $\boldsymbol{b} = (b^1, \ldots, b^m)$, and

$$\Psi = \left\{ (m, \boldsymbol{b}, \boldsymbol{\mathcal{C}}) : \sum_{j=1}^{m} b^j \leq B, \boldsymbol{\mathcal{C}} \text{ is a feasible partition of } \mathcal{S}, m \in \mathbb{Z}_+, \text{and } \boldsymbol{b} \in \mathbb{Z}^m \right\}.$$

In order to solve this problem, following Best et al. (2015), we first restrict the feasible region Ψ by focusing only on partitions obtained by making cuts along a fixed sequence $\mathcal{N}$ of specialties. The reason behind these "cuts" lies in the curse of dimensionality; if we were to consider all the possible partitions of a set of size 15, for example, we would need to consider approximately 1.05×10^{10} different possibilities. This would make computation time extremely long. Given a fixed sequence $\mathcal{N}$ of specialties, we therefore consider the restricted problem

$$Z_{\mathcal{N}} = \min_{(m,\boldsymbol{b},\boldsymbol{\mathcal{C}})} \left\{ \sum_{j=1}^{m} \phi(\mathcal{C}^j, b^j) : (m, \boldsymbol{b}, \boldsymbol{\mathcal{C}}) \in \{\Psi \text{ and } (\mathcal{C}^j \text{ are cuts in the sequence } \mathcal{N})\} \right\}.$$

We then solve the restricted problem using the DP approach proposed in Best et al. (2015), with the difference that expected reward for each state-action pair is evaluated using the CDOS heuristic explained in Section 5.2. We provide a brief overview of this approach here.

The state space for the DP formulation is represented by $\Theta = \{(j,k,l), j = 1,...,n; k = 0,...,B; l = j,...,n\} \cup \{(0,0,0)\}$, where j, k and l are indices for clusters, beds and specialties, respectively, and $(0,0,0)$ is an auxiliary state. For state $\theta = (j,k,l) \in \Theta, \mathcal{A}(\theta) = \{(b,t) : b = 0,...,B-k; t = l+1,...,n-l\}$ is the set of actions, where b is the number of beds to be allocated to the next cluster, and t is the number of specialties to be allocated to the next cluster. For $\theta = (j,k,l) \in \Theta$, the action $a = (b,t) \in \mathcal{A}(\theta)$ takes the system to the new state $\theta' = (j+1, k+b, l+t)$, representing the formation of a new cluster $\mathcal{C}^{j+1}$ with b beds and specialties indexed as $l+1, l+2, ..., l+t$ in the sequence $\mathcal{N}$. The associated expected cost will be $r(\theta, a) = \phi(\mathcal{C}^{j+1}, b)$, evaluated using the heuristic approach explained in Section 5.2.

The reward-to-go function $V(\theta)$ for $\theta = (j,k,l) \in \Theta$ gives the minimum expected daily cost that can be achieved by solving the partitioning and inter-cluster bed allocation sub-problem for the remaining $B-k$ beds and the remaining $n-l$ specialties in the sequence $\mathcal{N}$. The optimality equation for the DP problem is therefore

$$V(\theta) = \begin{cases} 0, & \text{if } \theta = (.,.,n), \\ \min_{a \in \mathcal{A}(\theta)} \{r(\theta, a) + V(\theta')\}, & \text{otherwise,} \end{cases}$$

which can be solved recursively starting from $V(0,0,0)$.

5.4 Conclusions

In this chapter, we investigated the feasibility of using the methodologies already proposed in the literature for solving our optimal partitioning and bed allocation problem. In particular, through a sample of numerical experiments we demonstrated a heuristic gradient-free search

method called CDOS that produces reliable results for our intra-cluster bed allocation problem. We also illustrated how a DP approach facilitated by fixing a sequence of specialties, proposed in Best et al. (2015), can be adapted for solving our partitioning and inter-cluster bed allocation problem.

There are different approaches to forming a sequence $\mathcal{N}$ of specialties. In their utility-maximization wing-formation problem, Best et al. (2015) suggest the sequence obtained by sorting the specialties based on the ratio of the utility of each specialty to its mean LOS. This would form wings including specialties with similar daily utilities. They also propose other considerations, such as the similarity of the diseases and their care procedures, in forming this sequence. Izady and Muhamed (2019) propose sorting the specialties based on their mean LOSs, suggesting that it is aligned with the results in the pooling literature which indicates merging customer classes is more likely to create advantages when LOSs are similar. We follow a similar approach in our case study in the next chapter.

6 Case Study

6.1 Introduction

We started a collaborative project with Royal Surrey County Hospital (RSCH) in January 2019. RSCH is a general NHS hospital located in Surrey, UK, providing emergency and general hospital services to a population of more than 330,000 people living across south west Surrey. As we shall demonstrate in Section 6.2, the hospital was experiencing a high level of bed occupancy even before the start of the COVID-19 pandemic. The aim of our collaboration was therefore to find innovative ways to reduce the bed pressure on hospital with a focus on reconfiguration of inpatient services. Operational and clinical knowledge and expertise was provided in this project by the hospital Director of Transformation, Mr. Nick Sands, and the Clinical Lead for Transformation, Dr. James Adams, to whom we refer as our hospital partners.

In this chapter, we report the interim results of our collaboration with RSCH. In Section 6.2, we provide a holistic overview of RSCH inpatient services, including a detailed analysis of patients' journeys in the current configuration. In Section 6.3, we apply the models developed in this dissertation to find a better configuration of inpatient beds, and compare its performance with the current situation as well as the other feasible configurations using a discrete-event simulation model. In Section 6.4, we provide our conclusions.

6.2 Current Status

Data Coverage. Admission data covering a three-year period starting from 01/10/2015 is provided by the hospital. The data captures a wide range of information for each patient admitted to the hospital. In particular, it contains information about different episodes of care within each hospital spell, including (but not limited to) their specialty, primary and secondary diagnoses, and the procedures conducted. In the NHS jargon, a hospital spell is defined as "... the total continuous stay of a patient using a hospital bed on premises controlled by a health care provider during which medical care is the responsibility of one or more consultants ...", while an episode of care means "...the time a patient spends in the continuous care of one

consultant..." (NHS Data Model and Dictionary 2021). For each episode of care, in turn, the data captures the sequence of beds the patient has visited, the start and end dates and times of each bed visit, and the corresponding ward. For example, the spell illustrated in Table 6.1 has 4 episodes of care, 2 of which belong to Geriatric Medicine, one to Rheumatology, and one to Respiratory specialties. The first episode, for example, involves 4 bed visits, three of which occur in the same ward.

Episode No	Specialty	Bed Visit	Visit Start Date & Time	Visit End Date & Time	Ward
1	Geriatric Medicine	1	01/10/2015 02:58	01/10/2015 10:07	EAU
		2	01/10/2015 10:07	01/10/2015 17:58	EAU
		3	01/10/2015 17:58	01/10/2015 18:19	M2
		4	01/10/2015 18:19	02/10/2015 10:40	EAU
2	Rheumatology	5	02/10/2015 10:40	02/10/2015 19:24	EAU
		6	02/10/2015 19:24	02/10/2015 21:30	EAU
		7	02/10/2015 21:30	03/10/2015 07:50	M2
3	Geriatric Medicine	8	03/10/2015 07:50	05/10/2015 15:10	M2
4	Respiratory Medicine	9	05/10/2015 15:10	07/10/2015 23:00	M1
		10	07/10/2015 23:00	10/10/2015 14:40	M1

TABLE 6.1: An example of episodes of care and corresponding bed visits within a hospital spell at RSCH.

There exists a total of 394 inpatient beds in the hospital divided among 13 inpatient wards, including 8 medical wards (215 beds), 5 surgical wards (136 beds), an escalation ward (12 beds), and an emergency assessment unit (EAU; 31 beds). The escalation ward is mainly used during Winter when demand for inpatient beds is at its peak. EAU is a short stay specialist assessment and admission facility specifically for patients whose LOSs are expected to be less than 48 hours. Episodes of care spent entirely in a paediatric, rehabilitation, mental health, or intensive care ward are excluded from our analysis as these wards often have dedicated resources, and so are not included in the reconfiguration exercise. We also exclude all episodes spent entirely in a day-case ward, e.g., in the Cardiac Day Ward or the Endoscopy Unit, so as to keep the focus on inpatient care. This leaves a total of 73,466 unique hospital spells with at least a period of care spent in an inpatient ward. The rest of our analysis is restricted to these spells.

Bed Occupancy. The numbers of beds and average occupancy levels of inpatient wards at RSCH are given in Table 6.2. The occupancy figures given in this table demonstrate the significant strain on hospital beds. This is highlighted further in Figures 6.1 and 6.2, which represent the daily occupancy levels over the last year of our data coverage period (Occupancy rates in excess of 100% in these figures indicate times at which beds in non-inpatient wards were used to accommodate patients.) In particular, we observe in these figures that daily occupancy exceeds 85% for significant periods of time for both medical and surgical wards, and that EAU and Escalation wards' occupancies hit 100% frequently.

Wards*	M1	M2	M3	M4	M5	M6	M7	M8	S1	S2	S3	S4	S5	E	EAU
No. Beds	30	30	16	30	24	30	31	24	28	19	30	30	29	12	31
Occupancy	93%	96%	87%	96%	92%	90%	92%	95%	80%	85%	88%	87%	92%	88%	76%

Note: M for medical wards, S for surgical wards, E for the escalation ward, and EAU for the emergency assessment unit.

TABLE 6.2: Bed numbers and average occupancy levels of different inpatient wards at RSCH.

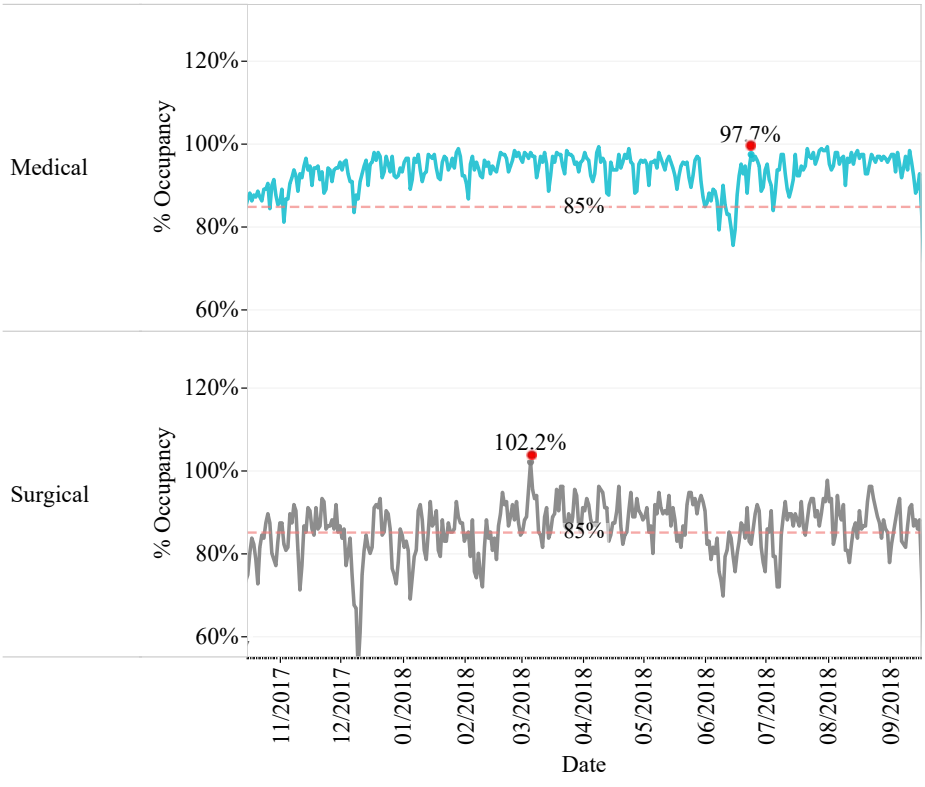


FIGURE 6.1: Daily occupancy levels for medical and surgical wards over the period 01/10/2017-01/10/2018 at RSCH.

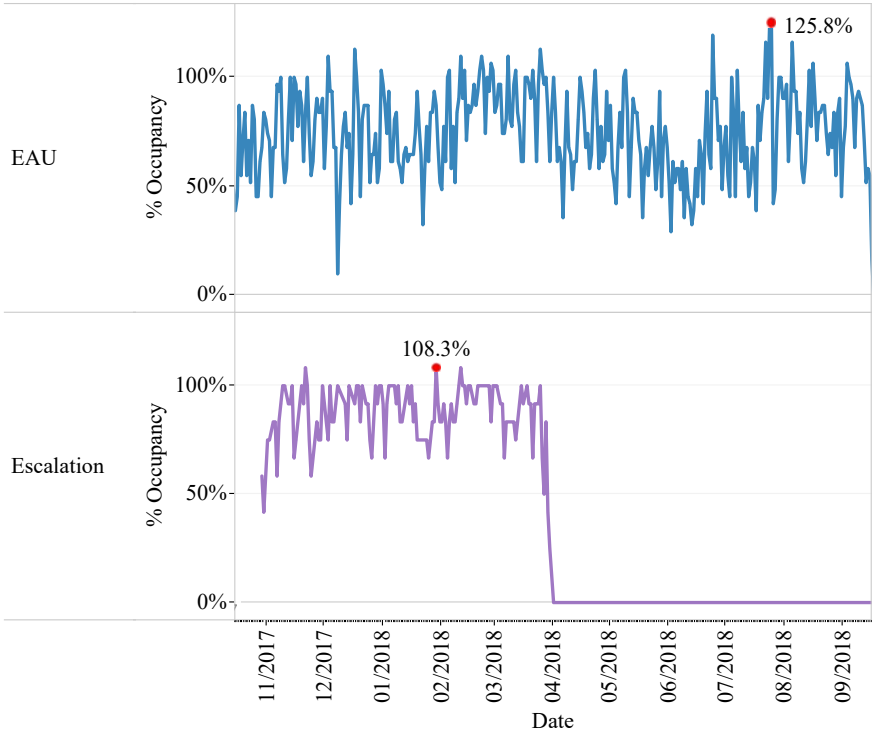


FIGURE 6.2: Daily occupancy levels for EAU and escalation wards over the period 01/10/2017-01/10/2018 at RSCH.

Re-defining Specialties. For the purpose of reconfiguration, we need to identify the specialty for each episode of care as accurately as possible. This is also important for assessing the extent as well as impact of patient outlying in the hospital. However, the "Specialty" field provided in the hospital data (see Table 6.1) represents the specialty of the consultant in charge of the corresponding care episode, which may not necessarily represent the clinical requirement of the patient, in particular when the patient is outlied. To identify the correct specialty for each episode of care, we decided in consultation with our hospital partners to use the FCE-HRG (Finished Consultant Episode - Healthcare Resource Group) codes provided in the data. HRG codes "... are standard groupings of clinically similar treatments which use common levels of healthcare resources ...", and consist of 5 parts each referring to a specific characteristic (NHS Data Model and Dictionary 2021). The FCE-HRG code in our data identifies the HRG code for each episode of care. The specialty of each episode can be derived from the first two letters of this code as illustrated in Table 6.3. We set the specialty of each episode of care to the specialty obtained from this table with the following exceptions:

- If the specialty of the consultant in charge of an episode is Geriatric Medicine, and the corresponding patient age is sixty or above, we set the specialty to Geriatric Medicine. This is because Geriatric Medicine does not have a designated HRG code.

- For episodes with an HRG code suggesting a Gastroenterology specialty, we set the specialty to General Surgery when the "Reason for Admission" field (provided in the data) is "Surgical Treatment".
- For episodes with an HRG code starting with letter "W", we set the speciality to the specialty of the consultant in charge. This is because the codes starting with "W" are related to infectious diseases and immune system disorders, which may affect different parts of the body.
- If the "First Diagnosis" or "Second Diagnosis" fields of an episode of care (provided in the data) is Palliative Care, we set the specialty to Palliative Medicine.
- For the very small number of episodes which do not have an HRG code, we set the specialty to the specialty of consultant in charge.

First two letters of FCE-HRG code	Specialty
A	Neurology
BZ	Opthalmology
CA, CB, CD	ENT
CZ	Oral & Maxillo Facial Surgery
DZ	Respiratory
EA, EB, EC, ED, EY	Cardiology
FZ	Gastroenterology
GA, GB, GC, QZ, JA, JB, JC	General Surgery
HA, HB, HC, HE, HN, HR, HT, VA	Trauma & Orthopaedics
HD	Rheumotology
JD	Dermatology
KA, KB, KC	Endocrinology
LA, LB	Urology
MA	Gynaecology
NZ	Obstetrics
RC	Radiology
SA	Haematology
SB, SC	Oncology
WA,WD, WH, WJ	Infectious Diseases and Disorders
YA, YD, YF, YG, YH, YJ, YL, YR	Interventional Radiology

Table 6.3: FCE-HRG codes and their corresponding specialties

As an example, consider the spell illustrated in Table 6.1. The HRG codes for all the episodes within this spell start with letters "DZ", suggesting the Respiratory specialty. However, since the patient age (provided in the data) is above 60 and the consultant specialty is Geriatric Medicine for the first and third episodes, we set the specialty for these episodes to Geriatric Medicine. The specialties of the second and fourth episodes are set to Respiratory.

Setting the specialties in the way explained above, we observe that RSCH provides inpatient care for a total of 18 specialties. Removing Radiology specialty, and combining Gynaecology

and Obstetrics as well as Haematology and Oncology, we obtain 16 specialties, including 9 medical and 7 surgical ones, as listed along with their abbreviations in Table 6.4.

Specialities	Acronym	Division
Cardiology	CRD	Medical
Endocrinology	END	Medical
Ear, Nose and Throat	ENT	Surgical
Gastroenterology	GAS	Medical
General Surgery	GSR	Surgical
Geriatric Medicine	GRT	Medical
Gynaecology & Obstetrics	GYN	Surgical
Neurology	NRO	Medical
Oncology & Haematology	ONC	Medical
Ophthalmology	OPL	Surgical
Oral & Maxillo Facial	ORM	Surgical
Palliative Medicine	PAL	Medical
Respiratory Medicine	RSP	Medical
Rheumatology	RUM	Medical
Trauma & Orthopaedics	ORT	Surgical
Urology	URO	Surgical

TABLE 6.4: Specialties at RSCH, and their corresponding divisions and acronyms.

To gain a better understanding of patients' journeys in inpatient services, a breakdown of a hospital spell is provided in Figure 6.3. This diagram shows that each hospital spell contains one or more *specialty spells*, which we define as the continuous amount of time a patient spends within one specialty. The change in specialty during a hospital spell is often due to different medical needs of a patient. Each specialty spell, in turn, includes one or more episodes of care, each of which contains one or more *bed visits*. We define a bed visit as a period of time a patient occupies a specific bed in a specific ward. The data provided by hospital already includes timings and the other relevant information for hospital spells, episodes of care, and bed visits (see Table 6.1). We create specialty spells by linking together the episodes of care within a hospital spell that have the same specialty (with specialties set using HRG codes as explained earlier). Our analysis of 73,466 hospital spells show that each spell includes an average (maximum) of 1.08 (5.00) specialty spells, each specialty spell includes an average (maximum) of 1.29 (8.00) episodes of care, and each episode of care includes an average (maximum) of 2.79 (25.00) bed visits. We obtain the length of each specialty spell by adding up the lengths of its constituent episodes of care, and refer to it as the specialty LOS.

Patient Outlying and its Impact. To estimate the extent of patient outlying at RSCH, we measure the percentage of specialty spells admitted to a non-primary ward. To do this, we first identify the primary ward(s) of each specialty as given in Table 6.5. This table is provided by our hospital partners based on their perception of what happens on the ground as well as

FIGURE 6.3: The breakdown of a hospital spell

the skill-set of nursing teams in different wards, and is verified by our data analysis. We then count a specialty spell as an outlying spell if it has at least one bed visit in a non-primary ward of the corresponding specialty. The results indicate that about 49% of all specialty spells are outlying, with the breakdown given in Figure 6.4.

Specialities	Primary Wards
CRD	M5
END	M3
ENT	S2
GAS	M6
GSR	S5
GRT	M2,M8,M3,M4
GYN	S3,S5
NRO	M8
ONC	M7
OPL	S2
ORM	S2
PAL	M1,M7
RSP	M1
RUM	M5,M7
ORT	S1,S4
URO	S3

TABLE 6.5: The primary wards of different specialties at RSCH.

To estimate the contribution to workload by outlying patients, we count a bed visit as an outlying visit if it occurs in a non-primary ward of its corresponding specialty. We then add up the lengths of outlying bed visits and divide it by the sum of all bed visits. The corresponding percentages are presented in Figure 6.5 for different wards. This figure suggests that the contribution of outlying patients varies significantly from ward to ward, with M6 (dedicated to GAS) having the highest contribution and M2 (dedicated to GRT) the lowest. Overall, outlying patients account for 27.7% of inpatient workload at RSCH.

As stated in Chapter 1, studies such as Stowell et al. (2013) report that outlying patients

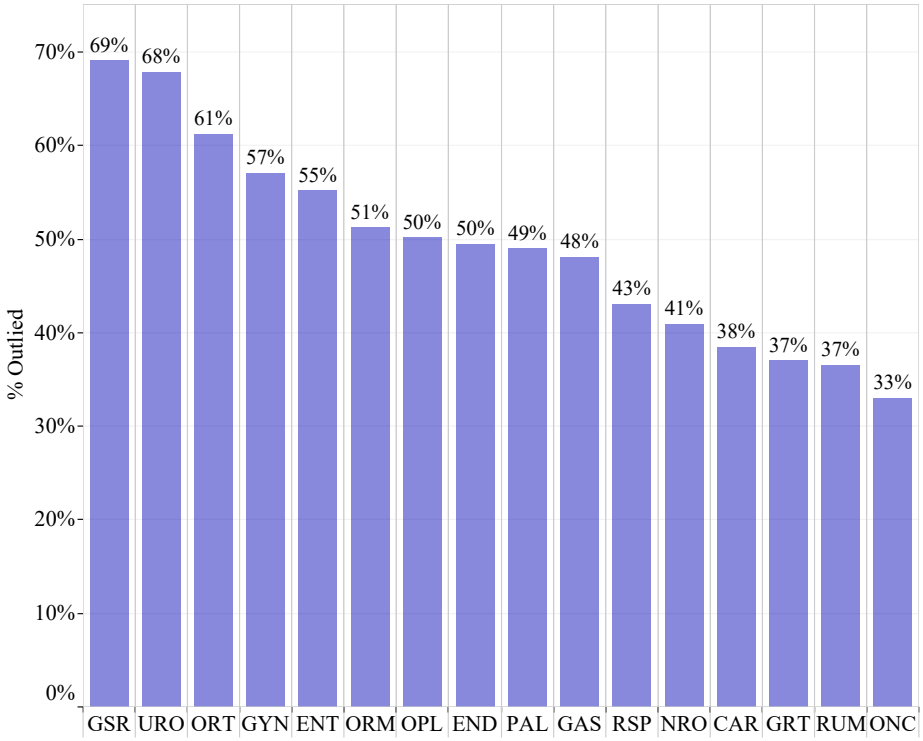


FIGURE 6.4: Percentage of outlying specialty spells for each specialty.

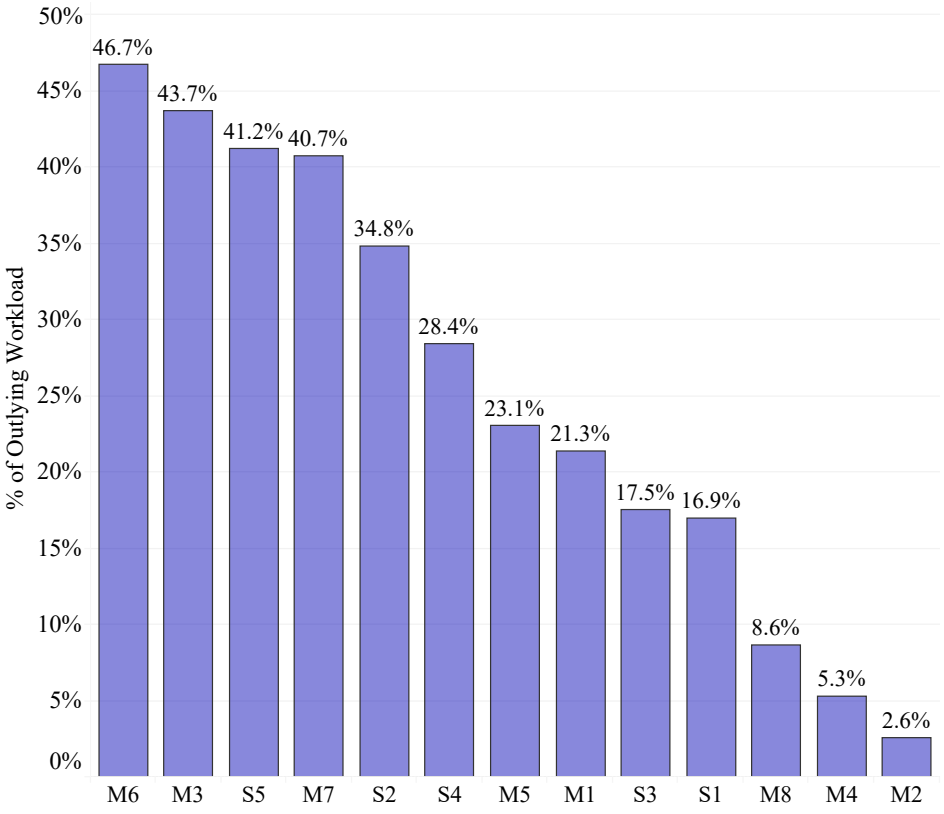


FIGURE 6.5: Contribution to workload by outlying patients for each ward

usually have a longer LOS. We investigate this on GSR and URO specialties which have the largest percentage of outlying spells. We observe that outlying GSR and URO specialty spells are 0.92 and 2.05 days, respectively, longer than the corresponding non-outlying spells in our data. Given the large number of outlying specialty spells at RSCH, it would also be interesting to see their impact on non-outlying spells, i.e., those occurring in primary wards of corresponding specialties. In a recent study, Lim et al. (2021) demonstrate empirically that mean LOS of non-outlying patients is longer in wards that receive a larger number of outlying patients. To investigate this impact on RSCH data, we focus on GRT and ORT specialties, which account for 15% and 11% of specialty spells, respectively, at RSCH. According to Table 6.5, wards M4 and M2 are both primary wards for GRT specialty. However, our analysis show that mean LOS of this specialty in ward M2, in which outlying spells account for 26.0% of all spells, is 1.25 days longer than in ward M1, wherein outlying spells stand at 16.4%. Similarly, wards S4 and S1 are both primary wards for ORT. But mean LOS of this specialty in ward S4, with 73% outlying spells, is 2.72 days longer than in ward S1, with 30% outlying spells. These analyses highlight the impact of outlying patients on both outlying and non-outlying spells.

Ward Change and its Impact. While it is normal to assume that patients may change ward as a result of a change in their specialty, we observe that in many cases ward changes occur under the same specialty. Our analysis indicate an average (maximum) of 1.94 (8) ward changes within a specialty spell, with the breakdown illustrated in Figure 6.6 for different specialties. As illustrated in this figure, GSR has the highest average, and GRT and ORT have the joint-highest maximum number of ward changes. To evaluate the impact of ward changes on specialty LOS, in Figure 6.7 we illustrate the mean LOS for GAS specialty as a function of the number of ward changes and patient age band as an example. This figure suggests that as the number of ward changes increases, the mean LOS of the specialty increases too, with the impact being more significant for elderly patients. Hence, the ward change not only is an inconvenience for patients, but also exacerbates the bed shortage by increasing the LOS, in particular for the elderly patients who are the most common users of beds.

Overall Situation. The analyses conducted above portrays a hospital under immense pressure throughout the year. While a dedicated configuration was originally intended for the inpatient services at RSCH, a wing formation configuration with overlapping clusters is currently operating in the hospital. This is evident from the partitioning of specialties, derived

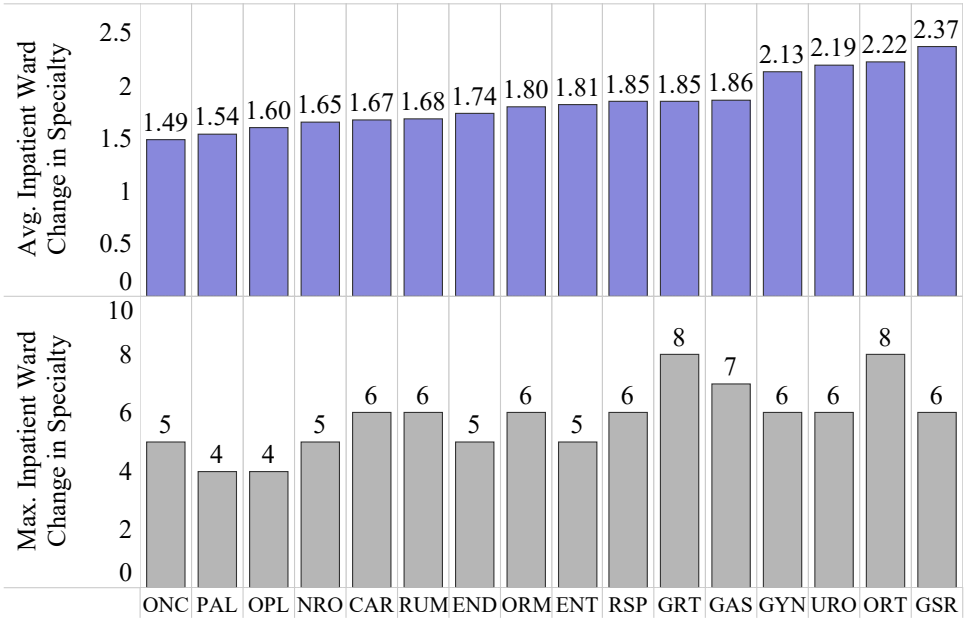


FIGURE 6.6: Average and maximum number of wards visited for different specialities at RSCH.

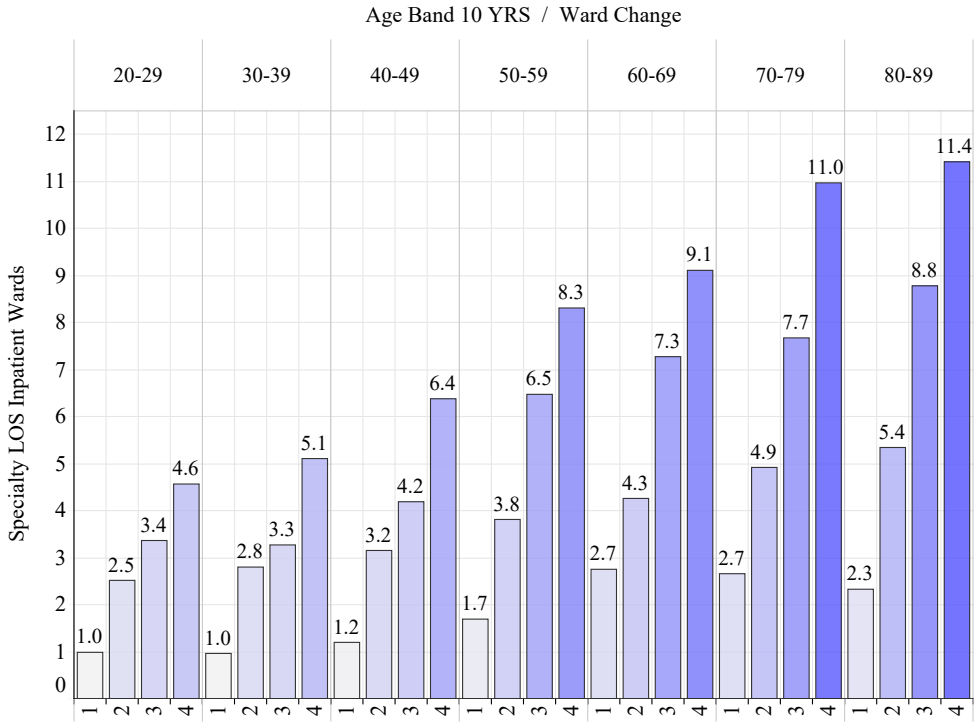


FIGURE 6.7: Mean LOS as a function of age band and number of ward changes for GAS specialty

from Table 6.5, as below

$$\mathcal{C}_{current} = \{\{ORT\}, \{ENT, OPL, ORM\}, \{URO, \boldsymbol{GYN}\}, \{GSR, \boldsymbol{GYN}\}, \{\boldsymbol{PAL}, RSP\}, \{\boldsymbol{GRT}\}, \{END, \boldsymbol{GRT}\}, \{GRD, RUM\}, \{GAS\}, \{\boldsymbol{PAL}, \boldsymbol{RUM}\}, \{\boldsymbol{GRT}, NRO\}\}$$

wherein highlighted specialties are allocated to more than one cluster. There also exists a substantial amount of patient outlying and ward changes in the hospital. Outlying patients not only experience a longer LOS, but may also negatively influence the LOS of other patients. This creates a vicious circle, wherein some patients are admitted to non-primary wards due to bed unavailability. This results in a longer LOS for both outlying and non-outlying patients, exacerbating the bed shortage problem, which in turn leads to more patients being outlied. A new configuration of inpatient services, in which beds are pooled in a structured way so as to reduce the number of patients outlying while minimizing the negative impacts of losing focus and increase in mix variability, is therefore likely to create some improvements. We investigate this in the next section.

6.3 Applying Reconfiguration Models

In this section, we apply the models developed in Chapters 3, 4, and 5 of this dissertation on RSCH data to find a better configuration of inpatient services. We first explain in Section 6.3.1 how different parameters required for our reconfiguration models are estimated from RSCH data. We then identify the best configurations obtained from our models under IWTT assumption in Section 6.3.2. In this section, we also compare the performance of the configurations obtained from our model with the current configuration of the hospital. Since patient-transfer rarely happens at RSCH (and other NHS hospitals), the IWTT assumption would be appropriate for comparison with the current situation. In Section 6.3.3, we obtain the best configurations from our models under FWTT assumption. A sensitivity analysis on cost parameters is conducted in Section 6.3.4.

6.3.1 Parameter Estimation

For estimating the input parameters of our models, we focus on specialty spells occurred within the last year of our data coverage period, i.e., from 01/10/2017 to 01/10/2018. The arrival rate, λ_i, of each specialty is evaluated by dividing the total number of spells of that specialty within this period by 365. The mean LOS of each specialty, m_i, is obtained by adding up the LOSs of that specialty within the one year period divided by the number of corresponding

spells. Nurse-to-patient ratio of each specialty, f_i, is provided by our hospital partners. See Table 6.6 for the set of specialties and corresponding λ_i, m_i, and f_i values.

Specialty	λ_i	m_i	f_i
CRD	5.29	3.50	0.15
END	1.55	3.35	0.15
ENT	4.47	2.59	0.15
GAS	9.29	3.78	0.20
GSR	6.22	5.28	0.20
GRT	11.33	8.90	0.20
GYN	3.18	1.93	0.15
NRO	2.29	4.68	0.15
ONC	1.98	5.54	0.20
OPL	0.21	2.29	0.20
ORM	0.13	2.82	0.15
PAL	0.08	7.73	0.20
RSP	6.33	6.36	0.20
RUM	1.90	5.22	0.15
ORT	6.99	5.80	0.15
URO	5.84	3.08	0.16

Time unit is one day.

TABLE 6.6: Input parameters for our reconfiguration models.

To apply our methodology for bed reconfiguration, we set the total number B of beds to 394. Dividing the total offered load, obtained as the sum-product of λ_i and m_i values given in Table 6.6, by this number of beds yields an overall traffic intensity of 86.7%. Note that the beds in the EAU and escalation wards are included in our total number of beds as advised by our hospital partners (Without these beds, the traffic intensity would rise to 97.5% making the queues extremely long.) We use the functional relation

$$\nu_i(d, \mathcal{A}) = \left(1 - \frac{\Delta \left(1 - \frac{|\mathcal{A}|}{n} \right)}{1 + e^{-\beta \left(\sum_{i \in \mathcal{A}} \frac{\lambda_i \tau_i}{d} - \epsilon \right)}} \right) \tau_i, \tag{6.1}$$

as proposed in Best et al. (2015) for estimating the mean LOS of specialty i patients admitted to a $d-$bed ward shared by a subset $\mathcal{A} \ni i$ of specialties. In Equation (6.1), $|x|$ represents the cardinality of set x, τ_i is the mean nominal LOS for specialty i patients (excluding the impact of focus and workload), Δ controls the impact of focus, and β and ϵ control the impact of workload as evaluated by $\sum_{i \in \mathcal{A}} \lambda_i \tau_i / d$ (see Figure 3 and the corresponding description in Best et al. 2015, p. 166). The substantial amount of patient outlying in the current configuration suggests that the impact of focus is negligible on the existing mean LOSs. As such, we set τ_i in Equation (6.1) to mean LOS values obtained from the data, m_i, given in Table 6.6. Following Best et al. (2015), we also set $\epsilon = 0.9$ and $\beta = 20.0$.

As stated in Section 3.3, our formulation and solution methodology are based on the assumptions that admission requests follow a Poisson process and LOS's are Exponentially distributed. We test these two assumptions for each specialty at RSCH. For admission requests, we estimate the coefficient of variation (CV) of daily arrival counts for each specialty and compare it with the corresponding theoretical CV under Poisson assumption, i.e., $1/\sqrt{\lambda}$ with λ the arrival rate of the specialty. The results presented in Table 6.7 show that the admission requests are nearly Poisson for all specialties except for GRT, which shows a highly variable arrival process.

Specialties	Admission Requests	
	Theoretical CV	Empirical CV
CRD	0.43	0.55
END	0.80	1.02
ENT	0.47	0.49
GAS	0.32	0.39
GSR	0.40	0.50
GRT	0.29	0.60
GYN	0.56	0.59
NRO	0.66	0.67
ONC	0.71	0.78
OPL	2.16	2.13
ORM	2.81	2.89
PAL	3.43	3.72
RSP	0.39	0.51
RUM	0.72	0.88
ORT	0.37	0.47
URO	0.41	0.48

TABLE 6.7: Empirical and Theoretical CV under Poisson assumption for arrival counts

For LOS's, we estimate the CV for each specialty and compare it with the theoretical CV under Exponential assumption, i.e., 1.0. The results presented in Table 6.8 show that LOSs are nearly Exponential for GRT, ORM, PAL, RSP, and ORT, but more variable than exponential for the remaining specialties. These analysis highlight the need for expanding our methodology to non-Poisson admission requests and non-Exponential service times for future research.

We continue with estimating the rest of the input parameters in our model. To estimate the daily cost $r(\mathcal{A})$ of a nurse working in a ward admitting patients of specialties in $\mathcal{A}$, we consider the average daily salary of a band 5 nurse equal to £103.03 per day plus a 10% additional

Specialties	Length of Stay Empirical CV
CRD	1.88
END	1.65
ENT	2.10
GAS	1.81
GSR	1.92
GRT	1.43
GYN	2.13
NRO	1.77
ONC	1.71
OPL	2.59
ORM	1.27
PAL	1.16
RSP	1.36
RUM	1.60
ORT	1.40
URO	1.83

TABLE 6.8: Empirical CV for LOSs.

payment for each additional specialty the nurse cares for in order to represent the higher value of multi-skilled nurses to the hospital. Three additional constraints are included to make the configurations proposed by our methodology viable. The first constraint is that medical and surgical specialties cannot be mixed in a cluster. The second constraint requires that the RSP must not be mixed with any other specialty to reduce the risk of in-hospital transmission of respiratory diseases. The last constraint is that for privacy reasons, GYN must also not be mixed with any other specialty. These three constraints are implemented by returning large cost values for clusters involving the specialties that cannot be mixed together in the intra-cluster allocation model. Note that the number of beds for RSP and GYN specialties must still be determined by our methodology.

The sequence $\mathcal{N}$ of specialties is created as follows: i) medical specialties excluding RSP are sorted in terms of their mean nominal LOS; ii) surgical specialties excluding GYN are sorted in terms of their mean nominal LOS, and inserted at the end of the sequence created in step (i); and iii) RSP and GYN are added to the end of the sequence created in step (ii). These steps ensure that specialties are sorted in terms of their mean nominal LOS as proposed in Izady and Muhamed (2019), while taking the three constraints mentioned above into account.

6.3.2 Best-Found Configurations with IWTT Assumption

Following discussion with our hospital partners, we set $c_w = 1030$, i.e., 10 times larger than the daily salary of a nurse. This figure is an illustrative value, reflecting the high cost of waiting for admission to patients. A sensitivity analysis is conducted in Section 6.3.4 to investigate

its impact on results. The best configurations obtained from our methodology are illustrated in Figures 6.8 and 6.9 for $\Delta = 0.0$ and $\Delta = 0.1$, respectively, representing negligible and high impacts of focus. These figures show that the best-found configuration is a COF configuration with either value of Δ as there exists at least one cluster with both dedicated and overflow wards. For $\Delta = 0.0$, as illustrated in Figure 6.8, the best-found configuration has 4 clusters, including one cluster for each of GYN and RSP specialties, one cluster for all surgical specialties, and one cluster for all medical specialties. The cluster involving all the surgical (medical) specialties, i.e., cluster 3 (cluster 4), has a total of 121 (214) beds allocated to it, 55 (71) of which are overflow beds. Overall, there exists a total of 268 dedicated beds and 126 overflow beds in the best-found configuration. We also observe that ORM, ENT, URO, OPL, and END specialties have not been allocated a dedicated ward and are treated in overflow wards of their clusters. As Δ increases to 0.1, as illustrated in Figure 6.9, a new cluster is formed by separating PAL from the cluster of medical specialties and allocating 3 dedicated beds to it. Also, URO is allocated a dedicated ward with 13 beds. Further, the number of overflow beds in clusters 3 and 4 decrease by 17 each as compared to the scenario with $\Delta = 0.0$.

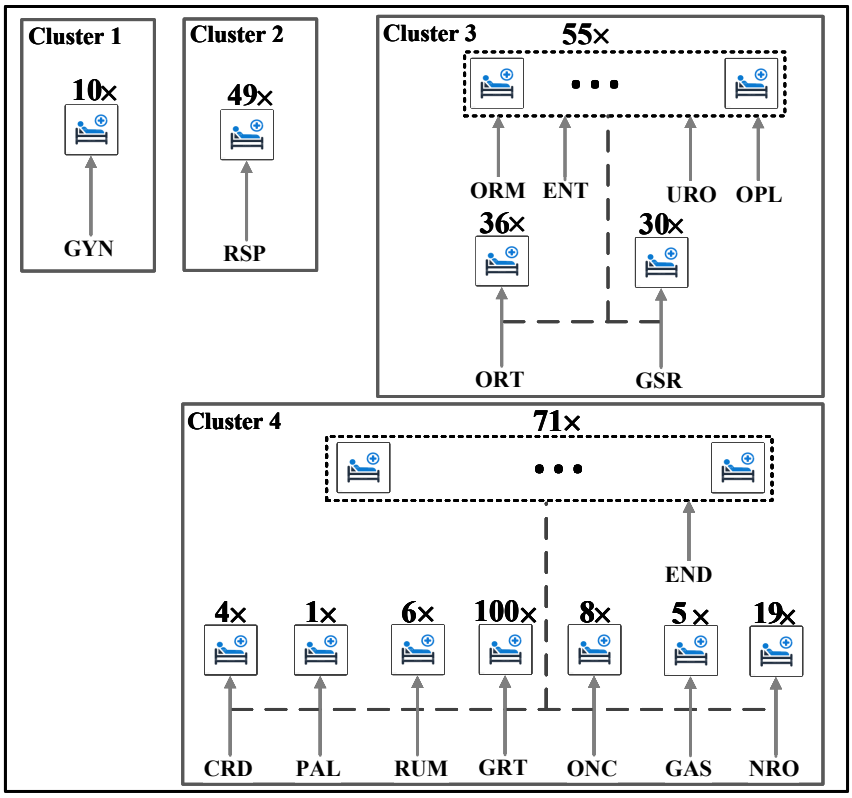


FIGURE 6.8: The best-found configuration under IWTT assumption with $\Delta = 0.0$

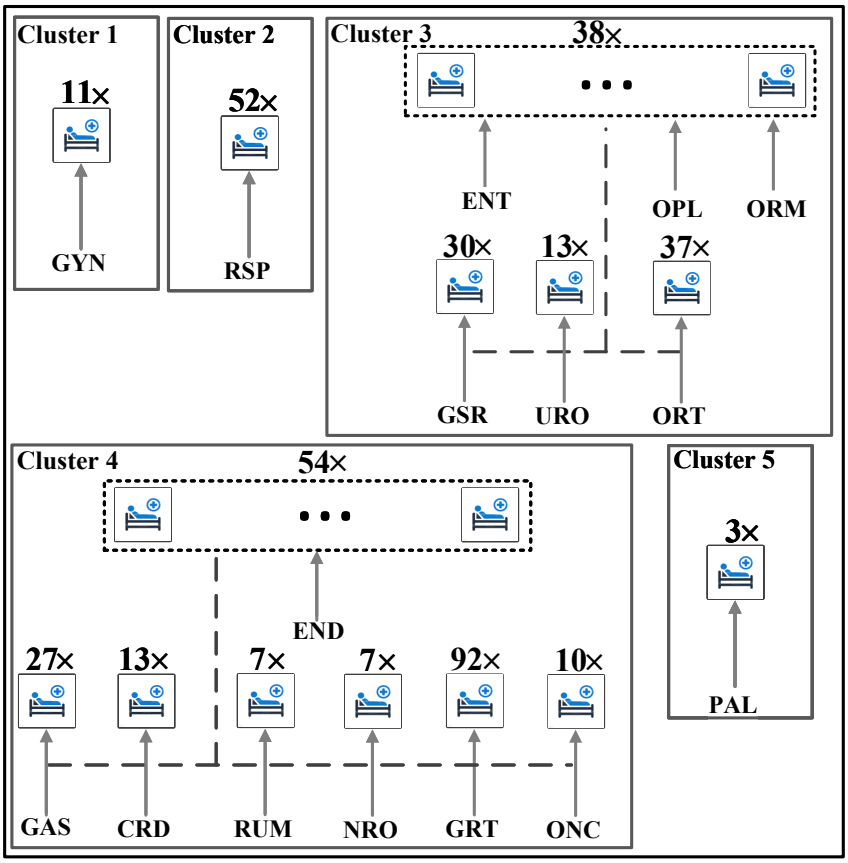


FIGURE 6.9: The best-found configuration under IWTT assumption with $\Delta = 0.1$

To compare the best configurations obtained from our methodology with the current configuration, we evaluate the performance of each configuration using a discrete-event simulation model. For the configurations generated by our models, we simulate each cluster, independently from other clusters, using the specialty parameters given in Table 6.6 and Equation (6.1). For the current configuration, we simulate each ward, independently from other wards, using the arrival rates and mean LOSs of different wards as given in Table 6.9 instead of those of specialties. This is because there exists a significant level of patient outlying in the current configuration as demonstrated in Section 6.2. Hence, simulating the wards with their overall arrival rates and LOSs would enable us to represent the current situation more accurately without having to capture the complex flows of patients between wards. The inter-arrival times and LOSs are assumed to be exponential in all of our simulation models. We replicate each simulation model 10 times, with each replication running for 100000 days, to estimate mean numbers of patients waiting in the queues and mean numbers of occupied dedicated and overflow beds, using which cost functions are evaluated.

We obtain the mean total daily costs of £11,840 (£7,840) for our best-found configurations for $\Delta = 0.0$ ($\Delta = 0.1$) as compared to £109,183 for the existing configuration. These figures

suggest substantial amounts of saving with our proposed configurations. These savings are achieved by a 97% (99%) reduction in the numbers waiting for admission at the expense of a 38.5% (19.1%) increase in staffing cost for $\Delta = 0.0$ ($\Delta = 0.1$).

Wards	M1	M2	M3	M4	M5	M6	M7	M8	S1	S2	S3	S4	S5	E	EAU
Arrival Rate	3.40	2.52	2.47	2.31	3.21	3.90	3.96	1.29	4.69	5.36	9.80	4.21	4.58	0.80	41.00
Mean LOS	8.11	11.51	5.61	12.52	6.72	6.74	7.20	18.02	4.78	3.00	2.68	6.18	5.78	5.27	0.60

Note: time unit is one day.

TABLE 6.9: Arrival rates and mean LOSs for inpatient wards at RSCH

We now compare the performance of our best-found configurations (presented in Figures 6.8 and 6.9) with those of fully dedicated (DED), fully flexible (FLX), wing formation (WNG), and earmarking (ERM) configurations using simulation. To obtain the best allocation of beds and specialties for each of these configurations, we amend the inter- and intra-cluster allocation models in the following ways. For DED, we revise the inter-cluster allocation model so that only one specialty is allocated to each cluster. There will be no need to change the intra-cluster allocation model for this configuration as there is only one specialty in each cluster, the performance of which is evaluated using the exact results for $M/M/s$ queues. There is no need to revise the intra-cluster allocation model for WNG configuration either as all the beds are given to the overflow ward of the cluster, whose performance is also evaluated using $M/M/s$ results. The partitioning and inter-cluster model for this configuration works in the same way as for COF. For ERM, the inter-cluster allocation model is revised so that only 4 clusters, one involving all medical specialties except RSP, one involving all surgical specialties except GYN, one involving only RSP, and one involving only GYN, are considered. The intra-cluster allocation model for this configuration works in exactly the same way as for COF. For FLX, the inter-cluster model is changed in the same way as for ERM. In addition to this, no intra-cluster bed allocation is conducted as all beds are given to the overflow ward of the cluster, whose performance is evaluated using $M/M/s$ results.

The savings in total, waiting, and staffing costs obtained from using the best-found configurations as compared to the other configurations are illustrated in Table 6.10. The results in this table suggest that WNG, FLX, ERM, and DED (ERM, WNG, FLX, and DED) rank 2nd, 3rd, 4th, and 5th, respectively, in terms of mean total daily cost following our best-found configuration, i.e., the COF configuration, for $\Delta = 0.0$ ($\Delta = 0.1$). DED has the lowest staffing cost but also the highest waiting cost for both values of Δ. WNG and FLX, on the other hand, have the joint-lowest waiting cost but also the highest staffing costs for $\Delta = 0.0$. For $\Delta = 0.1$, COF has the lowest waiting cost and second-lowest staffing cost following DED.

Scenario	Configuration	Total Cost	Staffing Cost	Waiting Cost
$\Delta = 0.0$	DED	60.6%	−25.7%	85.3%
	ERM	10.7%	1.8%	27.4%
	FLX	6.5%	20.7%	−72.3%
	WNG	6.4%	20.6%	−72.3%
$\Delta = 0.1$	DED	56.3%	−15.2%	95.5%
	FLX	31.1%	29.8%	44.6%
	WNG	25.4%	8%	79.5%
	ERM	11%	3.1%	58.2%

TABLE 6.10: The savings obtained from the best-found configuration (COF) as compared to other configurations under IWTT assumption.

6.3.3 Best-Found Configurations with FWTT Assumption

We run our methodology with $\gamma^{-1} \in \{1, 15, 30\}$ days in this section, assuming the cost of a patient abandoning the queue $c_a = 1030$. The sensitivity of results to this cost value is investigated in Section 6.3.4. Figures 6.10 and 6.11 depict the best-found configurations obtained for $\Delta = 0.0$ and $\Delta = 0.1$, respectively, with $\gamma^{-1} = 1$ day. In comparison with best configurations obtained under IWTT, we observe that the numbers of clusters increase to 7 with both focus scenarios; medical and surgical specialties are each split into two clusters. The total number of overflow beds has also reduced (risen) to 79 (93) for $\Delta = 0.0$ ($\Delta = 0.1$) as compared to the corresponding scenarios under IWTT. Further, no bed is allocated to PAL given its small arrival rate and long LOS.

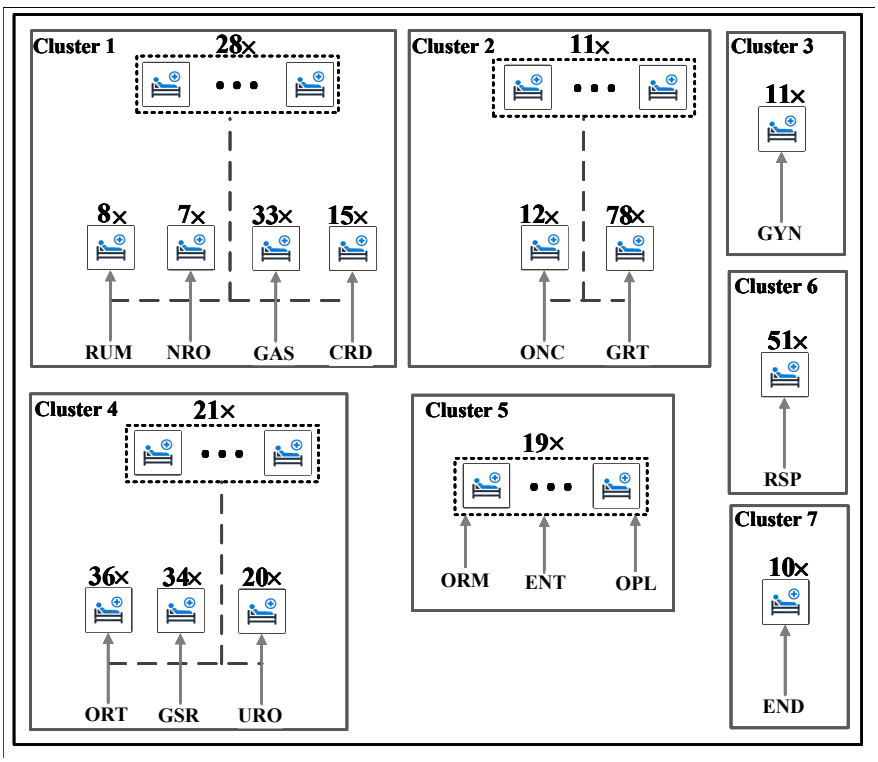


FIGURE 6.10: The best-found configuration under FWTT with $\gamma^{-1} = 1$ day and $\Delta = 0.0$.

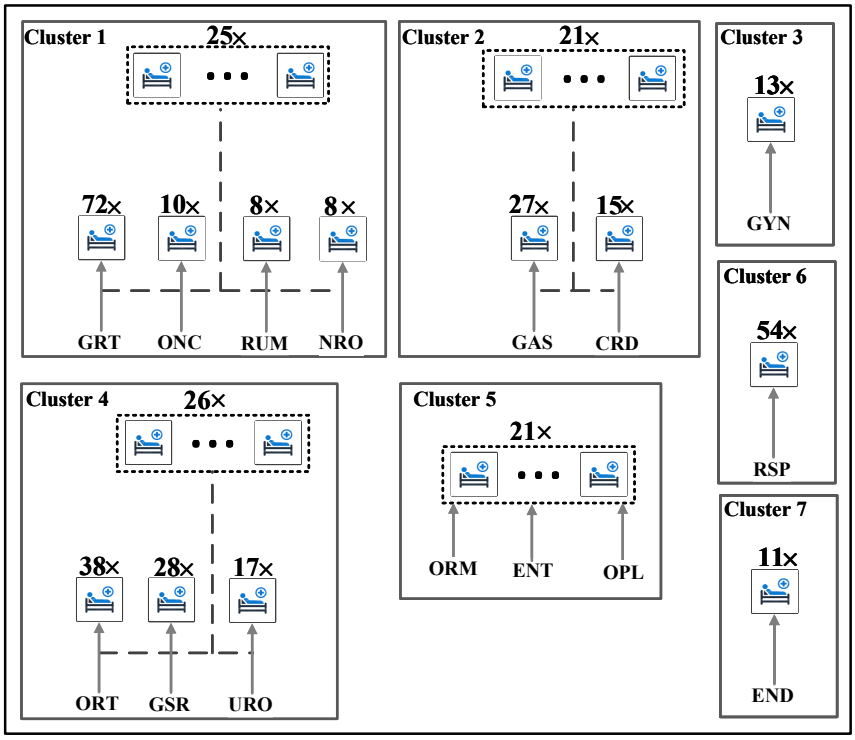


FIGURE 6.11: The best-found configuration under FWTT with $\gamma^{-1} = 1$ day and $\Delta = 0.1$.

In Table 6.11, we compare the performance of our best-found configuration with the other configurations under FWTT assumption with $\gamma^{-1} = 1$ day using simulation. The results in the table show that WNG and ERM are the 2nd-best configurations for $\Delta = 0.0$ and $\Delta = 0.1$, respectively, in terms of mean total daily cost. On the other hand, FLX is the worst configuration for both focus scenarios. DED provides the lowest staffing costs and the highest abandonment costs for both values of Δ. ERM, on the other hand, produces the lowest abandonment cost with $\Delta = 0.0$, and the same abandonment cost as COF with $\Delta = 0.1$.

Scenario	Configuration	Total Cost	Staffing Cost	Abandonment Cost
$\Delta = 0.0$	FLX	26.5%	31.6%	0%
	DED	9.5%	−4.2%	64.9%
	ERM	4.0%	7.6%	−82.1%
	WNG	3.5%	1.0%	26.4%
$\Delta = 0.1$	FLX	33.3%	34.3%	0%
	DED	14.9%	−1.5%	84.5%
	WNG	9.4%	4.0%	64.4%
	ERM	1.4%	4.2%	0%

TABLE 6.11: The savings obtained from the best-found configuration (COF) as compared to other configurations under FWTT assumption with $\gamma^{-1} = 1$ day.

The best-found configurations for $\gamma^{-1} = 15$ days are presented in Figures 6.12 and 6.13 for $\Delta = 0.0$ and $\Delta = 0.1$, respectively. As illustrated in these figures, the numbers of clusters increase to 8 and 9, while total numbers of overflow beds reduce to 77 and 84 for $\Delta = 0.0$ and

$\Delta = 0.1$, respectively, as compared to the corresponding scenarios with $\gamma^{-1} = 1$ day. Cost comparisons with other configurations are provided in Table 6.12. The main observation is that COF remains the lowest cost configuration, followed by WNG, DED, ERM and FLX in the same order.

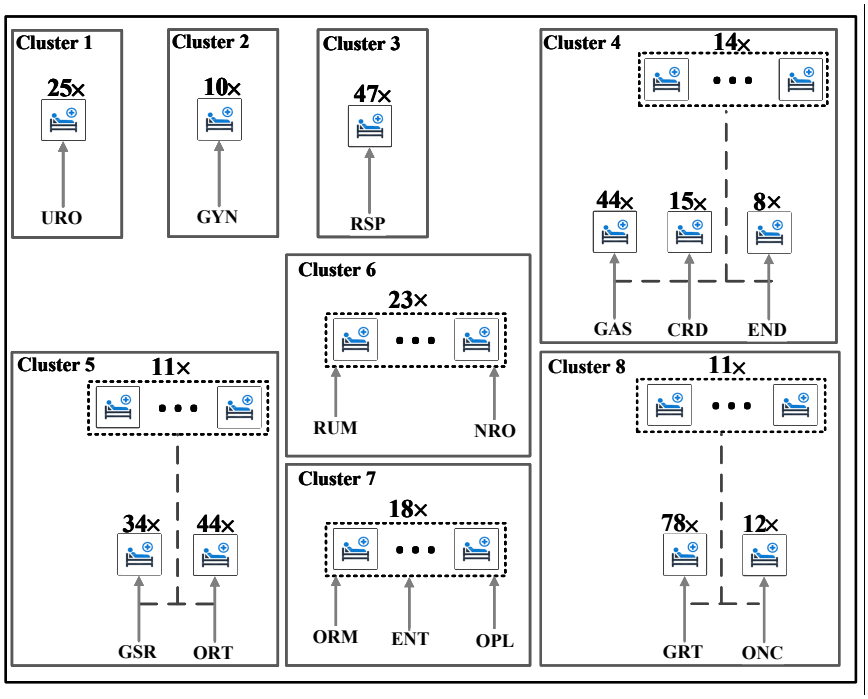


FIGURE 6.12: The best-found configuration under FWTT with $\gamma^{-1} =$ 15 days and $\Delta = 0.0$.

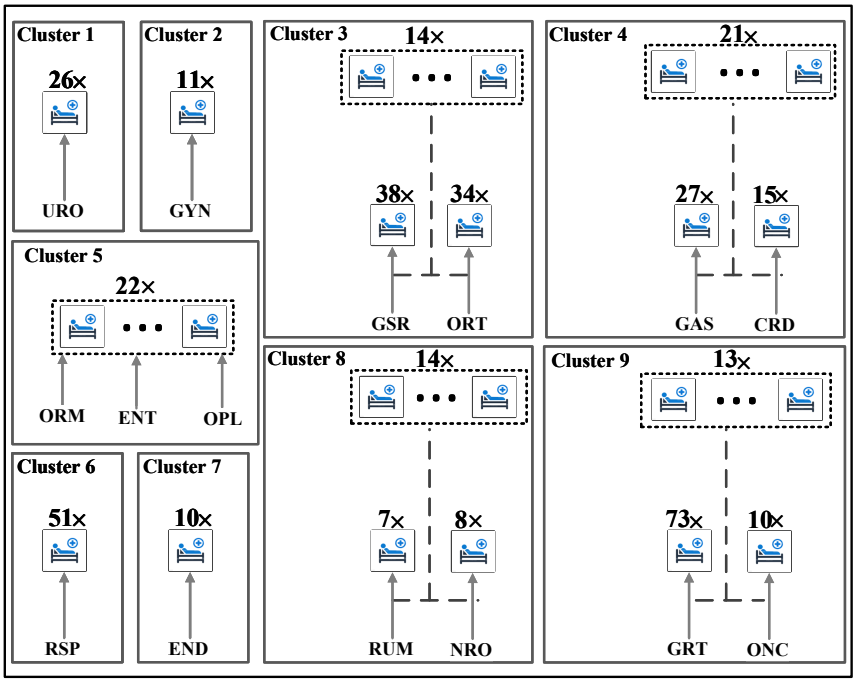


FIGURE 6.13: The best-found configuration under FWTT with $\gamma^{-1} =$ 15 days and $\Delta = 0.1$.

Scenario	Configuration	Total Cost	Staffing Cost	Abandonment Cost
$\Delta = 0.0$	FLX	32.1%	35.2%	0%
	ERM	9.0%	12.1%	0%
	DED	3.3%	1.8%	25.8%
	WNG	2.4%	2.8%	−5.2%
$\Delta = 0.1$	FLX	34%	35.2%	0%
	ERM	8.3%	9.9%	0%
	DED	7.3%	3.2%	67%
	WNG	4.4%	1.2%	60.7%

TABLE 6.12: The savings obtained from the best-found configuration (COF) as compared to other configurations under FWTT assumption with $\gamma^{-1} = 15$ days.

Finally, the best-found configurations for $\gamma^{-1} = 30$ days are presented in Figures 6.14 and 6.15 for $\Delta = 0.0$ and $\Delta = 0.1$, respectively. We observe in Figure 6.14 that the best-found configuration is in fact a WNG configuration for $\Delta = 0.0$ as no dedicated wards are included in clusters involving two or more specialties. For $\Delta = 0.1$, on the other hand, Figure 6.15 shows that COF is the best-found configuration. Figures 6.14 and 6.15 also show that, with $\gamma^{-1} = 30$ days, the number of clusters increases to 13 for $\Delta = 0.0$, while it remains at 9 for $\Delta = 0.1$, compared with the corresponding scenarios with $\gamma^{-1} = 15$ days. Further, the total numbers of overflow beds reduce to 46 and 72 for $\Delta = 0.0$ and $\Delta = 0.1$, respectively, when compared with the corresponding scenarios with $\gamma^{-1} = 15$ days. The saving figures provided in Table 6.13 show that DED, ERM and FLX are the 2nd, 3rd and 4th lowest cost configuration for $\Delta = 0.0$, while ERM, WNG, DED, and FLX are the 2nd, 3rd, 4th and 5th lowest-cost configurations for $\Delta = 0.1$. Note that it is not possible to force our models to produce a clustered-overflow configuration. As such, we could not present the saving results for this configuration in Table 6.13 for $\Delta = 0.0$.

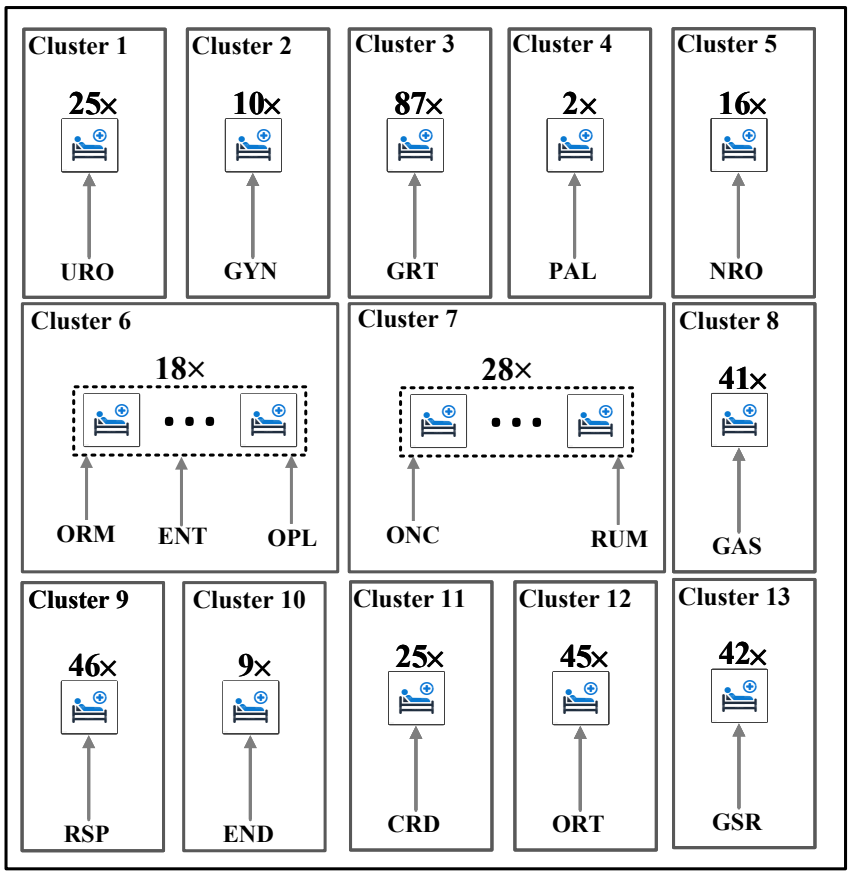


FIGURE 6.14: The best-found configuration under FWTT with $\gamma^{-1} = 30$ days and $\Delta = 0.0$.

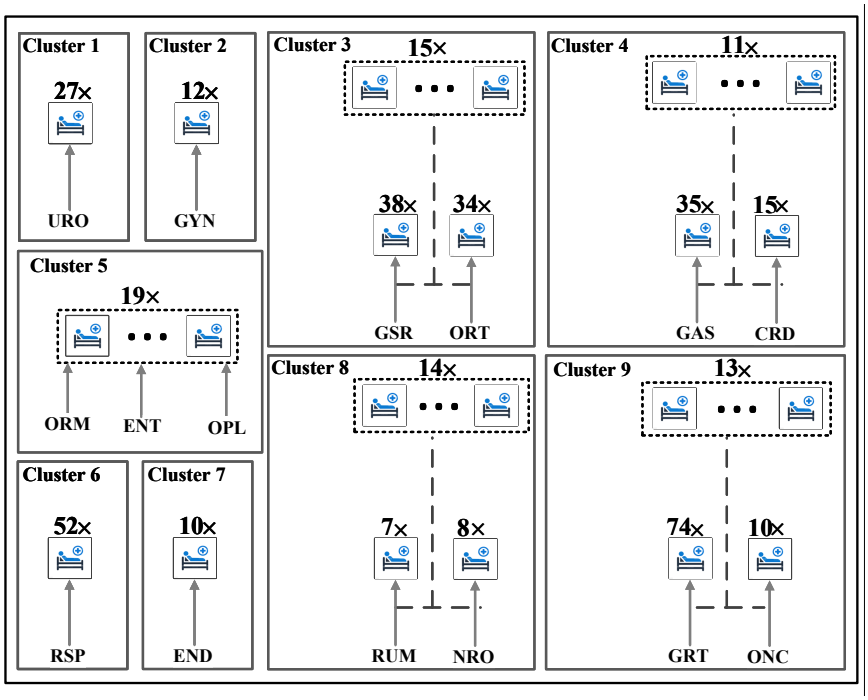


FIGURE 6.15: The best-found configuration under FWTT with $\gamma^{-1} = 30$ days and $\Delta = 0.1$.

Scenario	Configuration	Total Cost	Staffing Cost	Abandonment Cost
$\Delta = 0.0$	FLX	31.4%	33.4%	0%
	ERM	5.3%	5.0%	11.6%
	DED	4.6%	3.4%	28.9%
$\Delta = 0.1$	FLX	35.6%	36.6%	0%
	DED	6.6%	3.4%	65.1%
	WNG	3.7%	1.4%	56.1%
	ERM	3.1%	6.3%	0%

TABLE 6.13: The savings obtained from the best-found configurations as compared to other configurations under FWTT assumption with $\gamma^{-1} = 30$ days. For $\Delta = 0.0$ ($\Delta = 0.1$), WNG (COF) is the best-found configuration.

6.3.4 Sensitivity Analysis

In this section, we perform a sensitivity analysis on the impact of cost parameters. In particular, we investigate the impact of reducing nurse skill cost from the original 10% considered in Sections 6.3.2 and 6.3.3 to 2% per additional specialty, and the impact of increasing waiting/abandonment cost from the original £1030 considered in Sections 6.3.2 and 6.3.3 to £2060, on the best configurations obtained from our models. We consider a smaller nurse skill cost as in some hospitals a wider skill-set may not lead to a substantial increase in salary. A larger waiting/abandonment cost is considered to reflect the higher cost associated with keeping patients waiting/turning them away for some specialties or hospitals.

The number of clusters as well as the numbers of overflow and dedicated beds for the configurations obtained from our models are given in Tables 6.14 and 6.15 for IWTT and FWTT assumptions, respectively, for the two different values of nurse skill cost. We observe that, as the skill cost reduces, the model tends to form fewer clusters with a larger number of overflow beds and a smaller number of dedicated beds. This is to take advantage of the benefits of pooling when nursing costs do not vary much by the range of skills nurses posses.

Cross-Training Cost	No. Clusters	No. Dedicated Beds	No. Overflow Beds
10%	4	268	126
2%	4	123	271

TABLE 6.14: Sensitivity analysis on nurse skill cost under IWTT assumption.

γ^{-1}	Skill Cost	No. Clusters	No. Dedicated Beds	No. Overflow Beds
1	10%	7	315	79
	2%	5	263	131
15	10%	8	317	77
	2%	7	251	143
30	10%	13	348	46
	2%	8	93	301

TABLE 6.15: Sensitivity analysis on nurse skill cost under FWTT assumption.

The number of clusters as well as the numbers of overflow and dedicated beds for the configurations obtained from our models are given in Tables 6.16 and 6.17 for IWTT and FWTT assumptions for the two different values of waiting cost and abandonment cost, respectively. We observe in these tables that, as the abandonment or waiting cost increases, the model tends to allocate a larger number of overflow beds and a smaller number of dedicated beds. The number of clusters, however, does not show a monotone trend.

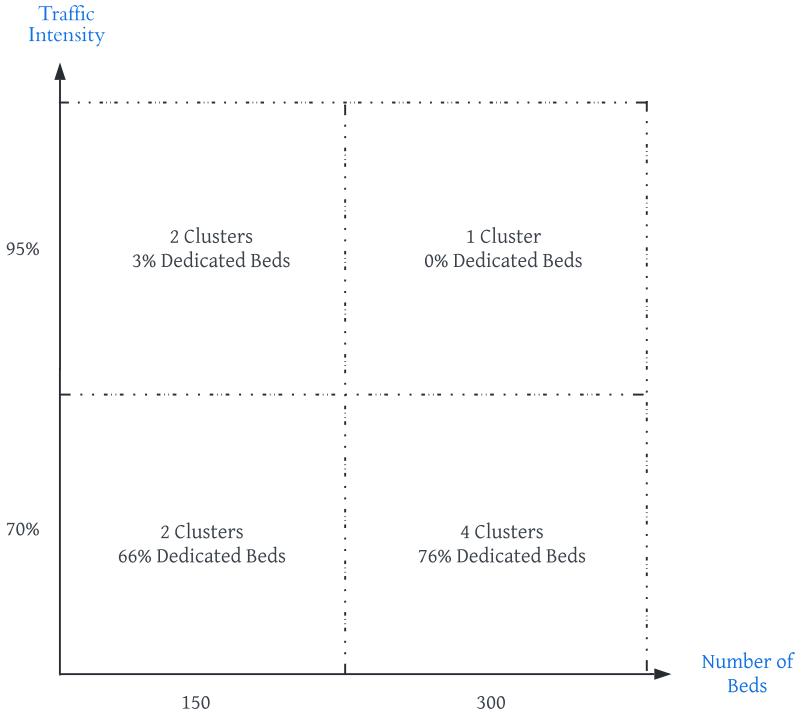
Cost of Waiting	No. Clusters	No. Dedicated Beds	No. Overflow Beds
1030	4	268	126
2060	4	197	197

TABLE 6.16: Sensitivity analysis on daily cost of waiting under IWTT assumption.

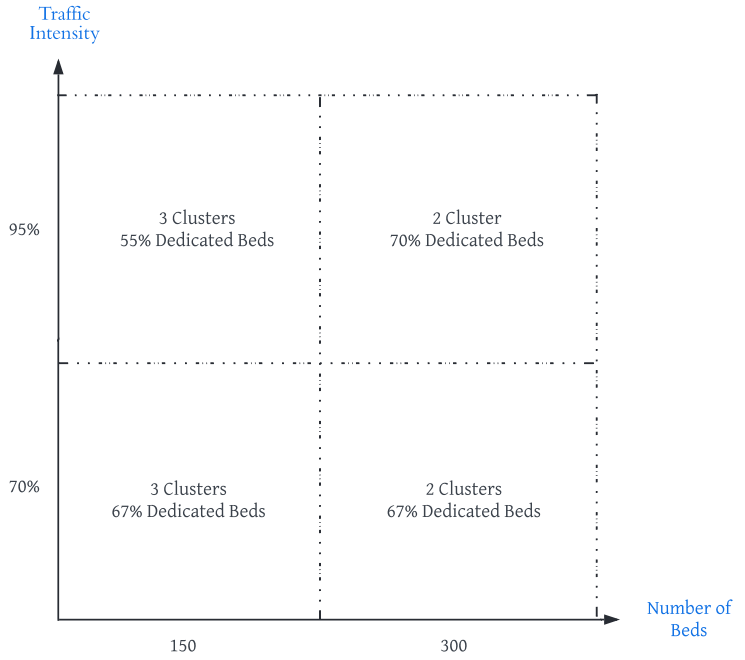
γ^{-1}	Cost of Abandonment	No. Clusters	No. Dedicated Beds	No. Overflow Beds
1	1030	7	315	79
	2060	7	244	150
15	1030	8	317	77
	2060	9	305	89
30	1030	13	348	46
	2060	10	309	85

TABLE 6.17: Sensitivity analysis on cost of abandoning under FWTT assumption.

For the final part of our sensitivity analysis, we investigate the impact of bed numbers and traffic intensity on the clustering of specialties and corresponding bed allocation. We focus on medical specialties, which have a total of 214 beds allocated to them and experience an overall traffic intensity of 89% in the current system. We consider two new values for total bed numbers, 150 and 300, and adjust the admission rates to achieve two new values for the overall traffic intensity, 70% and 95%. This results in four scenarios. We use our methodology to find the best configuration for each of the scenarios. The number of clusters and percentage of dedicated bed numbers are presented in Figure 6.16, panels (a) and (b), for $\Delta = 0.0$ and $\Delta = 0.1$, respectively.



(a)



(b)

FIGURE 6.16: Sensitivity analysis on impact of bed numbers and traffic intensity on clustering of medical specialties under IWTT assumption for $\Delta = 0.0$ (a) and $\Delta = 0.1$ (b)

The results presented in panel (a) of Figure 6.16 show that, when both traffic intensity and bed numbers are small (large), the model tends to create more (fewer) clusters with a

larger (smaller) percentage of dedicated beds. A large percentage of overflow beds is allocated when bed numbers is small and traffic intensity is high, whereas the largest number of clusters with the largest percentage of dedicated beds is formed when beds numbers is large and traffic intensity is low. The results in panel (b) illustrate that it is much harder to predict the number of clusters and percentage of dedicated beds when the impact of focus is positive. In general, the percentage of dedicate beds increase with the impact of focus.

6.4 Conclusions

In this chapter, we reported the interim results for our collaborative project with RSCH on inpatient bed reconfiguration. We started with analysing the data covering a three-year period to understand the current status of inpatient services in the hospital. This analysis showed that daily occupancy figures exceeded 85% for significant periods of time for both medical and surgical wards. We also observed a significant level of patient outlying in the hospital data. More specifically, we observed that around half of specialty spells identified in RSCH data were outlying ones, and that 28% of total workload in the hospital is attributed to outlying patients. We further demonstrated the impact of patient outlying on outlying and non-outlying patients. In particular, we illustrated through examples that the specialty LOS of outlying spells are longer than those of non-outlying spells, and that specialty LOS of non-outlying patients in wards with a larger percentage of outlying spells is longer than those in wards with a smaller percentage of outlying spells. In addition to patient outlying, we examined the extent and impact of ward changes within a specialty spell. This analysis showed an average (maximum) of 1.94 (8) ward changes within a specialty spell. Focusing on GAS as one of the specialties with a high average number of ward changes, we observed that specialty LOS increases with the number of ward changes, and this impact is more dramatic for elderly patients. These analyses indicated the high level of pressure on hospital inpatient beds and the immediate need for reconfiguration of services.

A critical step for assessing the extent and impact of patient outlying in the hospital was accurate identification of specialties for different episodes of care. We proposed using the HRG codes for identifying the specialties instead of relying on the specialties given in hospital data. This is because the specialty field in the NHS datasets typically reflects the specialty of the consultant in charge of the episode which may not match the clinical specialty required for the patient due to patient outlying. In contrast, the HRG codes are related to patient conditions and so are a better identifier for the patient specialty. The specialties defined using HRG codes were also used for our reconfiguration models.

Having estimated the input parameters from the data, we run our models to obtain a better configuration of inpatient services than the existing one. Following advice from our hospital partners, we added constraints to our models so that surgical and medical specialties were not mixed with each other in the same cluster, and that RSP and GYN each be allocated individual clusters. We run our models with both IWTT and FWTT assumptions. For both assumptions, we considered two different value of Δ, representing no impact and a high impact of focus. For FWTT assumption, we considered three different value of γ, representing low, medium, and high thresholds for patients' waiting time. This resulted in a total of eight scenarios for our experiments.

We observed in our experiments' results that the best-found configuration from our model is a COF configuration in seven scenarios and a WNG configuration in one scenario. To investigate this further, we obtained the best partitioning and bed allocation under the other four main configurations through a revised version of our model for each of our scenarios. We then compared the mean total daily cost of the best-found configuration with that of the other four configurations using a discrete-event simulation model. The simulation results confirmed that COF has the lowest mean total daily cost in seven scenarios, and WNG in one scenario. We also compared the mean total daily cost of the best configurations obtained from our models under IWTT assumption with that of the existing configuration at RSCH using the simulation model. This illustrated the potential for significant cost savings with the configurations obtained from our models.

We observed in our numerical experiments that our models in general tend to form a larger number of clusters under both IWTT and FWTT assumptions as the impact of focus increases. Under FWTT, the number of clusters tends to increase while the number of overflow beds tends to decrease as patients' waiting time threshold increases. We also conducted some sensitivity analysis on the impact of cost parameters. We observed in this analysis that, as the patients' abandonment or waiting cost increases, the model tends to allocate a larger number of overflow beds and a smaller number of dedicated beds. As the nurses' skill cost increases, on the other hand, the model tends to allocate a smaller number of overflow beds and a larger number of dedicated beds.

7 Conclusions and Future Research

7.1 Introduction

Many inpatient departments around the world were already experiencing a high amount of pressure when they were hit by the COVID-19 pandemic. These pressures were amplified during the pandemic as many more patients required acute and intensive care. This led to the cancellation and postponements of a large number of non-urgent procedures. Despite widespread vaccination campaigns, the pandemic is still not over two years later. The impact of the pandemic, combined with an ageing population and prevalence of long-term conditions, would only increase the pressure on inpatient services in the future. This motivated our research into inpatient services, aiming to make more efficient use of existing resources by applying principles of operations management and operations research.

In Chapter 1, we elaborated on the causes and consequences of shortage in inpatient beds. The common solutions for addressing this shortage were also discussed, and the reconfiguration of inpatient services was identified as a cost-effective solution pursued in this research. The configuration of inpatient services determines the subsets of specialties treated in each ward as well as the number of beds allocated to each ward. We explained that, once a decision is made to admit a patient, the primary ward(s) of the patient is decided based on her clinical needs. If no bed becomes available in the primary ward(s) of the patient before her waiting time threshold is reached, the patient would have to be transferred to another hospital. The waiting time threshold varies depending on the urgency of patient needs and the tolerance to wait in different healthcare systems. The major configurations of inpatient services, including fully dedicated, fully flexible, earmarking, wing formation, and clustered overflow configurations, were explained next. Their pros and cons were also discussed in terms of the level of focused care, the amount of slack capacity and mix variability, and cross-training costs. We further explained that patient outlying typically happens in hospitals as a short-term solution for accommodating patients whose primary wards are full. This will, however, negatively affect the quality of care and system performance. It is also an indication that a reconfiguration

of inpatient services is needed. We finally emphasized that the COF configuration acts as a meta-model capturing the other four configurations as special cases.

In Chapter 2, we first reviewed the literature of inpatient bed allocation. This literature was divided into three main streams, with the last stream being the closest to our research. This stream concentrates on papers that, given a set of specialties and a total number of beds, seek to find the optimal clustering of specialties and the corresponding allocation of beds simultaneously. The prominent papers in this stream were identified as those of Best et al. (2015) and Izady and Muhamed (2019), with the former focusing on the wing formation configuration and the latter on the clustered overflow configuration. It was discussed that the methodology proposed in Best et al. (2015) does not capture earmarking and clustered overflow configurations, works only with an FWTT assumption, and does not take nursing costs into account. The methodology proposed in Izady and Muhamed (2019), on the other hand, does consider all the major configurations and takes nursing costs into account, but it only works with a ZWTT assumption. Further, its application has only been investigated using a small dataset from a paediatric department with only seven specialties. A gap was therefore identified in the bed allocation literature for a methodology that determines the optimal configuration of inpatient services in an existing hospital under different levels of waiting time threshold. It was also highlighted that potential savings of applying such methodology must be estimated using real data.

Next, we explained the role of queueing performance evaluation models in clustering and bed allocation methodologies. The relevant literature was reviewed in nine different categories, specified based on the level of interaction between different wards and the waiting time threshold of patients. Given our focus on the COF configuration, two of these categories were found relevant to our research: (i) the category of systems with a hierarchical interaction between wards and an FWTT for patients; and (ii) the category of systems with a hierarchical interaction between wards and an IWTT for patients. The first (second) category is useful for performance evaluation of a cluster in the COF configuration with an FWTT (IWTT) assumption. We discussed that no analytical methodology is proposed in the literature for performance evaluation of systems in the first category, which we refer to as overflow abandonment systems. For systems in the second category, which we refer to as overflow delay systems, an analytical methodology is proposed by Chevalier and Van den Schrieck (2009). However, they assume that mean service times in dedicated and overflow wards are the same. It is important to capture situations with different mean service times in dedicated and overflow wards since empirical evidence in the literature (see, e.g., Best et al. 2015) shows that dedicated wards are likely to have a shorter LOS than overflow wards as they enjoy a higher level of focused care. A gap was therefore identified for a performance evaluation model for overflow abandonment

systems and overflow delays systems accounting for mean LOSs being different in dedicated and overflow wards.

Based on the literature review in Chapter 2 as outlined above, proposing a methodology for finding the best configuration of inpatient services given an overall number of beds, a set of specialties, and a given finite or infinite waiting time threshold was defined as our major objective. We also set out to evaluate the effectiveness of our proposed methodology by testing it on real data from a major hospital.

Chapter 3 started off with a mathematical representation for the COF configuration, which is used as the basis for developing our methodology given its versatility. We then outlined our main assumptions as follows: A(i) the admission requests for each specialty follow a Poisson process; A(ii) the LOSs of each specialty are i. i. d. random variables following an Exponential distribution; A(iii) the mean LOS of each specialty in a ward is a function of the workload and the number of specialties treated in the ward; A(iv) the waiting time thresholds of patients in a system with an FWTT assumption are i. i. d. random variables following an Exponential distribution with the same rate for all specialties; and A(v) the arrival, service, and abandonment processes are mutually independent. A cost-minimization problem, similar to that of Izady and Muhamed (2019), was formulated next. We proved that this problem can be decomposed into a partitioning and inter-cluster bed allocation problem and an intra-cluster bed allocation problem.

In Chapter 4, we first developed an exact methodology for performance evaluation of an overflow delay system with only two customer classes. Our experiments with this method illustrated that the computation time grows exponentially with the number of customers in the system. This highlighted the need for developing an approximation. Next, building on the approximation proposed by Chevalier and Van den Schrieck (2009) for overflow delay systems and the approximation proposed by Izady and Muhamed (2019) for overflow loss systems, we proposed approximation methodologies for overflow delay systems as well as overflow abandonment systems with potentially different mean service times in dedicated and overflow facilities. We examined the accuracy of our approximations by conducting a set of numerical experiments which indicated a practical level of accuracy for main performance metrics. They also showed that our approximation for the overflow delay system is more accurate than that of Chevalier and Van den Schrieck (2009). Further, timing of our experiments showed that our approximations work very fast, returning the required metrics for a given system in a fraction of a second. This makes them suitable for application within an optimization framework for finding the optimal bed allocation.

In Chapter 5, using examples we illustrated that the objective function of the intra-cluster bed allocation model is neither convex nor differentiable under either the IWTT or FWTT

assumption. This implies the need for a gradient-free heuristic optimization approach. A heuristic search called CDOS is applied in Izady and Muhamed (2019) for systems with a ZWTT assumption. We tested the accuracy of this heuristic for our problem by comparing its results with that of the complete enumeration for a series of scenarios with a hypothetical cluster with only two specialties. These tests indicated a high level of accuracy of CDOS heuristic under both IWTT and FWTT assumptions for the scenarios considered. For the partitioning and inter-cluster bed allocation problem, we adapted the dynamic programming and restrictive sequencing approach proposed in Best et al. (2015).

In Chapter 6, we investigated the application of our reconfiguration methodology using real data from inpatient wards at RSCH. Following a comprehensive data analysis, we concluded that RSCH is operating a wing formation configuration with overlapping clusters under a high level of daily occupancy. There also exists an outlying frequency of about 50%, which creates about 30% of the total workload in the hospital. This level of patient outlying prolongs the LOS of all patients in the hospital, including the non-outlying ones. There also exists a large number of ward changes within each specialty spell, which in addition to inconvenience for patients, results in longer specialty spells in particular for older patients.

We then applied our reconfiguration methodology to obtain the best configuration of inpatient services under both IWTT and FWTT assumptions. We illustrated that, for two scenarios, one with a high impact of focus and the other with no impact of focus, substantial savings can be made by using the configurations proposed by our methodology as compared to the existing configuration. We also showed how our methodology can be modified so that it produces the best partitioning and bed allocation for each the other four major configurations. We then compared the performance of the five major configurations (including COF) for eight different scenarios, created by considering all combinations of $\Delta \in \{0.0, 0.1\}$ and $\gamma^{-1} \in \{1, 15, 30, \infty\}$ ($\gamma^{-1} = \infty$ implies an IWTT system), using a discrete-event simulation model. We observed that COF configuration produces the least mean total daily cost in seven out of eight scenarios. The minimum savings obtained with the COF configuration in these seven scenarios vary between 1.4% (for the scenario with $\Delta = 0.1$ and $\gamma^{-1} = 1$ day) to 11.0% (for the scenario with $\Delta = 0.1$ and $\gamma^{-1} = \infty$). The only scenario for which COF is not the least cost scenario is the one with $\Delta = 0.0$ and $\gamma^{-1} = 30$ days. The wing formation configuration turns out to be the least cost configuration for this scenario.

We outline the major findings and contributions of the thesis in Section 7.2. Some areas for future research is proposed in Section 7.3.

7.2 Contributions

We proposed the most versatile methodology to date for finding a good configuration of inpatient services in a hospital. Given a total number of beds, a set of specialties and a specific finite ir infinite waiting time threshold, this methodology uses two search algorithms combined with novel performance evaluation approximations to find a configuration with a low total mean daily cost, considering all major configurations proposed in the literature. Our approach works for systems with both IWTT and FWTT assumptions, whereas the approach proposed in Izady and Muhamed (2019) is restricted to systems with a ZWTT assumption. Our methodology captures all the five major configurations proposed in the literature, whereas the one proposed in Best et al. (2015) does not capture two major configurations, namely the COF and earmarking configurations.

We demonstrated how our methodology can be revised so that the best partitioning and bed allocation of a given configuration other than COF, i.e., dedicated, flexible, wing formation, or earmarking, is obtained. This is useful for situations where the hospital manager prefers to optimize the service delivery while keeping the overall structure of inpatient services.

We proposed novel performance evaluation approximations for overflow delay and overflow abandonment systems with potentially different mean service times in dedicated and overflow servers. These approximations produce practically accurate results in a short time. In addition to inpatient bed planning, these can be applied to other hierarchical systems such as those observed in telecommunication and computer networks.

We made the following contributions to inpatient data analysis.

- We proposed identifying the specialty of each care episode based on the HRG code of that episode instead of relying on the specialty field provided in the data. This is because the specialty field in NHS data often represents the specialty of consultant in charge of the episode and thus could be misleading in case of patient outlying. We argued that defining specialties with our proposed way provides a more accurate representation of patients' clinical needs, and thus lead to a more realistic estimate of patient outlying in the hospital. It would also make the configurations obtained from our models more reliable.

- We proposed linking together the episodes of care within a hospital spell that have the same specialty to create what we referred to as specialty spells. Each hospital spell will therefore be made up a number of specialty spells, each of which includes a number of care episodes, each, in turn, encompassing a number of bed visits.

- We proposed counting a specialty spell as an outlying spell if it has at least one bed visit in the non-primary ward of the corresponding specialty. The percentage of outlying

spells can then be evaluated for each specialty as well as the hospital as a whole. This would give an indication of the frequency at which outlying occurs in the hospital.

- We proposed measuring the workload contribution of outlying patients in addition to their frequency. For this measurement, we proposed counting a bed visit as an outlying bed visit if it occurs in a non-primary ward of the corresponding specialty. The workload contribution of outlying patients in a ward and/or the entire hospital can then be evaluated by adding up the lengths of outlying bed visits and dividing it by the sum of all bed visits in that ward and/or the entire hospital. This measure provides a more accurate picture of patient outlying than the frequency measure defined in the previous point as it takes the duration of outlying spells into account.
- We proposed comparing the mean length of outlying and non-outlying specialty spells for each specialty to assess the impact of outlying on mean LOS of outlying patients.
- We proposed comparing the length of specialty spells in different primary wards of a specialty (assuming they have a different percentage of outlying spells) to evaluate the impact of patient outlying on non-outlying spells.
- We proposed measuring the number of ward changes within a specialty spell, and assessing its impact on mean LOS of each specialty.

The analysis outlined above are novel and we believe they are essential, not only for applying our methodology, but also for performance evaluation and performance improvement of inpatient services.

Using inpatient data from inpatient services at a large hospital, we illustrated how our reconfiguration methodology can reduce the bed pressure on hospital without expanding the bed base, and with a moderate increase in daily nursing costs. Using simulation experiments, we illustrated that the savings obtained from our methodology could be significant, and that COF configuration is likely to produce the best results in a variety of circumstances.

7.3 Future Research

Improving the accuracy of our performance evaluation approximations, as well as the accuracy and speed of the intra-cluster bed allocation heuristic, are clear directions for future research. Such improvements would improve the overall accuracy and speed and thus applicability of our methodology. In addition to these, we have suggested below further opportunities for future research.

7.3.1 Incorporating Size and Skill-Set of the Existing Nursing Teams

A major assumption of our reconfiguration methodology is that the hospital would be able to fill the shortages in size and skill-set of its nursing teams, as compared to the size and skill-set of the configurations proposed by our methodology, through a combination of cross-training of existing nurses and recruitment of new nurses. This is because the size and skill-set of the existing nursing teams are not taken into account in our methodology. A future direction for research is therefore extending our methodology so as to incorporate the constraints imposed by the size and skill-set of existing nursing teams, as well as the costs of cross-training them or recruiting new teams. We explain the foundations of this methodology below.

In order to reflect the current size and skill-set of nursing teams in the new methodology, a skill-set matrix can be developed with rows representing individual permanent nurses and columns showing different specialties. The (i, j)th element of this matrix will then be set to 2 if specialty j is the major expertise (by training) of nurse i, 1 if specialty j is not the major expertise of nurse i but she has learned the main skills of this specialty through experience and training on the job, and 0 if nurse i has no training or experience in dealing with specialty j patients. The cross-training cost can also be represented using a matrix, with specialties repeated in both rows and columns. The (i, j)th element of this matrix will be set to the cost of cross-training a nurse whose major expertise is specialty i to learn the skills required for specialty j. The cost of recruiting new nurses should also be reflected in the new methodology. This would involve the cost of advertisement and selection, as well as the cost of filling in the positions by temporary nurses during the recruitment period. The costs of cross-training existing nurses and recruitment of new nurses can then be discounted daily over a five to ten year period (depending on how long in average the nurses remain in the same hospital) and added up to the daily wages in the objective function. Alternatively, a constraint can be added to the model restricting the total spending on cross-training and recruitment.

It is important to understand the nature and scope of cross-training to be able to fill the skill-set and cross-training matrices defined above. The goal of cross-training is defined in the literature as enhancing "... knowledge of inter-personal activities by introducing team members to the roles and responsibilities of their teammates ..." (Marks et al. 2002, p. 4). By this definition, cross-training occurs in reality most times without the nurses even noticing it. However, cross-training nurses of one specialty to learn skills of another specialty is more complex. For example, the learning exercise of a nurse equipped with knowledge of specialty x to learn skills of specialty y is different from that of a nurse equipped with knowledge of specialty y to learn skills of specialty x. As an example, nursing teams of Oncology can be trained to take care of Palliative Care or Geriatric Medicine patients with an on-site training

programme from the senior members of the nursing team. A Geriatric Medicine nurse, on the other hand, needs a university degree to be able to take care of the Oncology patients. Based on this, the cross-training costs differ substantially depending on the major expertise of the nurse and the specialty she needs to be cross-trained for.

7.3.2 Considering Different Waiting Time Thresholds

Another assumption in our methodology is that all patients have the same threshold for waiting. This would not capture situations in which a mix of elective and emergency patients are admitted as the threshold for former type of patients is typically much longer than that of the latter. Extending our methodology to capture such situations would therefore be another direction for future research. This would require generalizing our performance evaluation models for the overflow abandonment queue to allow for different classes of customers having different waiting time thresholds. Patients of each specialty would then be split into two classes, the emergency class with a shorter threshold and the elective class with a longer threshold.

7.3.3 Considering Non-Exponential Inter-arrival Times and LOSs

Another assumption of our methodology is that for each specialty, the admission requests follow a Poisson process and LOSs follow Exponential distributions. Using RSCH data, we tested these two assumptions in Chapter 6. The results illustrated the need to extending our methodology to non-Poisson admission requests and non-Exponential LOSs. This requires extending our performance evaluation models. This extension has been achieved in Izady and Muhamed (2019) for overflow loss systems. Developing this extension for overflow delay and abandonment queues remains for future research.

Bibliography

Guide to reducing long hospital stays, 2018. URL https://improvement.nhs.uk/documents/2898/Guide_to_reducing_long_hospital_stays_FINAL_v2.pdf.

E. Akcali, M. J. Co^té, and C. Lin. A network flow approach to optimizing hospital bed capacity decisions. *Health Care Management Science*, 9(4):391–404, 2006.

M. Asaduzzaman and T. J. Chaussalet. Modelling and performance measure of a perinatal network centre in the united kingdom. In *2008 21st IEEE International Symposium on Computer-Based Medical Systems*, pages 506–511. IEEE, 2008.

M. Asaduzzaman, T. J. Chaussalet, and N. J. Robertson. A loss network model with overflow for capacity planning of a neonatal unit. *Annals of Operations Research*, 178(1):67–76, 2010.

A. Bassamboo, R. S. Randhawa, and J. A. V. Mieghem. Optimal Flexibility Configurations in Newsvendor Networks: Going Beyond Chaining and Pairing. *Management Science*, 56 (8):1285–1303, 2010.

R. Bekker and A. M. de Bruin. Time-dependent analysis for refused admissions in clinical wards. *Annals of Operations Research*, 178(1):45–65, 2010.

R. Bekker, G. Koole, and D. Roubos. Flexible bed allocations for hospital wards. *Health Care Management Science*, pages 1–14, 2016. ISSN 13869620. doi: 10.1007/s10729-016-9364-4.

S. Belciug and F. Gorunescu. Improving hospital bed occupancy and resource utilization through queuing modeling and evolutionary computation. *Journal of biomedical informatics*, 53:261–9, feb 2015. ISSN 1532-0480. doi: 10.1016/j.jbi.2014.11.010.

T. J. Best, B. Sandıkçı, D. D. Eisenstein, and D. O. Meltzer. Managing Hospital Inpatient Bed Capacity Through Partitioning Care into Focused Wings. *Manufacturing & Service Operations Management*, 17(2):157–176, may 2015. ISSN 1523-4614. doi: 10.1287/msom.2014.0516.

BMA. State of the health system beds in the nhs: Uk. *British Medical Association*, 2017. URL https://www.basw.co.uk/system/files/resources/basw_53535-7_0.pdf.

R. Boaden, N. Proudlove, and M. Wilson. An exploratory study of bed management. *Journal of management in medicine*, 1999.

D. Campbell. Thousands of patients die waiting for beds in hospitals – study, 2019.

P. Chevalier and J.-C. Van den Schrieck. Computing the performance of multiclass queueing systems based on equivalent loss systems. Unpublished Manuscript, 2009.

J. R. Clark and R. S. Huckman. Broadening Focus: Spillovers, Complementarities, and Specialization in the Hospital Industry. *Management Science*, 58(4):708–722, apr 2012. ISSN 0025-1909. doi: 10.1287/mnsc.1110.1448.

J. K. Cochran and A. Bharti. A multi-stage stochastic methodology for whole hospital bed planning under peak loading. *International Journal of Industrial and Systems Engineering*, feb 2006.

S. Conroy and T. Dowsing. What should we do about hospital readmissions? *age and aging*, 41(6):702–704, 2012.

R. B. Cooper. Queueing theory. *Handbooks in Operations Research and Management Science*, 2:469–518, 1990.

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

A. M. de Bruin, A. C. van Rossum, M. C. Visser, and G. M. Koole. Modeling the emergency cardiac in-patient flow: an application of queuing theory. *Health care management science*, 10(2):125–37, jun 2007. ISSN 1386-9620.

A. M. de Bruin, R. Bekker, L. van Zanten, and G. M. Koole. Dimensioning hospital wards using the Erlang loss model. *Annals of Operations Research*, 178(1):23–43, oct 2009. ISSN 0254-5330. doi: 10.1007/s10479-009-0647-8.

J. Dong, P. Shi, F. Zheng, and X. Jin. Off-service placement in inpatient ward network: Resource pooling versus service slowdown. *Available at SSRN*, 2019.

M. B. Dumas. Simulation modeling for hospital bed planning. *Simulation*, 43(2):69–78, aug 1984. ISSN 0037-5497. doi: 10.1177/003754978404300201.

M. B. Dumas. Hospital bed utilization: an implemented simulation approach to adjusting and maintaining appropriate levels. *Health services research*, 20(1):43, 1985.

G. J. Franx, G. Koole, and A. Pot. Approximating multi-skill blocking systems by Hyper-Exponential Decomposition. *Performance Evaluation*, 63(8):799–824, aug 2006. ISSN 01665316. doi: 10.1016/j.peva.2005.09.001.

A. A. Fredericks. Congestion in Blocking Systems-A Simple Approximation Technique. *Bell System Technical Journal*, 59(6):805–827, jul 1980. ISSN 00058580. doi: 10.1002/j.1538-7305.1980.tb03034.x.

J. Goldman, H. A. Knappenberger, and J. Eller. Evaluating bed allocation policy with computer simulation. *Health services research*, 3(2):119, 1968.

N. D. Goldstein, B. C. Ingraham, S. C. Eppes, M. Drees, and D. A. Paul. Assessing occupancy and its relation to healthcare-associated infections. *infection control & hospital epidemiology*, 38(1):112–114, 2017.

F. Gorunescu, S. I. McClean, and P. H. Millard. Using a Queueing Model to Help Plan Bed Allocation in a Department of Geriatric Medicine. *Health Care Management Science*, 5 (4):307–312, 2002a. ISSN 1572-9389. doi: 10.1023/A:1020342509099.

F. Gorunescu, S. I. McClean, and P. H. Millard. A queueing model for bed-occupancy management and planning of hospitals. *Journal of the Operational Research Society*, 53(1): 19–24, 2002b.

L. V. Green. How many hospital beds? *Inquiry*, 39(4):400–412, jan 2002. ISSN 00469580. doi: 10.5034/inquiryjrnl_39.4.400.

L. V. Green. Capacity Planning and Management in Hospitals. *Operations Research and Health Care*, pages 15–41, 2006. doi: 10.1007/1-4020-8066-2_2.

L. V. Green and V. Nguyen. Strategies for cutting hospital beds: the impact on patient service. *Health services research*, 36(2):421–442, jun 2001. ISSN 0017-9124.

D. Gross and D. R. Miller. The randomization technique as a modeling tool and solution procedure for transient markov processes. *Operations Research*, 32(2):343–361, 1984.

D. Gross, J. F. Shortle, J. M. Thompson, and C. M. Harris. *Fundamentals of Queueing Theory*, volume 627. John Wiley & Sons, 2011.

R. Hall, editor. *Handbook of Healthcare System Scheduling*, volume 168 of *International Series in Operations Research & Management Science*. Springer US, Boston, MA, 2012. ISBN 978-1-4614-1733-0. doi: 10.1007/978-1-4614-1734-7.

P. R. Harper and A. Shahani. Modelling for the planning and management of bed capacities in hospitals. *Journal of the Operational research Society*, 53(1):11–18, 2002.

G. Hesselink, M. Zegers, M. Vernooij-Dassen, P. Barach, C. Kalkman, M. Flink, G. Ön, M. Olsson, S. Bergenbrant, C. Orrego, et al. Improving patient discharge and reducing hospital readmissions by using intervention mapping. *BMC health services research*, 14(1):389, 2014.

L. B. Holm, H. Lurås, and F. A. Dahl. Improving hospital bed utilisation through simulation and optimisation: with application to a 40% increase in patient volume in a norwegian general hospital. *International journal of medical informatics*, 82(2):80–89, 2013.

X. Huang. Decision making support in reshaping hospital medical services. *Health Care Management Science*, 1(2):165–173, 1998. ISSN 1572-9389. doi: 10.1023/A:1019050720311.

F. Iravani and B. Balcioglu. On priority queues with impatient customers. *Queueing Systems*, 58(4):239–260, 2008.

N. Izady. *On queues with time-varying demand*. PhD thesis, Lancaster University, 2010.

N. Izady and I. Muhamed. A Clustered Overflow Configuration of Inpatient Beds in Hospitals. *Manufacturing & Service Operations Management*, 2019. ISSN 1523-4614. doi: 10.1287/msom.2019.0820.

N. Izady and D. Worthington. Approximate analysis of non-stationary loss queues and networks of loss queues with general service time distributions. *European Journal of Operational Research*, 213(3):498–508, 2011.

D. L. Jagerman. Some Properties of the Erlang Loss Function. *Bell System Technical Journal*, 53(3):525–551, mar 1974. ISSN 00058580. doi: 10.1002/j.1538-7305.1974.tb02756.x.

S. Jones, C. Moulton, S. Swift, P. Molyneux, S. Black, N. Mason, R. Oakley, and C. Mann. Association between delays to patient admission from the emergency department and all-cause 30-day mortality. *Emergency Medicine Journal*, 2022. doi: 10.1136/emermed-2021-211572.

W. C. Jordan and S. C. Graves. Principles on the Benefits of Manufacturing Process Flexibility. *Management Science*, 41(4):577–594, apr 1995. ISSN 0025-1909. doi: 10.1287/mnsc.41.4.577.

W. C. Jordan, R. R. Inman, and D. E. Blumenfeld. Chained cross-training of workers for robust performance. *IIE Transactions (Institute of Industrial Engineers)*, 36(10):953–967, 2004. ISSN 0740817X. doi: 10.1080/07408170490487713.

K. Kaier, N. Mutters, and U. Frank. Bed occupancy rates and hospital-acquired infections—should beds be kept empty? *Clinical microbiology and infection*, 18(10):941–945, 2012.

E. P. C. Kao and G. G. Tung. Bed Allocation in a Public Health Care Delivery System. *Management Science*, 27(5):507–520, may 1981. ISSN 0025-1909. doi: 10.1287/mnsc.27.5.507.

D. S. KC and C. Terwiesch. The Effects of Focus on Performance: Evidence from California Hospitals. *SSRN Electronic Journal*, apr 2011. ISSN 1556-5068. doi: 10.2139/ssrn.1815529.

King's Fund. Nhs hospital bed numbers: past, present, future, 2020a. URL `https://www.kingsfund.org.uk/publications/nhs-hospital-bed-numbers`.

King's Fund. What's going on with a&e waiting times?, 2020b. URL https://www.kingsfund.org.uk/projects/urgent-emergency-care/urgent-and-emergency-care-mythbusters.

A. Kokangul. A combination of deterministic and stochastic approaches to optimize bed capacity in a hospital unit. *Computer methods and programs in biomedicine*, 90(1):56–65, apr 2008. ISSN 0169-2607. doi: 10.1016/j.cmpb.2008.01.001.

X. Li, P. Beullens, D. Jones, and M. Tamiz. An integrated queuing and multi-objective bed allocation model with application to a hospital in china. *Journal of the Operational Research Society*, 60(3):330–338, 2009.

J. M. Lim, H. Song, and J. Yang. The spillover effects of capacity pooling in hospitals. *Available at SSRN 3800351*, 2021.

D. Maguire. Premature discharge: is going home early really a christmas gift? *The King's Fund*, 2015. URL https://www.kingsfund.org.uk/blog/2015/12/premature-discharge-from-hospital.

A. Mandelbaum and S. Zeltyn. The palm/erlang-a queue, with applications to call centers. *Faculty of Industrial Engineering & Management, Technion, Haifa, Israel*, 7:8, 2005.

M. A. Marks, M. J. Sabella, C. S. Burke, and S. J. Zaccaro. The impact of cross-training on team effectiveness. *Journal of Applied Psychology*, 87(1):3, 2002.

W. A. Massey and W. Whitt. An analysis of the modified offered-load approximation for the nonstationary erlang loss model. *The Annals of applied probability*, pages 1145–1160, 1994.

B. A. Mateen, H. Wilde, J. M. Dennis, A. Duncan, N. Thomas, A. McGovern, S. Denaxas, M. Keeling, and S. Vollmer. Hospital bed capacity and usage across secondary healthcare providers in england during the first wave of the covid-19 pandemic: a descriptive analysis. *BMJ Open*, 11(1), 2021.

I. Mohamed, I. El-Henawy, and R. Z. El-Din. Estimating bed requirements for a pediatric department in a university hospital in egypt. *Modern Management Science & Engineering*, 4:101, 2016.

S. Moiseev. Universal derivative-free optimization method with quadratic convergence. *arXiv preprint arXiv:1102.1347*, 2011.

NHS Data Model and Dictionary. Healthcare resource group. *NHS Data Model and Dictionary*, 2021. URL https://datadictionary.nhs.uk/supporting_information/healthcare_resource_group.html.

NHS Digital. Hospital admissions hit record high as population ages, 2018. URL https://digital.nhs.uk/news-and-events/news-archive/2016-news-archive/hospital-admissions-hit-record-high-as-population-ages.

J. S. Olshaker and N. K. Rathlev. Emergency Department overcrowding and ambulance diversion: the impact and potential solutions of extended boarding of admitted patients in the Emergency Department. *The Journal of emergency medicine*, 30(3):351–6, apr 2006. ISSN 0736-4679. doi: 10.1016/j.jemermed.2005.05.023.

M. Petkovsek, H. S. Wilf, and D. Zeilberger. *A=B*. 1997.

M. Powell. Quadratic termination properties of minimization algorithms i. statement and discussion of results. *IMA Journal of Applied Mathematics*, 10(3):333–342, 1972.

C. Propper, G. Stoye, and B. Zaranko. The wider impacts of the coronavirus pandemic on the nhs. *Fiscal Studies*, 41(2):345–356, 2020.

R. Roberston, L. Wenzel, J. Thopson, and A. Charles. Understanding nhs finanical pressure. how are they affacting patient care?, 2017.

F. Serafini, G. Fantin, R. Brugiolo, O. Lamanna, A. Aprile, and F. Presotto. Outlier admissions of medical patients: prognostic implications of outlying patients. the experience of the hospital of mestre. *Italian Journal of Medicine*, pages 299–302, 2015.

M. Sheikhzadeh, S. Benjaafar, and D. Gupta. Machine Sharing in Manufacturing Systems: Total Flexibility versus Chaining. *International Journal of Flexible Manufacturing Systems*, 10(4):351–378, 1998. ISSN 1572-9370. doi: 10.1023/A:1008057504351.

P. Shi, M. C. Chou, J. G. Dai, D. Ding, and J. Sim. Models and Insights for Hospital Inpatient Operations: Time-Dependent ED Boarding Time. *Management Science*, page 150422112841002, apr 2015. ISSN 0025-1909. doi: 10.1287/mnsc.2014.2112.

P. C. Sprivulis, J.-A. Da Silva, I. G. Jacobs, G. A. Jelinek, and A. R. Frazer. The association between hospital overcrowding and mortality among patients admitted via western australian emergency departments. *Medical Journal of Australia*, 184(5):208–212, 2006.

A. Steventon, S. Deeny, R. Friebel, T. Gardner, and R. Thorlby. Briefing: emergency hospital admissions in england: which may be avoidable and how? *The Health Foundation*, 2018.

A. Stowell, P.-G. Claret, M. Sebbane, X. Bobbia, C. Boyard, R. Genre Grandpierre, A. Moreau, and J.-E. de La Coussaye. Hospital out-lying through lack of beds and its impact on care and patient outcome. *Scandinavian journal of trauma, resuscitation and emergency medicine*, 21:17, jan 2013. ISSN 1757-7241. doi: 10.1186/1757-7241-21-17.

N. M. van Dijk and E. van der Sluis. To Pool or Not to Pool in Call Centers. *Production and Operations Management*, 17(3):296–305, may 2008. ISSN 1059-1478. doi: 10.3401/poms.1080.0029.

J. T. van Essen, M. van Houdenhoven, and J. L. Hurink. Clustering clinical departments for wards to achieve a prespecified blocking probability. *OR spectrum*, 37(1):243–271, 2015.

G. Vassilacopoulos. A simulation model for bed allocation to hospital inpatient departments. *SIMULATION*, 45(5):233–241, nov 1985. ISSN 0037-5497. doi: 10.1177/003754978504500502.

F. d. Véricourt and O. B. Jennings. Nurse staffing in medical units: A queueing perspective. *Operations Research*, 59(6):1320–1331, 2011.

X. Wu, J. Li, and C.-H. Chu. Modeling multi-stage healthcare systems with service interactions under blocking for bed allocation. *European Journal of Operational Research*, 278(3):927–941, 2019.

Z. Zhang. Call centres with balking and abandonment: from queueing to queueing network models, 2010.