Time-Dependent Stochastic Modelling for Predicting Demand and Scheduling of Emergency Medical Services

Julie Vile
School of Mathematics
Cardiff University

A thesis submitted for the degree of
Doctor of Philosophy

2013

Summary

As the prominence of the service sector is increasing in developed nations, new and exciting opportunities are arising for operational researchers to develop and apply models which offer managers solutions to improve the quality of their services. The development of time-dependent stochastic models to analyse complex service systems and generate effective personnel schedules are key to this process, enabling organisations to strike a balance between the provision of a good quality service whilst avoiding unnecessary personnel costs. Specifically within the healthcare sector, there is a need to promote efficient management of an Emergency Medical Service (EMS), where the probability of survival is directly related to the speed of assistance.

Motivated by case studies investigating the operation of the Welsh Ambulance Service Trust (WAST), this thesis aims to investigate how operational research (OR) techniques can be developed to analyse priority service systems subject to demand that is of an urgent nature, cannot be backlogged, is heavily time-dependent and highly variable. A workforce capacity planning tool is ultimately developed that integrates a combination of forecasting, queueing theory, stochastic modelling and optimisation techniques into a single spreadsheet model in order to predict future demand upon WAST, set staffing levels, and optimise shift schedules and rosters. The unique linking together of the techniques in a planning tool which further captures time-dependency and two priority classes enables this research to outperform previous approaches, which have generally only considered a single class of customer, or generated staffing recommendations using approximation methods that are only reliable under limited conditions.

The research presented in this thesis is novel in several ways. Primarily, the first section considers the potential of a nonparametric modelling technique known as Singular Spectrum Analysis (SSA) to improve the accuracy of demand forecasts. Secondly, the main body of work is dedicated to adapting numerical queueing theory techniques to accurately model the behaviour of time-dependent multi-server priority systems across shift boundaries and evaluate the likelihood of excessive waits for service for two customer classes. The final section addresses how shifts can be optimally scheduled using heuristic search techniques. The main conclusions are that in addition to offering a more flexible approach, the forecasts generated by SSA compare favourably to those obtained using traditional methods, and both approximate and numerical modelling techniques may be duly extended to set staffing levels in complex priority systems.

Declaration

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed Date

STATEMENT 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD.

Signed Date

STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references. The views expressed are my own.

Signed Date

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed Date

Acknowledgments

First and foremost, I would like to thank my team of supervisors Prof. Paul Harper, Dr Jonathan Gillard and Dr Vincent Knight for their time, excellent guidance and for the outstanding support they have offered throughout the course of my PhD. I am grateful to them for introducing me to such a motivating area of research, and for their continued sharing of their expert knowledge and advice, to assist my development in becoming an independent researcher.

I wish to also thank the entire OR research group at Cardiff University for making my time as a PhD student so enjoyable, especially Dr Janet Williams, without whom I would never embarked on this academic route. A special thanks goes to Penny, Izabela, Lisa, Daniel and Chris who have been through the PhD experience with me every step of the way; and to Leanne, Angelico and Lizzie for sharing numerous courses together.

My thanks are also extended to the team at WAST who have enabled this research to go ahead, all of whom have been friendly and approachable throughout; and I am grateful to the EPSRC, who provided the financial support for this thesis via the LANCS Initiative.

My sincere thanks go to my parents who did everything they could to bring me up in a loving, fun and stimulating environment. Their unwavering support and encouragement has been outstanding, and their motivation to look for the solution rather than the problem in every aspect of life has been key in enabling me to reach this milestone.

I would like to express my deepest gratitude to my husband, James, without whose support, love and encouragement this work would have seemed much harder to complete! I thank you for being so understanding during the many months of juggling wedding planning, daily life and continuing my research; for cooking me daily delicious meals to keep me going; and for sharing so many amazing experiences with me over the past few years. Thanks must also be given to my Uni 'Aberdare' girls for keeping me sane with our monthly girly weekends away!

Finally, I would like to thank God for His constant love, undescribable grace and for opening all the doors along the path to bring me to the place I have reached today. I pray that my heart will continually be open to your guidance in my future career.

Contents

Summary i

Declaration ii

Acknowledgements iii

Contents iv

List of publications ix

List of tables x

List of figures xi

List of abbreviations xiii

List of notation xv

1 Introduction 1

1.1 Introductory remarks 1
1.2 Background 1
1.3 The research problem 3
1.4 WAST 8
1.5 Thesis structure 14
1.6 Summary 18

2 Data description and preliminary analyses 19

2.1 Introductory remarks 19
2.2 The data source 20
2.3 The response process 23
2.4 Exploratory demand analysis 29

2.5 Summary . . . 32

3 Literature review (part 1): Forecasting demand 35
3.1 Introductory remarks . . . 35
3.2 Current Practice . . . 36
3.3 Early exploratory analysis . . . 38
3.4 Time series models . . . 40
3.5 SSA . . . 43
3.6 Summary . . . 45

4 Demand forecasting 46
4.1 Introductory remarks . . . 46
4.2 SSA Theory . . . 47
4.2.1 Decomposition and reconstruction . . . 47
4.2.2 Forecasting . . . 49
4.3 Model comparison . . . 50
4.3.1 Measures of accuracy: RMSE . . . 50
4.3.2 SSA model formulation . . . 51
4.3.3 Conventional models . . . 53
4.3.4 Results . . . 54
4.4 Summary . . . 59

5 Literature review (part 2): Queueing theory 62
5.1 Introductory remarks . . . 62
5.2 Preliminaries . . . 62
5.3 Time-dependent queueing theory . . . 69
5.3.1 Approximation methods . . . 72
5.3.2 Numerical methods . . . 76
5.4 Priority queueing theory . . . 82
5.5 Summary . . . 90

6 Computing service levels in $M(t)/M/s(t)/FIFO$ systems 92
6.1 Introductory remarks . . . 92
6.2 Approximation and numerical methods . . . 93
6.2.1 Approximation methodology . . . 93
6.2.2 Numerical methodology . . . 96
6.3 Application to WAST data (SE region) . . . 109

6.3.1 Numerical requirements . . . 112
6.3.2 Approximate requirements . . . 116
6.4 Summary . . . 126

7 Computing service levels in $M(t)/M/s(t)/NPRP$ systems 130
7.1 Introductory remarks . . . 130
7.2 Approximation and numerical methods . . . 132
7.2.1 Approximation methodology . . . 132
7.2.2 Numerical methodology . . . 138
7.2.3 Hybrid methodology . . . 152
7.3 Application to WAST data (Cardiff area) . . . 153
7.3.1 Numerical requirements . . . 157
7.3.2 Approximate requirements . . . 158
7.3.3 Hybrid Approach . . . 165
7.4 Summary . . . 169

8 Literature review (part 3): Scheduling and rostering 172
8.1 Introductory remarks . . . 172
8.2 Scheduling review . . . 175
8.3 Rostering review . . . 181
8.4 Summary . . . 184

9 Scheduling and rostering 186
9.1 Introductory remarks . . . 186
9.2 WAST shift scheduling model . . . 187
9.2.1 The IP model . . . 188
9.2.2 Solving the IP heuristically . . . 191
9.2.3 Solving the IP optimally . . . 193
9.2.4 Evaluation of heuristic approaches . . . 199
9.3 WAST crew allocation model . . . 201
9.3.1 The IP model . . . 202
9.3.2 Solving the IP heuristically . . . 206
9.3.3 Solving the IP optimally . . . 210
9.3.4 Evaluation of heuristic approaches . . . 212
9.4 Summary . . . 215

10 Workforce capacity planning tool **217**
10.1 Introductory Remarks . . . 217
10.2 Functions offered by the planning tool . . . 219
10.2.1 List of adjustable parameters/variables . . . 223
10.2.2 VBA Code . . . 228
10.3 Summary . . . 230

11 Conclusions and future research **232**
11.1 Introductory remarks . . . 232
11.2 Key findings and conclusions . . . 232
11.3 Novel contributions . . . 241
11.4 Research limitations and directions for future research . . . 243
11.4.1 Errors associated with demand modelling techniques . . . 243
11.4.2 Time-dependent and priority queueing theory . . . 244
11.4.3 Improving the quality of the roster . . . 245
11.4.4 WAST specifics . . . 246

References **249**

Appendices **262**

A Supporting documents **263**
A.1 Annual response performance 2010/11 . . . 263
A.2 Changes to National Ambulance Performance Standards . . . 264
A.3 Proof: Calculation of δ_c . . . 264
A.4 Cardiff EA shifts . . . 265

B Publication: Predicting ambulance demand using SSA **266**
B.1 Introductory remarks . . . 266
1 Introduction . . . 267
2 Previous research . . . 268
3 Data analysis . . . 270
4 SSA theory . . . 272
4.1 Decomposition: Embedding . . . 273
4.2 Decomposition: Singular value decomposition (SVD) . . . 273
4.3 Reconstruction: Grouping . . . 273
4.4 Reconstruction: Diagonal averaging . . . 274
4.5 Forecasting . . . 274

4.6 Measures of accuracy: root mean squared error 275
5 Model comparison 275
5.1 SSA model formulation 276
5.2 ARIMA model 278
5.3 Holt-Winters (HW) forecasting method 278
5.4 Results 279
6 Conclusions 284

C Publication: Staffing a mathematics support service 286
C.1 Introductory remarks 286
1 Introduction 287
2 MSS as a finite source queue 288
2.1 The mathematical model 288
2.2 Application to Cardiff University's MSS 289
2.3 Results 291
3 Optimisation of Staffing and Skill mix 292
3.1 Mathematical program 292
3.2 Finding a roster 294
3.3 Results 295
4 Conclusions 296

List of publications

i. Vile, J., Gillard, J., Harper, P., and Knight, V. (2012). Predicting ambulance demand using singular spectrum analysis, *Journal of the Operational Research Society*, **63**(11): 1556-1565.

ii. Williams, J.[1], Gillard, J., Harper, P., and Knight, V. (2010). Forecasting Welsh ambulance demand using singular spectrum analysis, in A. Testi, E. Tanfani, E. Ivaldi, G. Carello, R. Aringhieri, and V. Fragnelli (ed.), *Operations Research for Patient-Centered Health Care Delivery. Proceedings of the XXXVI International ORAHS Conference, July 2010*, pp. 196-208.

Submitted

i. Gillard, J., Knight, V., Vile, J., and Wilson, R. (2012). Staffing a mathematics support service.

[1]Maiden name

List of Tables

1.1 NHS re-organisational structure 11

2.1 Variables recorded in database 21
2.2 Number of unique incidents by nature of emergency 22
2.3 Summary statistics for response and service times 28

4.1 Comparison of model forecasts for daily demand (Jul - Dec 2009) . . . 56
4.2 Comparison of model forecasts for daily demand (Jul & Dec 2009) . . . 57

5.1 Terminology used to describe queue disciplines 70

6.1 SIPP, Lag Avg and SIPP Mix accuracy 120
6.2 SIPP Revised Reliability (RMSEs) 125

7.1 Pri SIPP, Pri Lag Avg and Pri SIPP Mix accuracy 159
7.2 Priority SIPP Revised Reliability (RMSEs) 164

8.1 Glossary of terms used in scheduling and rostering 173

9.1 Optimal shift assignments, July 2009 197
9.2 An example schedule for Cardiff EA staff, first week of July 2009 211

10.1 Run times required to execute programs for various forecasting horizons 222
10.2 Default values for response targets and average service rate 225
10.3 Default values for numerical methodology 226
10.4 Default values for shift preference weights 227
10.5 Default parameter values for shift scheduling heuristic 227
10.6 Default parameter values for crew constraints 228
10.7 Default parameter values for rostering heuristic 228

A.1 WAST performance statistics, by month 263
A.2 Pool of potential shifts to be assigned to Cardiff EA crews 265

List of Figures

1.1 Map of LHBs across Wales 11
1.2 Thesis Structure 15

2.1 Pie charts of demand, by incident type 23
2.2 The components of the Category A response process 24
2.3 Distribution of first response times to Category A and B/C incidents . 25
2.4 Time flow of ambulance response to incidents 27
2.5 Cumulative proportion of first responses to Category A incidents 28
2.6 WAST daily demand 30
2.7 WAST average monthly demand 30
2.8 Box plots of demand volumes for each month and each day of week . . 31
2.9 Mean number of incidents reported per hour, by weekday 32

3.1 Unique EMS incidents reported, Hour 13, Monday 36

4.1 Principal components related to the first 2 eigentriples 52
4.2 Scatter plot and periodogram of paired eigenvectors (2-3) 53
4.3 28-day forecasts beginning on 1st July 2009 58
4.4 Confidence intervals for 28-day SSA forecast by the bootstrap method . 58
4.5 Original time series with 7-step-ahead daily SSA forecasts for July 2009 59

5.1 The fundamental diagram of queueing theory 63
5.2 State transition diagram for a birth-death process 67
5.3 Urgent demand for ambulance transportation 69
5.4 Graphical illustration of Euler methodology 78
5.5 Urgent and routine demand for ambulance transportation 83
5.6 Markov chain representation of $M/M/2$ queue with 2 priority classes . 85

6.1 Representation of a shift boundaries 100
6.2 Waiting time formulae for various shift boundaries and server numbers 109

6.3 Minimum staffing levels suggested by numerical methodology 114
6.4 Minimum staffing levels recommended per hour (06/07/2009) 117
6.5 Proportion of patients reached within acceptable waiting time 118
6.6 Weighted RMSE for various levels of $\tau \in [0,1]_{\mathbb{R}}$ 122
6.7 Weighted RMSE for various levels of $\tau \in [0,1]_{\mathbb{R}}$ at 00:00 and 08:00 . . . 123

7.1 A schematic diagram of the priority queueing system 131
7.2 State spaces that could give rise to $S = (1,0,0)$ 143
7.3 Approximate and numerical staffing requirements, 1st July 2009 153
7.4 Weighted RMSE for various levels of $\tau \in [0,1]_{\mathbb{R}}$ 161
7.5 Initial staffing levels suggested by approximate approaches 168

9.1 EA coverage using original Cardiff shifts, July 2009 195
9.2 EA oversupply arising from scheduling original Cardiff shifts 196
9.3 EA oversupply arising from scheduling revised Cardiff shifts 196
9.4 Total cost for various shift pools, July 2009 198
9.5 Rate at which heuristics converge to optimality 200
9.6 Rate at which heuristics converge to optimality 213

10.1 A screenshot of the main menu options 219

List of abbreviations

A & E	**A**ccident & **E**mergency
AMPDS	**A**dvanced **M**edical **P**riority **D**ispatch **S**ystem
ANN	**A**rtificial **N**eural **N**etwork
AOF	**A**nnual **O**perating **F**ramework
AR	**A**uto**R**egressive
ARIMA	**A**uto**R**egressive **I**ntegrated **M**oving **A**verage
CLSM	**C**ontrolled **L**abour **S**cheduling **M**odel
CP	**C**onstraint **P**rogramming
DLSM	**D**antzig's **L**abour **S**cheduling **M**odel
DLL	**D**ynamic **L**ink **L**ibrary
DTM	**D**iscrete **T**ime **M**odelling
EA	**E**mergency **A**mbulance (Fully Equipped)
EMS	**E**mergency **M**edical **S**ervices
FIFO	**F**irst-**I**n **F**irst-**O**ut
FILO	**F**irst-**I**n **L**ast-**O**ut
HP	**H**igh **P**riority
HW	**Holt**-**W**inters
iid	**i**ndependent and **i**dentically **d**istributed
IP	**I**nteger **P**rogram
ISA	**I**terative **S**taffing **A**lgorithm
KLSM	**K**eith's **L**abour **S**cheduling **M**odel
LHB	**L**ocal **H**ealth **B**oard
LP	**L**ow **P**riority
LRF	**L**inear **R**ecurrent **F**ormulae

LST	**L**aplace-**S**tieltjes **T**ransform
MA	**M**oving **A**verage
MDCTMC	**M**ixed **D**iscrete-**C**ontinuous **T**ime **M**arkov **C**hain
MFNN	**M**ultilayer **F**eedforward **N**eural **N**etwork
MOL	**M**odified **O**ffered **L**oad
MSSA	**M**ultivariate **S**ingular **S**pectrum **A**nalysis
NHS	**N**ational **H**ealth **S**ervice
OR	**O**perational **R**esearch
PCS	**P**atient **C**are **S**ervices
PTS	**P**atient **T**rasnport **S**ervice
PSA	**P**ointwise **S**tationary **A**pproximation
PRP	**PR**emptive **P**riority
RA	**R**elative **A**mplitude
RDR	**R**ecursive **D**imensionality **R**eduction
RIRO	**R**andom-**I**n **R**andom-**O**ut
RMSE	**R**oot **M**ean **S**quare **E**rror
RRV	**R**apid **R**esponse **V**ehicle
SA	**S**imulated **A**nnealing
SD	**S**tandard **D**eviation
SE	**S**outh **E**ast
SIPP	**S**tationary **I**ndependent **P**eriod by **P**eriod
SSA	**S**ingular **S**pectrum **A**nalysis
SVD	**S**ingular **V**alue **D**ecomposition
UA	**U**nitary **A**uthority
VBA	**V**isual **B**asic for **A**pplications
WAST	**W**elsh **A**mbulance **S**ervice **T**rust
WTD	**W**orking **T**ime **D**irectives

List of notation

Forecasting

X_i	ith L-lagged vector of time series
$X = [X_1, \ldots, X_K]$	Trajectory matrix with columns X_i
X^T	Transposed matrix of X
$d = \text{rank}(X)$	Rank of matrix X
Z	Hankel matrix of matrix X
λ_i	i-th eigenvalue of the matrix XX^T
U_i	Orthogonal eigenvector of the matrix XX^t corresponding to λ_i
$V_i = \frac{X_i^T U_i}{\sqrt{\lambda_i}}$	i-th factor vector of the SVD of the matrix X
L	Window length
N	Length of observed time series
$K = N - L + 1$	Number of L Lagged vectors of Y_N
b	Number of disjoint subsets used to reconstruct the matrix X
q	Order of LRF used to forecast future values of time series
Y_N	Time series of length N
y_n	Observed n-th value of time series
e_n	Estimated n-th value of time series

Queueing theory

λ	Mean arrival rate (stationary system)
λ_H	Mean arrival rate of HP customers
λ_L	Mean arrival rate of LP customers
$\lambda(t)$	Mean arrival rate at time t
μ	Mean service rate
$\rho = \frac{\lambda}{s\mu}$	Server utilisation rate
$\rho_H = \frac{\lambda_H}{s\mu}$	Server utilisation rate for HP customers
$\rho_L = \frac{\lambda_L}{s\mu}$	Server utilisation rate for LP customers
$r = \frac{\lambda}{\mu}$	Offered load
$r_H = \frac{\lambda_H}{\mu}$	Offered load for HP customers
$r_L = \frac{\lambda_L}{\mu}$	Offered load for LP customers
k	Number of priority classes in a priority queueing system
n	Total number of customers in the system
$\{N(t), t \geq 0\}$	Continuous time stochastic process
t_z	Epochs at which number of servers and arrival rates may change
p_n	Steady-state probabilities of n customers in the system
(p_n)	Probability state vector tracking number of customers n present in the system
$p_n(t)$	Probability n customers are present in the system at time t
$p_n(t)^-$	Probability n customers are present in the system immediately before the shift boundary
s	Number of servers on duty (constant level)
$s(t)$	Number of servers on duty at time t
$s(t)^-$	Number of servers on duty immediately before the shift boundary
$s(t)^+$	Number of servers on duty immediately after the shift boundary
$s_c = s(t)^+ - s(t)^-$	Change in the number of servers over the shift boundary
x	Maximal acceptable waiting time
x_H	Maximal acceptable waiting time for HP customers
x_L	Maximal acceptable waiting time for LP customers

L_q	Expected number of customers in the queue
W_q	Time a customer waits in the queue before commencing service
W_q^n	Time a customer that arrives to find n people ahead in the system waits in the queue before commencing service
$P(W_q > x)$	Probability a customer waits greater than time x in the queue
a	Mean number of departing customers over a given time interval
δ_{pp}	Length of planning period
δ_c	Length of calculation period
G	Limit on number of customers considered in system
δs	Number of departing servers over shift boundary
δn	Number of customers ejected from system over shift boundary
δi	Number of HP customers ejected from system over shift boundary
$B(t)$	Probability matrix applied to map the probability state vector over shift boundaries
i	Number of HP customers in service
j	Number of LP customers in service
h	Number of HP customers in the queue
l	Number of LP customers in the queue
$\tilde{n}$	Cumulative number of LP customers in service and HP customers in the system
$S = (i, j)$	State space vector denoting i HP and j LP customers in service and at least one idle server
$S = (i, h, l)$	State space vector denoting i HP and $(s - j)$ LP customers in service, with h HP and l LP customers in the queue

Scheduling and rostering

D	Set of days of the week
P	Set of periods in a day
I	Set of planning intervals
J	Set of ambulance crew
S	Set of allowable shifts
T	Set of allowable tours
l_s	Length of shift s
p_s	Preferability weight assigned to shift s
overtime_j	Number of overtime hours assigned to crew j
crew_j	Binary variable denoting if crew j is assigned to at least one shift
w_{jsd}	Binary variable denoting if crew j works shift s on day d
a_{sp}	Binary variable denoting if shift s includes period p
r_{pd}	Desired crew requirement in period p of day d
x_{sd}	Number of crews working shift s on day d
c_s	Cost of assigning a crew to work shift s

KLSM Variables:

b_i^{β}	Limit on the bounded shortage of employees in interval i
b_i^{Π}	Limit on the bounded surplus of employees in interval i
α_i	Unbounded shortage of employees in interval i
β_i	Bounded shortage of employees in interval i
Θ_i	Unbounded surplus of employees in interval i
Π_i	Bounded surplus of employees in interval i
c^{α}	Cost of the unbounded shortage of employees, per employee-interval
c^{β}	Cost of the bounded shortage of employees, per employee-interval
c^{Θ}	Cost of the unbounded surplus of employees, per employee-interval
c^{Π}	Cost of the bounded surplus of employees, per employee-interval

Chapter 1

Introduction

1.1 Introductory remarks

This thesis primarily aims to investigate how operational research (OR) techniques can be developed to analyse priority service systems subject to demand that is of an urgent nature, cannot be backlogged, is heavily time-dependent and highly variable. Motivated by case studies investigating the operation of the Welsh Ambulance Service Trust (WAST), it demonstrates how efficient management of such service systems can be promoted through developing techniques to generate optimised staffing profiles that allow a minimum service quality to be consistently provided. The research compliments ongoing work at Cardiff University in collaboration with WAST focussing on interrelated issues such as maximal survival modelling; and the use of simulation models to assist Emergency Medical Service (EMS) planning for the location, capacity and deployment of emergency vehicles.

This chapter sets the general research context and provides the background to the role of OR in the service sector. Section 1.2 deals with the use of OR to set staffing requirements, and the particular research questions are addressed in Section 1.3. A brief description of WAST for whom the research was primarily conducted is contained in Section 1.4, followed by an overview of the thesis structure in Section 1.5.

1.2 Background

Over the last few decades, the importance of the service sector has increased in many developed countries. In the UK, the service sector now accounts for more than three

quarters of total gross domestic product; covering a range of activities concerning the provision of services to other businesses as well as consumers in a wide variety of industries such as government, healthcare, education, banking, insurance, tourism and retail sales (Office for National Statistics, 2011). As the service industry has evolved and greater competition has emerged between organisations, managers have become ever keen to gain more information about their own services with an aim of enabling them to be in a position to deliver what their customers want as quickly and inexpensively as possible.

In response to this, OR techniques have developed substantially. While the mathematical frameworks surrounding the techniques of stochastic modelling, heuristics and optimisation are all well-established, there has recently been a major effort to develop the theory with applications of the models (see, for example, Erdogan et al. (2010), Ingolfsson et al. (2010) and Izady and Worthington (2012)); thus providing businesses with improved operations to increase their opportunities. In particular the branch of queueing theory, which helps managers to make judicious decisions regarding the resources they need to provide an efficient service through revealing analytical details of system quality under a range of scenarios, has moved away from basic analyses of single server queues with consistent random arrival and service rates (as first proposed by Erlang (1918)), to analysing systems with time-dependent arrival rates, multiple servers and different service priorities for distinct customer classes. Whilst queueing models may be used to inform decisions relating to the resources needed to provide a service, their effectiveness is dependent on accurate forecasts of demand upon the system. Thus thorough system evaluation involves an amalgamation of a large range of statistical and OR techniques, ranging from forecasting future demand levels, translating the forecasts into employee requirements and advising ways the service could be modified to become more efficient through producing staff rosters, ensuring appropriate numbers of employees are present at appropriate times.

Whilst there is a wealth of literature providing guidance on advising staff requirements (summarised in Chapters 5 and 8), further work is still required in the area of time-dependent queueing theory to evaluate more complex systems that do not exhibit assumed traditional features such as consistent Poisson arrival and service rates; and especially multi-server priority systems where the complexities of time-dependent demand and priority routing occur simultaneously. Fildes et al. (2008) comment that OR has made many substantial contributions to forecasting as practitioners continue

to recognise that the accuracy of predictions is important to their organisations; yet the authors note that major research opportunities in this area still remain, as the quantity of study invested in this topic is relatively small to date. They advise that future study should shift away from traditional statistical analyses and into methods that can appropriately deal with the stochastic nature of demand, and that are assessed via organisational performance measures. This thesis responds to the call by the authors to improve the accuracy of forecasts through the use of Singular Spectrum Analysis (SSA) (introduced as a novel method in Chapter 4) to predict future demand. The technique is demonstrated to not only produce superior forecasts to other methods, but to also benefit from being flexible in approach, model-free, able to capture periodicities in the data, not reliant upon artificial assumptions and easily implemented within Excel. Furthermore, this research devotes particular attention to the development, solution and validation of sufficiently detailed stochastic models for time-dependent multi-server systems with varying service types, which can be ultimately be employed to optimise resource allocation.

1.3 The research problem

This thesis investigates how OR techniques may be applied to promote effective and efficient management of EMS, which is widely recognised as a significant challenge in many developed nations (Channouf et al., 2007; Setzler et al., 2009). A particular difficulty for EMS planners is to allocate often limited resources, whilst managing increasing demand for services in a way to ensure high levels of geographical coverage and to improve key performance targets. To aid with the decision of the number of ambulances and paramedics to be deployed, much OR study has been invested in the strategies to optimise optimal fleet sizes and vehicle deployments; yet for these deployment schemes to be effective, the values use to forecast future EMS demand must obviously be accurate (Setzler et al., 2009).

This research begins by responding to the need to produce accurate forecasts of demand, investigating methods that adequately account for nonstationarities; and subsequently considers the staffing problem of emergency response vehicles to provide sufficient coverage and efficient responses to patients requesting assistance with varying degrees of urgency. Whilst the problem of staffing a multi-class multi-type system is recognised as notoriously hard even when demand rates are perfectly predictable (Gurvich et al., 2010), this thesis describes a practical method for finding

staffing requirements in such systems, while simultaneously selecting shifts that cover these requirements, that minimise costs and achieve pre-defined performance standards.

The methodology developed throughout the thesis is evaluated in several case studies which test the applicability of the techniques to predict and simulate data provided by WAST, who provide the real-life context and needs for developing the operational models. The service system lends itself to advanced forecasting and queueing theory analysis through its requirement to provide a consistent service quality, to respond to different categories of calls within set time targets and to operate around-the-clock, while responding to widely fluctuating levels of demand for assistance. WAST is primarily interested in determining the minimum number of ambulance officers required to ensure that given proportions of calls are reached within set time frames, as outlined by governmental targets.

Since the rostering of employees using low-costs shifts that match stochastic demand levels requires the investigation of several inter-related techniques, employee scheduling has received a great deal of attention in the literature (Atlason et al., 2008; Ingolfsson et al., 2010). The process begins with the conversion of demand profiles to coverage requirements, and then progresses to generate an optimised shift schedule. The resulting shift schedule can be subsequently used as input to a rostering system to provide low-cost working schedules for each member of the workforce. Most current practice to optimise personnel scheduling follows the general approach originally presented in Buffa et al. (1976). The paper recommends that the following steps be taken to roster employees:

i. Forecast demand

ii. Convert demand forecasts into staffing requirements

iii. Schedule shifts optimally

iv. Assign employees to shifts

Although the integration of the four processes may allow the creation of the best overall roster, the methodology described allowing the decomposition of the task into several distinct parts makes the problem more tractable. Therefore to allow the generation of a roster in computationally efficient manner, the task is generally approached in a step by step procedure. Step (i) is often assumed known as a precursor in scheduling models

(Setzler et al., 2009), or achieved by extracting historical data from a database on the number of calls received within hourly periods in the past, and estimating the arrival rates for future periods based on the average number of calls received for that period for a given number of previous weeks. Several potential errors can however arise from estimating the arrival rate for a future period in this fashion, including forecasting errors resulting from estimation error associated with taking the average of a finite number of random variables, failure to detect nonstationarities that could be present in the data, and the failure to account for the presence of a random arrival rate which may be a function of external factors e.g. weather conditions (see Setzler et al. (2009)).

The presence of a time-dependent arrival rate makes step (ii) extremely difficult, hence it is often accomplished through the use of simple queueing models and approximation techniques (see Chen and Henderson (2001)). These consider each period of the day independently of other periods, the arrival rate to be stationary within each period and evaluate performance within each period using steady-state measures. The complicating factor however is that the staffing level in one period can considerably impact the service level in subsequent periods, and approximation techniques do not take this into account. Numerical methods may be employed to allow transient analysis of the system in such scenarios, but these provide accurate results at the expense of computation speed. Ingolfsson (2002) recently developed appropriate extensions to allow for the numerical analysis of time-dependent systems; but the techniques have not previously been extended to time-dependent systems which simultaneously deal with more than one class of customer. The implementation of approximation methods is further complicated in such systems when the performance measure of interest is a function of the customer waiting time distribution, since steady-state expressions are not available for this metric (Chen and Henderson, 2001).

Step (iii) is commonly solved through a set-covering problem (see Hari et al. (2011)) and step (iv) is often considered outside the scope of papers investigating the staffing problem. Whilst an entire field of literature is devoted specifically to this topic (especially to the nurse rostering problem, see Burke et al. (2004)), the rostering of ambulance officers to satisfy a large set of working time directives has specifically been approached using an integer programming model in Li and Kozan (2009).

This thesis considers an overview of the techniques required to promote service efficiency and ultimately develops a workforce capacity planning and scheduling tool

which amalgamates several of the techniques into a single integrated model. The self-contained workforce tool is designed with a user-friendly interface and contains several parameters which may be flexibly adjusted by the user to provide staffing recommendations for various scenarios that satisfy the response time targets (set by the Welsh Government). Several features embedded in the tool make it more likely to outperform previous approaches, as whilst prior research works have generally used multi-step procedures to roster staff and restricted in-depth analysis of systems to those with a single class of customer, this research presents a macro view of all the techniques linked together in an integrated planning tool which further captures both time dependency and two priority classes. Yet since the main aim of the thesis is to utilise the wealth of real-life data supplied by the Trust to evaluate the potential of the models developed as workforce planning tools, rather that to accurately model the entire WAST service system; the final model is based on a simplified representation of the EMS division of WAST. In effect, it recommends the number of ambulances and paramedics necessary to provide acceptable first responses to emergencies (based on the government targets), whilst in reality additional resources may also be required to aid the first responders at the scene of the incident.

Specific objectives and questions

While taking into account the importance of accurately estimating future demand, the need to develop OR methodology to evaluate service quality in time-dependent priority multi-server systems, and generate efficient shift schedules, this research aims to address each of the complexities discussed above and integrate the techniques into a self-contained workforce capacity planning tool that:

(a) Incorporates time-series methods that adequately account for the stochastic nature of demand to produce accurate forecasts of future demand;

(b) Provides both accurate and approximate evaluations of system performance over time;

(c) Permits a certain service quality to be met as inexpensively as possible by generating an efficient staffing function that accurately matches resources to fluctuating demand levels;

(d) Generates an optimised shift schedule;

(e) Assigns staff to shifts in an efficient manner, whilst adhering to governmental regulations and working time directives (WTD);

(f) Is user-friendly and practical; so it could be used to inform WAST staffing decisions and readily adopted by planners to optimise resources independently.

Seven main research questions arise from these objectives:

(I) Is it possible to improve the accuracy of demand forecasts, by adequately accounting for seasonality in the data?

(II) Can time-dependent queueing theory techniques be extended to appropriately model system behaviour as servers enter and leave the system in differing fashions across shift boundaries?

(III) To what degree do staffing levels in one period affect another? Can guidelines be provided regarding situations under which it is appropriate to approximate time-dependent behaviour, how accurate the approximations are, and if steps can be taken to increase their accuracy?

(IV) Can time-dependent and approximate queueing theory be extended to compute waiting-time probabilities in time-dependent multi-class, multi-server systems?

(V) Is it possible to increase the efficiency of numerical methods to accurately evaluate system performance?

(VI) Is it possible to develop suitable heuristics to optimise shift schedules and rosters that minimise cost and ensure satisfactory customer service?

(VII) Can the individual forecasting, modelling and optimisation techniques be combined into a generic integrated workforce planning tool to optimise staffing schedules in stochastic environments that must consistently deliver a certain service quality?

In addressing these problems, the research presented in this thesis is novel is several ways. Firstly, it specifically improves the accuracy of demand forecasts using a novel modelling technique known as SSA, and further adapts queueing theory techniques to model a complex time-dependent multi-server priority system. Moreover the unique feeding of novel time-dependent modelling and forecasting techniques, together with links to crew rostering, allows resource allocation to be optimised in an efficient manner.

1.4 WAST

As mentioned above, the methodology developed throughout the thesis is evaluated in several case studies which test the applicability of the techniques to predict and simulate data provided by WAST, who provide the real-life context and needs for developing the operational models. Akin to many developed nations, demand for ambulances in Wales is increasing year on year and is growing at a faster rate than the UK average (Lightfoot Solutions, 2009). Calls made to the service have tripled over the last 20 years (Welsh Government, 2012a), and with above average proportions of patients requiring transportation to hospital, it is unsurprising that WAST has recently struggled to meet key performance targets set by the Welsh Government.

WAST is the only organisation that provides urgent care services on a day to day basis across the whole of Wales and is thus a critical part of the national healthcare system. Whilst this research concerns the role of the EMS arm of WAST which provides urgent and unplanned services to emergency requests, the Trust also has a Patient Care Services (PCS) branch that provides non-emergency transport to care facilities. Operated from 86 ambulance stations, 3 control centres and 3 regional offices, the Trust endeavours to improve the health of patients by delivering a range of effective and appropriate healthcare services which aim to improve clinical outcomes for patients. Facing ever increasing pressures to provide rapid responses, WAST is keen to develop new initiatives to overcome the wide range of challenges presented throughout the country, such as poor transport networks impeding efficient responses in sparsely populated areas of Mid Wales, and the presence of high demand pressures in the South (Welsh Ambulance Services NHS Trust, 2010).

When 999 calls are received, they are immediately categorised into one of three categories of urgency by the calltaker who uses a triage system known as the Advanced Medical Priority Dispatch System (AMPDS) (see Lightfoot Solutions (2009)) as follows:

- Category A - immediately life threatening condition/injury.
- Category B - serious but not life threatening condition/injury.
- Category C - neither life threatening or serious condition/injury.

The Welsh Government specifies a set of response time targets that are considered as

'acceptable' times to reach emergencies within. Whilst these are continually revised and updated, the set of Annual Operating Framework (AOF) targets for Category A calls, reported by the Welsh Government (2011), coinciding with the period of data supplied for this thesis were as follows:

- To achieve a monthly minimum performance of 60% of first response to Category A calls (immediately life-threatening calls) arriving within 8 minutes in each Local Health Board (LHB) area;
- To attain and maintain a month on month all-Wales average performance of at least 65% of first responses to Category A (immediately life-threatening) calls arriving within 8 minutes.
- Performance in all geographical areas needs to reflect continuous improvement in achieving the overall target.

There were additional standards for response times for Category A calls (where the first response was not a fully equipped ambulance), Category B and C calls, and urgent journeys; namely:

- To follow up first responder vehicles to Category A calls with a fully equipped ambulance to a level of 95% within 14, 18 or 21 minutes respectively in urban, rural or sparsely populated areas;
- To respond to all other emergency calls to a level of 95% within 14, 18 or 21 minutes respectively in urban, rural or sparsely populated areas;
- For 95% of responses to doctors urgent calls to arrive no later than 15 minutes after the requested arrival time.

Since WAST's establishment in 1998, it has been scrutinised in respect of performance issues (Welsh Ambulance Services NHS Trust, 2007). In 2006, the Wales Audit Office was commissioned to produce a report into the problems that WAST was facing. It identified the primary areas of concern as (i) the mismatch of the levels of people, vehicles and equipment on the road with those outlined in the plan, (ii) poor management of the 999 cycle process from call receipt to availability for the next job and (iii) driving times in rural areas; and accordingly made a number of recommendations for improvements. In an attempt to revitalise the service, a modernisation plan for both the ambulances services and NHS Direct Wales named 'Time to make a difference'

(Welsh Ambulance Services NHS Trust, 2007) was introduced.

The plan consisted of two key elements: firstly to focus on 'getting the basics right' in terms of the strategy, people, processes and systems; and secondly to aim to reduce bureaucracy and develop new ways of delivering patient care in the long term. Following a significant number of these recommendations being introduced, the follow up review published in 2008 found that WAST's performance relating to the target to respond to 65% of Category A calls within 8 minutes had significantly improved; but there was little improvement in the Category B, Category C and Category A 14/18/21 minute standards, outlined above. This underperformance was deemed to be due to insufficient staffing levels with a high reliance on overtime to fill core shifts, and the estimated costs of employing additional staff to meet the performance targets surpassed £3,000,000 (Lightfoot Solutions, 2009).

As a new policies are being continually devised and implemented, WAST has become a much improved organisation and achieved its best ever responses to Category A calls in 2009/10 (see Appendix A.1, Welsh Ambulance Services NHS Trust (2010)). There is however some way to go in terms of achieving consistency both across the various LHBs in Wales and over time (Welsh Government, 2012b). Whilst the latest released statistics are not directly comparable to the data and standards that are used within this thesis (since new rules have recently been placed on the ambulance service - see Appendix A.2 for further information), they show that WAST has failed to meet the target of responding to 65% of the most urgent calls within eight minutes for the last four successive months analysed at the time of writing this thesis, from June - September 2012.

The 'Time to make a difference' report additionally recommended a re-organisation of NHS Wales, which came into effect on 1st October 2009. This has established seven integrated LHBs throughout the country in place of the previous twenty-two Unitary Authorities (UAs) that were previously responsible for delivering all healthcare within the segregated geographical areas (see Figure 1.1 and accompanying Table 1.1).

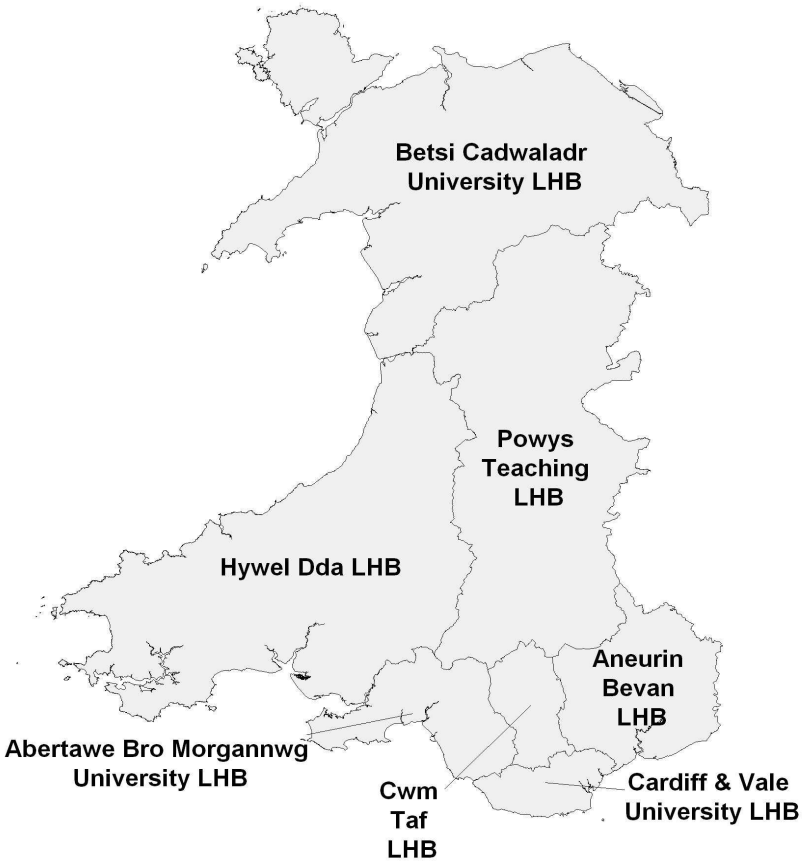

Figure 1.1: Map of LHBs across Wales

Table 1.1: NHS re-organisational structure

Current LHB	Ambulance Region	Previous UA
Powys Teaching LHB	Central & West Wales	Powys Teaching UA
Abertawe Bro Morgannwg University LHB	Central & West Wales	Bridgend UA
		Neath Port Talbot UA
		Swansea UA
Cwm Taf LHB	Central & West Wales	Merthyr Tydfil UA
		Rhondda Cynon Taf UA
Hywel Dda LHB	Central & West Wales	Carmarthenshire UA
		Ceredigion UA
		Pembrokeshire UA
Aneurin Bevan LHB	South East Wales	Blaenau Gwent UA
		Caerphilly UA
		Monmouthshire UA
		Newport UA
		Torfaen UA

Cardiff & Vale University LHB	South East Wales	Cardiff UA
		Vale of Glamorgan UA
Betsi Cadwaladr University LHB	North Wales	Anglesey UA
		Conwy UA
		Denbigshire UA
		Flintshire UA
		Gwynedd UA
		Wrexham UA

Data relating to various LHBs are reported on in chapters throughout this thesis and may be described in the following way. The data discussed for the purpose of demand modelling in Chapter 4 relates to data for the whole of Wales (i.e. accumulation of all seven LHBs), and the models are applied to smaller regions as more in-depth analysis is performed in the latter part of this thesis. The data utilised in the regional analysis in Chapter 6 relates to the data from the South East (SE) Region (accumulation of two LHBs) that allows direct comparisons to be made with current demand predictions and rostering patterns used by WAST. This region has been selected as the principal region for the purpose of comparison, as it covers the largest area in Wales in terms of population and has readily available data that covers a wide range of topographies. Finally the data employed in Chapters 7 and 9 relates to the Cardiff UA only (part of the Cardiff & Vale LHB), as the government response targets are not consistent for the other areas of the South East Region (i.e. 95% of Category B calls should be reached within 14 minutes in the Cardiff region, but 18 minutes outside).

The rules governing the dispatch of ambulances and targeted performance measures are continually changing; the latest being amended to comply with standards that focus more upon clinical outcomes for patients (see Appendix A.2). Despite the fact that the standards fluctuate, in order not to constantly recalculate scenarios, the research conducted within this thesis is based upon the initial criteria, provided in conjunction with the data from WAST in December 2009. The assumptions are based on communication with WAST officials and documented guidance available at this time (see Lightfoot Solutions (2009)). The two targets considered throughout this thesis strictly relate to EMS demand, and may be summarised as follows:

- **Target 1**: To attain and maintain a month on month performance of at least 60% of first responses to Category A calls arriving within 8 minutes in each LHB area; and to follow up with a fully equipped emergency ambulance to a level of 95% within 14, 18 or 21 minutes respectively in urban, rural or sparsely populated

areas.

- **Target 2**: To send a fully equipped emergency ambulance to all other emergency calls (Category B and Category C) to a level of 95% within 14, 18 or 21 minutes respectively in urban, rural or sparsely populated areas.

A large fleet of vehicles may be called upon by WAST to respond to an emergency request for assistance; however the primary vehicles used are Rapid Response Vehicles (RRVs) and fully equipped Emergency Ambulances (EAs). RRVs cannot be used to transport patients as they are typically small vehicles operated by a single health worker; however they offer the advantage that they can rapidly reach the scene of the incident. EAs can be used to transport patients and are typically manned by two crew members (at least one of whom must be a fully trained paramedic). This research assumes the following rules apply when dispatching vehicles to emergencies:

- **RRVs**: A single RRV is sent to every Category A incident. RRVs are reserved for this purpose only, and 65% should arrive within 8 minutes to attain Target 1.
- **EAs**: A single EA is sent to all emergency calls (as a first responder to Category B and C calls; and in conjunction with an RRV to category A calls). In order to achieve Target 2, 95% of EAs should arrive within 14, 18 or 21 minutes in urban, rural or sparsely populated areas respectively.

It should be noted that whilst the situation may be more complex in reality, and ad-hoc decisions may be made on a day-to-day basis, the above assumptions are required in order to allow consistency in the investigations. Since the workforce capacity planning tool developed in conjunction with this thesis is based on the above assumptions, it essentially provides a simplified version of the WAST service system; as in reality additional resources may also be required to aid the first responders at the scene of the incident. Devotion has however been given in the development stage of the models to ensure that they have been constructed in a generic fashion, so they may be readily adapted to incorporate various information, and consider different parameter values to reflect revised performance standards.

In order to ensure the that assumptions made by the models are reasonable and the ultimate results they provide are highly beneficial for WAST, a strong working relationship has been developed with the Trust over the course of the research period. WAST

is keen to develop new initiatives to improve their performance, and the analytics team has accordingly been enthusiastic to maintain contact throughout the project, obtain research updates, provide ideas of direction for future research and answer various queries as they have arisen. Since the receival of the data from the Health Informatics Department in January 2010, several meetings, conference calls and site visits have been made to WAST to gain an appreciation of the day-to-day operations and ensure the forthcoming models reflect the service as accurately as possible. The research has benefitted from discussions with Nicki Maher, Andrew Rees and Jason Weall in the analytics team in North Wales, the Research and Development Manager Richard Whitfield in Cardiff and a site visit to an Ambulance Control Centre in Cwmbran. By holding regular meetings and maintaining close contact with WAST, it has been possible to gain answers to various queries as and when they have arisen, obtain an indication of the main challenges facing the service (and how they have changed over the course of the investigation), and provide regular research updates to management who have been keen to receive progress updates, in anticipation of the opportunity to implement the tools developed to support current practice.

1.5 Thesis structure

The work described is thesis is presented in 11 chapters with appendices, which may be segmented into six primary parts that are designed to follow a logical sequence through the subject matter. A graphical representation of the thesis illustrating the interrelation of the parts and chapters is given in Figure 1.2.

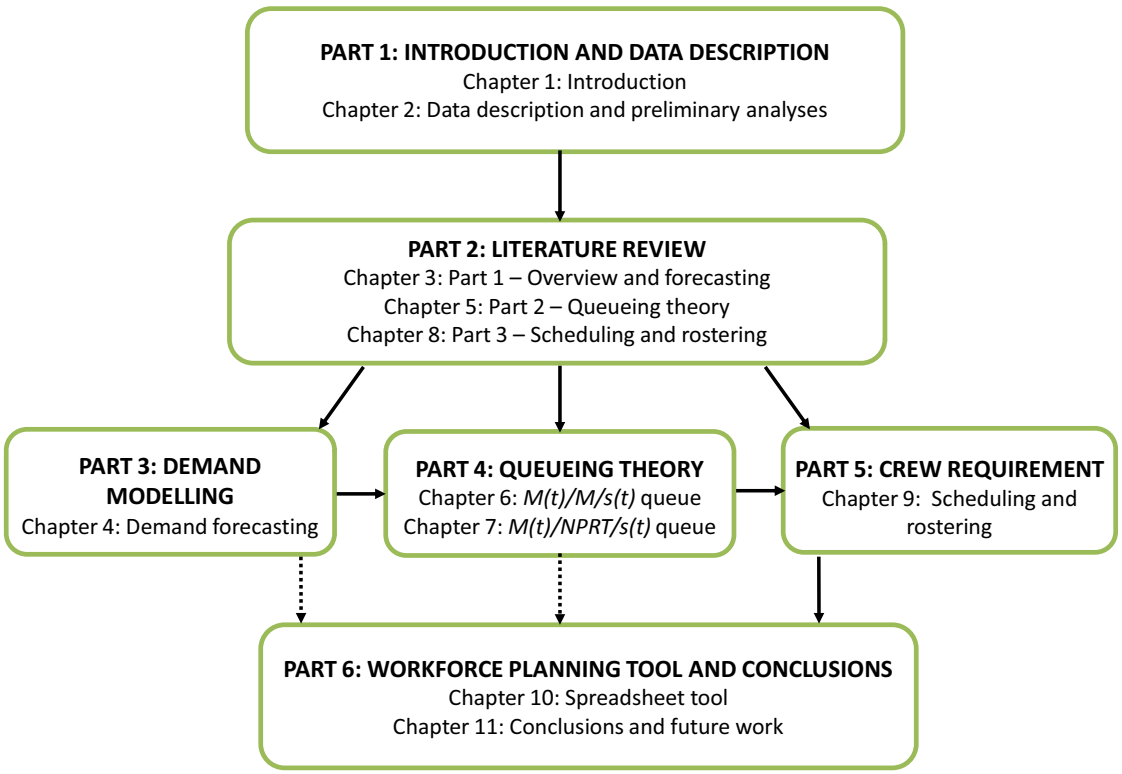

Figure 1.2: Thesis Structure

Part 1 is used to set the context of the work and accordingly comprises the introduction and a description of the data used to test the application of the methods developed throughout this thesis. Part 2 provides a brief overview of the previous research performed in the literature related to the topics investigated. Due to the wide scope of literature relevant to discuss, it is formed of three chapters (3, 5 and 8) which are presented at various points throughout the thesis in conjunction with the investigation of each of the relevant fields. The discussions surrounding the key concepts aim to equip the reader with a wider knowledge of the general approach taken by operational researchers, and to justify the methods selected for investigation. Parts 3 through 5 describe the research work itself; and part 6 consists of two chapters (chapters 10 and 11) which draw together the key parts of the thesis. It includes a summary of the workforce capacity planning tool, followed by a discussion of the key findings, limitations and contributions of this work. The appendices supplied at the end of this document contain further information, supporting tables and research papers that summarise some of the major contributions of this thesis.

The chapters are introduced below in turn. A brief synopsis of each is included, with a description of it's links to the other chapters.

Chapter 1 describes the general importance of the issues discussed in this thesis and sets the research objectives. It additionally provides the motivation for the work and describes the background of WAST (for whom the research is primarily conducted).

Chapter 2 provides a description of the data provided by WAST for analysis. The structure of the data provided for analysis is explained, together with preliminary investigations that summarise the main patterns and trends exhibited in the data.

Chapter 3 examines the literature specifically related to demand modelling and forecasting. It provides a review of conventional forecasting methods, the SSA modelling technique and outlines the development of models devoted to the prediction of demand for emergency response purposes. This chapter informs the research conducted in Chapter 4.

Chapter 4 investigates the potential of SSA to produce superior forecasts of demand to the conventional methods proposed in the literature. For this purpose, a detailed description of the theory underpinning the SSA technique is provided prior to the execution of a comparative study of the ability of several methods to forecast held-back demand data obtained from WAST. These experiments show that in addition to being more flexible in approach, SSA produces superior longer-term forecasts that are especially helpful for EMS planning, and comparable shorter-term forecasts to well established methods.

Chapter 5 provides the literature review of queueing theory that is divided into four main segments: preliminaries, time-dependent approximation methods, time-dependent numerical methods and priority queueing theory. The literature on these topics is wide-ranging, thus the review is highly selective. The chapter aims to give the reader a flavour of some of the general methods used in the literature to deal with non-stationary and priority queues, and to justify the methods investigated and developed in chapters 6 and 7.

Chapter 6 evaluates the potential of various approximation and numerical methods

commonly used by operational researchers to determine staffing levels in service systems subject to time-varying demand. The chapter includes the proposal of novel revisions to SIPP approach which potentially allow the generation of improved forecasts, and explains how the numerical methodology can be employed to accurately track the system behaviour over shift boundaries. So as to illustrate the practical application of the methodology, the chapter applies the models to WAST data to determine staffing requirements for RRVs within the SE region. The case study clarifies the way in which the methods can be used and applied in a practical way.

Chapter 7 focusses on the development of the time-dependent methods discussed in Chapter 6 to enable their application within time-dependent priority systems. This study represents the first time that the techniques have been developed to allow accurate analysis of time-dependent systems with two priority classes; hence some of the most noteworthy contributions of this thesis lie within this chapter. The chapter is composed of identical structure to Chapter 6, but the methodology is applied to generate staffing requirements for EAs which are required to attend to all types of emergency requests, giving priority to the more serious incidents (in place of determining staffing requirements for RRVs which are only assumed to respond to life-threatening requests).

Chapter 8 details the array of different methods that have been used by operational researchers to tackle the problems of scheduling and rostering. The chapter represents the final component part of the literature review and is used to inform Chapter 9.

Chapter 9 explores several issues associated with the staffing of time-dependent priority queues. Drawing upon existing analytical models for staffing systems with time-varying demand, the first section of this chapter reports on the development of integer programming models with heuristic search techniques to effectively schedule shifts for RRV crews in the Cardiff region. The second section proposes a simple rostering algorithm that assigns a hypothetical set of staff at WAST to low-cost shifts. The chapter examines the potential for the optimised shift schedule developed in the first section of the chapter to be subsequently inputted into the rostering system, concluding that it is beneficial to integrate the shift scheduling and the rostering task into a single problem, so the selection of shifts and assignment of shifts to staff can be optimised simultaneously.

Chapter 10 contains a description of the workforce capacity planning and scheduling

tool to optimise resource allocation at WAST. The methodology required to forecast future demand, generate staffing requirements, optimise shift schedules and design a roster that is developed throughout this thesis, are all embedded within the model, and an explanation of the parameters that may be adjusted in the model is given.

Finally, Chapter 11 summarises the contents of this thesis, in addition to describing some potential further work arising from the investigations undertaken. The findings ultimately suggest that the workforce and capacity planning and scheduling tool (which includes an accumulation of the techniques discussed and developed throughout this thesis) may be readily revised and adopted by ambulance service trusts to efficiently and effectively allocate future resources. In particular, it highlights the scenarios under which each of the functions programed in the tool provide reliable results, and highlights the major contributions of this thesis to allow for the accurate tracking of system behaviour across shift boundaries and development of appropriate formulae to calculate the probability that patients are responded to within acceptable time frames.

1.6 Summary

In summary, this chapter has set the general research context, outlined the methodology and discussed the structure of this thesis. In identifying distinct research questions, it has highlighted the need to develop, solve and validate sufficiently detailed stochastic models for planning and managing EMS resources.

Before reporting upon the methodology to be investigated and developed in the main body of this thesis, Chapter 2 first provides a brief description of the data supplied for analysis by WAST.

Chapter 2

Data description and preliminary analyses

2.1 Introductory remarks

This chapter provides an overview of the data that shall be used to test the application of the methods developed within this thesis. It details trends observed in the data and identifies the primary characteristics of the real-life system to be modelled using mathematical techniques.

All the data for this piece of work have been generously supplied by WAST. The primary database provided for analysis is sizeable with 2,500,000 data records; thus for practical purposes, various portions of this data set are used in case studies that test the application of individual models, before they are ultimately integrated in a single workforce capacity planning tool. This chapter aims to inform the reader of the selection of data used for investigations in later chapters, the attributes related to each incident recorded in the data set, and of underlying structures in the data that must be accounted for by the following forecasting and scheduling models.

The chapter is structured as follows. Section 2.2 overviews the information provided in the primary database. It details the most common types of injuries reported to WAST, lists the attributes associated with each incident that are recorded in the database and includes a description of the type of incidents considered to represent a true case of 'unique demand' for the purpose of this thesis. Section 2.3 explains the major stages of the response process, how further categories of service (such as response and

service times) can be computed from the time-stamps recorded in the database, and details corresponding summary statistics relating to WAST's current performance. A summary of the temporal variations observed in demand, focussing on periodicities that can be attributed to the hour-of-day, day-of-week and month-of-year effects, is given in Section 2.4. Finally the chapter ends with a summary in Section 2.5.

2.2 The data source

The primary database corresponds to around 2,500,000 data records from 1st April 2005 to 31st December 2009. Each of these records corresponds to either a submission of request for WAST assistance, the dispatch of a response vehicle, or both. Exploratory analysis reveals that around 7% of incidents reported are instantly discarded by WAST e.g. in cases where the injury is seen as too trivial to justify the dispatch of an ambulance. Over the reported period, 2,335,352 responses were made to 1,754,455 unique incidents, giving a vehicle to incident ratio of about $\frac{4}{3}$. Due to the disparity between the injuries classed as real emergencies by the AMPDS system and those seen as sufficient to justify a 999 call by the general public, there is no singular agreed definition as to what constitutes true demand for WAST assistance. For the purpose of this thesis, demand is considered to be the **number of *unique* emergencies reported to the service that require the deployment of *at least one* emergency response vehicle.**

Table 2.1 contains the names and descriptions of the 24 variables recorded in the database for each data entry: variables 1-10 relate to attributes of the injury (e.g. time, location, incident type, etc); variables 11-16 list the vehicles and hospitals used to treat the patient; variables 17-22 detail important time-stamps constituting marker points of the response operation; and variables 23-24 indicate the reason a response vehicle is stood down, if applicable.

Table 2.1: Variables recorded in database

Field	Field header	Description of header
1	ID	Unique incident identifier
2	Incident Date	Date reported of the incident
3	Incident Time	Time reported of the incident
4	Hour Of Call	Hour that the call originated
5	Incident Weekday	Weekday that the incident originated
4	Postcode Area	Postcode area of the scene of the incident
7	PCT Code	Unitary Authority Code
8	Nature	Coded medical nature of the incident
9	Incident Type	Type of call requiring service
10	AMPDS	Category of incident based on the priority
11	Dispatch Code	Coded information of the dispatch including region
12	Vehicle Order	Specific allocation order of vehicle to the incident
13	Vehicle Type	Type of vehicle allocated
14	Vehicle Station	Base station the vehicle is assigned to
15	Hospital Attended	Hospital the vehicle transports the patient to
16	Units	Vehicle call sign/identifier
17	Time Allocated	Time vehicle assigned/allocated to the incident
18	Time Mobile	Time the vehicle goes mobile
19	Time At Scene	Time vehicle arrives at scene
20	Time Left Scene	Time vehicle leaves scene
21	Time At Hospital	Time vehicle arrives at hospital
22	Time Clear	Time vehicle becomes clear for another call
23	Stood Down	Yes and No for vehicle stepping down
24	Reason Stopped	Reason why the vehicle is stood down

One of the first issues to address is the large range of incidents that prompt a request for WAST assistance (variables 8-10). Table 2.2 displays the percentage of unique incidents requiring the mobilisation of at least one emergency response vehicle, reported by clinical category. For clarification purposes, all injuries that were reported less than 1,500 times for the investigated period are grouped into a single category 'other' (representative of less than 0.1% of the total), whilst those classed as 'missing' relate to instances where no entry is recorded for this field in the database. More than half of the calls requiring emergency transportation relate to traumatic falls/back injuries, breathing problems, chest pain, unconscious/passing out or the general 'urgent admission' category.

Table 2.2: Number of unique incidents by nature of emergency, Apr 2004 - Dec 2009

Nature	Number of unique records	Percentage
Falls/Back Injuries - Traumatic	224,943	12.82%
Urgent Admission	205,237	11.70%
Breathing Problems	184,223	10.50%
Chest Pain	157,913	9.00%
Unconscious/Passing Out	104,739	5.97%
Sick Person - Specific Diagnosis	98,324	5.60%
Overdose/Ingestion/Poisoning	72,755	4.15%
URGENT TRANSFER	63,131	3.60%
MISSING	***62,938***	***3.59%***
Convulsions/Fitting	58,916	3.36%
Haemorrhage/Lacerations	56,381	3.21%
Traffic Accidents RTA	55,567	3.17%
Abdominal Pain/Problems	52,988	3.02%
Assault/Rape	49,334	2.81%
Traumatic Injuries - Specific	44,693	2.55%
Unknown Problem - Collapse-3rd Pty	30,879	1.76%
Stroke - CVA	28,755	1.64%
Psychiatric/Suicide Attempt	26,581	1.52%
Pregnancy/Childbirth/Miscarriage	24,409	1.39%
Cardiac/Respiratory Arrest	21,863	1.25%
Diabetic Problems	21,119	1.20%
Heart Problems	15,153	0.86%
Use code UKN	14,036	0.80%
Burns/Explosion	14,015	0.80%
Back Pain - Non-Traumatic	12,851	0.73%
Headache	11,096	0.63%
Call From Other Ambulance Service	9,547	0.54%
OTHER	***8,621***	***0.49%***
Allergies/Rash/Med Reaction/Stings	7,411	0.42%
Choking	5,449	0.31%
Stab/Gunshot Wound	3,795	0.22%
Running Call	3,242	0.18%
Eye Problems/Injuries	1,871	0.11%
Carbon Monoxide/Inhalation	1,680	0.10%
Total	**1,754,455**	**100%**

Whilst the majority of injuries listed in Table 2.2 are life-threatening, some are less severe (although potentially still serious); and WAST vehicles are also required to these incidents. In fact, the injuries recorded in the database correspond to three different types of demand: EMS demand (i.e. Category A/B/C calls (see Chapter 1) arriving by means of 999 calls), urgent demand (cases where an ambulance is

requested by a doctor to transport the patient to a hospital within a specified time frame) and routine demand (any incident within the alert system that falls outside EMS and urgent). Variable 9 defines the type of demand that relates to each record, and variable 10 further partitions all EMS demand into Category A, B and C injuries. Figure 2.1 shows that EMS calls accounted for 80% of all incidents reported to WAST over the 5 year period. The second pie chart shows that of these calls, there were approximately even number of Category A and B incidents reported, and a smaller proportion of Category C calls.

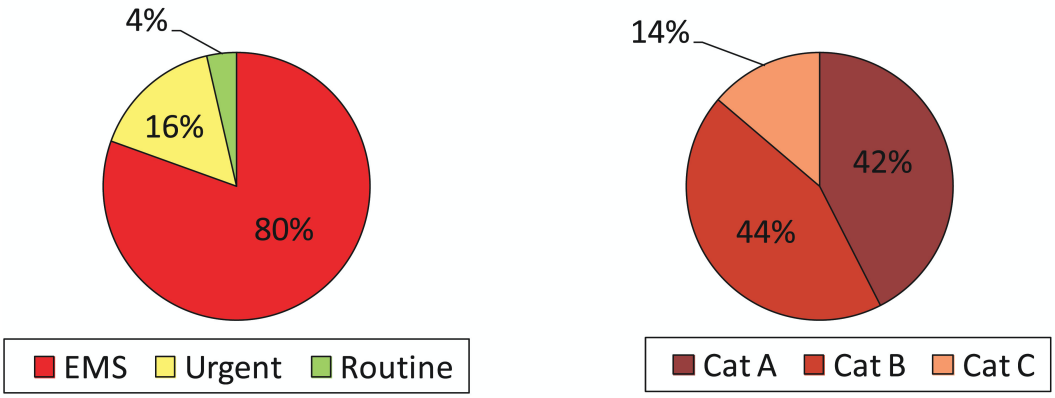

Figure 2.1: Pie charts of demand, by incident type (April 2005 - December 2009)

The demand analysed in Section 2.2, Section 2.4 and Chapter 4 corresponds to all three types of demand; but since EMS demand is the major object of study in this thesis, only the data contained in the second pie chart is used to test the application of the methodology in the latter chapters (as the targets outlined in Chapter 1 specifically relate to particular categories of EMS demand). Urgent and routine demand can generally be viewed as a separate entity, and can sometimes be dealt with by other responders such as a high dependency service vehicle or a 24 hour Patient Transport Service (PTS) vehicle designated for the non-emergency transport of passengers for medical purposes.

2.3 The response process

The resource requirements per incident are of the order of a few minutes for call evaluation and dispatch, but around an hour for an ambulance and its crew. The

time required for crew members to deal with each incident is increasing in many locations due to lengthy patient handover times at hospitals (Channouf et al., 2007). As mentioned in Section 1.4, WAST is expected to attain and maintain a month on month all-Wales average performance of at least 65% of first responses to Category A calls arriving within 8 minutes. The purpose of this is to give a patient who has suffered a heart attack a reasonable chance of resuscitation, so the response may consist of any suitably trained person who has access to a defibrillator which they have been trained to use. In the case of WAST, this 'response time' includes the time taken to mobilise the vehicle and for it to travel to the scene of the incident, plus an additional 15 seconds to allocate the vehicle - see Figure 2.2 (Lightfoot Solutions, 2009). The figure demonstrates that assuming best practice processes and technology, the time that is available for the responder to travel to the scene is 7 minutes.

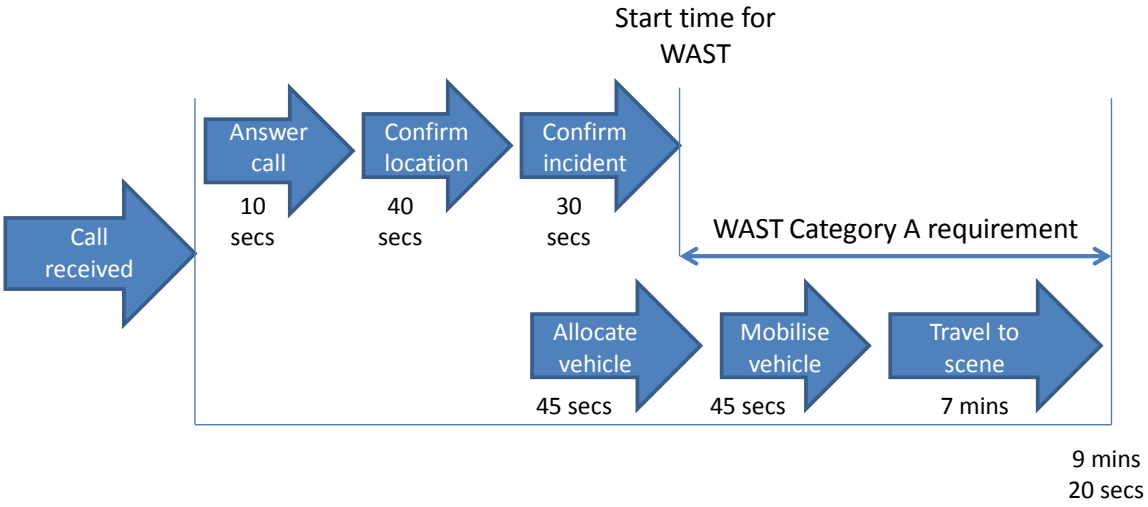

Figure 2.2: The components of the Category A response process

In response to Category B and Category C calls, WAST is expected to send an EA to the scene within 14, 18 or 21 minutes, depending on the area; thus the response times to such calls are generally more variable.

Direct analysis of WAST data reveals that the actual response times vary considerably from those given as guidelines in Figure 2.2. The distribution of first response times to all Category A calls, and first EA responses to Category B/C calls calculated from the database are displayed in Figure 2.3. (Recall from Chapter 1 that only EA's are sufficient to achieve the Category B/C target; thus for such incidents, the first

response time has been calculated for this vehicle type specifically and any response vehicles arriving at the scene of such incident before the first EA have been excluded from the analysis). There is a long-tail in the distribution of precise response times, which extend to over 10 hours in extreme cases, but for clarification purposes the number of responses with times above 30 minutes have been grouped into a single category in the chart.

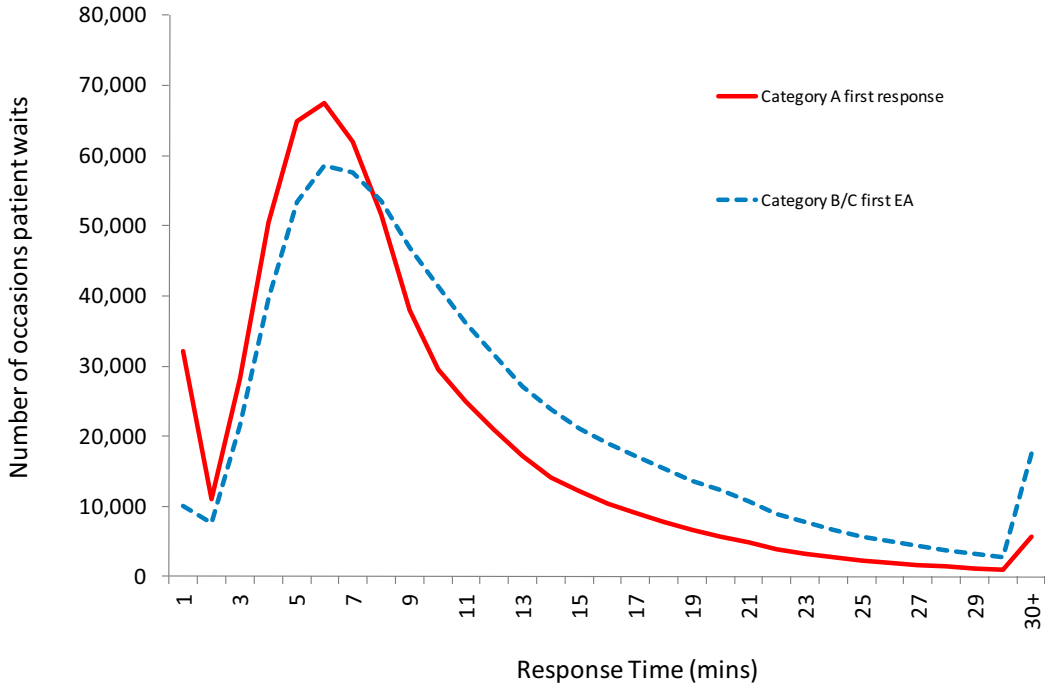

Figure 2.3: Distribution of first response times to Category A and B/C incidents

Notably high volumes of incidents are responded to in 1 minute or less (common in settings where an ambulance is placed at the scene of a large event as a precautionary measure). Although the mode response time for the relevant first responders is around 5 minutes (for both categories), the long tails increase the average response times considerably. Whilst both distributions are positively skewed, the degree of skewness and kurtosis is higher for Category A responses, and the slower degree of decay for the Category B/C curve plotted in Figure 2.3 demonstrates that higher proportions of B/C calls experience longer response times. On the whole, WAST did not quite achieve the Category A target at the all-Wales level over the investigated

period, with an average of 63% of responses to Category A incidents occurring within 8 minutes; although notable improvements have been made in recent years (see Appendix A.1). The response targets for Category B calls are dependent on the area of the country in which the incident was reported in, and thus there is no directly comparable all-Wales target. Yet even if the target that applies to sparsely populated areas (i.e. to respond to 95% of calls within 21 minutes) was treated as the all-Wales standard, WAST's true performance (90%) would be classed as sub-standard.

The response times plotted in Figure 2.3 were calculated using the interval defined in Figure 2.2, from the fields listed in the database (i.e. response time = time at scene - time allocated + 15 seconds allocation time). As the targets only concern the time that the *first* responder/EA arrives at the scene, the time lag was calculated between the earliest non-empty arrival time at scene and the corresponding vehicle allocation time, for each incident. Several cases were removed in the data cleaning process prior to the analysis, as some incorrect time entries were observed that provided infeasible negative response times. Immediate responses were also recorded for numerous cases (i.e. 0 minute response times); but these cases have been retained for analysis as WAST officials have verified that such cases are feasible in cases such as sports events, where vehicles maybe placed on standby at the scene of the event.

Correspondingly, service times have been calculated as service time = time clear - time mobile. The calculations are summarised in Figure 2.4, which portrays the main stages of the WAST service system using the time-stamps recorded for each journey in the database.

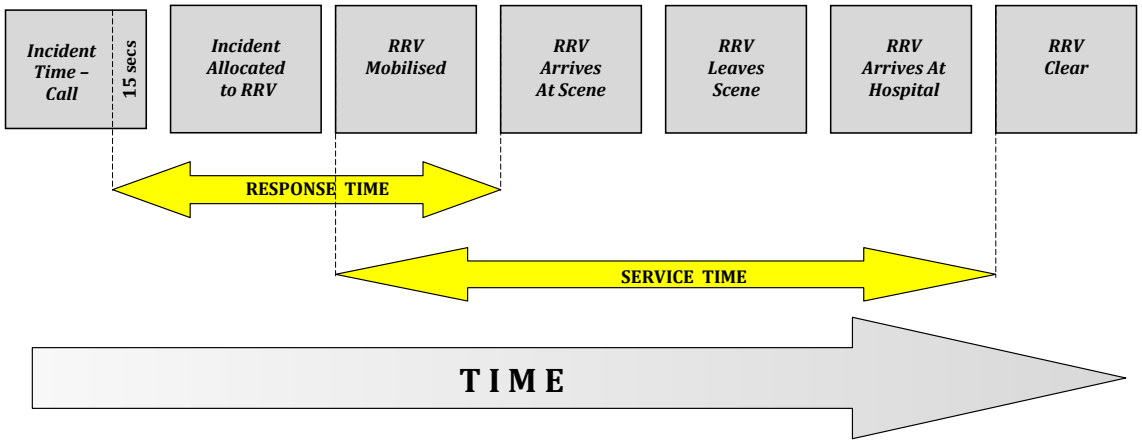

Figure 2.4: Time flow of ambulance response to incidents

In addition to the main target to attain a monthly all-Wales average performance of 65% of first responses to Category A calls arriving within 8 minutes; secondary standards for response times are also specified for such calls. The targets state that WAST should achieve a monthly all-Wales average performance of 70% and 75% of first responses to Category A calls arriving within 9 and 10 minutes respectively. Figure 2.5 shows that WAST came close to upholding these standards between April 2004 and December 2009; achieving the guidelines in 68% and 73% of cases respectively, although it should be noted that the performance levels varied considerably throughout the country. The chart highlights that response times in excess of 20 minutes applied to around 5% of calls, and 99% of calls were responded to within 30 minutes. Due to the long tail of response times extending to over 10 hours, the response times for the remaining 1% of calls have been excluded from the chart.

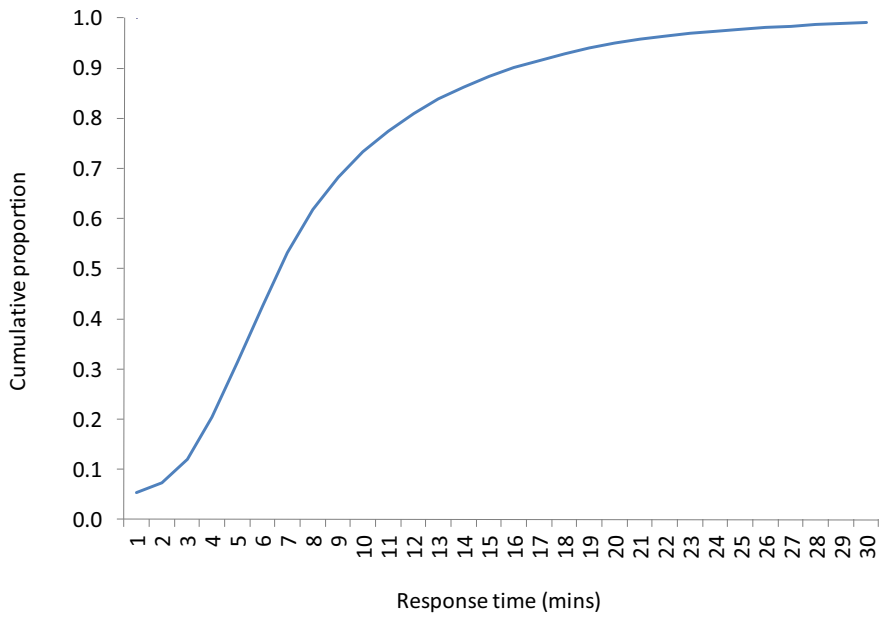

Figure 2.5: Cumulative proportion of first responses to Category A incidents

Both response and service times are critical inputs required by the scheduling models developed in the latter section of this thesis. A summary of the differences between the response and service times for to all unique Category A and Category B/C incidents recorded in the database is given in Table 2.3. Since the target for Category B/C incidents refers to EAs, the times recorded in the table refer specifically to the first EA to arrive at the scene of the incident, whilst the times recorded for Category A incidents relates to the first vehicle to reach the scene of the incident, regardless of its type.

Table 2.3: Summary statistics for response and service times, Apr 2004 - Dec 2009

Statistic	Response time (mins)		Service time (mins)	
	Category A	Category B/C	Category A	Category B/C
Mean	8.22	10.89	49.17	52.96
SD	6.18	7.88	45.60	40.75

The average response times for Category A and B/C incidents appear as expected. It is logical that a patient exhibiting Category B/C symptoms waits longer for service than a Category A patient since these injuries are of lower priority, and RRVs (which are generally able to travel at quicker speeds) do not count as sufficient responses

for the Category B/C target. This factor may also contribute to the marginally longer service times reported for Category B/C first EA responders, since if the first responder to a Category A incident is not sufficient to transport the patient to hospital then an additional vehicle may be called upon for such purposes, if required. The RRV is often released from its current duty and deemed free to respond to another call at the time it leaves the scene of the incident, but if the patient requires transportation to hospital, the RRV sometimes follows the EA, and will thus be busy for a longer period of time in such a scenario.

Whilst an exact arrival time is recorded for each call in the database, this research primarily considers the total number of calls received each day, or in latter chapters the number of calls within each hour of each day, to facilitate the application of time series and scheduling models. The next section explains that demand is more volatile during weekends and winter months, and that there are daily, weekly and monthly cycles in the data that must be accounted for by the stochastic models. Similarly to the data used for the forecasting investigation in Chapter 4, the data analysed here relates to all EMS, urgent and routine demand at an all-Wales level; and a reported incident is defined as an event if one or more emergency medical vehicle is deployed in response.

2.4 Exploratory demand analysis

Figure 2.6 is the time series of the total number of unique calls arriving in each day throughout the whole of Wales over the 57 month period. The series indicates that the total number of requests for assistance fluctuates greatly from 697 to 1,485 each day. Preliminary analysis reveals daily, weekly and yearly periodicities; special-day effects; autocorrelations and a positive trend. Linear regression of daily demand against time yields a slope coefficient of 0.045 ($p = 0.003$) and a distribution analysis reveals that the daily demand level exerted upon WAST follows a normal distribution ($p < 0.0001$) with an average of 1,011 requests (SD= 68.43). The four high extreme values displayed in the chart all occur on January 1st, representing the repeating pattern of extreme demand for the service following annual New Year's Eve celebrations. Whilst forecasting models are not expected to explain these extreme values, they have been retained in the analysis in Chapter 4 so as to maintain the weekly structure inherent in the data. The notable troughs occur on 21st March 2006, 31st October 2007 and 18th May 2009. There however appears no obvious reason for these low counts.

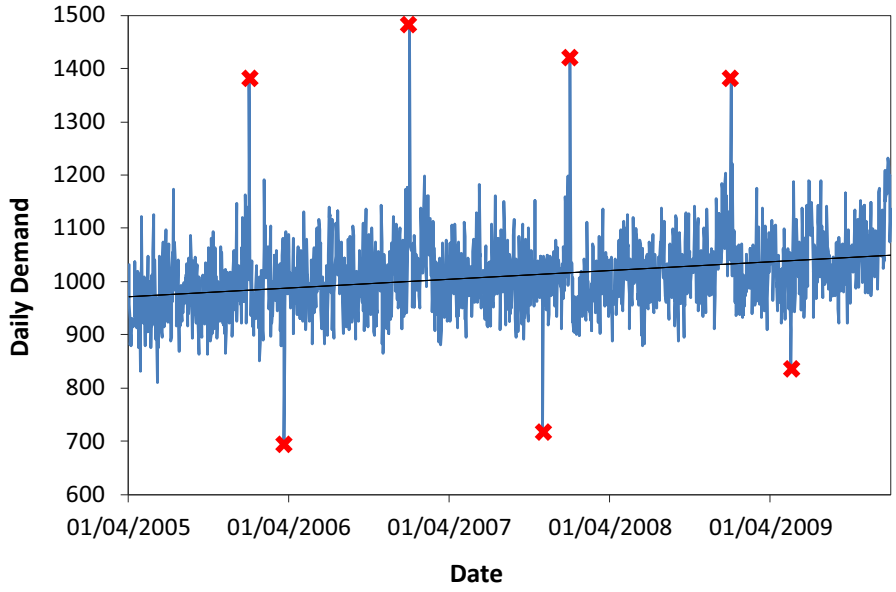

Figure 2.6: WAST daily demand, April 2004 - December 2009

Figure 2.7 shows the average volumes, by month, over the yearly cycle. The graph reveals a clear increase in demand levels over the five-year period and one can observe a marked peak for ambulance services in December.

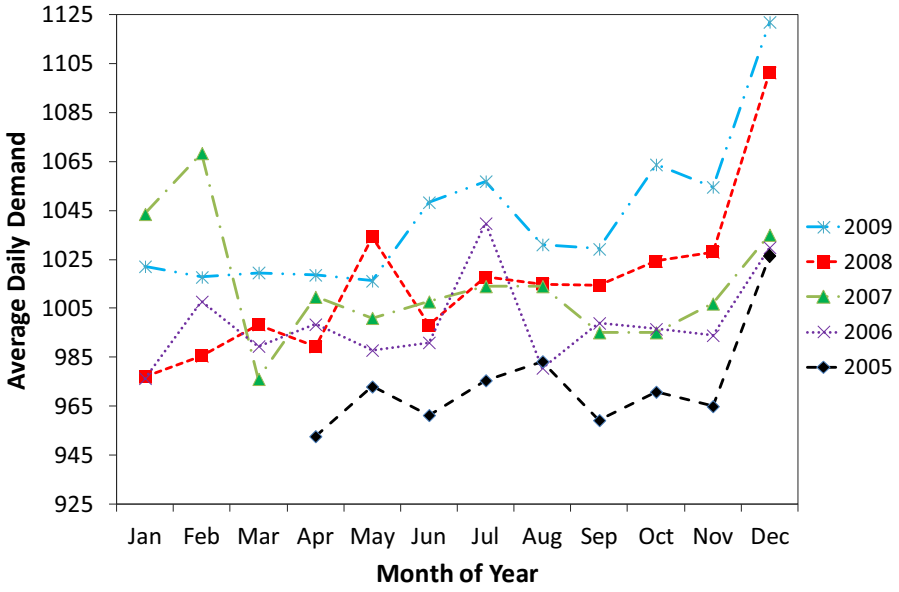

Figure 2.7: WAST average monthly demand, April 2005 - December 2009

The box plots in Figure 2.8 highlight the marked differences in demand volumes by month of the year and day of the week. December is the busiest month with a median of 1,063 incidents requiring WAST mobilisation a day. With the exceptions of January and July, higher demand is generally demonstrated during the winter months. Clear weekday effects are notable with larger volumes of incidents observed on Fridays and Saturdays. All such observations will become of key importance when designing schedules for ambulance crews.

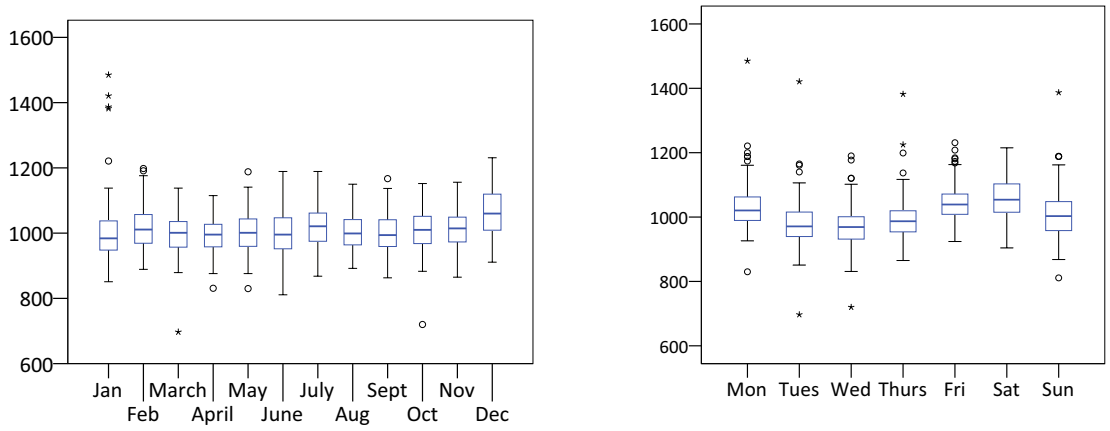

Figure 2.8: Box plots of demand volumes for each month and each day of week

Assuming that the latent call arrival intensity function can be well approximated as being constant over one hour periods, Figure 2.9 displays the average number of requests for WAST assistance for each hour on each day of the week. The chart shows that the pattern of call arrivals has a distinct shape over a typical weekday, but the volumes are notably different on weekends. For weekdays, call volumes generally increase in the late morning and peak around the middle of the day - lunchtime perhaps - about 11am to 1pm, before falling again throughout the afternoon and early evening. Substantially lower volumes are seen overnight with a trough around 5-6am. One can also observe increased activities over Friday and Saturday evenings and overnight, as may have been expected. Recall from Figure 2.8 that larger values are also observed on these days with respect to daily volume.

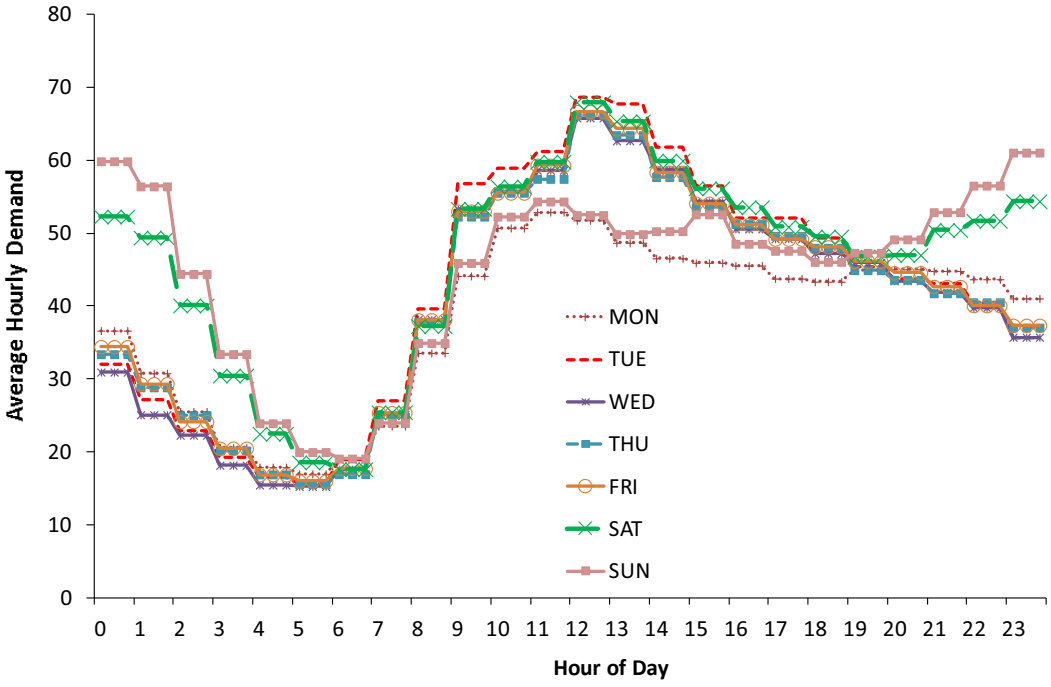

Figure 2.9: Mean number of incidents reported per hour, by weekday, April 2005 - December 2009

Due to the heavily time-dependent nature of demand throughout the day, discussions with WAST officials have revealed that three separate time periods are generally considered for the purpose of scheduling staff and generating resource allocation plans for each day of the week: Morning (6am-12pm), Afternoon (12pm-7pm) and Night (7pm-6am). The research contained to generate staffing profiles in Chapter 6 accordingly generates outputs that satisfy these standard periods, but additional shifts are considered as the shift pattern is scrutinised in the investigation of scheduling and rostering techniques in Chapter 9.

2.5 Summary

This chapter has provided an overview of the data made available by WAST, defined key intervals of interest and concluded with an exploratory analysis revealing the main trends and patterns in the data. The provision of such a rich data set containing so many variables has awarded the researcher with great potential to test the application of a wide range of statistical and OR techniques; and the data will be utilised

for such purposes in later chapters of the thesis. Yet for practical purposes, the models developed require several assumptions and information provided by different sub-sections of the data. This chapter has attempted to provide the reader with such key definitions and assumptions before the analysis was performed. In summary of the key points outlined, the research performed in the remainder of this thesis relates to all emergencies reported to WAST between April 2004 and December 2009 for which at least one ambulance was dispatched (and for Chapter 6 onwards, the analysis concerns only the EMS division).

It is good practice for all statistical modelling processes to start with an exploratory statistical analysis of the data in order to provide insights into the data set, uncover underlying structures and detect outliers; and the last section of the chapter was devoted to such preliminary analysis. It revealed that large variation is apparent in the daily arrival rates; with increased demand in recent years, on weekends, during peak travel times and throughout certain months. This is not surprising since certain accidents are linked with activities that occur on cyclical bases; and such patterns must be incorporated in forecasting and scheduling models.

The next few chapters will look at mathematical techniques that can be used to forecast call arrival rates and model the operations of service systems subject to time-dependent arrivals, with differing priority levels of requests. Once the techniques have been established, the data discussed in this chapter shall be used to form models and evaluate their potential to inform real-life scheduling decisions. In the subsequent modelling process, sufficient attention will be devoted to ensure that the main system characteristics highlighted in this chapter are accounted for in the following ways:

i. Seasonal and stochastic variations in demand:

   - The potential of a relatively modern time series technique, SSA, shall be considered to produce forecasts that adequately account for the seasonal and stochastic variations in the data.

ii. Time-dependent demand:

   - The demand for WAST assistance is highly non-stationary; thus it is the intention of this thesis to develop a suite of methods that adequately deal with non-stationary data. The SSA forecasting technique is an innovative flexible data-adaptive method for analysing non-stationary time series, and

time-dependent queueing theory techniques shall further be investigated in conjunction with sophisticated rostering patterns to provide staffing profiles that accurately match the fluctuating demand levels.

iii. Distinction between Category A and Category B/C emergencies:

- Priority queueing theory techniques shall be developed (in conjunction with the time-dependent models) to account for the priority awarded to life-threatening emergencies.

iv. WAST's accountability to meet key performance targets:

- Through devoting sufficient attention to capture the aforementioned characteristics of WAST, the integrated workforce capacity planning tool that is summarised in Chapter 10, amalgamates several of the advanced stochastic modelling techniques that are developed throughout this research in a single spreadsheet tool. It ultimately outputs staffing requirements and rosters that allow the performance targets to be met.

Chapter 3

Literature review (part 1): Forecasting demand

3.1 Introductory remarks

This chapter is dedicated to research specifically related to forecasting demand. The review focusses on the major milestones presented in the development of forecasting of ambulance demand, but is not restricted to ambulance demand alone since such models are comparable to those designed for call centres, and police and fire services (see Holcomb and Sharpe (2007)).

This chapter begins with a description of the current forecasting methodology used by WAST (in Section 3.2) and a summary of the main trends and patterns discovered in ambulance demand by the early exploratory papers (in Section 3.3). Section 3.4 overviews conventional statistical models that have been researched and developed in the literature to generate predictions, followed by Section 3.5 which focusses on the applications and theory of the SSA technique (used to generate forecasts for the queueing models employed in the latter chapters of this thesis). Whilst SSA has been shown to be successful in many diverse areas to date (see for example Rodo et al. (2002) (climatology), Weare and Nasstrom (1982) (meteorology) and (Hassani et al., 2010) (socio-economic science)), it has not previously been applied to ambulance demand.

3.2 Current Practice

Despite the potential of advanced statistical models to offer accurate demand forecasts, current practice to forecast call arrival rates is often rudimentary. For example, WAST currently use a demand pattern analysis technique known as average peak demand, that is intended to provide sufficient capacity to respond to requests for aid during periods with peak demand levels. For each hour of each day, the number of requests made for an ambulance at that hour on that day for the 50 weeks previous is calculated and then the maximum number in each 10-week period is selected to provide 5 'peak' demand values. The average of these values is selected as the average peak demand value for this hour of this day, and the number of ambulances deployed during this hour in future weeks is based on the concept that there must be a sufficient number to cope with such demand. A numerical example of this methodology is contained in Figure 3.1. It reports the number of EMS requests arising within the SE Region between 13:00-14:00 for each Monday for a 50 week period, with peak call rates highlighted for each 10 week block.

Weeks 1-10	35	40	41	40	**43**	31	30	35	34	40
Weeks 11-20	**42**	38	32	40	40	39	38	31	26	25
Weeks 21-30	**51**	50	33	29	22	45	30	39	35	27
Weeks 31-40	37	31	29	35	41	40	34	**45**	35	29
Weeks 41-50	40	27	40	30	35	39	33	39	42	**45**

Figure 3.1: Unique EMS incidents reported, Hour 13, Monday, 50 weeks beginning 21st September 2009

Taking an average of all the data in Figure 3.1 would give $1,807 \div 50 = 36.1$ requests per hour, but averaging out the peaks gives a higher value of $226 \div 5 = 45.2$. Thus the rosters constructed for future Monday 13:00 periods are designed to provide sufficient coverage to respond to 45.2 incidents, as from the data sample above this would only leave 2 days out of the 50 which would be over the 'peak' demand.

The methodology followed by WAST is quite common in the industry, although

different trusts use varying numbers of historic observations in their formulations (e.g. Setzler et al. (2009) document a EMS agency in North Carolina which consider the twenty previous observations). Noting that EMS system managers commonly utilise historical data to calculate their anticipated call volumes via a method of prediction known as demand pattern analysis, Brown et al. (2007) analysed three variations of the method: average peak demand, smoothed average peak demand and 90th percentile rank; to predict demand for each hour of each day for 52 weeks, based on known values for an initial 20 week period. Whilst average peak demand makes predictions for each hour based upon the average of the highest number of calls recorded for that hour in the first 10 weeks and last 10 weeks of the initial 20 week period; smoothed average peak demand takes a weighted average of the average peak prediction for the hour in question along with the preceding and following hour; and 90th percentile rank distinctly considers the demand for each of the 20 weeks in the initial period, and takes the value at the 90th percentile from the rank list. The authors concluded that demand pattern analysis generally either accurately estimated or overestimated call volume; making it a reasonable predictor for ambulance staffing patterns. In particular, they discovered that 90th percentile rank accurately predicted call volume more consistently than the other methods. Yet even though this was found to be the best of the three methods, it was shown to accurately predict call volumes ($\pm$ 1 call) 19% of the time. It underestimated demand 7% of the time, and overestimated the remaining 74% of the time. It is at the discretion of individual communities to determine if these over- and under-estimation rates are acceptable for their situation.

Although the demand pattern analyses take into account some stochastic variation of the inputs, they do not necessarily take into account seasonal variations and other stochastic effects that might arise; and a great deal of information is lost through using summary measures to inform the forecasts in place of the data itself (Gillard and Knight, 2012). Furthermore, the predictions generated by the demand pattern analysis techniques are infrequently updated in practice, meaning the coverage requirements are often repeated on cyclical bases, and the forecasts can provide inaccurate forecasts if the 'peak' demand values are extreme.

It is of course imperative to develop appropriate estimates of future demand since excessively large estimates lead to over-staffing and unnecessarily high costs, while low estimates lead to under-staffing and slow response times. In recognition of the shortfalls of the elementary methods to produce accurate demand forecasts, numerous statistical

models have been investigated and developed by operational researchers to provide advanced forecasts by simultaneously dealing with trend, seasonal fluctuations, and random error. Sections 3.4 and 3.5 describe several such models commonly employed to achieve this goal in the literature; together with their potential, drawbacks and applications. When developing planning models for EMS systems, researchers have used both regression models to explain the spatial variation of demand, and time series models to account for variations over time. Before considering the potential of methods to explain variation in demand over time, Section 3.3 firstly overviews the early papers investigating EMS demand which focus upon exploratory analysis of trends, patterns and spatial variations in demand, which must be accounted for in the scheduling models.

3.3 Early exploratory analysis

In the 1960's authors such as Weinerman et al. (1966), King and Sox (1967), Lavenhar et al. (1968) and Alpert et al. (1969) recognised the necessity for knowledge of the nature and distribution of emergency cases following a period of increased numbers of people attending emergency facilities over the previous few decades. The main reasons for the analyses were to provide basic statistics relating to the number and characteristics of patients attending the facilities, to allow investigations into the established systems and evaluation of proposed changes. King and Sox (1967) gathered data relating to ambulance usage in the San Francisco area between April 1963-April 1964 to primarily determine the nature and distribution of emergencies, and Lavenhar et al. (1968) analysed data obtained through a screening procedure at Yale-New Haven hospital (the largest hospital in Connecticut) to develop more appropriate indices for the prediction of use of the hospital emergency service. The reports suggested that much of the increase related to non-urgent usage and was due to dependence on the hospital for medical needs by the inner-city population, as they relied on the hospital for general care and often failed to seek prompt attention from private physicians. Common findings were that people were more likely to use an emergency facility if they were young, male, single, non-white, living in an inner-city area, of lower socio-economic class or unemployed.

The study by King and Sox (1967) discovered that approximately half of the ambulance runs were made between 4pm and midnight, with peak loads occurring on Saturdays and Sundays, and that no patient was delivered to an emergency facility

in about one third of cases. A few specific studies were performed alongside the exploratory papers, such as that by King (1968) who exclusively analysed highway accidents, recognising that almost all such accidents involving injury required the assistance of the ambulance service.

In the early 1970's, several authors began to tackle the problem of 'ambulance demand prediction'; attempting to explain the demand in small regions of the USA using a limited number of demographics, alongside land use and socio-economic variables as discussed in Aldrich et al. (1971), Kvalseth and Deems (1979) and Kamentzky et al. (1982). The earliest models, based on multiple regression, were often performed on incomplete data sets with outdated socioeconomic and population data; nevertheless they generated models capable of predicting total yearly demand to a high degree of accuracy due to the significant effects of sociodemographic characteristics.

Aldrich et al. (1971) used 31 independent variables to predict per capita EMS demand in 157 areas of Los Angeles; in addition to per capita demand for 6 different incident types (namely traffic accidents, other accidents, dry runs, cardiac cases, poison cases and other illnesses). Socioeconomic variables were found not to be as powerful in the analysis as other variables, but housing density and land use had some effect. Despite the problems with outdated data and incomplete records of emergency calls, the linear model predicting total per capita demand achieved an adjusted R^2 value of 0.93. The regression models developed for the distinct incident categories were reasonably accurate, but did not achieve as high R^2 values as total demand. The paper concluded that a linear model employing sociodemographic variables was capable of predicting total ambulance demand to a high degree of accuracy and found that areas with high densities of low-income families, non-whites, elderly people or children tended to use the public ambulance service more often than others. Although Aldrich noted in his work that the method described in his paper would be applicable to any region, he highlighted that the specific regression model would not.

Aldrich's work was extended by several authors such as Siler (1975), Kvalseth and Deems (1979) and Kamentzky et al. (1982), who recognised that yearly demand for public ambulances appeared to be highly predictable within a certain city or country. By simply adjusting the variables considered, they developed regression models capable of predicting demand variations for specific ambulance trusts. Siler (1975) produced a simplified model using linear and non-linear forms of

only four socioeconomic explanatory variables selected through stepwise and multiple regression analyses; and Kvalseth and Deems (1979) concentrated on specifically predicting demand in the city of Atlanta by experimenting with first- and second-order regression models.

Kamentzky et al. (1982) further successfully explained the variation in demand in Southwestern Pennsylvania using only four independent variables selected using both stepwise and forced entry procedures. The final model to explain total demand provided an excellent fit to the data ($R^2 = 0.90$); and those developed to separately predict the number of life-threatening cases, minor-moderate cases and trips in which there was no patient obtained adjusted R^2 values ranging from 0.68 to 0.91. More recent work using regression techniques has been performed by McConnel and Wilson (1998) who gave particular focus to the age distribution of the population. Their research found that pattern of utilisation of EMS services associated with age was tri-modal, with rates rising geometrically for individuals aged 65 and over, to the extent that the utilisation rate was 3.4 times higher ($p < 0.001$) for those aged 85 years and over compared to those aged 45 - 64.

3.4 Time series models

A new collection of models was established at the end of the 1980's as authors considered classical time series models such as Autoregressive Integrated Moving Average (ARIMA) models, which were successful in overcoming some of the shortfalls of regression techniques such as multicollinearity, autocorrelation and the difficulty of selecting covariates. The optimisation technique of goal programming (which extends general linear programming methods to consider multiple objectives), was used as part of a multistep approach taken by Baker and Fitzpatrick (1986) to choose the optimal smoothing parameters in a Holt-Winters exponential smoothing model (HW) to forecast daily ambulance demand in South Carolina. The specific time series they investigated was an aggregate of two separate series recording emergency and non-emergency calls, which the researchers first disaggregated to separately forecast the distinct types of demand using the HW model. The models for each of the demand types included terms to account for average, trend, seasonal and day-of-week effects, with goal programming subsequently used to select the best parameter estimates for the overall model, giving higher priority to the forecasting of emergency demand. By means of forecasting the distinct types of demand separately and aggregating the

results, the model allowed a prioritised forecast to be generated and provided more accurate forecasts for total demand levels. This model however was not primarily selected by the researchers because it was the 'best' model, but because it could be easily explained to local planners, and required a minimal amount of technical expertise.

Few studies have focussed specifically on the use of conventional time series models to forecast EMS call arrival rates, but ARIMA and HW methods have received much more study in the closely related problem area for forecasting call centre demand. ARIMA models, originally described by Box and Jenkins (1970), provide a class of models to approximate a time series using a large class of autocorrelation functions after allowing the time series to be stationarised through transformations such as differencing and logging. These models account for temporal dependencies using autoregressive (AR) terms, which are lagged observations of the dependent variable and moving average (MA) terms, which are lagged error terms, as explanatory variables. Practitioners have long recognised the potential of ARIMA models to model time series data with strong seasonal patterns. For example, Tomasek (1972) used the methodology to forecast aggregate telephone installations and disconnects, Nijdam (1990) demonstrated the capability of the models to account for seasonal trends in call volumes, Bianci et al. (1993) illustrated their effectiveness to model calls to AT&T call centres and Andrews and Cunningham (1995) demonstrated their ability to be applied to time series to ultimately develop efficient schedules for telephone agents two weeks in advance. The ability of other autoregressive and ARIMA models to specifically predict ambulance demand levels has been recently analysed by Channouf et al. (2007) who developed and compared time series models to generate daily and hourly forecasts of EMS calls in Calgary, Alberta. They concluded that autoregressive models were able to forecast a few days into the future with a higher degree of accuracy than ARIMA.

HW models offer an alternative methodology to generate forecasts with both trend and seasonal variations through predicting future demand using a set of simple recursions that rely on a weighted average of historical data values, with the more recent values carrying more weight. When selecting an appropriate model to predict future estimations, both ARIMA and HW are commonly applied to the same test time series, to determine the most appropriate to handle the specific data in question. Both methods have been applied to specifically estimate call volumes to the Cleveland

City Police Department in Holcomb and Sharpe (2007).

The time series approaches discussed above are all parametric in nature and require restrictive distributional and structural assumptions, such as stationarity of the data. The empirical studies performed in Taylor (2008) additionally claim that there is no clear 'winner' amongst the univariate methods to forecast call volumes because they perform differently under various lead times and different workloads. Whilst these traditional methods prove useful for upper-level capacity planning and budgeting, recent advances in location analysis allowing ambulance deployment strategies to become more flexible and dynamic in nature, call for more responsive predictions of demand and model-free methods to predict call volumes (Brotcorne et al., 2003; Smith et al., 2009). Some more recent methods developed to produce forecasts have proven to be successful, such as those based on the Singular Value Decomposition (SVD) (explained in Chapter 4.2). A statistical model to forecast call volumes based on the use of SVD, which additionally allowed intraday forecast updating was proposed in Shen and Huang (2008). With minor technical modifications, with particular reference to forecasting, this SVD approach is essentially SSA (discussed in Section 3.5); and the case studies included within the paper that test the approach against standard industry practice and a multiplicative effects model yield very promising results. Artificial Neural Networks (ANNs) may alternatively be used to forecast demand volumes for specific areas at various times of the day by first segregating a large county or region into smaller areas (commonly through fitting a data grid over the given region, and obtaining summary statistics for a number of characteristics to use as input data for each of the individual grid squares) and then using a training algorithm to calibrate the model and discover the underlying relationships. However, whilst they do not require assumptions about the error inherent in the data and are not affected by multicollinearity, they still have many limitations since they cannot be easily generalised to new data sets, disregard the temporal nature of the data and require considerable computational effort to calibrate. In order to perform well, ANNs are further dependent upon the availability of a large diversity of data for training purposes.

In conjunction with the investigation of the conventional time series methods to predict future demand levels, this thesis analyses the capability of the model-free technique of SSA to account for both trend and seasonality patterns exhibited in the data; and to subsequently estimate future demand levels, weighting the importance of over- and under-estimating demand accordingly. SSA generates forecasts using a SVD (observed

to generate high quality forecasts by Shen and Huang (2008)) and the problems inherent with the traditional methods are not present in SSA as it is able to expose important characteristics of the time series without requiring either a parametric model, or assumptions concerning the signal or noise (Golyandina et al., 2001).

3.5 SSA

SSA has been shown to be a powerful and effective technique for nonparametric time series analysis and forecasting in many diverse areas. It is essentially a model-free technique that combines elements of classical time series analysis, multivariate statistics, multivariate geometry, dynamical systems and signal processing (Golyandina et al., 2001). From its origins associated with the papers of Broomhead and King (1986) and Broomhead et al. (1987), SSA has been applied to many practical problems ranging from physics and meteorology to economics. Within the physical sciences, SSA has already become a standard tool in the analysis of climatic, meteorological and geophysical time series; see for example, Yiou et al. (1996), Ghil et al. (2001) (climatology), Weare and Nasstrom (1982) (meteorology) and Colebrook (1978) (marine science). SSA has also recently been used to model of Cholera outbreaks in accordance with the El Niño cycle (Rodo et al., 2002). In the socio-economic sciences, SSA has been used to predict daily exchange rates and the volatility of the financial market (Thomakos et al., 2002; Hassani et al., 2010).

The Basic SSA method comprises of two operations known as 'decomposition' and 'reconstruction', which are fully explained in Section 4.2. In the decomposition phase the original series is decomposed into a number of individual time series (each representing either a slowly varying trend, an oscillatory component or 'structureless' noise) and in the reconstruction phase a certain number of these time series are selected to reconstruct the main series so it may be used for forecasting. The technique is based on the SVD of a specific matrix constructed upon the time series. There are two parameters to choose in the basic version: the window length L and the number of individual time series selected in the reconstruction stage. Guidelines on the optimal values to assign to these parameters are contained in Golyandina et al. (2001), but the choice is somewhat dependent on the task SSA is being used for. In addition to forecasting, Basic SSA can also be used to find trends of different resolution, extract seasonal components, fill in missing values, find structure in short time series, for change-point detection and for smoothing data. Technical details concerning SSA may

be found in Chapter 4.

The basic method has several extensions. 'Caterpillar-SSA' is similar to the Basic SSA method in terms of the algorithmic details, but differs in the assumptions necessary to apply SSA. In the Basic model, the time series should be interpretable in the form 'signal plus noise'. In Caterpillar-SSA, the main emphasis in the methodology is placed on the separability of one series from another and neither a parametric model nor stationarity-type conditions need to be assumed for the time series; allowing the technique to be employed in a large range of scenarios: the singular characteristic necessary for the time series to possess is a potential structure (Zhigljavsky, 2010). For further information on the developments of Caterpillar-SSA and the software, see www.gistatgroup.com.

Other recent developments in SSA focus on aspects of classical time series analysis, classical signal processing and classical statistics. A method known as Toeplitz SSA has been developed which slightly outperforms the Basic SSA if the original time series is stationary (Golyandina, 2010); whereas Monte-Carlo SSA may be used to test the hypothesis of the presence of a signal based on simulations in situations where the noise is not 'white' (i.e. a random signal with a sequence of uncorrelated random variables with zero mean and finite variance) but 'coloured'. If the signal is weak then it can only be detected, but not extracted (Allen and Smith, 1996). Cadzow iterations which are repeats of the Basic SSA, have also been investigated but Gillard (2010) found no perceived advantage in running several iterations over the Basic SSA, although there were subtle differences in the forecasts produced.

The multivariate version of SSA (MSSA) is particularly useful for analysing several series with similar structures as it allows simultaneous extraction of the time series and may additionally be used to improve forecasts through considering the influence of causality between time series (Patterson et al., 2011; Hassani et al., 2009). Also, the same ideas used in MSSA are utilised for change-point detection in time series, outlined by Golyandina et al. (2001).

SSA has been shown to have superior performance when compared to traditional time series modelling methods; see for example Hassani et al. (2009) and Hassani et al. (2010). Whilst many successful applications have been made of SSA, to the best of the researcher's knowledge this study represents the first time the technique has been

applied to emergency demand.

3.6 Summary

The reviews contained throughout this chapter identify issues related to the first main research focus of this thesis: to produce accurate forecasts of demand, and highlight the importance of inputting reliable demand predictions to resource planning models. When decision models assume a longer term perspective, hourly and daily fluctuations in demand are generally overlooked; yet staffing models that aim to optimise resources on an hourly basis require more precise estimations of demand, and this chapter has summarised several advanced models developed by operational researchers to achieve this goal.

The current 'average peak demand' methodology used to predict ambulance demand by WAST is based on the concept that there must be a sufficient number of emergency response vehicles to cope with peak demand levels at all times, but it is calculated in an elementary fashion. More formal time series approaches have been shown to be capable of accounting for possible differences from week to week and allow consideration of the expected demands in the hours immediately preceeding/following the hour in question.

This review has demonstrated that researchers have used both regression models to explain the spatial variation of demand, and time series models to account for variations over time. With regards to modelling time variations in EMS demand, ARIMA and HW methods have been presented as two general approaches employed to capture the fluctuations. However, most conventional time series techniques are based on Gaussian type models of stationary series, which may provide highly inaccurate estimates when the call volume is low (typical of EMS calls at the hourly level). The model-free technique of SSA has been proposed to overcome the restrictive distributional and structural assumptions of such models.

In light of the review which demonstrates the successful application of SSA in several areas, this thesis progresses to investigate the potential of the technique to improve the quality of WAST forecasts in Chapter 4. The potential of several traditional time series models to generate forecasts is also considered, providing benchmark statistics that allow for the evaluation of SSA performance.

Chapter 4

Demand forecasting

4.1 Introductory remarks

Whilst Chapter 3 demonstrates that intensive research has been conducted in the field of demand forecasting; relatively little work has been initiated in incorporating these forecasts in vehicle deployment and staffing models (Gans et al., 2003), as these models often assume that demand is known as a precursor (sometimes based on coarse ad hoc estimates (Matteson et al., 2011)). Yet for these deployment schemes to be effective, it is essential that the values in the demand forecasts are accurate (Setzler et al., 2009). Use of inaccurate parameter estimates in these models can result in poor allocation of resources, and within emergency response settings, this can lead to unacceptable response times to emergency calls. This chapter aims to generate improved forecasts of call volumes by considering the potential of a relatively new nonparametric technique for time series analysis known as SSA (as outlined in Chapter 3), along with two conventional time series models (namely ARIMA and HW) to predict the daily demand for ambulances in Wales. Through exploring the prediction power of a relatively new nonparametric technique, this research further responds to the call by Fildes et al. (2008) to improve the forecast quality by considering techniques with a shift away from traditional statistical analysis. Furthermore, the SSA technique allows the count-valued arrivals per hour to be directly modelled, and thus avoid dependence upon the artificial assumption of normality.

This chapter is structured as follows. Section 4.2 is devoted to the SSA technique: it begins with a description of the theory, and outlines the forecasting methodology which is based on linear recurrent formulae (LRF). The formulations of the conventional time series methods are given in Section 4.3, followed by a summary of the results gener-

ated from the case study investigating the potential of all three methods to predict the total number of unique incidents requiring WAST assistance each day. The model performances are evaluated by inspecting the root mean square errors (RMSEs) associated with the predictions, and SSA is motivated as a suitable tool for WAST demand prediction. The chapter ends with a short summary in Section 4.4.

4.2 SSA Theory

SSA decomposes a time series into a sum of time series. Each component within this sum might be a trend component, periodic component, quasi-periodic component or noise. The main stages of SSA are as follows:

$$\text{Stage 1: Decomposition} \begin{cases} \text{Step 1: Embedding} \\ \text{Step 2: Singular value decomposition (SVD)} \end{cases}$$

$$\text{Stage 2: Reconstruction} \begin{cases} \text{Step 1: Grouping} \\ \text{Step 2: Diagonal averaging} \end{cases}$$

This Section will outline these stages for a real-valued nonzero time series with N observations $Y_N = (y_0, \ldots, y_{N-1})$.

4.2.1 Decomposition and reconstruction

4.2.1.1 Decomposition: Embedding

The first step of SSA is to map the given time series Y_N to a multidimensional series $X_1, \ldots, X_K$. Here $X_i = (y_{i-1}, \ldots, y_{i+L-2})^T$ for $i = 1, \ldots, K$ where $K = N - L + 1$. The parameter L is known as the window length and is an integer such that $2 \leq L \leq N$. Usually L is selected so that it is proportional to the periodicity within Y_N and lies between $\frac{N}{3}$ and $\frac{N}{2}$. Some advice is given to the choice of L in Golyandina et al. (2001) and Hassani (2007).

The trajectory matrix X is formed:

$$X = [X_1, \ldots, X_K] = \|y_{i+j-2}\|_{i,j=1}^{L,K}$$

X is a Hankel matrix as all elements along the anti-diagonals are identical.

4.2.1.2 Decomposition: Singular value decomposition (SVD)

Let $\lambda_1, \ldots, \lambda_L$ denote the eigenvalues of XX^T (ordered by magnitude such that $\lambda_1 \geq \ldots \geq \lambda_L \geq 0$) and $U_1, \ldots, U_L$ denote the orthogonal system of the eigenvectors of the matrix XX^T corresponding to $\lambda_1, \ldots, \lambda_L$.

If we denote $V_i = \frac{X_i^T U_i}{\sqrt{\lambda_i}}$ for $i = 1, \ldots, d$ then the SVD of the trajectory matrix can be written as

$$X = X_1 + \ldots + X_d \tag{4.1}$$

where $d = rank(X) = max(i : \lambda_i > 0)$ and $X_i = \sqrt{\lambda_i} U_i V_i^T$. The matrices $\{X_i, i = 1, \ldots, d\}$ have rank 1. The collection $(\sqrt{\lambda_i}, U_i, V_i)$ is called the i-th eigentriple of the matrix X.

4.2.1.3 Reconstruction: Grouping

Carefully selecting sets of the matrices within $\{X_i,\ i = 1, \ldots, d\}$ will give various trend or periodic components of Y_N. The grouping procedure partitions the set of indices $\{1, \ldots, d\}$ (obtained in expansion (4.1)) into b disjoint subsets $I_1, \ldots, I_b$.

Let $I = \{i_1, \ldots, i_p\}$. The resultant matrix X_I is defined as X_I as $X_I = X_{i_1} + \ldots + X_{i_p}$. This is computed for $I = I_1, \ldots, I_b$ and leads to the decomposition $X = X_{I_1} + \ldots + X_{I_b}$. For example, let $d = 10$ and $b = 2$. Then the set of indices is $\{1, \ldots, 10\}$. Let $I_1 = \{1\}$ and $I_2 = \{3, 4\}$. Then $X = X_{I_1} + X_{I_2}$ where $X_{I_1} = X_1$ and $X_{I_2} = X_3 + X_4$.

Auxiliary information may help the researcher to select particular components. For example, if it is known that there is a monthly periodicity within our time series, one may wish to identify the component(s) that reflect this. A plot of the singular values identifies the number of components to be taken (in a similar manner to principal component analysis, see (Jolliffe, 2008)). Explicit plateaux in the singular value spectra indicates pairs of components that are likely to be important. Pairwise scatter plots of components allow the visual identification of the components corresponding to harmonic elements of Y_N. Analysis of the periodograms from the original series, and of its components, will inform of the frequencies that need to be considered to reconstruct the time series. The "art" of SSA is in the selection of the subsets $I_1, \ldots, I_b$, and further details are provided in Golyandina et al. (2001). As more and more indices from $\{1, \ldots, d\}$ are selected, then more of the original signal is reconstructed. If too few indices are selected, then the reconstructed signal might not adequately explain

the variation in Y_N (this might be sufficient to describe the overall trend of the series, however). If Y_N is a noisy time series, then taking too many indices would result in the noise forming part of the reconstructed signal, or we would 'over-fit' the original series.

4.2.1.4 Reconstruction: Diagonal Averaging

Selecting $I_1, \ldots, I_b$ and computing $X = X_{I_1} + \ldots + X_{I_b}$ results in a matrix that is not of Hankel structure. In order to find the approximated time series, X must be transformed into a Hankel matrix. This may be done via diagonal averaging, described as follows.

If z_{ij} is an element within a matrix Z, the k-th term of the resulting time series is obtained by averaging z_{ij} over all i, j such that $i + j = k + 2$. This diagonal averaging operates on an $L \times K$ matrix Z ($L \leq K$) in the following way. For $i + j = s$ and $N = L + K - 1$ the element $\widetilde{z_{ij}}$ as a result of the diagonal averaging of Z is given by:

$$\widetilde{z_{ij}} = \begin{cases} \frac{1}{s-1} \sum_{l=1}^{s-1} z_{l,s-l} & 2 \leq s \leq L, \\ \frac{1}{L} \sum_{l=1}^{L} z_{l,s-l} & L+1 \leq s \leq K+1, \\ \frac{1}{K+L-s+1} \sum_{l=s-K}^{L} z_{l,s-l} & K+2 \leq s \leq K+L. \end{cases}$$

4.2.2 Forecasting

SSA uses LRF in order to forecast future time series points. LRFs are extremely flexible; if a series is representable by a LRF then it may also be represented as a product of exponentials, polynomials and harmonics (and vice versa). Technical details are provided in Golyandina et al. (2001). Y_N satisfies a LRF (of order q) if

$$y_{i+q} = \sum_{k=1}^{q} a_k y_{i+q-k} \qquad 1 \leq i \leq N - q + 1$$

The eigenvectors of XX^T provided in the SVD step yield the coefficients $a_1, \ldots, a_q$.

Confidence intervals for such forecasts can be obtained by bootstrapping (for further information see Efron and Tibshirani, 1993).

4.3 Model comparison

In this Section, the ability of the SSA technique to forecast daily demand levels at an all-Wales level is evaluated and compared with the well-established ARIMA and HW forecasting methods based on the precision of the model fits as reflected by the RMSE. All the models are formulated using the first 51 months of data (1st April 2005 - 30th June 2009) and the forecasting error is measured using the data from the following six months. Figures 2.7 and 2.8 illustrated that July and December are volatile months, and are thus expected to be harder to forecast; for this reason the errors associated with these distinct two months are also individually investigated and reported. The results show that whilst all methods produce reasonably accurate results for certain time periods, SSA is superior for the longer forecasting horizons. Section 4.3.1 proceeds to define the RMSE which is used to evaluate the goodness of fit of the models, Section 4.3.2 provides the formulation of the particular SSA model developed to predict ambulance demand in Wales, and Section 4.3.3.2 briefly outlines the well-known forecasting algorithms of the conventional models, which are used to benchmark the forecasting accuracies.

4.3.1 Measures of accuracy: RMSE

As a measure of prediction accuracy and to compare the goodness of fit of the models, the RMSE is calculated for various models and for different forecast lags, defined as:

$$\text{RMSE} = \sqrt{\frac{\sum_{n=0}^{N-1}(y_n - e_n)^2}{N}}$$

where y_n is the observed value, e_n is the estimated value and there are N fitted points in the time series. The RMSE is a commonly used forecast-accuracy metric in time series analysis to report how close forecasts or predictions are to the known data (Channouf et al., 2007; Matteson et al., 2011). Similar conclusions may be drawn if the results are evaluated using the Mean Absolute Error (MAE) or Mean Absolute Percentage Error (MAPE), but the RMSE is reported for this investigation as in addition to overcoming the common problem encountered with the MAPE (that the percentage error may become inflated if the actual value y_n in the denominator is relatively small compared to the forecast error); this performance measure also gives relatively higher weight to large errors, which are particularly undesirable for ambulance services.

This evaluation criteria will now be applied to benchmark the ability of the SSA, ARIMA and HW methods to forecast to the all-Wales daily demand levels for WAST assistance using the data described in Section 2.4.

4.3.2 SSA model formulation

As discussed in Section 4.2, the choice of the number of components to retain in the methodological process requires careful consideration. As a series of pure noise generally produces a slowly decreasing sequence of singular values, further guidance may be obtained by checking for breaks in the plot of logarithms of the eigenvalues. Explicit plateaux in the plot correspond to the components representative of clear periodicities within the data (because harmonic components with different frequencies produce two eigentriples with close singular values) and elbows in the chart demonstrate points at which retaining a larger number of components in the data explains little extra variation within the original series. When the methodology is applied WAST daily demand data (truncated at various points of the post-sample period for investigation purposes), an elbow in the chart is commonly seen to correspond to the 14th eigenvalue; hence it is the decision of the researcher to construct the SSA model using the first 14 components and a window length L = 581 (between $\frac{1}{3}$ and $\frac{1}{2}$ of the series). All of the results and figures in the following application are obtained by means of Caterpillar-SSA 3.30 software, available from www.gistatgroup.com.

Figure 4.1 displays the eigentriples of the first two components generated from the SVD of the trajectory matrix, with the proportion of the variation accounted for by each component printed in brackets after the component number. The majority of the variation (99.6%) is captured in the trend element accounted for by the first component, whilst the second plot reveals a periodic component in the data.

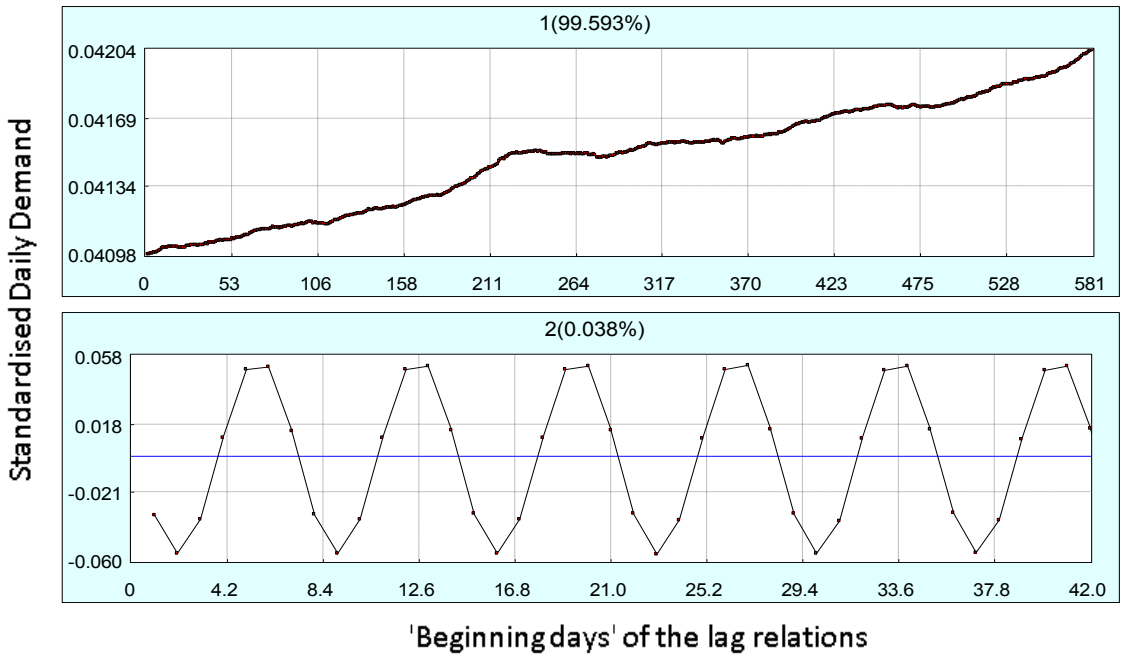

Figure 4.1: Principal components related to the first 2 eigentriples. [The second graph depicts up to the 42nd lagged vector only, to allow the 7 days periodicity to be clearly visible].

Analysis of the pairwise scatter plots and their corresponding periodograms allows visual identification of the harmonic components. Two components that form clearly defined polygons when the points of intersection of their eigenvectors are plotted against each other in scatter plots are indicative harmonic components, since their eigenvectors will intersect at regular intervals; and the periodicity accounted for by the component is represented by the number of sides of the polygon. Equivalently, periodograms of the eigenvectors with sharp peaks around some period may be used to identify harmonic components. Figure 4.2 depicts the scatter plot and periodogram corresponding to the second and third components. The 7-sided polygon and spike at $x = 7$ in the periodogram demonstrate that these components account for the 7 day (weekly) periodic effect. Other periodicities are accounted for by the remaining components. By means of retaining all significant components in the series reconstruction, one can account for the main periodicities in the data and build an accurate representation of demand.

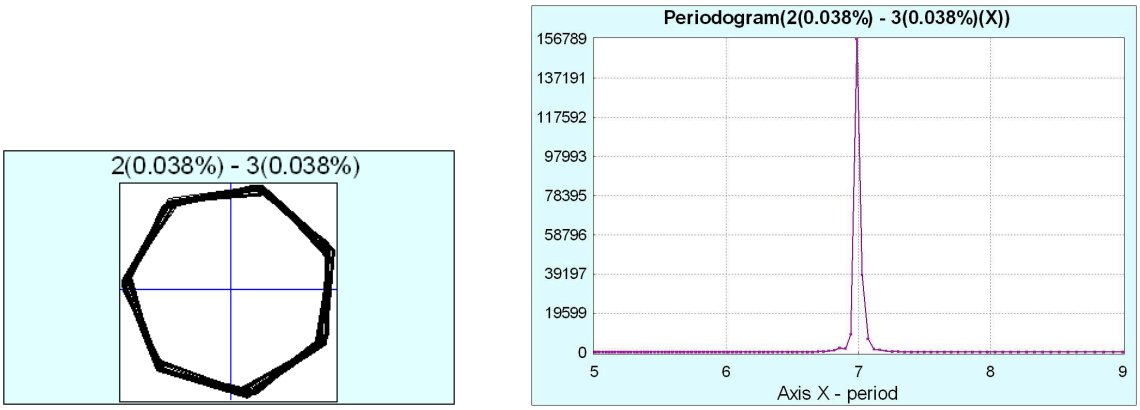

Figure 4.2: Scatter plot and periodogram of paired eigenvectors (2-3)

4.3.3 Conventional models

This section compares the SSA technique with two well-known methods, namely the traditional Box-Jenkins ARIMA model and the Seasonal HW algorithm. Hassani (2007) applied these methods to forecast six future data points of a well-known time series data set reporting monthly accidental deaths in the USA, and discovered that SSA provided a more accurate forecast than the conventional methods. These methods are described below and their forecasting power to predict WAST demand is compared against the SSA method in Section 4.3.4.

4.3.3.1 ARIMA model

ARIMA models, originally described by Box and Jenkins (1970), provide a class of models to approximate a time series after allowing the time series to be stationarised using transformations such as differencing and logging. These models account for temporal dependencies using autoregressive (AR) terms, which are lagged observations of the dependent variable and moving average (MA) terms, which are lagged error terms, as explanatory variables. Selection of the appropriate AR, MA and differencing terms to include in the model is usually considered subjective, but it does not have to be (Hyndman and Khandakar, 2008) as many attempts have been made to automate the ARIMA process. To select the optimal parameters, we use the 'forecast' package in R as described in Hyndman and Khandakar (2008). For all models built to forecast demand starting on the 1st day of the month between July-December 2009, the optimal model is ARIMA(1,0,1)(1,0,1).

4.3.3.2 HW forecasting method

Exponential smoothing is a simple, but one of the most widely used, techniques for adaptive time series forecasting (Gardner, 1985). The model generates forecasts using a set of simple recursions and relies on a weighted average of historical data values, with the more recent values carrying more weight. Given the seasonality in the data, this research considers the HW extension of the basic exponential smoothing model which includes additional terms to account for the linear trend and seasonality exhibited in the data (see Chatfield, 2001; Brockwell and Davis, 2002). For the data in this case study, one may observe that the HW additive model predicts the historic data more accurately, and thus this version of the model is selected to forecast forward. The optimal model parameters are selected using the time series forecasting system within SAS, which includes a completely automatic forecasting model selection feature that selects the best fitting model for a time series and reports diagnostic check results. When forecasting for each month post June 2009, the optimal parameter values selected for the model vary slightly.

4.3.4 Results

In this Section, the models outlined in Sections 4.3.2 - 4.3.3 are evaluated in terms of their quality of fit and their forecasting performances are compared using the RMSE. For identification and estimation of the models, the first 1,552 daily counts (from 1st April 2005 - 30th June 2009) are entered as input data to the models, and the last 6 months of daily counts from July - December 2009 are held back to measure the forecasting performance of the models. The accuracy of the forecasts are evaluated by:

i. Performing a series of 1-through-28 day step-ahead forecasts beginning with the first 'unknown' observation on 1st July 2009;

ii. Updating the within-sample period by one observation and again performing a series of 1-through-28 day step-ahead forecasts;

iii. Repeating step 2 until less than 28 post-sample observations are available for evaluation purposes;

iv. Combining all forecasts to calculate the average RMSE for each forecasting method.

To allow a comprehensive evaluation of the model capabilities, the average RMSE is calculated individually for 7-day, 14-day, 21-day and 28-day forecasting horizons to provide a number of comparisons between the time series methods. It is desirable for EMS planners to have an estimate of forecasts one month in advance so rosters may be finalised.

The following results correspond to the performance of the models when estimating daily counts on a rolling basis over the entire post-sample six month period, starting with the first 'unknown' observation on 1st July 2009 and ending with the first 'unknown' observation on 4th December 2009 (to retain a 28-day post-sample period of known data for comparison purposes). The ARIMA and HW models are permitted to re-parameterise at the start of each new month to ensure that the parameters selected in the models are optimised for every forecast horizon, but whilst the HW parameters are discovered to vary slightly from month to month (using SAS software), the optimal ARIMA model (found using the 'forecast' package in R as described in Hyndman and Khandakar (2008)) remains consistently ARIMA(1,0,1)(1,0,1). For a fair comparison, 14 components are consistently used in the SSA model as the plot of logarithms of the eigenvalues suggests that this is a reasonable number of components to retain for all months. However, SSA does also provide flexibility for different components to be selected; hence the SSA model may be fine-tuned if necessary.

Table 4.1 reports the RMSE for each forecasting horizon, averaged over the 184 model runs between July and December. The first line of the table reports the retrospective error, representing the closeness of fit of the model predictions with the initial true data used for the model construction (the period from 1st April 2005 - 30th June 2009). One may observe that the predicted values are very close to the data for all models considered.

Table 4.1: Comparison of model forecasts for daily demand (July - December 2009). Standard deviations are included in brackets.

Average RMSE	SSA	ARIMA	HW
Retrospective	6.19 (35.32)	6.11 (35.26)	6.37 (36.34)
7-day forecast	42.20 (12.92)	41.55 (13.69)	45.46 (15.79)
14-day forecast	42.86 (8.71)	44.06 (9.18)	47.47 (13.85)
21-day forecast	43.87 (7.14)	46.16 (9.50)	48.32 (12.74)
28-day forecast	45.46 (8.66)	48.75 (13.86)	51.14 (14.11)

The second part of Table 4.1 summarises the model forecasting performances. Values generated using the SSA technique generally follow the data more accurately than those predicted by the standard models, especially for the longer-term forecasts. The standard deviations show that the forecasts are additionally of consistent high-quality across all model runs.

Tables 4.2(a) and 4.2(b) display the segregated results when the rolling forecasts are individually computed for the first month (July) and last month (December). Recall from Figure 2.8 that these months have the highest demand levels and are thus the most volatile months to forecast. Whilst the models were updated and re-run 31 times throughout July, lower numbers of runs were possible for December (e.g. the 28-day forecast could only be updated and re-run on 4 occasions as the true demand is only known until 31st December 2009).

Table 4.2: Comparison of model forecasts for daily demand (July & December 2009). Standard deviations are included in brackets.

(a) July 2009

Average RMSE	SSA	ARIMA	HW
7-day forecast	44.77 (11.03)	44.69 (13.20)	60.12 (15.87)
14-day forecast	44.25 (4.80)	48.96 (7.84)	63.52 (13.83)
21-day forecast	45.04 (3.31)	50.75 (4.63)	60.87 (12.20)
28-day forecast	45.76 (3.02)	50.74 (3.86)	62.50 (10.73)

(b) December 2009

Average RMSE	SSA	ARIMA	HW
7-day forecast	70.38 (35.52)	69.63 (38.26)	52.58 (29.80)
14-day forecast	70.96 (25.22)	86.32 (33.01)	63.20 (27.43)
21-day forecast	73.87 (10.87)	97.24 (13.00)	71.40 (6.98)
28-day forecast	80.85 (0.72)	105.47 (1.25)	90.19 (4.62)

Both tables show that SSA often produces improved forecasts, especially for the longer forecasting horizons, but remains comparable at the least to other well-established methods for shorter periods. Further investigations have found that by selecting 200 components for December, far superior forecasts can be generated (this is an example of how SSA may be modified to produce even better forecasts for precise months). Yet the researcher has chosen to display the results for the simpler 14-component model in this thesis, as although the selection of a higher number of components can prove useful for forecasting the more volatile months, fewer components provide higher quality results over the greater part of the year.

An illustration of the 1-month SSA forecast beginning on 1st July is given in Figure 4.3. All 1,552 daily counts up to 30th June 2009 are used in the estimation of the SSA model shown, although only the within-sample data for the month of June 2009 is displayed in the chart (to aid with clarity). In addition to the RMSE, visual inspection of Figure 4.3 shows that the HW method captures some element of the periodic nature of demand, but not the full variation of peaks and troughs throughout July. In contrast the SSA and ARIMA forecasts follow the true demand values reasonably closely, but of the two methods, the SSA forecast maintains the lowest RMSE when the rolling forecast is computed until December, as shown in Table 4.1. 95% confidence intervals for the SSA forecast are shown in Figure 4.4.

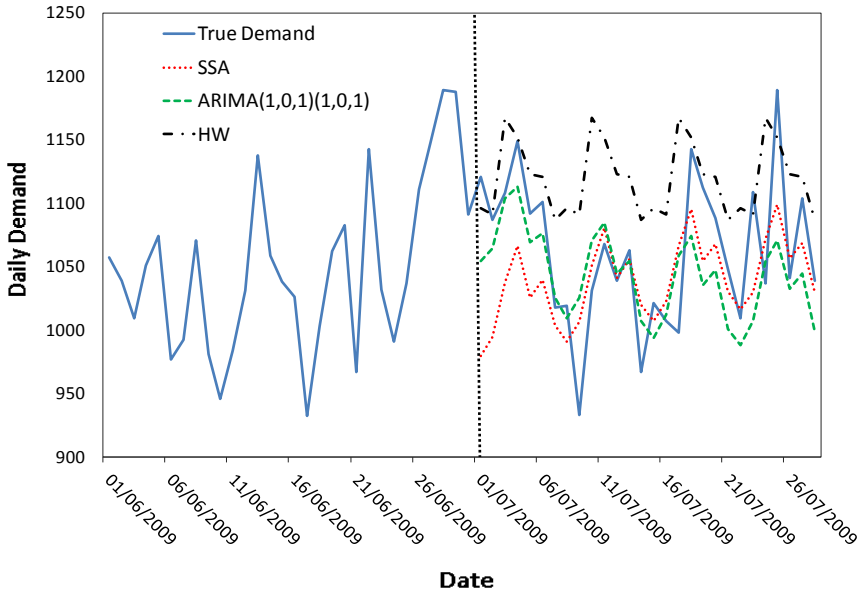

Figure 4.3: 28-day forecasts beginning on 1st July 2009

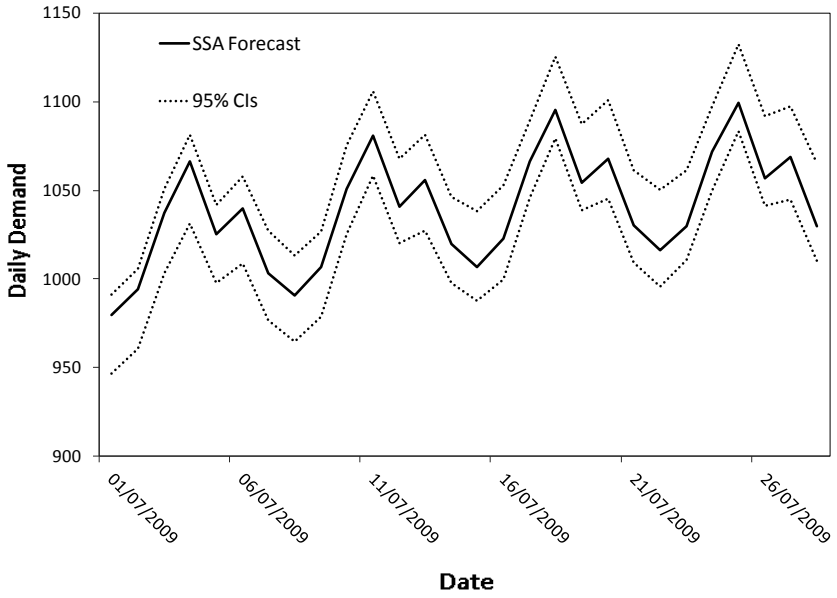

Figure 4.4: Confidence intervals for 28-day SSA forecast by the bootstrap method (1,000 repetitions) beginning on 1st July 2009

For practical purposes, forecasts would additionally be updated as new demand levels are obtained and inputted in to the system, to finalise rostering and scheduling

plans. Figure 4.5 illustrates the 7-step-ahead daily forecasts where the total demand levels for the current day are used to predict the demand levels for the same day one week ahead. In the same way, any "n-step" ahead forecast could be produced as required to allow WAST to update forecasts as and when required, leading to the researcher's decision to evaluate all 7-day, 14-day, 21-day and 28-day forecasts in the preceding tables. One may observe the immeasurable value of updating forecasts, as the predicted SSA values forecasted one-week in advance in Figure 4.5 follow the SSA trend far more closely than those in the 28-day ahead forecast in Figure 4.3. However, the clear benefit that the SSA model provides more accurate predictions for longer forecasting horizons is a major advantage of the technique, as it is a costly operation to change staff rosters at the last minute.

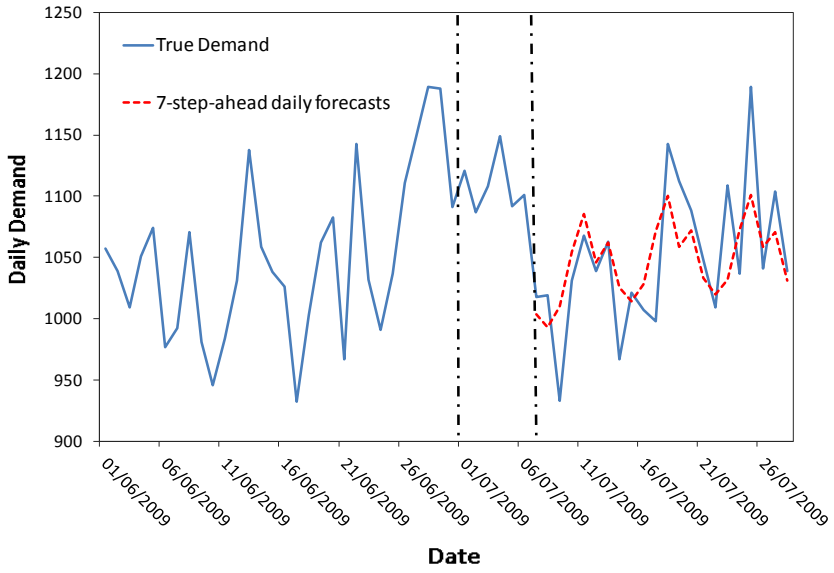

Figure 4.5: Original time series with 7-step-ahead daily SSA forecasts for July 2009, beginning with forecast when demand is known until 30th June

4.4 Summary

This chapter has outlined the theory of three time series methods whose capability to model and forecast demand has been evaluated in a case study investigating ambulance demand levels in Wales. Whilst two of the time series methods (ARIMA and HW) are well-established in the literature, the investigation of the nonparametric SSA method responds to the call by Fildes et al. (2008) to improve the accuracy

of forecasts through considering novel methods. Motivation for the utilisation of SSA as a tool to accurately predict Welsh ambulance demand has been provided, with empirical results demonstrating that it produces superior longer-term forecasts (which are especially helpful for EMS planning), and comparable shorter-term forecasts to well-established methods. The benefit of the SSA technique is however not only in its ability to forecast as it produces superior, or in the least, comparable forecasts to other methods; but in its capability to recognise periodicities in the data and be flexible in approach, with the advantage that it may be easily implemented.

As demand for EMS assistance is rising in Wales, it is becoming ever more critical to ensure accurate demand forecasts are input to WAST scheduling models as use of inaccurate parameter estimates in these models can result in poor resource allocation, leading to low performance. The workforce capacity planning and scheduling tool discussed in Chapter 10 illustrates how the SSA technique may be ultimately embedded into a spreadsheet model, and directly used to inform scheduling functions that are also integrated within it. Precise details regarding the optimisation of the SSA technique for such purposes, run times, the default values programmed within the model and details regarding the adaption of the technique to produce forecasts for two categories of patients requesting WAST assistance are included within Chapter 10. It is however worth noting here that the technique can produce forecasts fairly efficiently: for example it takes around 5 minutes to produce demand estimates at an hourly level for two categories of emergencies in the SE region of Wales for a 3 month period (based on 4 years worth of known historic demands) on a 3GHz machine with 2.96GB RAM. Through generating estimates of the total number of emergency calls per period that are highly accurate, straightforward to implement, and have the potential to simultaneously lower operating costs while improving response times, the tool illustrates how the advanced forecasts can be readily embedded into OR methodologies to determine the minimum number of ambulances to be deployed at any given time. The necessary SSA software to execute the program may further be readily obtained and provided to WAST employees to allow automatic revisions of the forecasts as new data becomes available, so the Trust can prepare more appropriately for future demand.

The idea of this chapter was to provide the reader with the methodology of the SSA technique and to demonstrate its prediction power. Following on from the motivation provided for SSA as a suitable tool to accurately predict demand; future chapters will consider the ability of the technique to predict WAST demand at regional levels, for

different shift periods and for different call priorities, in order to provide forecasts that can ultimately be input into the workforce capacity planning tool. Where SSA is used to generate different forecasts in various case studies in the remainder of this thesis, the precise data used has been clearly outlined in the methodological sections. When forecasts are required at finer levels of granularity, it appears to be more fitting to forecast on a shift, rather than hourly, basis in order to maintain an accurate SSA performance through allowing for a rostering algorithm to be more readily embedded in the forecasts, retaining a large degree of seasonality in the data and reducing the number of 'zero' counts in the time series. To estimate hourly counts, the shift forecasts may subsequently be apportioned to provide the expected number of requests for emergency assistance per hour, and full details of this process are provided in Chapter 10.

Chapter 5

Literature review (part 2): Queueing theory

5.1 Introductory remarks

Queuing theory techniques are utilised in Chapters 6 and 7 to analyse the EMS division of WAST, which is modelled as a time-dependent multi-server priority service system. Before the implementation and evaluation of such methods, this review intends to provide the reader with a comprehensive background of the literature related to the topic and is organised as follows. The chapter begins with some preliminaries (including the general assumptions and main terminology used in queueing theory), and then progresses to review several approximation and numerical methods that are employed to evaluate service quality in time-dependent systems in Sections 5.3.1 and 5.3.2. Section 5.4 summarises the major milestones presented in the research of priority queues, and a summary of the chapter is provided in Section 5.5.

5.2 Preliminaries

Queueing theory is the mathematical study of how systems distribute their resources to customers, who sequentially arrive at a service facility in order to obtain a service. Queues commonly form because resources are not always readily available, and queueing theory attempts to model the system performance and evaluate the service quality expected to be provided to customers under different scenarios. Much of the theory is devoted to the derivation of performance measures evaluating characteristics such as the throughput, probability of delay, number of queueing customers and the

expected waiting time of customers in the queue (Bhat, 2008). Where the goal is to strike a balance between service quality and economic considerations, queueing theory may be utilised to optimise resource allocation and recommend staffing levels, ensuring that queues do not build up excessively, whilst servers are active a reasonable proportion of time.

In the context of queueing theory, one may think of a service system as comprising of two elements: (i) the service facility itself, which may be staffed by a number of servers; and (ii) a queue for service (except in specific cases where it may be specified that queueing is not permitted). At each facility, customers arrive and queue for some activity. Such a situation is depicted in Figure 5.1:

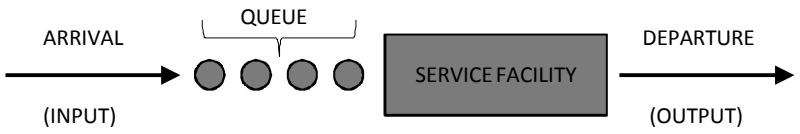

Figure 5.1: The fundamental diagram of queueing theory

Every queueing network is characterised by two major components: the arrival process and the service process. Queueing theory involves setting up mathematical models corresponding to Figure 5.1, analysing the system, and evaluating various performance measures. The system performance will depend on a number of characteristics and since these processes are usually stochastic by nature, queueing theory is based on probabilistic analysis. The main characteristics are outlined below, and further details are given in Gross and Harris (1998).

The **arrival process** defines how customers arrive at the service facility (e.g. singly or in groups) and how these arrivals are distributed in time. The Poisson process generates uniformly distributed arrival times with an arrival rate λ and is thus often selected to model random arrivals of customers. Throughout this thesis it is assumed that customers arrive at random, in a Poisson fashion, so the inter-arrival times are independently and exponentially distributed.

The **queue discipline** describes the order in which customers enter and leave the queue. This may be on a "first-in, first-out" (FIFO) basis, "first-in, last out" (FILO)

basis, "random-in, random-out" (RIRO) basis or in terms of priority. In the situation of a priority queueing system, it is also necessary to define whether preemption is allowed. **Preemptive priority** (PRP) means that a customer of higher priority is able to replace a customer of lower-priority already in service, whilst the terms **head-of-the-line** discipline or **non-preemptive priority** (NPRP) are used to represent the non-preemptive priority case where the service in currently in operation must first be completed.

The **service mechanism** outlines the resources needed for service to occur. The service time distribution defines how long the service will take, whilst other parameters such as the number of servers available, and whether the servers are in series (each server has a separate queue) or in parallel (one queue for all customers), must all be known before meaningful analysis of the system may be performed. In systems where the exponential distribution is assumed to provide an accurate representation of the distribution of service times, it's Markovian (memoryless) property allows one to map the system to a continuous-time Markov chain which can be solved analytically. Similarly to the Poisson distribution, the exponential distribution is defined by a single parameter. To distinguish between the mean arrival and mean service rate, it is common to denote the mean service rate by μ, so that $\frac{1}{\mu}$ represents the mean service time.

Additional notation commonly used in the literature to analyse queueing systems, and that shall be followed throughout this thesis, may be outlined as follows:

- p_n, $n = 0, 1, ...$: the probability that there are n customers in the system
- s: the number of servers on duty
- x: the maximal acceptable waiting time permitted for a customer
- W_q: the time that a customer waits in the queue before commencing service
- W_q^n: the time that a customer that arrives to find n people ahead in the system waits in the queue before commencing service

The quantity $\rho = \frac{\lambda}{s\mu}$ which may be referred to as **server utilisation rate**, **traffic intensity** or **load** per server is a common measure of interest that represents the behaviour of the queue over time, whilst $r = \frac{\rho}{s}$ quantifies the **offered load** to the

system i.e. the amount of traffic in the queue (Gross and Harris, 1998). Thus the relationship between these quantities and the system capacity gives meaningful insight as to the performance of the system, provided that all servers have the same service rate. Essentially, if $\rho \leq 1$ then the servers are able to process customers faster than the rate at which they arrive, on average, so the queue will not grow infinitely long. If the system runs with $\rho \leq 1$ for an adequate period of time with stable mean service and inter-arrival rates, then all system characteristics (such as number of customers in the system, the number in the queue and expected waiting times) will eventually settle down and the system will run at a consistent level, considered as 'stable' or 'stationary'. When the system reaches this point of time, it is said to be operating in a **steady-state** fashion. It is this steady-state behaviour which has been intensely researched and is well-understood in the literature, since closed-form formulas have been derived to evaluate performance measures under these stationary conditions. The analysis of systems with non-stationary arrival rates is however far more complex (Green et al., 2007) and the literature on this topic is overviewed in Sections 5.3.1 and 5.3.2.

In order to describe the various elements that form specific queueing systems, Kendall devised a notation to represent six components of a queueing system using six characters 1/2/3/4/5/6 (Gelenbe and Pujolle, 1998), where:

(1) The first character describes the arrival process:

- M represents exponential, independent and identically distributed (iid) inter-arrival times
- D represents deterministic inter-arrival times
- E_k represents Erlang (with parameter k) iid inter-arrival times
- GI represents general iid inter-arrival times

(2) The second character specifies the distribution of the service times (again commonly categorised as M, D, E_k or GI. Phase-type (PH) distributions may also be used to specify systems with inter-related Poisson processes occurring within phases)

(3) The third character lists the number of parallel servers, s

(4) The fourth character specifies the queue discipline e.g. FIFO, FILO, RIRO, ...

(5) The fifth character details the maximum allowable number of customers in the system

(6) The sixth character indicates the size of the population from which the customers are drawn

Kendall's notation is commonly simplified to list only the first three characters 1/2/3. In this format it is assumed that the queue discipline is FIFO, and no limits are imposed on the system capacity or the size of the population.

The $M/M/s$ model (see for example Ulukus (2011), and references therein) is one of the most widely researched models in the classic queueing literature since it is simultaneously capable of capturing randomness in arrival and service times, permits the number of servers to be greater than one, and possesses the appealing benefit of a tractable steady-state solution. It represents a system with a single queueing facility in which customers arrive at, and possibly queue, before being served by one of s identical servers. Arrivals occur according to a time-homogenous Poisson process with a constant rate, and the service time has an exponential distribution with a constant mean. Such a system may be modelled as a basic birth-death process as described below (Gross and Harris, 1998).

A birth-death process can be considered as a continuous time stochastic counting process $\{N(t), t \geq 0\}$. Letting $p_n(t) = \text{Prob}\{N(t) = n\}$ be the probability that the system is in state n at time t, the transition diagram of the birth-death process may be depicted as in Figure 5.2. When a birth occurs, the system goes to state n to $n+1$ and when a death occurs it conversely goes from state n to $n-1$. The process is specified by birth rates $\{\lambda_i\}_{i=0,..,\infty}$ and death rates $\{\mu_i\}_{i=1,..,\infty}$.

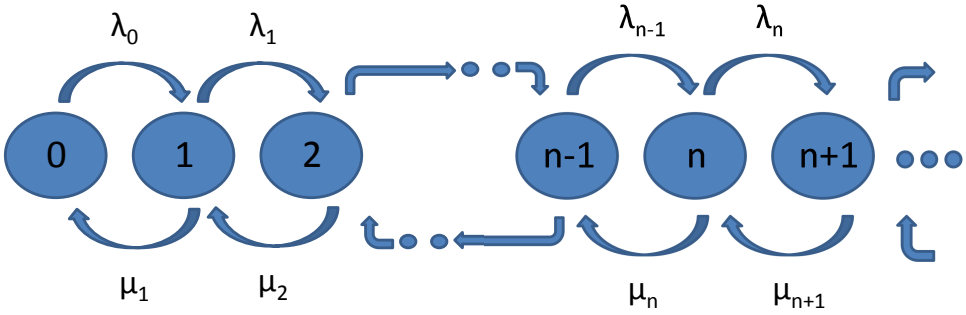

Figure 5.2: State transition diagram for a birth-death process

In queueing systems that directly model the behaviour of people, who arrive at a service facility requiring a specific service to be performed, the number of customers in the system is an appropriate state variable; thus $p_n(t)$ can be used to denote the probability that there are n customers in the system at time t. If λ_n does not depend on the number of customers in the system, then λ can be used to represent the mean arrival rate of customers. If μ represents the mean service rate provided each of the s identical servers at all points in time regardless of the number of customers in the system, then $\mu_n = s\mu$ for $n \geq s$ and $\mu_n = n\mu$ for $0 \leq n < s$. Under these conditions the state probabilities evolve according to the following differential-difference equations (see, for example, Gross and Harris (1998)):

$$\begin{aligned}
\tfrac{dp_o(t)}{dt} &= -\lambda p_0(t) + \mu p_1(t), \\
\tfrac{dp_n(t)}{dt} &= \lambda p_{n-1}(t) + (n+1)\mu p_{n+1}(t) - (\lambda + n\mu)p_n(t), \qquad 1 \leq n < s, \\
\tfrac{dp_n(t)}{dt} &= \lambda p_{n-1}(t) + s\mu p_{n+1}(t) - (\lambda + s\mu)p_n(t), \qquad n \geq s.
\end{aligned} \tag{5.1}$$

The equations presented in (5.1) are often referred to as the **balance** or **Chapman-Kolmogorov forward differential equations**. As the behaviour of the system settles to steady-state (as $t \to \infty$) then $p_0(t)$ and $p_n(t)$ are independent of time, so $\frac{dp_o(t)}{dt} = \frac{dp_n(t)}{dt} = 0$, giving:

$$\begin{aligned}
&-\lambda p_0 + \mu p_1 = 0, \\
&\lambda p_{n-1} + (n+1)\mu p_{n+1} - (\lambda + n\mu)p_n = 0, \qquad 1 \leq n < s, \\
&\lambda p_{n-1} + s\mu p_{n+1} - (\lambda + s\mu)p_n = 0, \qquad n \geq s.
\end{aligned} \tag{5.2}$$

The steady-state probabilities defining the mean number of customers in the system are

given by equation (5.3) (for a derivation of the summary measures, see Gross and Harris (1998)).

$$\begin{aligned} p_0 &= \left[\sum_{n=0}^{s-1}\frac{\lambda^n}{n!\mu^n}+\sum_{n=s}^{\infty}\frac{\lambda^n}{s^{n-s}s!\mu^n}\right]^{-1} \\ p_n &= \begin{cases} \frac{\lambda^n}{n!\mu^n}p_0, & \text{if } 1\leq n\leq s-1, \\ \frac{\lambda^n}{c^{n-c}c!\mu^n}p_0, & \text{if } n\geq s. \end{cases} \end{aligned} \tag{5.3}$$

The first rigorous consideration of queueing theory is attributed to Erlang. In the early 1900's, he modelled a telephone exchange as an $M/M/s$ system and determined the optimal size of the exchange to ensure as few calls as possible were not connected because the exchange was busy (Erlang, 1918). The papers written by Erlang over the following few decades contain many of the fundamental concepts and techniques used in modern queueing theory, including formulas to calculate performance measures (Izady, 2010). The characteristics of $M/M/s$ systems permit relatively simple derivation of a number of performance measures relating to the queue, such as the expected number of customers in the queue (L_q), the expected waiting time in the queue (W_q). While these measures both give insights into the degree of congestion that exists within a system, the distribution of the queueing time, and especially the probability of waiting greater than time x in the queue $P(W_q > x)$, is often of greater interest, although more difficult to obtain analytically (Utley and Worthington, 2011). Since $P(W_q > x)$ directly relates to Targets 1 and 2 (outlined in Chapter 1.4), this thesis devotes a great deal of attention to the calculation of this probability, and a detailed discussion surrounding the methodology developed to allow its evaluation in more complex time-dependent and priority queueing systems is presented in Chapters 6 and 7.

Numerous authors such as Hershey et al. (1981) and Artalejo and Lopez-Herrero (2001) have since progressed Erlang's analysis of steady-state systems through deriving additional measures, including the moments of the length of a busy period and expected utilization for constrained network facilities. In service systems governed by targets that specify minimum required standards, models can be set up to simulate the performance of the system under various staffing levels and find the minimum number of staff required to ensure the expected measures exceed the threshold levels. Yet

since the steady-state formulas are only capable of giving a single recommendation of an optimal staffing level (as they can only be applied to situations where the arrival of customers is strictly stationary), the earlier papers tend to place greater emphasis on system insights than the use of performance measures for this type of exploratory investigation.

5.3 Time-dependent queueing theory

Whilst much literature is devoted to the analysis of a service system with constant arrival and service times (Green and Kolesar, 1991); most actual systems today are subject to time-varying demand, where arrival rates and the number of servers vary throughout the period of operation. Call centres, banks, airports and EMS systems are just a few examples of systems subject to both predictable and stochastic sources of variation. Figure 5.3, for example, depicts a typical profile of patients requiring urgent ambulance transportation to hospital in Wales over a 24-hour period. Unlike emergency demand, urgent demand is not heavily dependent on weekday; thus the demand plotted in the chart has been averaged over all week days.

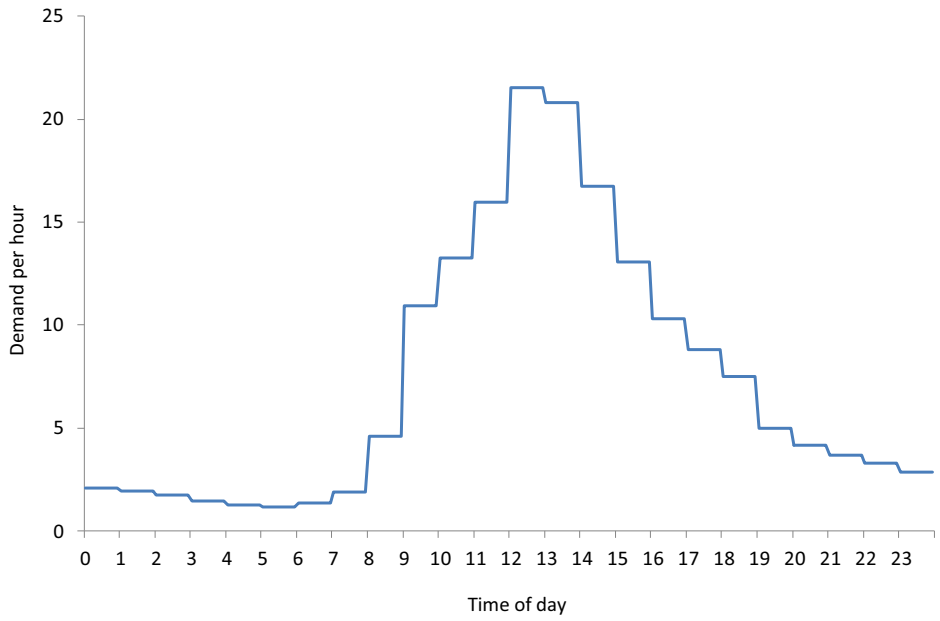

Figure 5.3: Urgent demand for ambulance transportation

Figure 5.3 highlights that the majority of requests for urgent ambulance transfers

occur during the hours of GP surgery operation, particularly between 10am-4pm, and substantially lower demand volumes are seen overnight. Given such demand levels, it is logical to vary the number of crews employed over time. These variations in the number of staff are intended to mirror the variations in the arrival process, so a consistent service quality can be provided throughout the course of the day. The investigation into the optimal staffing profile, referred to as the staffing problem, is a critical component of workforce scheduling of service systems subject to time-dependent demand.

If one allows the number of servers to change for certain shifts in response to the workload, then when a server who is scheduled to leave at the end of the shift is still busy, two situations are possible: the server may complete the service currently in operation and leave when it is complete, or they may leave the system instantaneously so the customer in service is re-routed to join the queue. The first case is called the **exhaustive** discipline and the latter is known as the **preemptive** or **non-exhaustive** discipline. (The term non-exhaustive is less common in the literature, but nevertheless used in this thesis to avoid confliction with the 'preemptive' discipline which additionally refers to the situation in priority queues where a customer of higher priority requesting service may replace a customer of lower priority currently in service (see Section 5.2)). For clarification purposes, Table 5.1 summarises the terminology used to represent the various disciplines in this thesis.

Table 5.1: Terminology used to describe queue disciplines

Term	Type of queue	Definition
Exhaustive	Time-dependent	A server scheduled to leave must first complete any service currently in operation.
Non-exhaustive	Time-dependent	A server scheduled to leave exits the system instantaneously, so the customer in service is re-routed to join the queue.
Head-of-the-line	Priority	A high priority customer can move ahead of all the low priority customers waiting in the queue, but cannot preempt a non-priority customer in service.
Preemptive	Priority	A high priority customer can move ahead of all the low priority customers waiting in the queue *and* can preempt a non-priority customer in service. The non-priority customer is re-routed to join the queue.

The motivation for the research in this thesis lies with the ambulance service where paramedics must remain with a patient until they have fully completed their duty, and life-threatening injuries are prioritised in the queue for response above those of less-serious natures, to be attended to by the next available responder. Hence the subsequent modelling processes considered in Chapters 6 and 7 shall focus simultaneously on the exhaustive and head-of-the-line queue disciplines.

Pollaczek (1934) noted that the steady-state theory developed by Erlang was inadequate for time-dependent systems, and began numerical investigations into the behaviour of the system during a finite interval. The study of time-dependent queues remains a vibrant area of research. The processes involved are far more complex and as a consequence more sophisticated mathematical procedures are necessary (Channouf et al., 2007; Holcomb and Sharpe, 2007; Feldman et al., 2008; Bekker and de Bruin, 2010; Caiado, 2010; Izady and Worthington, 2012). Analytical models for such situations are often intractable, but in addition to numerical approaches, approximation methods have been developed that provide reliable results in suitable scenarios (Green et al., 1991). Sections 5.3.1 and 5.3.2 outline several approximation and numerical methods that have been used in the literature to analyse such queueing systems. Whilst the approximation methods provide rapid solutions, the numerical approaches are able to offer solutions with a higher degree of accuracy at the expense of computation speed.

Before describing the time-dependent methods, it is important to note how the notation described for Kendall's model (see Section 5.2) may be extended to take into account this time-dependency and allow measures such as the arrival rate, service rate and number of servers, to vary over time. For the simplified 3-character 1/2/3 version of Kendall's notation:

(1) The first character describes the arrival process:

- $M(t)$ represents exponential time-dependent inter-arrival times
- $D(t)$ represents deterministic time-dependent inter-arrival times
- $E_k(t)$ represents Erlang (with parameter k) time-dependent inter-arrival times
- $GI(t)$ represents general time-dependent inter-arrival times

(2) The second characters specifies the distribution of the service times (again categorised as $M(t), D(t), E_k(t)$ or $GI(t)$)

(3) The third character gives the fluctuating number of parallel servers, $s(t)$

Hence the dependence on time is represented by the inclusion of '(t)' into the standard notation. For example, an arrival rate described as $M(t) = 2$ may be used to represent exponential time-dependent arrivals, with a rate equal to 2 at the given time t. As a more comprehensive example, $M(t)/M(t)/s(t)$ denotes a service system with time-dependent exponential arrival and service rates; and a fluctuating number of servers, $s(t)$.

5.3.1 Approximation methods

To estimate the behaviour of systems with non-stationary demand, several approximation methods have been proposed in the literature which use a series of tractable stationary models to estimate the time-dependent nature. The methods however only give reliable results under certain conditions as they do not consider non-stationary and transient effects, so will only be accurate if the rate of change of arrival rate relative to the throughput of the system is sufficiently slow to allow the system to quickly achieve the steady state associated with any arrival rate (Utley and Worthington, 2011).

Two methods which make use of compartmentalized steady state models to find the minimum number of servers required to meet a desired service target in each planning period are the stationary independent period-by-period approach (SIPP) and the pointwise stationary approximation (PSA). Whilst variants of these methods have been developed in the literature over the last four decades (Kolesar et al., 1975; Green and Kolesar, 1991; Green et al., 1991; Green and Kolesar, 1997; Green et al., 2001, 2006, 2007; Ingolfsson et al., 2007), an alternative method known has the modified-offered-load (MOL) has also been investigated (Massey and Whitt, 1994, 1997; Ingolfsson et al., 2007).

The SIPP approach works by segmenting the time period into distinct intervals, finding the average arrival rate in each period and inputting these average rates into a series of stationary $M/M/s$ finite-server system formulas (importantly assuming the system reaches steady-state within each period) to evaluate the performance measures and set staffing levels based on system performance. The PSA is effectively a limiting version of the SIPP approach when the period length approaches zero: the primary difference between the methodologies is that whilst the PSA is computed

by integrating over time (taking the expectation of) the formula for the stationary performance measure with the arrival rate that applies in each point in time, SIPP initially averages the arrival rate over the staffing period and thus restricts staffing changes to occur at the boundary of each interval (Green et al., 2007). It is important to remember that these methods however fail to account for the dependence between periods which occurs in practice, as the queue length at the start of each period is reliant on the number of customers remaining in the system at the end of the previous period.

Green and Kolesar (1991) comprehensively investigated the PSA approach and conclude that it provides a tight upper bound on key performance measures. They analysed its performance under varying conditions, with models based on variables spanning a broad spectrum of parameter values, including:

- Number of servers, s: varied from 1 to 12
- Service rate, μ: varied from 0.2 to 20
- Average traffic intensity, ρ: varied from 0.25 to 0.75
- Relative amplitude of the arrival rate, RA: varied from 0.1 to 1.0 ($RA = \frac{A}{\lambda}$, where A represents the amplitude of the demand curve, calculated as the average of the maximum and minimum displacement from the mean level)

The study revealed that the performance of PSA is primarily affected by the service rate, and it generally produces accurate results for systems with service rates exceeding 2 per hour, assuming a 24-hour cycle. Improved performance is also seen in systems with higher service rates and lower traffic intensities.

Green et al. (2001) similarly tested the accuracy of the SIPP methodology under various scenarios and found that the technique provides good approximations for systems comprised of short planning periods; and its predictions become more accurate as ρ and RA decrease. It is difficult to put a limit on the parameter values which should be considered as appropriate to enable PSA to perform well, since the interaction between all system components has a significant effect on the model's performance; but for example, if $RA = 0.1$ then SIPP provides accurate predictions for planning periods of 0.25 or 0.5 hours whenever $\mu \geq 4$; or equivalently whenever $\mu \geq 16$ for 1 hour planning periods. When put into context, these results are hardly surprising since although longer planning periods increase the likelihood of convergence to steady

state within each period, the individual periods tend to be subject to more variable arrival rates and higher relative amplitudes which can violate the assumption of independent periods. Furthermore if servers are not heavily utilised, then provided customers are served quickly, the system should rapidly converge to steady state so it is less likely that congestion will carry over from the previous period, as the time lag between peak arrival rate and peak congestion is smaller for higher service rates. In situations where this the arrival rate is decreasing and this lag is significant, SIPP is commonly found to understaff (Green et al., 2001).

In some cases, the accuracy of SIPP can be improved by adjusting the arrival rate function prior to its application. For a comprehensive summary of the effects arising from the application of a wide range of such transforms upon model performance, the reader is referred to Green et al. (2001). A popular revision to the arrival rate is known as Lag Avg (or 'Lag SIPP') (Green and Kolesar, 1991) which uses a modified arrival rate to account for customers who arrived in an earlier period, but receive service in a subsequent period. The technique works by first estimating the lag between peak arrivals and peak congestion, and subsequently taking the expected arrival rate in each period to be equal to the arrival rate for the period shifted backwards by the time lag. It is reliable and efficient when relative amplitude is low ($RA \leq 0.5$) and planning periods are short (0.25 or 0.5 hours).

Alternatively, SIPP Max is effective if there is a problem of understaffing. SIPP Max uses the maximum arrival rate over the entire planning period, rather than the average, and provides reliable results in situations where the service rate is high ($\mu \geq 8$) , the relative amplitude is small ($RA < 0.5$) or planning periods are long (for $\mu \geq 4$ when planning periods are 0.5 hours or longer). Results can be inaccurate when the service rate is low due to the lag between the arrival rate curve and delay rate curve. Due to this time lag which commonly occurs in practice, standard SIPP is more likely to understaff when the arrival rate is decreasing. An alternative way to overcome this problem is to employ SIPP Mix which uses the average arrival rates for phases where this is strictly increasing, and the maximum arrival rate otherwise. However, this is not reliable for large relative amplitudes. When $RA = 1$, even for shorter planning periods, SIPP Mix may be unreliable for large values of both λ and μ.

The SIPP and PSA approaches have been used to generate staffing requirements in

numerous settings; in some implementations workforce schedules are then constructed by managers without the benefit of additional models, while in others staffing schedules are generated using the requirements output by SIPP as constraints in an integer programming model (Green et al., 2003). Both SIPP and PSA have been successfully applied in several service establishments where managers need to adjust staffing levels in an attempt to provide a uniform level of service at all times, whilst customers arrive in a random but cyclic pattern (e.g. call centers (Quinn et al., 1991; Brigandi et al., 1994; Green et al., 2003), banking (Brewton, 1989) and airline services (Holloran and Byrne, 1986)); and the usefulness of the SIPP approach to aid in making emergency department scheduling decisions was confirmed in Green et al. (2006). The authors used a Lag Avg approach to account for the fact that peak congestion lags behind the peak arrival period in emergency departments, and improved performance rates by recommending that some provider hours should be switched from the night to much earlier in the day.

More difficult cases with very long service times and other complicated features, can often be treated by a MOL approximation. This technique recommends a staffing function by first numerically solving the $M(t)/M(t)/\infty$ queue, then subsequently using the approximation given in equation (5.5) to derive an arrival rate function based on the mean number of busy servers in a non-stationary infinite-server system (Massey and Whitt, 1994, 1997; Ingolfsson et al., 2007).

Applying Little's Law to a stationary $M/M/s$ system (see Gross and Harris (1998)) gives the expected number of busy servers, $B(t)$, as $\frac{\lambda}{\mu}$. Thus

$$\frac{\lambda}{\mu} = E[B(t)] \tag{5.4}$$

so:

$$\lambda = E[B(t)]\mu \tag{5.5}$$

Hence by approximating a stationary $M/M/s$ system with the arrival rate equal to the expected number of busy servers at each time point, service rate 1 and $s(t)$ servers, MOL is capable of generating a staffing function that satisfies the performance target level at all times. The approximation works particularly well when the utilisation is

low enough that the approximation of the expected number of busy servers in the system is accurate or when the system is overloaded for an extensive period of time (Ingolfsson et al., 2007).

Ingolfsson et al. (2007) compared the performance of the MOL and a lagged stationary approximation (essentially PSA with a revised arrival rate function), amongst other methods, in computing service levels for $M(t)/M/s(t)$ systems. Of the two aforementioned approximations, they concluded that the MOL approximation was most frequently the more accurate. Although the lagged stationary approximation was found to produce results at a quicker speed, the authors commented that the advantage was small in comparison.

5.3.2 Numerical methods

There are however many time-dependent queues, especially in health care, where the approximation approaches will not work well, so other methods are required to provide accurate insights of system behaviour. Whilst the $M/M/s$ model possesses the benefit of a tractable steady-state solution; the non-stationary nature of the arrival process in its time-dependent counterpart ($M(t)/M(t)/s(t)$) renders the queueing model analytically intractable i.e. there is no closed-form expression by which one can evaluate various performance metrics of interest over time. This section accordingly considers numerical methods which have been proposed in the literature to analyse time-dependent queues, and provide accurate insights in cases where the results provided by approximation methods do not necessarily hold.

When the assumption of a homogeneous arrival rate is relaxed and replaced by a piecewise function, the balance equations may be modified (replacing λ with $\lambda(t)$) and solved numerically to model the progression of the system over time. If one also allows the number of servers to become a time-varying function which changes in accordance with the workload (replacing s with $s(t)$), then the balance equations given in equation (5.1) are revised to:

$$\begin{aligned}
\tfrac{dp_o(t)}{dt} &= -\lambda(t)p_0(t) + \mu p_1(t), \\
\tfrac{dp_o(t)}{dt} &= \lambda(t)p_{n-1}(t) + (n+1)\mu p_{n+1}(t) - (\lambda(t) + n\mu)p_n(t), \quad 1 \leq n < s(t), \\
\tfrac{dp_o(t)}{dt} &= \lambda(t)p_{n-1}(t) + s(t)\mu p_{n+1}(t) - (\lambda(t) + s(t)\mu)p_n(t), \quad n \geq s(t).
\end{aligned} \tag{5.6}$$

Although it is also possible to replace the service rate μ with a time-varying rate $\mu(t)$, this would inadvertently imply that the service rate changes according to the queue length at time t. The implication arises since the service rate would affect the customers being *served* at time t in place of the more logical scenario where the mean rate of service provided to a customer would be related to the time at which they arrived (and therefore what type of service they desired). Thus for all time-dependent methods considered in this thesis, it is assumed that $\mu(t) = \mu$. The assumption is fairly realistic and commonly used in the literature, for the reason that in addition to the fact that the service rate more commonly varies in relation to the change of customers needs over time rather than the queue length, it also generally varies more slowly than the arrival rate (Ingolfsson et al., 2007).

The equations given in (5.6) hold for an $M(t)/M/s(t)$ system; thus customers are assumed to arrive according to an inhomogeneous Poisson process with rate $\lambda(t)$ and are served by the first available server (the number of which may fluctuate from one constant level to another throughout the service operation). Servers are assumed to have iid service durations that are exponentially distributed with mean $\frac{1}{\mu}$.

In a survey of the research, four numerical methods emerged as prominent techniques used to solve the balance equations in (5.6) and allow the subsequent analysis of the performance measures in time-dependent systems. These are the Euler method, the Runge-Kutta method, the randomization method and the discrete time modelling (DTM) method. Further detail regarding the theoretical underpinnings of the numerical methods used to solve the balance equations is provided in Fogiel (1983); Gross and Harris (1998) and Stewart (2009), and the main methods are discussed in turn below.

The Euler and Runge-Kutta methods are general approaches for solving ordinary differential equations. The Runge-Kutta method has the advantage that its calculation error may be limited to a desired level by choosing appropriate parameter values, and the solutions generated by the technique are often referred to as 'exact' since the only approximations required are the approximation of the infinite set of equations in (5.6) with a finite set, and those inherent in any numerical solution of ordinary differential equations. The Euler method however has the advantage that it may be implemented to provide solutions at a quicker rate and does not require an ordinary differential equation solver (Izady, 2010). For this reason, it is used as a benchmark in

a number of papers evaluating service quality, such as those by Davis et al. (1995) and Massey and Whitt (1997); and is utilised within this research to produce numerical solutions to specific WAST case studies. The Euler method approximates the solution by evaluating the equations at a starting value, and then at steps separated by small time intervals (between which the solution is not expected to have changed greatly). Smaller step sizes generate solutions with higher accuracies, but this requires greater computation time. Figure 5.4 geometrically illustrates how such periods may be constructed, and demonstrates the workings of the Euler method to consider the slope of the tangent line to approximate the solution at each interval.

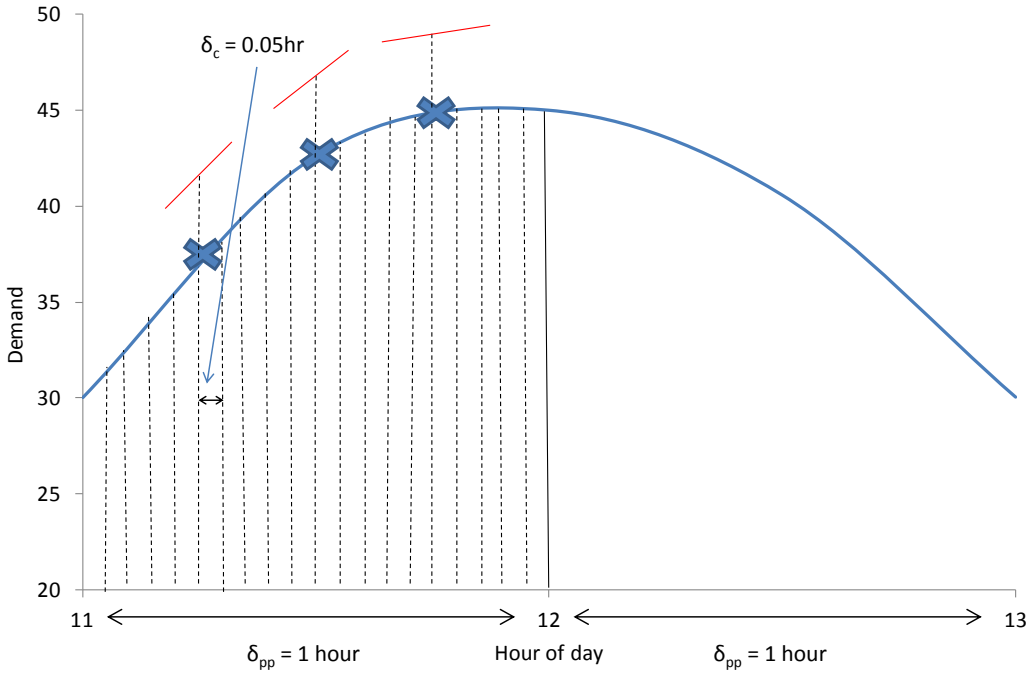

Figure 5.4: Graphical illustration of Euler methodology to solve the balance equations

When evaluating performance methods as a function of time over a period of service operation $(0, T]$, it is commonly assumed that the period $(0, T]$ is divided into planning periods of length δ_{pp} and the performance measure is evaluated at calculation periods separated by an interval of length δ_c, which is a divisor of δ_{pp} small enough to guarantee convergence to the actual solution. Figure 5.4 illustrates that for each planning period of length $\delta_{pp} = 1$ hour, the performance measure is evaluated (by calculating the slope of the tangent line) at at each step $\delta_c = 0.05$ hours, as illustrated

at three example points in the chart. In reality the value would be calculated at every step for δ_c, but for clarification purposes only three points are illustrated in Figure 5.4. The graph clearly demonstrates how choice of δ_c significantly effects the accuracy and speed of the method: as δ_c is decreased, both the accuracy and computation time increase (Izady, 2010). Further details regarding the computation of the Euler method are given in conjunction with the application of the method in Chapter 6.2.2.

The randomization method (described in Grassmann (1977)) was originally suggested for transient analysis of homogeneous, discrete-state, continuous time Markov chains. As transition rates from different states are not equal in most chains, this technique equalises them by adding fictional self-transitions i.e. it defines transitions for the fictitious case where system moves from a state to itself. Whilst the solutions generated by the method are not 'exact'; it was found to provide similar results as the Runge-Kutta method in Ingolfsson et al. (2007), but was more computationally efficient.

The DTM approach has also been shown to produce accurate results at a much faster speed than several approximation methods (Wall and Worthington, 2007) - especially in situations where the time-dependent performance is important. The approach uses discrete-time models to approximate the behaviour of continuous time queues by dividing the time of operation of the system into a set of non-overlapping intervals. The system state is only observed at interval boundaries and assuming that the probability distribution of the number of arrivals in a slot can be calculated independently of arrivals in other slots, the technique constructs a transition matrix (updated at each interval) to take account of the various states that occur at each time step and evaluate the probabilities associated with each state.

Comparative studies investigating the potential of the methods to compute service quality have been performed in Ingolfsson et al. (2007) and Izady (2010), revealing that approximately the same level of accuracy as the Runge-Kutta method can be obtained by the randomization method in shorter times. DTM was discovered to be the slowest method with modest accuracy, although some of its inaccuracy was attributed to variance mismatch resulting from approximating the service time with a discrete Geometric distribution. The Euler method was declared as the fastest method whose accuracy is not as good as randomization, but was very close and seems good enough for most practical purposes. For this reason, combined with the advantage

that the Euler method can be embedded into an Excel spreadsheet and incorporated into a workforce capacity scheduling tool in line with the ultimate aim of this thesis, its potential to compute WAST performance is investigated in the form of case studies in Chapters 6 and 7.

In situations where systems allow the arrival rate and number of servers to vary at the start of pre-specified shifts, but then remain constant for the entire duration of each shift, the system can be modelled as a mixed discrete-continuous time Markov chain (MDCTMC) (Ingolfsson, 2002). The difficult component to understand in such a system is the behaviour at shift boundaries and the effect of a departing server if they are providing service when they are scheduled to leave. Ingolfsson (2002) comprehensively investigated the influence of departing servers actions under the exhaustive discipline in a MDCTMC, further covering scenarios where servers may stop accepting customers Δt units before their shift is due to end. It was noted that the equations in (5.6) hold between each shift, but that the state probability vector (denoting the probabilities of various numbers of customers in the system) undergoes instantaneous transitions at shift boundaries as servers join/leave the system.

Kao and Wilson (1999) also commented that it is impractical to avoid truncation (in terms of the number of equations considered in the set of balance equations) in a numerical solution of the problem, since the infinite set must be reduced to a finite set to be solved numerically. Thus to ensure the dimension of the Markov chain is finite, a limit G is imposed on the number of customers considered in the system. Whilst no formal methodology exists to determine the value of G, it must be large enough to approximate an infinite capacity system with reasonable accuracy. Ingolfsson et al. (2007) selected the value so that it was the greater of 100 or $\lceil \max_{t\epsilon(0,T]} s(t) \rceil$ if the solution satisfied $P_G(t) \leq 10^{-6}$ for all t. If not, they increased G by 50% until the condition was satisfied.

Several summary measures can be computed as a direct function of the number of customers in the system. For example, the average number of customers in any queue is given by $\sum_{n=s(t)}^{\infty} nP_n(t)$, but this expression can only be converted to a closed-form formula for a limited number of queueing systems with specific characteristics. Discussions surrounding methods that allow computation of the time that a customer would have to wait for service if they arrived to a queueing system at instant t, referred to as the **virtual waiting time**, are included in Ingolfsson (2002); and

extensions to the standard formulation that allow the quantity to be computed in systems with a time-varying number of servers are proposed. Since closed-form expressions are unobtainable for such time-dependent systems, the methods are based on the logical concept of aggregating the products of the probability of n customers present in the queue and the probability of a wait longer than time x given n customers in the queue. The main challenges faced by WAST relate to limiting the proportion of patients with unacceptable waiting times; thus this thesis devotes considerable attention to develop the research contained in Ingolfsson (2002) to allow the virtual waiting time distribution to be computed in time-dependent systems that are additionally subject to varying levels of priority requests. Chapters 6 and beyond accordingly describe the proposed extensions to the methods, with a particular focus on the computation of the virtual waiting time over shift boundaries.

In cases where numerical methods are implemented to determine the minimum number of servers required to meet a given service level, the equations in (5.6) may be solved iteratively using a single dimensional search over the number of servers in each period, to find the minimum number required to meet the the desired performance target. The incorporation of a preemptive or exhaustive discipline however drastically affects the performance of the system under time-dependent conditions. Discussions surrounding this issue are contained in Ingolfsson (2002).

Whilst this thesis investigates the potential of the aforementioned approximation and numerical techniques to analyse the WAST service system, it is also worth noting that simulations have long been used to explore analytically intractable models. Drawing upon infinite-server models' results and a square root staffing law as well as the strength and stability of simulation models, Izady (2010) recently proposed a heuristic iterative approach for staffing emergency departments, based on the concept of time-stable performance. The square root safety-staffing principle recommends a number of servers s given by

$$s = r + \Delta = r + \beta\sqrt{r}, \quad -\infty < \beta < \infty, \tag{5.7}$$

where β represents the performance quality, which may be appropriately selected for each particular model (for example, as a ratio between hourly staffing and waiting time costs). The idea is that if r is large enough, then staffing the system with $r + \beta\sqrt{r}$ servers (for some parameter β) will achieve both short customer waiting times and high service utilisation (Koole and Mandelbaum, 2002). The form of s has

been shown to be extremely accurate for $M/M/s$ systems, and also robust for systems with multiple queues where agents have different skills (Askin et al., 2007).

The use of simulation is also explored in Defraeye and Van Nieuwenhuyse (2012) to analyse system performance under staffing levels proposed by an extension of the Iterative Staffing Algorithm (ISA). The ISA is a simulation-based approach to determine staffing requirements under time-varying arrivals, which targets a stable delay probability throughout the day. It operates by evaluating the delay probability for each period, and subsequently updating the staffing function with a more appropriate number of servers and re-analysing system performance, until the stopping criterion is met. The extended ISA specifically considers the probability of an excessive wait as the performance measure of interest; and is therefore more relevant to this study.

In many real life queueing systems, customers are not served in the direct order in which they arrive, but in terms of a pre-assigned 'priority' level. There is a wealth of literature devoted to the analysis of multi-server priority queues (Sleptchenko, 2005); however priority queueing theory in a multi-server setting with time-dependent arrivals is unsurprisingly less tractable to analyse. Section 5.4 outlines some of the existing research in the field of priority queueing theory, and notes the current gaps in the literature that the research contained in this thesis aims to narrow.

5.4 Priority queueing theory

Priority queueing is relevant to a range of real-life systems such as profit-making businesses which allow their users to pay additional capital for their jobs to be completed with priority, medical environments where patients suffering life-threatening conditions are attended to by paramedics before those exhibiting minor injuries, and hospital units which serve both emergency and elective patients. In priority systems higher priority customers are served before those with lower levels of priority, and only when there is no higher priority unit awaiting service is a lower-level priority unit taken into service. Figure 5.5, for example, depicts a typical profile of routine requests for ambulance transportation to Wales over a 24-hour period, which must be responded to in conjunction with urgent requests (previously discussed with Figure 5.3).

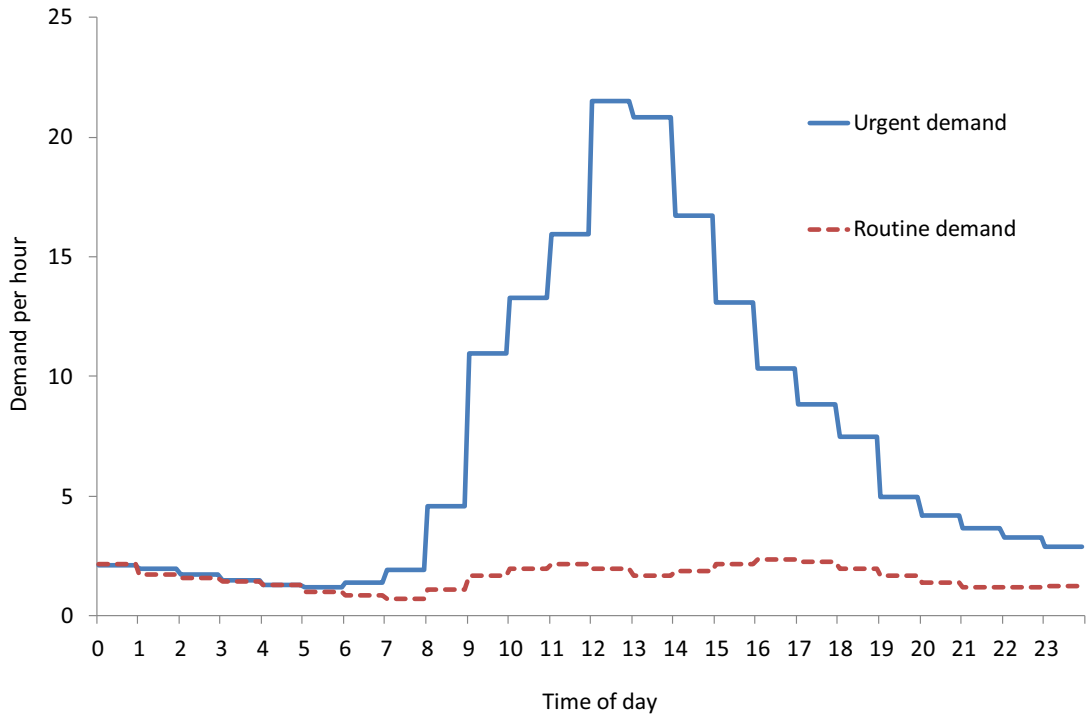

Figure 5.5: Urgent and routine demand for ambulance transportation

Figure 5.5 illustrates that the two sources of demand are time-dependent (especially urgent demand). Both are subject to stochastic and predictable sources of variation; yet the urgent requests must be prioritised over the routine requests at all times, even when the number of urgent requests far outweigh the routine demand, which varies little in comparison. Governed by the rules of priority queueing theory, an ambulance may only be sent to attend to a routine request if there are no urgent patients currently awaiting assistance. Thus to provide a constant service quality to both customer classes, the number of crews and ambulances on duty must rise along with the total number of requests; otherwise patients with routine requests would be significantly backlogged in periods with high demand volumes. Hence whilst the number of unacceptable waiting times for urgent requests may rise at such times, the proportion of routine patients experiencing excessive waiting times would rise significantly more.

For the purpose of this research, only two priority classes will be considered: high priority (HP) and low priority (LP) customers, and within each class of customer the

queue discipline is first come first served. When representing such a priority system as a queueing model, the following notation is commonly used in the literature:

- λ_H and λ_L denote the average arrival rates of HP and LP customers respectively
- μ denotes the mean service rate (assumed equal for HP and LP customers)
- $\rho_H = \frac{\lambda_H}{s\mu}$ and $\rho_L = \frac{\lambda_L}{s\mu}$ represent the server utilisation rates of each category of customer
- $r_H = \frac{\lambda_H}{\mu}$ and $r_L = \frac{\lambda_L}{\mu}$ represent the offered loads from each category of customer

Priority queueing theory falls into two branches according to the priority structure governing the order in which customers are served, as when a HP item arrives at a facility requesting service, two situations are possible. The HP customer arriving when no other customers of its class are present may either go directly into service and replace any LP customer in service (if one is present), or it may wait for the non-priority unit to complete service. Recall from Section 5.2 that the former of these is known as the **preemptive** priority discipline and the latter is called the **head-of-the-line** discipline. This thesis concerns the analysis of the the head-of-the line priority discipline, since once an ambulance is assigned to deal with a patient, it cannot be re-routed to attend another incident until it has completed its service.

The head-of-the line priority discipline for a single-server queue was first studied by Cobham (1954). Morse (1955) extended this work to obtain the probability generating function of queue lengths for systems with Poisson arrivals and exponential service times. More recently, Miller (1992) used the embedded Markov chain technique (see Gross and Harris (1998)) to analyse queues with Poisson arrivals and general service times, which was found to provide different solutions to the generating function approach.

In general, priority queueing is difficult to analyse in a multi-server setting because customers of different priority classes may be in service at the same time. In order to predict which customer will leave in a time interval, it is necessary to know the precise composition of all customers in the system within that interval and their position within the facility (i.e. the number of HP customers in the system, HP customers in service, LP customers in the system *and* LP customers in service). Thus the Markov chain representation of the multi-class, multi-server queue requires tracking the number

of jobs of each class through the system, meaning that a chain must be constructed which grows infinitely in k dimensions (where k is the number of priority classes) (Harchol-Balter et al., 2005). For smaller numbers of priority classes, it is possible to illustrate the Markov chain representation of the system in diagrammatic form. Figure 5.6 displays this chain for the case of an $M/M/2$ queue with two priority classes, which is simpler to analyse since it grows infinitely in only two dimensions. It demonstrates that HP customers are served according to an $M/M/2$ queue; but LP customers may be served according to an $M/M/2$ or $M/M/1$ queue, or may not have access to a any server. Thus whilst closed-form formulas exist to evaluate performance measures relating to HP customers (at least under steady-state conditions); the task of evaluating the service quality to LP customers is more complex (Harchol-Balter et al., 2005)).

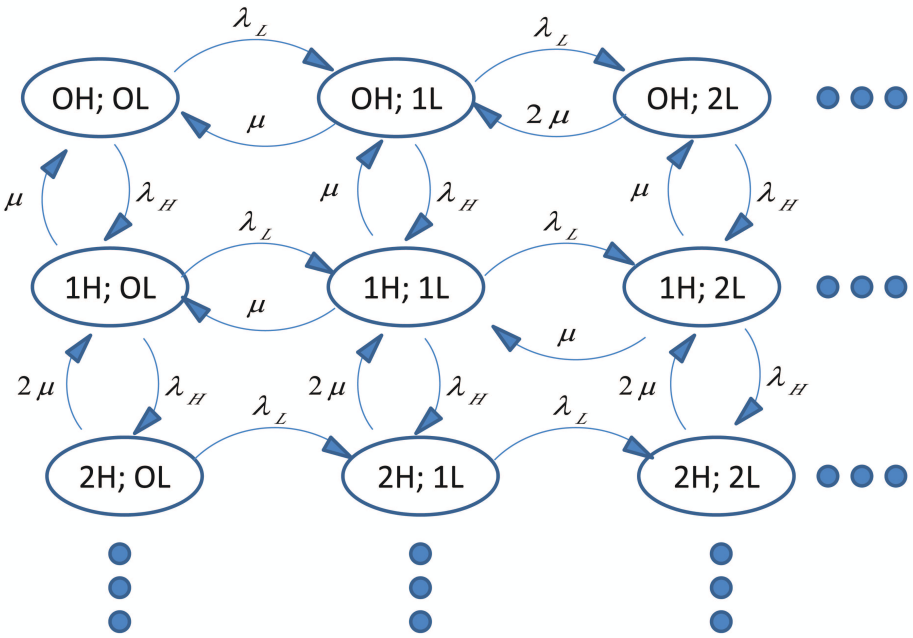

Figure 5.6: Markov chain representation of $M/M/2$ queue with 2 priority classes. The numbers presented for each state represent the number of HP (H) and LP (L) customers within the system

Of all the papers analysing multi-server priority queues, almost all are restricted to two priority classes with exponential service times. Other papers not restricted to two priority classes often use coarse approximations based on assuming that the multi-server behaviour is related to that of a single-server system (Bondi, 1984), or approximations based on aggregating the many priority classes into two classes, see Ngo and Lee (1990).

A survey of methods used in the literature until 2005 for solving priority queueing systems with two classes by Harchol-Balter et al. (2005) revealed that analytic methods may be divided into four main types:

- Approximations via aggregation or truncation
  - Several papers have investigated methods to simplify the representation of the Markov chain, which grows infinitely as the number of priority classes considered in the system is increased. For example, Kao and Narayanan (1990) truncate the chain by limiting the number of high priority jobs and Nishida (1992) aggregates states. However these methods provide poor approximations for large values of ρ.
- Matrix analytic methods
  - Matrix analytic methods model the birth-death process illustrated in Figure 5.2 using matrices that encode transitions from various states to track system behaviour over time. They have been shown to offer a numerical solution to the problem (see Wagner, 1997; Kao and Wilson, 1999), and although transition matrices can be constructed to track the probability of specific states in infinite Markov chains with $k = 2$, they are far more computationally efficient in infinite chains with $k = 1$. For this reason, many researchers choose to use additional approximations (e.g. by truncating the state space) alongside these methods.
- Generating function methods
  - In queueing systems with exponential service times, probability generating functions may be defined that provide the probability that the system is in different states (see Cohen, 1956). However, the solutions generated by the approach are susceptible to larger errors as the traffic intensity grows.
- Particular cases where the classes have the same mean service times
  - In systems where the service rate is identical for all priority classes, the resulting analysis may be simplified so closed-form formulas may be developed for certain performance measures, and Chen and Henderson (2001) explain how the probability of an excessive wait may be approximated when the mean service time is the same for all customers. Most papers assume that the

service times are exponentially distributed for all classes, but Sleptchenko (2005) further analyses systems with two hyper-exponential service distributions using a combination of generating functions and matrix analytic methods.

Whilst the methods discussed above provide accurate results under restricted circumstances, the first near-exact analysis of the $M/PH/s$ queue with several priority classes was recently proposed by Harchol-Balter et al. (2005). Their analysis developed a new technique known as Recursive Dimensionality Reduction (RDR). RDR allows the k-dimensional infinite Markov chain (resulting from tracking the k priority classes) to be reduced into a one-dimensional infinite Markov chain, which may be readily solved via numerical methods. It reduces the chain by firstly analysing the system only considering the highest priority (class 1) customers, and then using the analysis of the class 1 customers to analyse the performance of the class 2 customers etc, until all k priority classes have been considered. Since the behaviour of the highest priority customers is not influenced by the lower priority customers, the queue for the highest priority customers can be treated as an $M/M/s/FIFO/\infty/\infty$ queue. By recursively analysing the length of busy periods resulting from higher priority customers in the system, approximating the busy periods using PH distributions and using the PH distributions to construct one-dimensional Markov chains for lower priority customers, RDR ultimately allows the system to be represented as a one-dimensional infinite Markov chain.

RDR can handle systems with any number of servers, any number of classes and PH service times. In addition, RDR involves no truncation and results in only a small inaccuracy when compared to simulation ($< 2\%$ error for all the cases studied) and is extremely efficient. However, the complexity of the RDR method does increase with both the number of servers and the number of classes. As the method becomes less practical in situations where these numbers are high, a simpler (but slightly less accurate) approximation RDR-A has been developed which works by collapsing the k classes into two classes.

Many of the analytic methods researched in the papers discussed above have been used to provide insights into expected waiting times; and into how these vary among customers of different priority classes. In addition, Jaiswal (1962) considered the transient solution of two-class priority queues with Poisson arrivals and general service rates with two servers (by developing a Laplace transform of the time-dependent priorities for the general case, and deriving an explicit solution for queue length

probabilities for exponential service-time distributions with equal mean rates). The solutions found were different to the generating function approach; revealing that for complex queues, the asymptotic probability distribution obtained by considering the evolution of the queueing process may be different from the one obtained by considering the queue lengths at regeneration points (i.e. points at which the system is permitted to change state). Cohen (1956) provided analytical solutions for the delay probability, the average number of HP and LP customers in the queue, and the average delay experienced by both priority of customers in a $M/M/s$ system with two priorities and infinite waiting capacity; and also additional formulas to evaluate the number of LP customers 'lost' in systems where they are not able to wait if all servers are busy when they arrive. Whilst exact solutions (derived from probability generating functions) were provided for the case where the mean service time was identical for both categories of customers, only a general solution was provided for the case where the service time differed. Thus although the analysis of mean response time has been well understood in the case of a single-server priority $M/GI/1$ queue since the 1950's (Cobham (1954) and Cohen (1956)), the analysis of $M/GI/s$ and $M/M/s$ systems when services have different service rates has only been investigated in more recent years. A solution detailing the mean waiting time experienced by two classes of customers subject to exponentially distributed service times with different means was successfully provided by Gail et al. (1988), through solving a matrix equation to obtain the generating function of the steady-state probability distribution.

By means of applying matrix-analytic methods, Wagner (1997) further developed Laplace-Stieltjes Transforms (LSTs) to calculate the moments of the waiting time distribution for the different priority classes awaiting service (with equal average service times) for a head-of-the-line priority system with finite capacity. Whilst expressions for the steady-state customer waiting time distribution are not available for such models (Chen and Henderson, 2001), the LSTs can be used to calculate the probability that virtual waiting times for each class of customer exceed threshold levels. The use of these LSTs to set staffing levels in a call centre with priority customers was investigated in Chen and Henderson (2001). Whilst the authors noted that the LST for the call waiting time of the highest priority group could be easily inverted, allowing the probability of an excessive wait to be estimated for such customers, they observed that the same methodology could not be applied to lower priority customers. In order to avoid the calculations required by the algorithms given in Abate and Whitt (1995) to compute the tail probabilities of the waiting time distribution (that are not easily

implemented in a spreadsheet setting), the paper proposed easy-to-implement inequalities to obtain a lower-bound on the waiting tail probabilities for the lower priority calls.

Whilst the analysis of priority queueing systems remains a thriving area of research in OR today, only a few papers have been devoted to the analysis of such systems that are also subject to time-dependent demand. Through considering methods to set staffing call center levels in call centers with priority customers, the research contained in Chen and Henderson (2001) draws the most parallels to the key issue investigated within the following few chapters of this thesis. Chen and Henderson's analysis is however restricted to constructing staffing profiles based on the SIPP approximation technique, which has been shown to be inaccurate for many service systems, as its performance is heavily reliant upon the system possessing several necessary characteristics. The main insight provided by the research however relates to the scrutiny of the considerable error that can emerge when estimating performance for given staffing levels, resulting from estimating the arrival rates for future periods based on the average number of calls in a hourly periods throughout the week in the past. The paper identifies three main sources of potential error in estimating the arrival rate in this fashion: (i) estimation error arising from taking the average of a finite number of random variables, (ii) failure to detect nonstationarities that could be present in the data, and (iii) the presence of a random arrival rate (which may be a function of external factors e.g. weather conditions). In particular, the pilot study performed in Chen and Henderson (2001) revealed that the presence of a random arrival rate can lead to overpredictions of service performance, so if one ignores randomness, the risk of underestimating the number of staff required to achieve a given performance level is increased. Despite the importance of this fact, the need to account for random arrivals was found to be of second-order for the case study performed in the paper, as it revealed that for the few cases where the failure to account for random fluctuations produced different staffing requirements, the recommendation only differed by a single call taker. As far as queueing models are concerned, the paper provides strong motivation for considering the case where, conditional on the realisation of the random arrival rate, calls arrive according to a Poisson process. Whilst the shortfall to neglect randomness was deemed to be of second-order for the particular scenario investigated, the paper concluded by emphasising that it might be important in other setting and recommended future studies should consider different forecasting methods that are able to adequately deal with nonstationarities in the data to achieve lower forecasting errors.

More recent papers investigating priority systems have focussed on systems with various attributes such as impatient customers, the analysis of the impact of buffer finiteness, developing new techniques to track the number of HP and LP customers, and improving the accuracy of numerical solutions (Harchol-Balter et al. (2005) and Sleptchenko (2005)). Gurvich et al. (2010) recently considered solving the problem of staffing a multi-class multi-type call centre facing demand uncertainty by translating the problem of staffing with uncertain demand forecasts to one of solving a small set of problems with perfectly predictable demand-rate vectors. The paper also highlighted the importance of taking the uncertainty of arrival rates into account and addressed two main issues in call centre staffing: (i) that arrival rates are forecasted in advance (and are rarely precise), and (ii) different customer classes have different service requirements so cannot be considered as a single customer class. The researchers proposed a 'chance-constrained' formulation of the problem to set staffing levels in service systems to achieve pre-specified target service levels. This chance-constrained formulation is based on the concept that the performance measure should met with high probability (chosen by management) with respect to the uncertainty in the demand rate; and by means of considering a staffing frontier approach, it translates a complex staffing problem with uncertain rates into one of solving multiple problems with predictable rates. Whilst the paper does not directly consider time-varying arrival rates, the authors note that the method could be extended to consider such issues. However, although the procedure generates solutions that are nearly optimal for large call centers, the solutions are subject to greater errors in smaller service systems.

5.5 Summary

This chapter has outlined many preliminaries required for the analysis of time-dependent and priority queues. Beginning with a summary of the various components of a queueing system, it has described key assumptions required in queueing theory analysis, explained how performance measures may be computed in time-dependent systems and outlined the most common methods used in the literature to analyse priority queues.

Section 5.3 was specifically devoted to the analysis of time-dependent systems. It explained that whilst the non-stationarity of the arrival process renders queueing models analytically intractable, numerical and approximation methods have been proposed in the literature that allow analysis of performance measures. Numerical

methods can provide high degree of accuracy at the expense of computation speed, whilst approximation methods are fast, but not so accurate. Sections 5.3.1 and 5.3.2 outlined the main methods that have been investigated in the literature, together with the benefits and drawbacks of each approach. In particular, the SIPP and Euler approaches were presented as suitable techniques to evaluate performance measures in time-dependent systems, which shall be further studied and applied to WAST data in Chapter 6. In conjunction with the application of these methods, more specific details of the techniques will be provided, along with suggestions of potential improvements that could be applied to enhance the performance of the models.

The review of priority queueing systems highlighted that whilst detailed solutions have been provided for the mean waiting times experienced by two priority groups of customers, little research has been devoted to the analysis of other performance measures in time-dependent priority systems. Section 5.4 explained how a priority queue may be modelled as a Markov chain, and summarised the most popular methods that have been developed to analyse such queues in the literature. The work contained in Chapter 7 of this thesis aims to narrow the observed gap in the research relating to the evaluation of performance measures systems that are simultaneously subject to priority and time-dependent demand, through considering extensions that may be applied to the time-dependent methods discussed in Section 5.3, to enable their application within priority queueing systems. The analysis shall attempt to enhance the methods previously considered in the literature by investigating revisions that can be applied to the arrival rate in the SIPP approach, examining the behaviour of the system over shift boundaries to enable the numerical approach to be applied within time-dependent priority systems, and proposing a hybrid approach that employs both numeric and approximation methods.

Chapter 6

Computing service levels in $M(t)/M/s(t)/FIFO$ systems

6.1 Introductory remarks

The purpose of this chapter is to investigate the potential of the methods summarised in the queueing theory literature review (Chapter 5) to evaluate the performance of non-stationary systems, together with their drawbacks and possible extensions. For the approximation method, the focus is on the SIPP approach due to its capacity to restrict changes to occur at the boundaries of staffing intervals; and a Euler solver is employed to provide a numerical solution, as introduced in Chapter 5. For the numerical method, a detailed study examining the true behaviour of the system at shift boundaries is included, focussing on the adjustment of the waiting time formulas and distribution of the expected number of customers in the system at such epochs.

On the application front, the models are used to predict a portion of the full EMS problem: the minimum number of paramedics required to respond to Category A incidents arising within the SE Region, so that 65% of all such incidents are responded to within 8 minutes as specified by the AOF targets (Welsh Government, 2011). Model comparisons allow for empirical specification of the conditions under which each of the methods perform well, whilst consideration is also given to the effort required for their execution. The chapter serves as a precursor to Chapter 7, where the investigated approaches are extended to model a priority queueing system, so the minimum number of ambulances required to respond to Category A and Category B calls can be computed simultaneously.

The remainder of this chapter is organised as follows. Section 6.2.1 outlines the SIPP methodology and includes discussions surrounding the calculation of excessive wait probabilities (the system performance measure of interest to WAST) in stationary $M/M/s$ systems, since these systems are adjoined in a series by the approximation technique. Details of the numerical method follow in Section 6.2.2 with explanations clarifying how the standard queueing theory expressions may be extended to accurately evaluate performance measures in time-dependent systems. Particular attention is devoted to the behaviour of the system performance at shift boundaries, and descriptions of the extensions necessary to allow for meaningful analysis at such epochs are given. Section 6.3 provides an overview of the data used to test the application of the methods and demonstrates how the techniques may be employed to model particular scenarios before evaluating the potential of several variations of the SIPP approach to determine staffing requirements in an emergency response setting. Supplementary case studies are performed which reveal greater insights into the numerical methodology and illustrate the effects arising from the incorporation of the varying types of server behaviours at shift boundaries. A short summary of the chapter is ultimately provided in Section 6.4 which brings together the applied and theoretical aspects of the time dependent queueing theory techniques.

6.2 Approximation and numerical methods

This section provides further details regarding approximation and numerical methods that can be used to analyse systems subject to time-dependent demand.

6.2.1 Approximation methodology

Service systems with time-varying arrival rates commonly use the $M/M/s$ model (which assumes a constant arrival rate, service rate and number of servers) as part of a larger SIPP model to allow staffing requirements to change throughout the operation period (Green et al., 2001, 2006, 2007). Both SIPP and PSA methodologies are outlined in Chapter 5.3.1, and for the reason that the SIPP approach additionally restricts changes to occur at the boundaries of staffing intervals, this method is widely recognised as a practical technique that will be accordingly investigated to analyse the data within this thesis. Hence all references to the approximate methodology herein directly refer to the SIPP approach.

In order to allow a time-varying arrival rate that occurs in many service systems, the methodology followed by SIPP is to first divide the full operation period into distinct staffing intervals, find the average the arrival rate of customers in each interval, and to use this average as input to a series of stationary $M/M/s$ queueing models to estimate the number of servers required in each period. The advantage of this approach is that formulae for performance measures such as the probability of delay can be easily obtained for each independent $M/M/s$ queue and programmed in a spreadsheet tool, so recommendations of the minimum number of servers required to meet the service target in each period can be instantly generated for any forecasting horizon.

The methodology assumes that delays in consecutive periods are statistically independent (which is clearly a fundamental flaw of such models), that the system reaches steady state within each planning period (so the probability of the number of customers in the system follow the formulae in equation (5.2)), and that the arrival rate does not change within each planning period. Thus the accuracy of the results is strongly reliant upon a number of parameters possessing suitable properties outlined in Chapter 5.3.1: in particular the magnitude of the service rate, the amplitude of the arrival rate, and how these interact together. Thus despite its very widespread use, SIPP may not always provide appropriate staffing levels (Green et al., 2001). In cases where these assumptions are violated, the accuracy of the approach can be improved by adjusting the arrival rate function before the applications of the approximation methods (Thompson (1993); Green et al. (2001)), as described in Section 5.3.1.

Delay systems are service systems which operate with a finite number of agents and an infinitely large waiting space, so that all customers prepared to queue may eventually be served. A common performance measure of interest for evaluating service quality in such systems (e.g. WAST) is the probability of an excessive wait and several organisations have devised targets that specify that this wait should be kept below a given threshold for a given proportion of customers. For example, up until the first quarter of 2011, there was a target for 98% of patients in A & E departments to be discharged, transferred or admitted to inpatient care within 4 hours of arrival (Izady and Worthington, 2012), whilst call centers often aim to answer a given proportion of calls within 20 seconds of a connection being made (Atlason et al., 2008). Moreover this performance measure is used as the primary indicator of service quality of the ambulance service, and the specific targets for the different categories of incidents have been detailed in Section 1.4. As mentioned in Section 5.2, closed-form formulas

exist to calculate the probability of excessive waits in $M/M/s$ systems. The analysis contained in the following section explains how these formulas may be derived and applied to a series of $M/M/s$ systems adjoined by the SIPP technique, to ultimately approximate the probability of an excessive wait within each period.

6.2.1.1 Virtual waiting time distribution

Using the notation as described in Chapter 5.2, it is clear that the probability of an excessive wait (also known as the *waiting tail* probability) may be expressed as:

$$P(W_q > x) = \sum_{n=s}^{\infty} P(W_q^n > x) p_n \tag{6.1}$$

Thus if the values for the infinite dimensional vector $(p_n) = (p_0, p_1, p_2, ...)$ are known, the likelihood that a customer waits greater than time x before commencing service may be computed instantaneously. For stationary $M/M/s$ systems, the Poisson distribution may be employed to provide the values for this vector since it evaluates the probability of exactly r services finishing in time t. Observing that the wait in the queue will be greater than time x if less than n services are completed in the time x, equation (6.1) may be reduced to the following closed-form formula (see Gross and Harris (1998)):

$$P(W_q > x) = \left(\frac{(\frac{\lambda}{\mu})^s \; p_0}{s!(1 - \frac{\lambda}{s\mu})} \right) \quad \left(e^{-(s\mu - \lambda)t} \right) \tag{6.2}$$

Whilst equation (6.2) allows for the evaluation of this probability in a steady-state system, the computation of the same measure in a time-dependent system must capture the change in probability distribution of the number of customers in the system as servers join and leave the system at shift boundaries. Whilst the question of how to evaluate this probability for inhomogeneous time periods is considered in detail in Section 6.2.2.2, an approximate probability may be obtained by embedding equation (6.2) as part of a larger SIPP model: by dividing the period of operation into distinct staffing intervals, assuming steady-state conditions in each and inputting the relevant arrival rate for each shift, it is possible to iteratively compute the probability of an excessive wait with varying quantities of staff, s, and estimate the minimum required per shift to meet the target guidelines.

It is important to remember however that the results are strictly approximations as

when adjoining multiple $M/M/s$ systems, SIPP fails to account for the dependence between periods which occurs in practice as the queue length at the start of each period is reliant on the number of customers remaining in the system at the end of the previous period. In cases where the assumptions of SIPP are violated, more accurate predictions may be determined by appropriately adjusting the arrival rates input to equation (6.2), using techniques as discussed in Section 5.3.1.

Before applying SIPP to the ambulance data, a description is given of the particular numerical method that shall be employed in this work to provide exact probabilities of excessive waits and used to benchmark the performance of the approximation technique.

6.2.2 Numerical methodology

In order to evaluate the quality of the SIPP approach, the Euler method is used to provide a numerical solution of the balance equations for the reasons that it has been used as a benchmark in a number of other papers evaluating service quality (Davis et al. (1995); Massey and Whitt (1997)), is relatively straightforward to implement and provides accurate results since the only approximations required (after the substitution of the arrival rate function with one that is piecewise constant over the calculation periods) are the replacement of an infinite set of equations with a limited set and those to numerically solve a set of differential equations. Hence all references to numerical approaches from this point onwards assume that the solutions are those generated by a Euler solver.

Section 5.3.2 presented the balance equations for an $M(t)/M/s(t)$ system (see equation (5.6)), explained that a limit G must be imposed on the number considered in the system to reduce the infinite set of balance equations to a finite set, demonstrated how the Euler method may be used to solve the equations, and stated that the service quality function is generally evaluated at calculation intervals of length δ_c across the period of operation (divided into shifts of length δ_{pp}, assumed to be an integer multiple of the length of δ_c). When choosing a sensible calculation interval, it is desirable to select one such that it provides a fairly accurate representation of the system in a reasonable computation time. Until recently there has been no formal method used to control the accuracy of the Euler method, but Izady (2010) advises that this quantity be chosen such that $\delta_c = \frac{1}{2v}$ where v refers to the rate at which the process leaves its

current state (as a result of a customer arrival, departure or change in the number of servers), defined in Appendix A.3.

Systems which allow the arrival rate and number of servers to vary at the start of pre-specified shifts, and remain constant for the entire duration of each shift, can be modelled as MDCTMCs. Considering a birth-death process $\{N(t)\}$ (defined in Section 5.2), and letting $p_n(t) = \text{Prob}\ \{N(t) = n\}$ and $t_z, z = 1, 2, ...$ represent the times that shifts start and finish, when the number of servers and arrival rates are permitted to change; then at shift boundaries t_z the stochastic process {N(t)} behaves as a discrete time Markov chain, and like a continuous time Markov chain between these time points. Thus during the intervals $(0, t_1), (t_1, t_2)$ the probabilities $p_n(t)$ satisfy the balance equations given in (5.6). If $(p_n(t))$ represents the vector $(p_0(t), p_1(t), ...)$ then at shift boundaries (i.e. time points where $t = t_z$), the vector is subject to an instantaneous transition $(p_n(t)) = (p_n(t))^- B(t)$ where $(p_n(t))^- = \lim_{y \to t_z}(p_n(y))$ (i.e. the probability vector immediately before the shift boundary) and $B(t)$ is a probability matrix. Further information regarding the theory of the MDCTMC may be found in Ingolfsson (2002) who defines the model, and investigates the analysis of the Markov chain if servers stop accepting customers Δt units before they are due to finish duty. The research contained within this thesis extends the approach by Ingolfsson (2002) through converting the formulae to a format that may be applied to (i) various types of shift boundaries and (ii) priority systems; so they may ultimately be readily applied to the WAST service system in Section 6.3, where crew members cannot decline calls for assistance before their shift ends due to the severity of the incidents requiring their assistance.

6.2.2.1 Mappings of state probability vector across shift boundaries

In the representation of a system as a MDCTMC, it may be desirable to allow the arrival rate and number of servers to vary at set times throughout the duration of a long shift. For example, in situations where shifts are pre-defined (say morning (6am-12pm), afternoon (12pm-7pm) and evening (7pm-6am)), then if the arrival rate changes significantly across the course of a shift, a more appropriate staffing profile could be formulated by employing a consistent set of base staff who work for the entire duration of the shift (i.e. all hourly periods throughout the shift in succession), with a small number of additional staff who may start/stop work for certain hourly periods as the demand rises/falls throughout the shift to consistently match the demand for services.

In recognition of these scenarios, this research defines and analyses two types of shift boundaries:

- A **true shift boundary**: These boundaries mark the set times at which pre-defined shifts (e.g. morning, afternoon, night) finish. It is assumed that **all** staff leave at the end of this shift under exhaustive discipline guidelines, and are replaced by an entirely new set of staff in the following shift.
- A **dummy shift boundary**: This type of boundary occurs at epochs at which the number of servers is permitted to change, say at the end of every hourly period. At a dummy shift boundary, it is assumed that the same servers are employed if the equivalent number (or more) are required for the following period. If more staff are required, these work alongside the existing servers for as many hourly periods as needed. Since customers are assigned to servers at random, then if some servers are idle at a dummy shift boundary when staff are scheduled to leave, it is possible that the idle servers are those assigned to leave or remain within the system. All busy servers departing the system adhere to the rules as specified by the exhaustive discipline.

Under the exhaustive discipline, all servers that are busy at a true shift boundary continue serving the patients they are currently dealing with until they complete their service - thus all patients being attended to by such servers at the shift boundary are consequently '**ejected**' (i.e. removed) from the system as they no longer require the assistance of a resource scheduled to work. In reality, the customers continue to be served by the departing servers until their work is completed. The servers scheduled to work in the next period immediately begin attending to the patients waiting in the queue.

Since busy staff who are scheduled to leave the system at the shift boundary are assumed to continue working until they complete their current service under the exhaustive discipline, at the same time that new servers begin serving customers at the front of the queue, this research implicitly assumes that resource shortages never arise i.e. sufficient equipment is available for both crew sets to operate simultaneously. In the case of WAST, this implies that new crews may acquire separate RRVs/EAs to those deployed for the preceding shift, and that the the previous crews are not required to return to base to pass on any equipment to new staff commencing duty.

Motivated by the scenario where there are 3 busy servers and 4 customers in the queue awaiting service before the shift boundary, Figure 6.1 provides a visual representation the effect of applying each type of shift boundary on the number of customers in the queue, when various staffing levels are required after the boundary (see Figure 6.1(b - d)).

- If the number of servers is decreased from 3 to 2 (as in Figure 6.1(b)), then:
  - For a true shift boundary: 2 customers remain in the queue since the original 3 staff leave the system (continuing to serve the customers in service) and are replaced by a new set of 2 employees who immediately see the first 2 customers in the queue.
  - For a dummy shift boundary: 4 customers remain in the queue since only one server departs and the remaining two continue their duty as normal, so there is no effect on the number of customers in the queue.
- If the number of servers remains constant (as in Figure 6.1(c)), then:
  - For a true shift boundary: 1 customer remains in the queue since the original set of staff are replaced by an entirely new set so the queue length is effectively reduced by the original number of servers.
  - For a dummy shift boundary: 4 customers remain in the queue since the system continues to operate in a standard fashion, so the queue length remains unchanged.
- If the number of servers is increased from 3 to 4 (as in Figure 6.1(d)), then:
  - For a true shift boundary: 0 customers remain in the queue since the entire set of original staff leave in the case and are replaced by a new set of 4, who are able to deal with all customers awaiting service in the queue.
  - For a dummy shift boundary: 3 customers remain in the queue since no staff leave the system, but an extra employee is assigned works to work along side the existing servers and accordingly takes on the first customer in the queue.

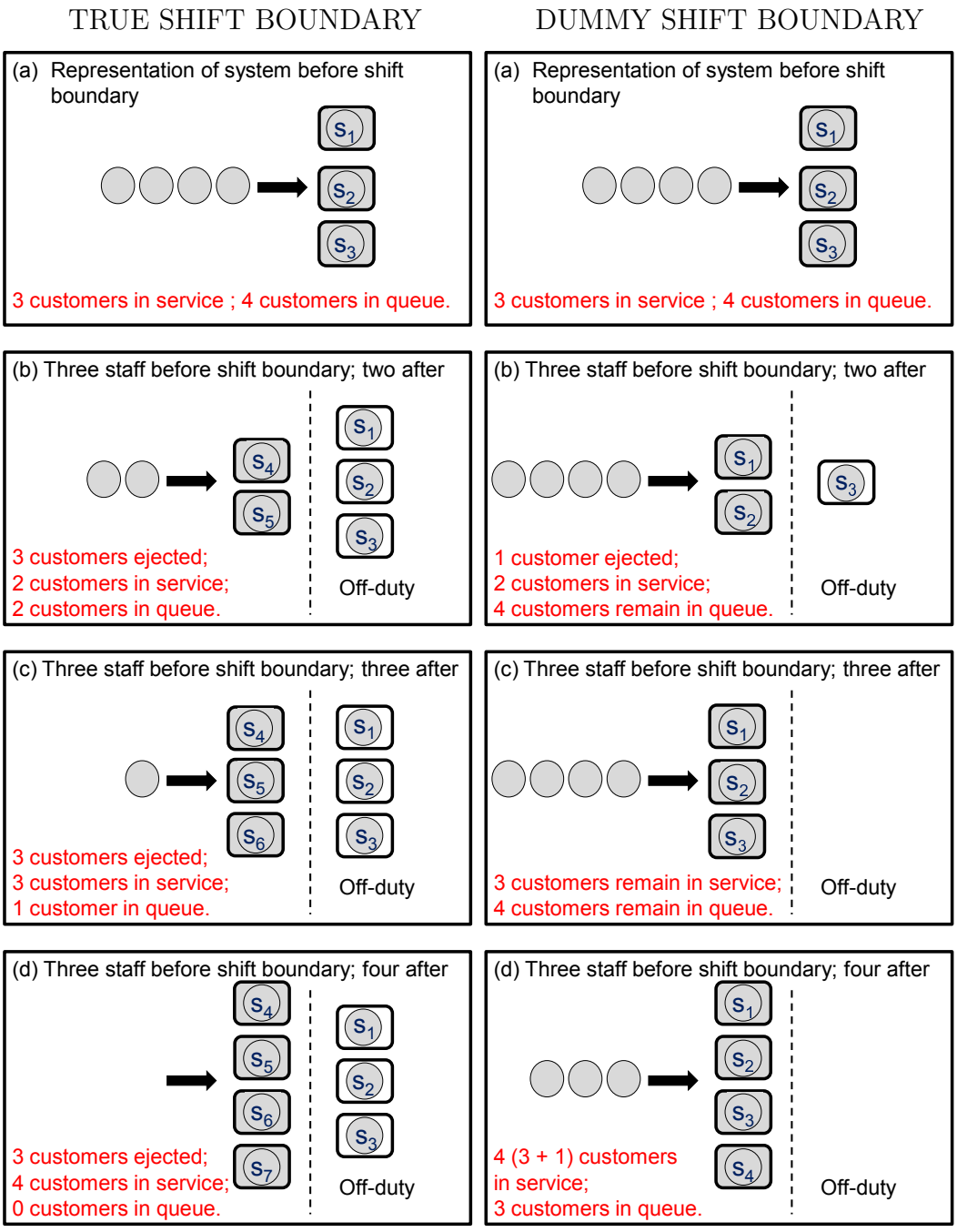

Figure 6.1: Representation of system before shift boundary (a), and movement of customers after true and dummy boundaries with varying number of servers (b - d)

In view of the above observations, it is necessary to apply instantaneous transitions to the state probabilities $p_n(t)$, $n = 0, 1, ...$ over each type of shift boundary to account for the effect of departing servers.

Case A: True shift boundary

At true shift boundaries, all servers exit the system, so the probabilities evolve according to the mapping:

$$\begin{aligned} p_0(t) &= \sum_{n=0}^{s(t)^-} p_n(t)^-, \\ p_n(t) &= p_{n+s(t)}(t)^-, \text{ for } 1 \leq n \leq G - s(t)^-. \end{aligned} \tag{6.3}$$

where $p_n(t)^- = \lim_{r \to t_z} p_n(r)$ and $s(t)^- = \lim_{r \to t_z} s(r)$ (i.e. the probability vectors and number of servers on duty immediately before shift boundaries, assumed to occur at time points $t = t_z$).

Case B: Dummy shift boundary

The transitions are more complex to define in the case of a dummy boundary, as it becomes necessary to account for the actions of departing servers. If no servers leave, then the state probability vector requires no modification over the boundary. However, if some servers depart the vector undergoes an instantaneous transition according to $(p(t)) = (p(t))^- B(t)$. The formulation of the matrix $B(t)$ for the exhaustive discipline where some servers stop accepting customers Δt units before they are due to finish duty, is presented in Ingolfsson (2002). With slight modifications to the assignment of the variables proposed in this paper, the same form of the matrix may be used to model the case of a dummy shift boundary, where the departing servers follow an exhaustive discipline. Whilst only a small modification is necessary to develop an matrix that appropriately accounts for the effect of departing servers at dummy shift boundaries, the workforce capacity planning tool provided in conjunction with this thesis establishes how it contributes greatly to the existing literature with a demonstration of a beneficial application of the formulae. In the workforce planning tool, shift schedules are developed around the hourly period requirements that are produced by the SIPP and Euler methodologies: thus since the exact type of shift boundaries are unknown at the stage that the hourly period requirements are generated, it is appropriate to apply a dummy shift boundary to every period; since shifts are flexible and will be selected to match the period requirements as closely as possible.

For the case of a dummy shift boundary in a standard time-dependent queue, the matrix $B(t)$ used to calculate the probability that δn customers are ejected from the

system is derived as follows.

If all servers are busy at a dummy boundary, the number of customers ejected will be equivalent to the number of departing servers, so a value of 1 is assigned to the entries in the relevant positions of the matrix $B(t)$ to reflect this fact.

Contrarily, if not all servers are busy before the dummy boundary, then the number of customers ejected can be less than the number of servers that leave (i.e. if an idle server is selected to depart). It is assumed that customers are randomly assigned to servers and the probability that each server is selected to leave at the boundary of an interval is independent of whether they are busy or idle; thus following a similar argument to that presented in Ingolfsson (2002), it is clear that that the number of customers ejected follows a hypergeometric distribution (Johnson et al., 1993). Letting δs represent the number of departing servers over the shift boundary and n represent the total number of customers in the system before the boundary, then the probability that δn customers are ejected from the system is:

$$\phi(\delta n; \delta s, s(t)^-, n) = \frac{\binom{n}{\delta n}\binom{s(t)^- - n}{\delta s - \delta n}}{\binom{s(t)^-}{\delta s}} \tag{6.4}$$

Thus the transition matrix $B(t)$ has the following nonzero entries:

$$\begin{aligned} &B_{n,n-\delta s} = 1 && \text{for } n = s(t)^-, s(t)^- + 1, ... \\ &B_{n,n-\delta n} = \phi(\delta n; \delta s, s(t)^-, n) && \begin{cases} \text{for } n = 0, 1, ..., s(t)^- - 1 \text{ and} \\ (n - (s(t)^- - \delta s))^+ \leq \delta n \leq \min(\delta s, n) \end{cases} \end{aligned} \tag{6.5}$$

where $(n - (s(t)^- - \delta s))^+ = \max(0, n - (s(t)^- - \delta s))$.

As a motivating example of the above, consider a service system with 3 servers on duty at a dummy shift boundary, where 1 server is scheduled to leave. If all servers are busy at the shift boundary, then one customer will certainly be ejected from the system; but if some servers are idle then it is possible that either 0 or 1 customer could be ejected. For this simple example it is possible to logically determine the mappings that provide the probabilities of varying numbers of customers in the system, before using $B(t)$ to confirm the results, since it is clear that:

- If no customers are in the system before the boundary, none can be present after.

- If 1 customer is present before the boundary, this customer must be in service so 1 server will be busy and 2 will be idle at the shift boundary. This gives a $\frac{1}{3}$ chance that the customer will be ejected from the system (if being served by the departing server), and a $\frac{2}{3}$ chance they will continue to receive normal service.
- If 2 customers are in the system before the boundary, there is a $\frac{2}{3}$ chance that 1 customer will be ejected and a $\frac{1}{3}$ chance that both will remain in the system.
- If 3 or more customers are in the system before the boundary, then all servers must be busy, so 1 customer will be ejected with certainty.

The matrix $B(t)$ may be used to formally provide the probability mappings to be applied to each of the state probability vector $(p_n(t))$ under such a scenario to account for the occurrences described above, as follows:

$$(p_0(t)\ p_1(t)\ p_2(t)\ p_0(3)\ \cdots) = (p_0(t)^-\ p_1(t)^-\ p_2(t)^-\ p_0(3)^-\ \cdots) \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots \\ \frac{1}{3} & \frac{2}{3} & 0 & 0 & \cdots \\ 0 & \frac{2}{3} & \frac{1}{3} & 0 & \cdots \\ 0 & 0 & 1 & 0 & \cdots \\ 0 & 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Multiplying out the matrix gives:

$$\begin{aligned} p_0(t) &= p_0(t)^- + \tfrac{1}{3}p_1(t)^- \\ p_1(t) &= \tfrac{2}{3}p_1(t)^- + \tfrac{2}{3}p_2(t)^- \\ p_2(t) &= \tfrac{1}{3}p_2(t)^- + p_3(t)^- \\ p_n(t) &= p_{n+1}(t)^- \ \ \forall n \geq 3 \text{ (i.e. } s(t)^-), \end{aligned}$$

as desired.

6.2.2.2 Virtual waiting time distribution

In time-dependent systems, the probability distribution of the number of customers in the system varies throughout the day, so the formula to calculate the probability of an excessive wait must capture this element of change. The performance measure may be calculated as a function of the state probability vector, but the expression must be refined to account for the transitions over shift boundaries. Ingolfsson (2002) states that the incorporation of a preemptive or an exhaustive discipline can have a considerable impact on performance predictions, especially when average service times are relatively

long and the number of servers changes quite drastically. This research hypothesises that the incorporation of a dummy or true shift boundary is also vitally important.

Green and Soares (2007) explain how the probability of an excessive wait may be computed under the non-exhaustive discipline, assuming that the infinite dimensional vector $(p_n(t))$ is known in their derivations as a result of solving the balance equations. However in extension to the work presented in this paper, as well as contrastingly dealing with a exhaustive discipline, this research recognises that the state probability vector experiences an instantaneous transition at the shift boundaries to account for the actions of a leaving server (as described above); and thus additionally allows for the application of the adjustments defined in Section 6.2.2 in conjunction with the waiting time formulae. For the case of a true shift boundary, it is observable that the virtual waiting time $W_q(t)$ of a customer in the queue is equivalent to the first passage of time that the continuous Markov chain tracking the number of customers in the system $N(t)$ reduces to match the number of servers available in the shift after the boundary, since if customers are served in a FIFO fashion then when there are the same number of customers as servers, all customers are able to be served. However, more complex formulations of the expression are required if the number of servers on duty change within the target waiting time (Green and Soares, 2007).

The calculation is surprisingly simpler under the exhaustive discipline, since it is not necessary to account for customers being preempted back into the queue. This chapter furthers the work of Ingolfsson (2002) on this subject (who define the case for the exhaustive discipline with allowances for servers to stop accepting customers Δt units before their shift is due to end) by defining how the tail probability of the waiting time may be computed for both dummy and true shift boundaries, in a format readily extendable to priority systems.

The derivation of the formula to calculate the probability of an excessive wait requires the notation defined in Chapter 5.2 to be extended to incorporate time-dependent behaviour, such that $W_q(t)$ denotes the virtual time that a customer arriving at time t must wait before commencing service. In particular, this research is concerned with limiting the waiting tail probability, $P(W_q(t) > x)$, to a pre-determined proportion. Letting $W_q^n(t)$ denote the waiting time in the queue for a customer that arrives to the system at time t to find n people in the system ahead and s servers on duty, then this

may be computed for a time-dependent system as:

$$P(W_q(t) > x) = \sum_{n=s}^{\infty} P(W_q^n(t) > x) p_n(t) \tag{6.6}$$

Queueing models for non-stationary $M(t)/M/s(t)$ systems are analytically intractable (Izady and Worthington, 2012); thus unlike the steady-state case, no closed-form expressions exist to provide the values of the infinite dimensional vector $(p_n(t))$. In such situations, $P(W_q(t) > x)$ can be evaluated by computing the state probability vector at small intervals throughout the operation period (during which the system is not expected to have changed much), and considering the minimum number of customers required to leave within the target time for the target measure to be met at each of these epochs.

The expression formulated to evaluate $P(W_q(t) > x)$ at a time t, when time t and $t + x$ lie within the same interval (e.g. within $(0, t_1)$, within (t_1, t_2) etc) forms the basis of the formula developed to compute the measure over shift boundaries. The formula will initially be derived for the first case; henceforth referred to as the case where the maximal allowed target wait time lies within a singular interval. Within such intervals, the arrival rate and number of servers remains constant so {N(t)} behaves like a continuous time Markov chain. One may exploit this property to compute the waiting tail probability using the fact that the departure process behaves as a nonhomogeneous Poisson process over the entire interval with rate μs (for periods when all servers are busy over t interval), so the mean number of departures over $[t, t+x]$ is given as below (for further details see Green et al. (2007)):

$$a = \mu s x \tag{6.7}$$

Thus the probability of the event $P(\text{“}n - s \text{ or fewer departures over } [t, t + x]\text{”})$ is equivalent to:

$$\sum_{b=0}^{n-s} \frac{a^b e^{-a}}{b!} \tag{6.8}$$

Hence,

$$P(W_q^n(t) > x) = \begin{cases} \sum_{b=0}^{n-s} \frac{a^b e^{-a}}{b!} & \text{if } n \geq s, \\ 0 & \text{if } n < s. \end{cases} \tag{6.9}$$

Yet if number of servers changes over $[t, t+x]$, the derivation is complicated by the fact that the event depends not only on $s(t)$, the number of servers available at time t, but also the number of servers in the period immediately after i.e. $s(t+x)$. In this case, the waiting time formula cannot be simply extended by replacing s with $s(t)$; as if the number of servers is increased exactly once over the interval for example, say at time $t+\Delta t$, where $\Delta t < x$, then fewer than $n - s(t)$ departures may result in an arriving customer waiting less than time x before being served (Green et al., 2007). This is because the additional staff starting at time Δt will each acquire a customer as soon as their shift begins, so fewer departures than service commencements need to occur across the interval, to meet the waiting time target. The following research adapts the approaches followed by Green et al. (2007) and Ingolfsson (2002) to derive distinct waiting time formulae for both dummy and true shift boundaries. The expressions presented are applicable to the case where the number of servers change at most once during the maximal allowed waiting time since this is a valid assumption in the ambulance service where the target response time is measured in minutes and staffing numbers may only change on an hourly basis; and the calculation of the waiting tail probability for situations where the number of servers change more than once can only be approximate, at best (Green et al., 2007).

For the case where the maximum allowed waiting time overlaps two intervals and the number of servers changes exactly once in $[t+\Delta t, t+x]$, there exists some $\Delta \leq x$ such that:

$$s(u) = \begin{cases} s(t)^- & \text{if } u \in [t, t+\Delta t], \\ s(t)^+ & \text{if } u \in [t+\Delta t, t+x]. \end{cases} \tag{6.10}$$

In this case a must be redefined as:

$$a = \mu \int_t^{t+x} s(u)\,\mathrm{d}u, u \in [t, t+x] = \mu s(t)^- \Delta t + \mu s(t)^+ (x - \Delta t) \tag{6.11}$$

to reflect the mean number of departures expected over an interval covered by two different staffing teams.

Note that when $n < \max(s(t)^-, s(t)^+)$, it will always be true that $P(W_q^n(t) > x) = 0$ because the $(n+1)$st customer will begin either begin service immediately (if $s(t)^- > s(t)^+$) or at time $t+\Delta t$ if $s(t)^- < s(t)^+$. When $n \geq \max(s(t)^-, s(t)^+)$, then $P(W_q^n(t) > x)$ will be dependent on the number of servers in time period $[t, t+x]$.

Case A: True shift boundary

At a true shift boundary where an exhaustive discipline applies, all $s(t)^-$ servers employed for the first shift continue to serve any customers they are dealing with at the shift boundary (working overtime), so all customers currently in service are ejected from the system. The $s(t)^+$ new servers concurrently commence work and begin serving customers in the queue immediately (assuming there is no resource shortage resulting from equipment needing to be acquired from a departing server). Thus greater than $n - s(t)^- - s(t)^+$ departures in $[t, t+x]$ would permit an arriving customer to wait less that time x before being served, such that:

$$P(W_q^n(t) > x) = \begin{cases} \sum_{b=0}^{n-s(t)^- - s(t)^+} \frac{a^b e^{-a}}{b!} & \text{if } n \geq s(t)^- + s(t)^+, \\ 0 & \text{otherwise.} \end{cases} \tag{6.12}$$

The formula presented in equation (6.12) remains consistent, no matter if the number of servers increase, decrease or remain the same over the boundary. The formulation of the expression is essentially equivalent to that in Ingolfsson (2002), but specifically concerns evaluating the number of customers in the system, rather than the number in the queue alone, to maintain consistency with the formulae derived in the remainder of this thesis.

Case B: Dummy shift boundary

For the case of the dummy shift boundary, it is necessary to define different formulae for each of the three scenarios (i.e. an increase, decrease or no change in staffing levels over the boundary). All formulae are based on the same expression, but are summed over different quantities, depending on the number of customers that need to be served within the acceptable waiting time threshold x to allow the performance target to be met. These formulations further the research by Ingolfsson (2002) and Green et al. (2007), to incorporate the behaviour emulated by servers at a dummy shift boundary.

Case B1: Number of servers remains consistent

First consider the case where the number of servers remains unchanged over the shift boundary. Since the same servers are employed either side of the boundary and their actions are not affected by its occurrence, the same formula applies as if no boundary was imposed:

$$P(W_q^n(t) > x) = \begin{cases} \sum_{b=0}^{n-s(t)^+} \frac{a^b e^{-a}}{b!} \text{ if } n \geq s(t)^+, \\ 0 \qquad\qquad\qquad \text{otherwise.} \end{cases} \tag{6.13}$$

Note, that since $s(t)^-$ and $s(t)^+$ are equal, they are interchangeable within the expression. Following the same argument, the expression given for a in equation (6.11) is also equivalent to that in equation (6.7).

Case B2: Number of servers is increased

Conversely, if the number of servers increases at time $t+\Delta t$, the waiting tail probability $P(W_q^n(t) > x)$ may be calculated by computing the probability that there are $n - s(t)^+$ departures over $[t, t + x]$, given by:

$$P(W_q^n(t) > x) = \begin{cases} \sum_{b=0}^{n-s(t)^+} \frac{a^b e^{-a}}{b!} \text{ if } n \geq s(t)^+, \\ 0 \qquad\qquad\qquad \text{otherwise.} \end{cases} \tag{6.14}$$

where a is defined as in equation (6.11).

Case B3: Number of servers is reduced

If the number of servers decreases at time $t + \Delta t$, then the event $P(W_q^n(t) > x)$ and the probability that $n - s(t)^-$ or fewer customers leave within time $[t, t + \Delta t]$ are equivalent, as the only way that a customer arriving at time t will wait longer than time x before being served is if there were not enough departures for the customer to have entered service even if the number of customers had not been reduced, with the adjusted departure rate $a = y_0 + y_1$ to account for the fact that there are less servers for time period $(x - \Delta t)$. The event that not enough departures occurred before the shift change and the number of departures after the shift change left more than $s(t)^+$ people in the system at time $t + x$ is impossible, as the additional $(s(t)^- - s(t)^+)$ customers that would not have completed service under the non-exhaustive discipline, can be considered to have been ejected from the system for the exhaustive case. Thus

$P(W_q^n(t) > x)$ may be computed as:

$$P(W_q^n(t) > x) = \begin{cases} \sum_{b=0}^{n-s(t)^-} \frac{a^b e^{-a}}{b!} \text{ if } n \geq s(t)^-, \\ 0 \qquad\qquad\qquad \text{otherwise.} \end{cases} \tag{6.15}$$

The tree diagram presented in Figure 6.2 summarises the results of the analysis presented in Section 6.2.2.2.

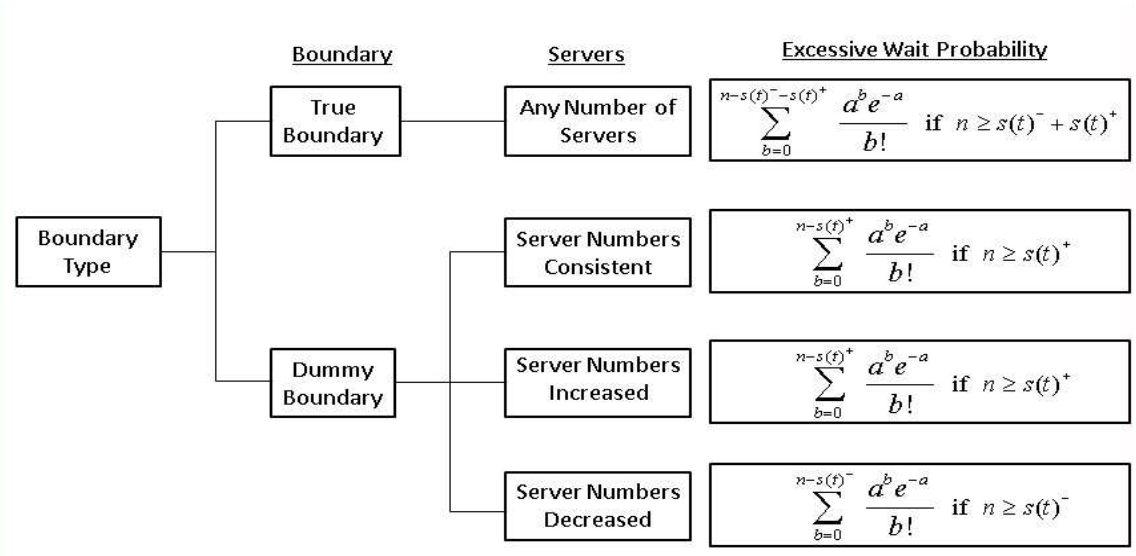

Figure 6.2: Waiting time formulae for various shift boundaries and server numbers

Section 6.3 presents an application of the above theory to the a subsection of the data described in Chapter 2. The SIPP methodology is used as the exemplary approximation approach, and the Euler solver is used to demonstrate the numerical approach throughout. By considering the results obtained from the numerical method as a 'gold standard', variations of the approximation approach are evaluated.

6.3 Application to WAST data (SE region)

This section examines how well SIPP and its variants perform on a model derived from empirical data, by benchmarking its performance against a Euler solver. In order to suitably test the application of the methods discussed above, which are relevant for $M(t)/M/s(t)$ queueing systems, the methods are applied to a subsection of the

full ambulance system at this stage; leaving scope for the body of data they consider to be increased as more complex queues are investigated in latter chapters of this thesis. For example, Chapter 7 introduces priorities to the system, so the quantity of vehicles required to respond to Category A and Category B/C calls may be computed simultaneously.

In addition to evaluating the SIPP approach in terms of its accuracy, case studies are included that reveal greater insights into the numerical methodology through examining the effect of applying different shift boundaries to hourly periods.

An overview of the data of the system that will be modelled by some of the methods developed in this chapter is provided below. The primary aim is to determine the minimum number of paramedics, s, required to operate RRVs throughout the day (assuming that each RRV is operated by a single paramedic), to provide acceptable responses to Category A emergency calls in the SE region of Wales. The precise problem, data inputs and constraints that are used to determine the optimum staffing levels are as follows:

- The investigation considers setting minimum staffing levels for RRVs that allow Target 1 to be attained at a minimum cost within the SE region of Wales for July and December 2009 (i.e. provide minimum coverage levels that enable responses to at least 60% of Category A calls to be achieved within 8 minutes).
- The expected number of Category A emergencies for each period of each day in the scheduling horizon are obtained from SSA forecasts (see Chapters 3 and 4) based on historic demand from 2005.
  - Following the guidance provided in Chapter 4, the first 20 components and a window length of 1,367 are selected to construct the forecasts, and to reduce the risk of understaffing, the predicted counts are uplifted by 10%.
  - Individual counts are predicted for the total number of unique incidents requiring emergency assistance for each shift, pre-defined by WAST as 6am-12pm, 12pm-7pm and 7pm-6am (see Chapter 2.4).
  - SSA is initially applied to the the time series recording all EMS calls, and in order to estimate the proportion of calls per shift that are strictly Category A, the technique is subsequently re-applied to the time series recording the proportion of Category A calls for the period of known historic demand

(with 6 components and a window length of 1367). The forecasted series specifying the expected number of Category A calls in each shift is ultimately obtained by applying the projected proportions to the projected counts.

  - Finally the expected counts per shift are converted to expected counts per hour by analysing the typical distribution of calls for each shift for each weekday throughout 2008 (to gain a relatively recent representation of the system), and applying the average determined proportions to each shift projection. Distinct proportions are specified for each weekday to improve the forecast quality, since a two-way ANOVA reveals that the expected hourly demands are not equal for all weekdays ($p < 0.05$).

- The queueing model is based on the assumption that exactly one RRV is required to attend each Category A incident reported to WAST, and that each RRV is operated by a single paramedic, who operates under the exhaustive discipline.

- Preliminary analysis of 2008 data finds that the average journey time (defined in Chapter 2.3) to Category A incidents in the SE region is notably lower than the average all-Wales 2005-2009 average presented in the same section, and equal to 6.65 minutes. Since the targeted response time for Category A incidents is 8 minutes, the maximum acceptable waiting time x before an RRV is mobilised is considered as $8 - 6.65 = 1.35$ minutes for 60% of emergencies.

- Similar analysis reveals that the average service time is 39.7 minutes.

The final issue to consider is the setting of an appropriate duration for the planning periods that average arrival rates are determined for and staffing requirements are produced. Green et al. (2001) find that the performance of SIPP improves when the planning periods are shortened. Preliminary investigations examining the performance of SIPP against the Euler solver on 2009 Category A data for the SE region for periods of 1-hour, 2-hour, 3-hour durations (in addition to the pre-defined shifts) confirm this theory, leading to the decision to **construct staffing requirements for shifts of 1-hour durations**; since requirements developed for periods of any shorter durations could lead to difficulties when developing practical schedules and assigning staff to shifts in future work. To achieve a fair comparison, the arrival rate function is input to both Euler and SIPP methodologies as a piecewise step function revised at hourly intervals with the average rate for that hour. Within each hour, it is assumed that the time between arrivals follows the exponential distribution with the corresponding mean.

The service rate μ may change as well. However, this is expected to change more slowly than the arrival rate (see Ingolfsson et al. (2007)). The models therefore assume that the service rate is constant to simplify the analysis and limit the number of varying parameters in the study.

Section 6.3.1 describes how the above system may be modelled as a $M(t)/M/s(t)$ system with exhaustive discipline, so the Euler methodology may be employed to find staffing requirements that, if followed, will provide a specified level of service, and the time-varying service level resulting from the given staffing schedule can be evaluated. An implicit assumption in this approach is that the minimum service level must hold for *every* time point in the scheduling horizon (rather than being considered as an aggregate service level that is to be achieved on average across all hourly periods in the planning horizon), to ensure a consistent quality of service is provided.

6.3.1 Numerical requirements

The underpinnings of the Euler methodology have been outlined in Chapter 5.3.2 and Section 6.2.2 above. Assuming that Category A incidents are reported to WAST with rate $\lambda(t)$ at time t, patients served by the first available server (selected at random if more than one server is available), servers have identical capabilities, service rates are independent and exponentially distributed with mean rate $\frac{1}{\mu}$, and that a busy server who is scheduled to leave will complete the current service before leaving, then the system may be modelled as a MDCTMC (see Section 6.2.2). Thus during the intervals $(0, t_1), (t_1, t_2)$ the probabilities $p_n(t)$ satisfy the balance equations given in (5.6), but the state-probability vector $(p_n(t))$ is subject to instantaneous transitions $(p_n(t)) = (p_n(t))^- B(t)$ at shift boundaries (i.e. times where $t = t_z$).

For computational efficiency, the infinite capacity $M(t)/M/s(t)$ system is approximated by a finite equivalent, with a cap of $G = 80$ imposed on the number of patients considered in the system at any specific time instance; and following the discussion surrounding the selection of appropriate calculation intervals presented in Chapter 5.3.2, calculation periods are selected as $\delta_c = 0.04$ hours (i.e. 2.4 minutes) which is a common divisor of the length of the planning periods $\delta_{pp} = 1$ hour. To ensure that dynamic steady state is reached, the service quality is also computed for a one-day warm-up period (using the forecasted demand data for the first day in the scheduling horizon) before meaningful analysis of the system is performed.

This investigation is concerned specifically with calculating the service level $P(W_q(t) > x)$ defined for various types of shift boundaries in Section 6.2.2.2, and limiting the fraction of customers experiencing a wait longer than time $x = 1.35$ minutes (before a RRV is mobilised) to 40%, as specified by the AOF guidelines. The ultimate goal is to find the minimum staffing levels, s, that allow the targeted service level to be met for each hourly period. Ingolfsson (2002) observes that the incorporation of a preemptive or exhaustive discipline drastically affects the performance of the system under time-dependent conditions. Since ambulances cannot be re-routed to attend other incidents until they have completed their current service, the analysis models the EMS using an exhaustive service discipline and develops appropriate formulae to calculate the time-varying virtual waiting time distribution, illustrating that customers receiving service from servers working past their scheduled end times do not influence the waiting times of customers that arrive later.

Whilst the approach followed by Ingolfsson (2002) is appropriate for the purpose of developing staffing levels for pre-defined shifts (at the end of which all crew members exit the system and are replaced by an entirely new set of staff), the period requirements provided for each hour here do not coincide with hourly shifts - they are simply used to construct minimum hourly requirements to be satisfied by a subsequently developed shift schedule. Suitable methods to develop desirable schedules around these requirements are later considered in Chapter 9. Since a large variety of shifts are permitted to be scheduled in this chapter, so paramedics may begin and end their duty at various times throughout the day, it appears most appropriate to apply the exhaustive discipline that arises over a dummy shift boundary for each hourly period. This implies that if the same (or more) staff are needed in the consecutive period, the paramedics will continue their duty as normal, but if less are required then the paramedics that finish their shift operate under the exhaustive discipline. Although this may occasionally provide an artificial representation of the number of ejections (for example, in cases where shifts cannot be selected to provide the exact hourly requirements, so some periods have more than the minimum quantity of staff on duty); it is certainly far more accurate than the application of a truly exhaustive shift boundary which would considerably overestimate the number of ejections by assuming that an entirely new set of staff is employed for every hourly period.

Figure 6.3 shows the effect arising from the application of the dummy and true of shift boundaries upon the recommended staffing levels for hourly periods and set shifts

(taken as the three defined shifts specified for the SSA forecasts at the start of Section 6.3 as Morning, Afternoon and Night) for a typical day in the scheduling horizon, selected as 1st July 2009.

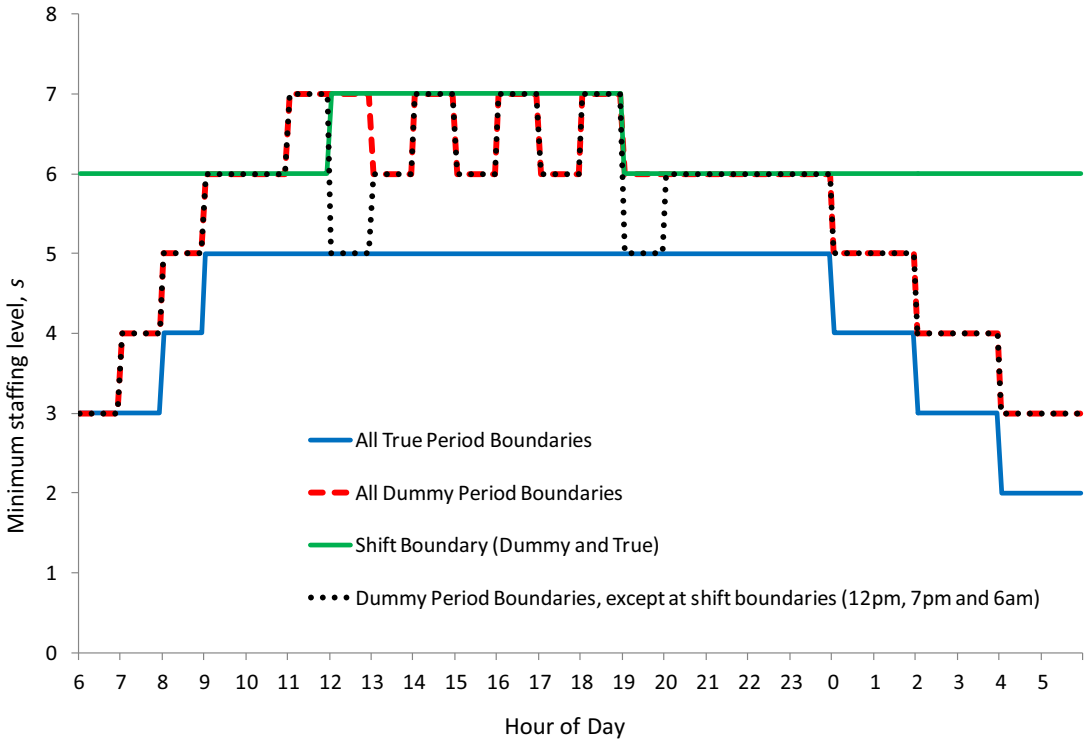

Figure 6.3: Minimum staffing levels suggested by numerical methodology, for various boundary types (01/07/2009)

Figure 6.3 shows the that the incorporation of the truly exhaustive shift boundary can greatly impact on a patients expected waiting time in the system for a short duration after the boundary, as the blue and red lines demonstrate that lower requirements are consistently suggested when a true boundary is applied to hourly periods. This is because a true shift boundary implies that all paramedics exit the system after their one hour stint, but continue to serve the patients they are currently serving until the service is complete, whilst the patients awaiting assistance at the front of the virtual queue are simultaneously served by a new set of staff. Since the mean service times (39.7 minutes) are relatively long compared to the period durations (1 hour), this considerably impacts on the number of servers needed to uphold the targets since the system fails to reach steady state conditions within each period. Lower staffing

quantities are needed primarily because a large proportion of the work is completed by servers working past their specified 'end' times.

However the incorporation of a dummy or true shift boundary has little effect on staffing levels suggested for set shifts which span several consecutive hours, as the impact of the initial conditions for each shift is reduced over time as the system settles down to operate under steady state conditions, resulting in the recommendation of identical staffing levels no matter if a dummy or true shift boundary is applied (shown by the green line). Hence, although it is correct to apply the true shift boundary to such shifts (assuming that the paramedics are each employed for the exact shift durations), the causal effect upon staffing levels is of secondary order.

The staffing levels recommended by the black dotted line in Figure 6.3 are appropriate for the scenario where a base set of staff are employed to work pre-defined shifts, but where additional staff may be added to the base set for certain hourly periods, to allow the demand profile to be matched as closely as possible. Figure 6.3 shows that the consideration of such an instance provides an identical staffing function to the case where dummy boundaries are applied to all periods, except for the period immediately following the application of the true shift boundary. Thus for the particular problem instance investigated, the application of a true shift boundary only impacts greatly upon the system characteristics for the period immediately following it.

The central insight provided by Figure 6.3 is that the incorporation of a dummy or true shift boundary impacts greatly on the resulting staffing profile, when the staffing requirements are produced for hourly periods (or in the more general case, periods that are not considerably longer than the average service time). Assuming that the minimum staffing levels constructed for the hourly periods are ultimately desired to inform the development of appropriate shift schedules, the dummy shift boundary will accordingly be applied to each period in the remainder of the investigations in this thesis, to most closely mirror the true action of paramedics.

The presence of a time-dependent arrival rate makes a numerical analysis of the system computationally intensive; hence approximation techniques are commonly used to analyse such problems, in place of numerical methods. One may note that the system described above possesses low presented loads (around 0.65) and relatively low amplitudes (around 0.5); thus the main elements required for SIPP to perform well, as

discussed in Section 5.3.1, are met. Its potential to construct period requirements for this problem instance is explored is Section 6.3.2; and enhanced with further investigations that examine the potential to improve the forecasts by modifying the arrival rate function and employing some of the model revisions discussed in Chapter 5 (e.g. Lag Avg).

6.3.2 Approximate requirements

This section uses the RRV staffing requirements recommended by the numerical method to benchmark the ability of SIPP and its variants to approximate state probabilities and generate hourly staffing requirements. The main simplifying assumption in the SIPP methodology is that staffing requirements for a particular period can be determined independent of staffing in previous periods, and further details of the technique have been given in Chapter 5 and Section 6.2.1. The case studies investigated in this section recommend the minimum number of paramedics to be employed for each hour of the day for two distinct 28-day periods: one beginning on 1st July 2009 and the other on 1st December 2009.

Prior to discussing the results generated by all methods for longer 28-day periods, the results generated for a specific day (chosen as 06/07/2009: the first Monday of July 2009 in this case) are focused upon, as an opportunity to discuss the suitability of the application of the approximation methods based on the magnitude of the parameter values on a typical day. The staffing levels suggested by SIPP compared to Euler for 6th July 2009 are shown in Figure 6.4. The average arrival rate of unique Category A incidents requiring assistance for this day is forecasted as 6.2 incidents per hour (varying from 2.8 to 7.8), with a service rate of $\mu = 1.5$, $0.57 \leq \rho \leq 0.74$ and $0.32 \leq RA \leq 0.88$ for each 1-hour interval throughout the day. The conclusions reported in Green et al. (2001) suggest that the staffing levels provided by SIPP should be reliable estimations due to the low service rate, low server utilisation and relatively low proportion (60%) of reported incidents that are required to be responded to within the target time. In fact, as the average rates for ρ and μ are so small, they are below the lowest considered by Green et al. (2001), but as the authors found that SIPP reliability is almost always nondecreasing as the service rate increases, the values should ensure an accurate performance. Although there is an undesirable high RA in some periods, the overall performance should still be relatively accurate due to the low service and server utilisation rates.

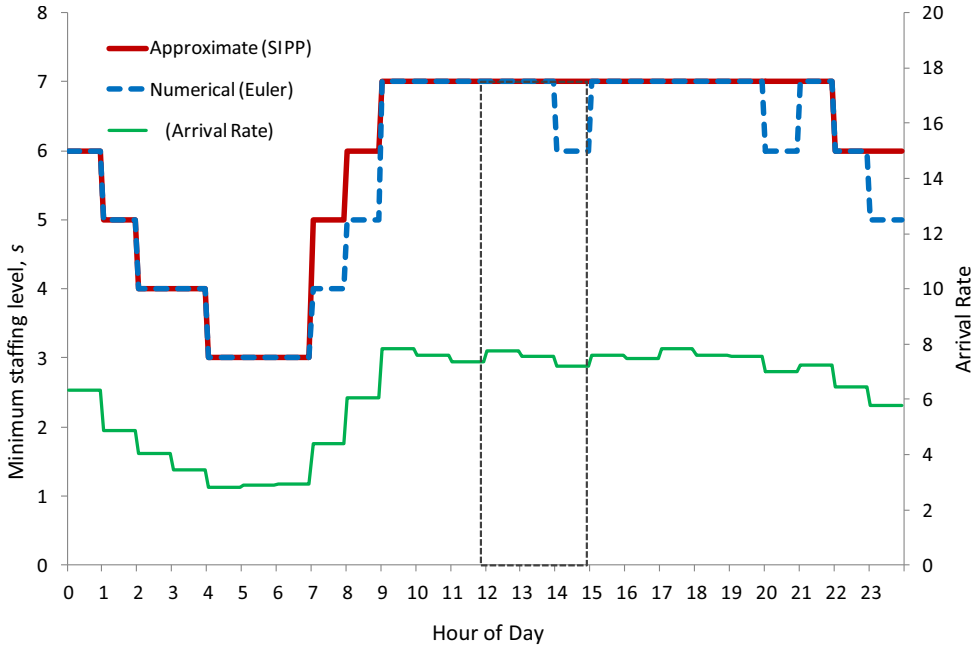

Figure 6.4: Minimum staffing levels recommended per hour (06/07/2009)

Figure 6.4 illustrates that whilst SIPP recommends the same staffing levels as the numerical method for a large proportion of the day, it overstaffs several periods by one paramedic. The main reason for the disparity is that SIPP does not account for the effect of the service level in the previous period upon the current period and assumes the system operates in a steady-state fashion throughout each one-hour period. As such, the methodology fails to recognise that in periods where demand is strictly increasing (e.g. between 7:00-9:00), it takes time for the queue to build up to a level great enough to justify the employment of additional staff. Moreover between 09:00-22:00 SIPP recommends that a constant level of 7 paramedics should be employed; but the exact method recognises that during the first two periods where 7 paramedics are on duty, considerably more than 60% of patients are responded to within the target response time, so less paramedics are required for the following period as there is less congestion in the system at its commencement. Figure 6.5 illustrates how one may compensate for the effective 'overstaffing' of one period (which may considerably affect system performance in practice as staff levels can only be incremented in discrete steps, meaning the minimum service level can be considerably exceeded in some periods), by employing less staff than would otherwise be required in the following period. Figure 6.5 displays the transient probability of an acceptable wait given the time-dependent arrival rates between 12:00-15:00 (the highlighted

period in Figure 6.4) when 7 staff are employed between and 12:00-14:00 and 6 staff between 14:00-15:00 (denoted 7, 7, 6) as recommended by the numerical approach, against the steady-state evaluations provided by the SIPP approach for the same employment of staff and when one additional paramedic is employed in the 14:00-15:00 period (7, 7, 7). The SIPP approach generates requirements independently for each period and thus recommends that 7 staff are employed for the full three hours, since if the system were to operate under steady-state conditions within each hour, only 51% of patients would be reached within the target response times between 14:00-15:00 with a reduced fleet of vehicles. Yet since the transient method recognises that there is little congestion in the system at the start of the 14:00 period, it finds that the service level may be kept just above the required standard throughout the 14:00-15:00 period with only 6 paramedics. Despite these shortfalls, the overall model fit of the SIPP approach is reasonably appealing, given the efficiency of the technique to rapidly approximate requirements. In fact, if the approximate approach is sufficient, considerable time savings can be made as investigations reveal that whilst SIPP can generate minimum hourly staffing requirements for a 3 month horizon within a few seconds (on a 3GHz machine with 2.96GB RAM), the Euler method requires around 6 minutes.

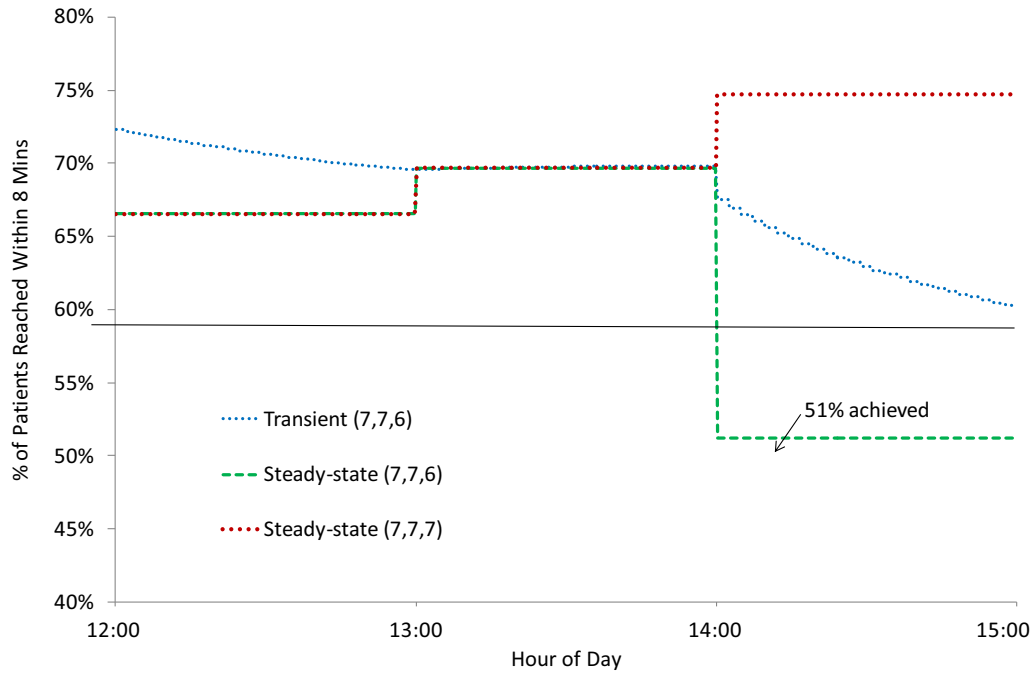

Figure 6.5: Proportion of patients reached within acceptable waiting time

The remainder of this section is dedicated to the consideration of methods used to adjust the arrival rate function prior to the application of the SIPP technique in attempts to improve the approximations.

Revisions investigated for different scenarios by Green et al. (2001) include Lag Avg which is reliable and efficient when relative amplitude is low and planning periods are short; or SIPP Max / Lag Max which provide reliable results in situations where the amplitude is relatively large or planning periods are long. Lag Avg attempts to account for the lag between the arrival rate curve and the probability of delay curve that commonly occurs in operational service systems when service times are longer (so customers being served in period t, may be those that arrived in period $(t-1)$). The idea is to estimate the required staffing level based on an average arrival rate calculated from shifting the arrival rate function by L units (estimated as the average service time) in an attempt to incorporate a suitable estimation of the lag. Green et al. (2001) state that through giving a more realistic view of the system, Lag Avg usually produces better results that standard SIPP, while using the same number of staff hours. Green et al. (2001) also comment that SIPP often leads to understaffing, and suggest that an alternative way to improve it's performance could be to use the maximum arrival rate over each period, rather than the average value in a method known as SIPP Max. Yet, as the small section of data considered in the example above shows that SIPP overstaffs as well as understaffs, this research investigates the SIPP Mix which is a compromise of the standard SIPP and SIPP Max methodologies. It uses the average planning period arrival rate in all periods where the arrival rate is strictly increasing, and the maximum arrival rate otherwise (calculated as 1.2 $\times$ average rate, based on preliminary investigations), to avoid the problem of understaffing.

Table 6.1 evaluates the potential of SIPP, Lag Avg and SIPP Mix methodologies to approximate staffing levels for each hour of each of the first 28 days of July and December i.e. the staffing levels for $24 \times 28 = 672$ hourly periods of each month. It reports the RMSE (defined in Section 4.3.1, where y_n is the requirement given by the exact approach, e_n is the approximation and we have N predicted values) for each of the variants of the standard SIPP technique. Cells where the calculated RMSE is greater than 1 are printed in bold text. All methods perform better for the month of July than December which is subject to more volatile demand. This coincides with the findings in the literature that SIPP performs better for lower

loads, although the planning period duration, service rate and waiting time targets are taken to be identical for both methods, and the relative amplitudes are very similar.

Table 6.1: SIPP, Lag Avg and SIPP Mix accuracy

		RMSE		
Hour	λ	SIPP July/Dec	Lag Avg July/Dec	SIPP Mix July/Dec
0	7.1 / 8.1	0.45 / **2.96**	0.77 / **2.85**	**1.33** / **2.78**
1	6.1 / 6.9	0.33 / 0.63	0.87 / **1.09**	**1.18** / **1.48**
2	4.9 / 5.6	0.27 / 0.33	0.80 / 0.78	0.89 / 0.87
3	4.0 / 4.5	0.42 / 0.00	0.80 / 0.73	0.78 / 0.94
4	3.1 / 3.5	0.46 / 0.57	0.82 / 0.93	0.91 / 0.98
5	2.9 / 3.3	0.27 / 0.27	0.27 / 0.60	0.50 / 0.71
6	3.1 / 3.8	0.63 / 0.38	0.57 / 0.00	0.65 / 0.53
7	4.2 / 5.2	0.71 / 0.50	0.00 / 0.76	0.71 / 0.50
8	5.9 / 7.2	0.80 / 0.98	0.46 / 0.46	0.80 / 0.98
9	7.4 / 9.1	0.78 / 0.93	0.27 / 0.19	0.78 / 0.93
10	7.7 / 9.4	0.63 / 0.60	0.60 / 0.50	0.98 / **1.07**
11	7.8 / 9.5	0.46 / 0.57	0.46 / 0.53	0.93 / **1.17**
12	8.0 / 9.1	0.60 / 0.00	0.38 / 0.46	0.68 / **1.68**
13	7.8 / 8.8	0.50 / 0.27	0.70 / 0.27	**1.32** / 0.96
14	7.8 / 8.8	0.42 / 0.19	0.42 / 0.19	**1.10** / 0.82
15	7.8 / 8.8	0.42 / 0.42	0.38 / 0.42	0.93 / 0.98
16	7.7 / 8.7	0.33 / 0.50	0.50 / 0.50	**1.15** / **1.15**
17	7.7 / 8.7	0.46 / 0.50	0.42 / 0.50	**1.04** / **1.02**
18	7.8 / 8.8	0.38 / 0.33	0.38 / 0.33	0.85 / 0.68
19	7.8 / 9.0	0.53 / 0.65	0.50 / 0.73	**1.00** / **1.21**
20	8.0 / 9.2	0.63 / 0.46	0.46 / 0.46	0.78 / 0.60
21	8.3 / 9.5	0.63 / 0.63	0.50 / 0.60	**1.00** / **1.20**
22	8.0 / 9.1	0.42 / 0.33	0.63 / 0.65	**1.04** / **1.41**
23	7.8 / 8.9	0.50 / 0.50	0.57 / **2.96**	**1.12** / **2.90**
MEAN	6.6 / 7.7	0.52 / 0.96	0.56 / **1.01**	0.96 / **1.23**

The standard SIPP approach provides the best results overall, with identical results to the numerical method in $\frac{489}{672} = 73\%$ of cases for July and $\frac{457}{672} = 68\%$ of cases for December. The periods for which it obtains larger RMSEs unsurprisingly correspond to the periods throughout which the demand steadily increases (around 6am-10am): since SIPP assumes the system reach steady state in each period, it fails to recognise that it takes time for the congestion in the system to build up to a high enough level to justify the employment of an extra paramedic. The technique conversely understaffs the hours near to midnight in December as it fails to account for the congestion in

the system arising between 22:00-00:00 (when it is not unusual to see higher demand levels, likely resulting from Christmas celebrations).

Whilst the SIPP Mix approach produces inferior results to the standard methodology, possibly due to the relatively long service times and tendency to overstaff the hours across which the demand rate falls (as SIPP is sufficient alone to avoid the problem of overstaffing for the majority of cases in this case study); the approximations generated using Lag Avg are very similar to standard SIPP. Lag Avg incorporates a time-lag to account for the disparity between peak arrivals and peak congestion, enabling it to successfully produce more accurate requirements throughout the morning interval, but this is compromised by inferior results for other periods. The reliability of the Lag Avg approach is reduced by the long service times, and it's failure to significantly improve upon the standard results may also be partly attributed to the fact that the relative amplitude is larger than the 'safe' quantity of 0.5 in some periods.

The RMSE penalises overestimation and underestimation equitably. In the EMS industry, it is of utmost importance to have a sufficient fleet of vehicles to allow rapid respond to life threatening emergencies. Thus the weighted RMSE for a given $\tau \in [0,1]_{\mathbb{R}}$ is defined as:

$$\text{RMSE}_\tau = \sqrt{\frac{2}{N}\left((1-\tau)\sum_{y_n<e_n}(y_n-e_n)^2 + \tau\sum_{y_n>e_n}(y_n-e_n)^2\right)} \tag{6.16}$$

This modifies the RMSE measure to weight underpredictions by a factor of τ, and overpredictions by a factor of $(1-\tau)$, so whilst a value between 0.5 and 0.8 may be most desirable for evaluation purposes in the EMS industry, substituting $\tau = 1$ in equation (6.16) only penalises underpredictions and $\tau = 0$ only penalises overpredictions. The standard RMSE may be recovered by taking $\tau = 0.5$ to normalise the measure.

Figure 6.6 presents the overall RMSE_τ for various levels of τ (calculated at 0.05 increments between 0 and 1) for the month of July. It demonstrates that the standard SIPP approach and Lag Avg perform better than SIPP Mix for all values of τ except $\tau = 1$ (i.e. when only underpredictions are penalised, since SIPP Mix inflates the estimated demand it avoids the error of a single underestimation for the entire 28-day prediction period throughout July). The fact that RMSE_τ is a strictly decreasing

function for increasing levels of τ shows that the approximation models overpredict the counts considerably more than they underpredict. The performance of the SIPP and Lag Avg approaches are very similar, but as Lag Avg attempts to account for the lag that exists between peak arrival rates and peak congestion, it understaffs less often. Thus if the goal is to minimise under-staffing the standard SIPP approach is marginally more favorable as it attains a lower RMSE_τ for $\tau < 0.7$.

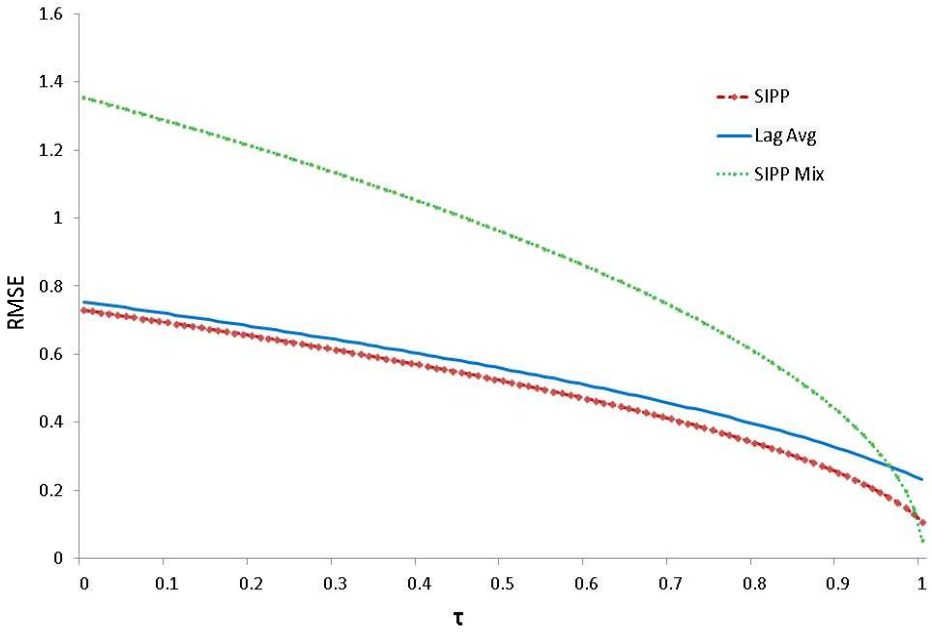

Figure 6.6: Weighted RMSE for various levels of $\tau \in [0, 1]_{\mathbb{R}}$

Figure 6.7 highlights how the behaviour of RMSE_τ changes for different periods, depending on whether SIPP is prone to overestimate or underestimate the staffing level in that period. The first chart refers to the approximations generated for the 00:00-01:00 period. All the approximation methods commonly overstaff this period in July (unlike December), but the methodology followed by SIPP Mix makes it more likely to overstaff, so it is always considered to be the inferior method. Since the approximation methods never understaff this period, they are only equivalent (with no error) when overstaffing is not penalised.

The second chart displays the values calculated for RMSE_τ for the 08:00-09:00 period. Since this period is commonly overstaffed by SIPP and SIPP Mix, the resulting error is decreased as the a more lenient penalisation is applied to overstaffing. However, since

Lag Avg attempts to incorporate for the time-lag that exists between peak demand and peak congestion, it sometimes understaffs this period, and is therefore more favourable if policy makers decide to penalise overstaffing more harshly than understaffing.

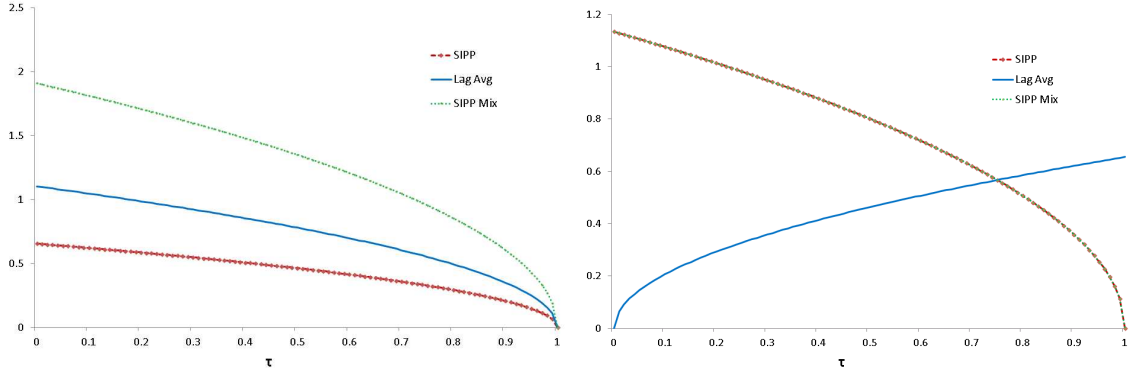

Figure 6.7: Weighted RMSE for various levels of $\tau \in [0,1]_{\mathbb{R}}$ at 00:00 and 08:00

A recognition that customers arriving in a particular period may affect subsequent periods motivated the development of the revisions suggested above. However, the inability of both revisions investigated to improve upon the results of the standard SIPP approach suggests that a slightly different adjustment is needed for the data employed in this case study. The above analysis demonstrates that queues often build up at a slower rate than SIPP expects (under the steady-state assumption) during the periods in which demand is significantly increasing, and that less congestion at the start of periods when demand is decreasing means that fewer staff than otherwise estimated can be required in the following period, especially when demand is decreasing. These observations suggest specific revisions related to such characteristics may be more appropriate.

In light of the analysis, the following revisions are investigated for their potential to improve the SIPP technique:

i. Producing a modified arrival rate for each period where the original rate differs by more than 20% to the rate expected in the preceding period. Two variants of the arrival rate are considered: one resulting from taking an average of the current and preceding period, and another calculated as a weighted average (25%/75% in favour of the current period). This revision aims to incorporate the effect of previous arrivals in each period, and avoid the problem of over/understaffing

where the approximation methods fail to recognise that it takes time for the queue to increase/decrease significantly.

ii. Given the staffing levels as output by the original SIPP model, scaling the arrival rate within each period where the expected proportion of patients seen within the acceptable time is considerably greater than the minimum proportion required, before re-running the model to provide revised requirements. For the data employed in this case study, the arrival rates are scaled by a factor of 0.9 for periods where over 70% of patients are seen within the acceptable time following preliminary investigations. Given the discussions surrounding Figure 6.5 it appears logical to only investigate if the reduced demand rate can be satisfied with one less paramedic if there is little congestion in the pervious period. However, adding this constraint to the model produces marginally inferior results for the test data.

iii. Reducing the staffing levels output by the original SIPP model by one member in periods where the expected proportion of patients seen within the acceptable time is considerably greater than the minimum proportion required. Again, for the case study data it appears sensible reduce the requirement by one paramedic if over 70% of patients are expected to be reached within the acceptable waiting time. Consideration is also given to the expected congestion in the system in the previous period, but find that this extra consideration fails to improve the model performance for the test data.

The results of the above revisions are summarised in Table 6.2. The parameters that are chosen for each of the revisions in this table are those that provide the best results, selected from a wider set of parameters initially considered. The 'SIPP' column serves as a reminder of the RMSEs achieved by the standard SIPP approach, and the remaining columns present the revised RMSEs for the variants of SIPP described above.

Table 6.2: SIPP Revised Reliability (RMSEs)

Hour of day	SIPP	If 20% change in λ		SIPP re-run		Modified SIPP	
		$\lambda' = 0.5\lambda(t-1) + 0.5\lambda(t)$	$\lambda' = 0.25\lambda(t-1) + 0.75\lambda(t)$	If $P(W_q(p) > x) > 0.7$, $\lambda' = 0.9\lambda$	If $P(W_q(p) > x) > 0.65$ and $P(W_q(p-1) > x) > 0.65$, $\lambda' = 0.9\lambda$	If $P(W_q(p) > x) > 0.7$, RRVs' = RRVs $-$ 1	If $P(W_q(p) > x) > 0.65$ and $P(W_q(p-1) > x) > 0.7$, RRVs' = RRVs $-$ 1
0	0.45 / 2.96	0.45 / 2.96	0.45 / 2.96	0.45 / 2.93	0.42 / 3.19	0.49 / 2.97	0.59 / 3.22
1	0.33 / 0.63	0.53 / 0.89	0.42 / 0.80	0.60 / 0.33	0.46 / 0.57	0.68 / 0.57	0.63 / 0.87
2	0.27 / 0.33	0.63 / 0.68	0.46 / 0.60	0.57 / 0.57	0.46 / 0.65	0.68 / 0.65	0.68 / 0.53
3	0.42 / 0.00	0.57 / 0.53	0.53 / 0.38	0.46 / 0.53	0.46 / 0.46	0.82 / 0.84	0.50 / 0.68
4	0.46 / 0.57	0.63 / 0.91	0.50 / 0.82	0.38 / 0.00	0.38 / 0.19	0.63 / 0.53	0.85 / 0.50
5	0.27 / 0.27	0.27 / 0.46	0.27 / 0.27	0.42 / 0.53	0.38 / 0.38	0.78 / 0.76	0.60 / 0.63
6	0.63 / 0.38	0.57 / 0.00	0.57 / 0.00	0.38 / 0.87	0.42 / 0.00	0.60 / 0.78	0.65 / 0.71
7	0.71 / 0.50	0.27 / 0.00	0.46 / 0.00	0.38 / 0.60	0.42 / 0.00	0.27 / 0.57	0.46 / 0.76
8	0.80 / 0.98	0.00 / 0.19	0.60 / 0.63	0.42 / 0.46	0.57 / 0.60	0.27 / 0.46	0.60 / 0.68
9	0.78 / 0.93	0.27 / 0.38	0.60 / 0.85	0.00 / 0.65	0.33 / 0.53	0.00 / 0.65	0.50 / 0.42
10	0.63 / 0.60	0.63 / 0.60	0.63 / 0.60	0.38 / 0.00	0.46 / 0.33	0.38 / 0.00	0.65 / 0.42
11	0.46 / 0.57	0.46 / 0.57	0.46 / 0.57	0.53 / 0.19	0.60 / 0.46	0.53 / 0.19	0.65 / 0.46
12	0.60 / 0.00	0.60 / 0.00	0.60 / 0.00	0.33 / 0.38	0.57 / 0.38	0.33 / 0.38	0.71 / 0.50
13	0.50 / 0.27	0.50 / 0.27	0.50 / 0.27	0.27 / 0.63	0.19 / 0.65	0.27 / 0.63	0.53 / 0.38
14	0.42 / 0.19	0.42 / 0.19	0.42 / 0.19	0.42 / 0.68	0.50 / 0.91	0.42 / 0.68	0.63 / 0.65
15	0.42 / 0.42	0.42 / 0.42	0.42 / 0.42	0.50 / 0.57	0.53 / 0.80	0.50 / 0.57	0.63 / 0.63
16	0.33 / 0.50	0.33 / 0.50	0.33 / 0.50	0.50 / 0.76	0.50 / 0.87	0.50 / 0.76	0.53 / 0.78
17	0.46 / 0.50	0.46 / 0.50	0.46 / 0.50	0.50 / 0.71	0.53 / 0.78	0.50 / 0.71	0.53 / 0.68
18	0.38 / 0.33	0.38 / 0.33	0.38 / 0.33	0.38 / 0.57	0.46 / 0.89	0.38 / 0.57	0.60 / 0.80
19	0.53 / 0.65	0.53 / 0.65	0.53 / 0.65	0.42 / 0.19	0.53 / 0.38	0.42 / 0.19	0.57 / 0.68
20	0.63 / 0.46	0.63 / 0.46	0.63 / 0.46	0.60 / 0.65	0.63 / 0.63	0.60 / 0.65	0.71 / 0.68
21	0.63 / 0.63	0.63 / 0.63	0.63 / 0.63	0.42 / 0.19	0.50 / 0.42	0.42 / 0.19	0.65 / 0.60
22	0.42 / 0.33	0.42 / 0.33	0.42 / 0.33	0.57 / 0.53	0.57 / 0.50	0.57 / 0.53	0.81 / 0.53
23	0.50 / 2.71	0.50 / 2.71	0.50 / 2.71	0.57 / 2.99	0.54 / 2.71	0.63 / 2.99	0.60 / 2.73
MEAN	0.52 / 0.96	0.49 / 0.96	0.50 / 0.96	0.45 / 0.98	0.48 / 1.02	0.52 / 1.03	0.63 / 1.06

Surprisingly, Table 6.2 demonstrates the consideration of the expected wait in the previous period fails to improve the results, and that the adjustment of the methodology based on the proportion of patients seen within the maximal allowed waiting time being greater than 70% (revision (ii) above) improves the SIPP approach most. However, method (i) also produces results of an identical/ marginally improved accuracy. The performance of all the methods appears strongly dependent on the particular dataset, and no revision succeeds to improve the accuracy of the predictions for December, so there is not strong enough evidence to support the use of a particular revision over another. Since the SIPP technique is an approximation method, it will always produce erroneous results for certain periods, and in particular those where the dependence on the behaviour of the system at the end of the previous period has a considerable impact on the performance of the current period.

The particular datasets investigated in this section possess low presented loads (around 0.65) and relatively low amplitudes (around 0.5); thus they exhibit the main characteristics required for SIPP to perform well (Green et al., 2001). Tables 6.1 and 6.2 demonstrate that the standard methodology provides fairly reliable requirements for the given demands, suggesting that if the main assumptions of SIPP are met, revisions of the technique are unlikely to generate significantly improved predictions. If more a more accurate staffing profile is desired under such a scenario, this research recommends consideration of numerical methodology.

6.4 Summary

Motivated by a problem facing WAST to set staffing levels that provide adequate responses to Category A incidents arising within SE Wales, this chapter has considered methods capable of finding staffing requirements for $M(t)/M/s(t)$ systems that, if followed, provide a specified level of service. The primary performance measure for responses provided to Category A emergencies at WAST is the fraction of calls reached within 8 minutes. Since the performance of this service level cannot be evaluated over time by means of a closed-form formula, the research has considered numerical (Euler) and approximation (SIPP) methods that are commonly used in the literature to analyse time-dependent systems. Numerical methods can provide high degree of accuracy at the expense of computation speed whilst approximation methods are fast, but often inaccurate. The accuracy challenges associated with each approach have been highlighted, and extensions suggested to the models that contribute to the

literature of time-dependent queues and promote improved analysis of such systems over time.

The chapter began with an overview of the SIPP methodology, which included an outline of the main assumptions required by the technique to approximate the behaviour of time-dependent systems. Despite its widespread use (see Kolesar et al. (1975); Green and Kolesar (1991); Green et al. (1991); Green and Kolesar (1997); Green et al. (2001, 2006, 2007) and Ingolfsson et al. (2007)), SIPP does not always produce reliable requirements as it assumes that staffing may be determined for consecutive periods independently of each other, the system reaches steady-state conditions within each planning period and the arrival rate remains consistent within the planning period. (This third assumption also holds for the way the Euler methodology has been implemented in this chapter; thus no errors highlighted in the empirical analysis should be attributed to this factor). Hence SIPP should only be used in restricted situations, such as to determine staffing levels in systems comprised of short planning periods with low presented loads and low relative amplitudes. However in situations where the assumptions are met, the benefit of the approximate approach is considerable as it can provide a fairly accurate staffing profile at a rapid rate; for example, whilst the Priority Euler method requires around 100 minutes to generate minimum hourly staffing requirements for a 3-month horizon on a 3GHz machine with 2.96GB RAM, Priority SIPP can offer an approximate solution in around 10 minutes.

The main contribution of this chapter lies in the in-depth analysis presented in relation to the numerical Euler methodology. The research defines novel shift boundaries that allows for the appropriate tracking of the system behaviour across dummy shift boundaries (where a proportion of the staff may exit the system) and true shift boundaries (where all staff end their duty and are replaced by an entirely new set of employees). Case studies are included that demonstrate the effect of the application of the various type of boundaries to hourly periods and set shifts, illustrating that the incorporation of the varying boundary types considerably impacts on the performance of the system immediately after the boundary, but that the effect weakens over time as steady state conditions are approached. Considerable research is also devoted to the calculation of the probability of an excessive delay, and adjustments of the formula needed to compute this measure are derived which allow the accurate tracking this performance measure over shift boundaries. The probability of an excessive delay is the primary measure of interest to many operational service

systems today, and whilst it has received much interest in the literature (Ingolfsson, 2002; Ingolfsson et al., 2007; Green et al., 2007; Izady and Worthington, 2012), this study represents the first time that appropriate revisions have been proposed for dummy shift boundaries where staff operate under the exhaustive discipline. Whilst the performance has been heavily studied in time-dependent systems, equivalent formulae has not been developed for commensurate purposes in priority queueing systems. Thus the methods discussed in this chapter serve as a precursor to Chapter 7 which details extensions that can be applied to each of the approaches to determine staffing levels in priority systems, with particular attention devoted to the evaluation of an excessive wait over shift boundaries; so the minimum number of ambulances required to respond to Category A and Category B calls can be computed simultaneously.

When selecting an appropriate method to construct period requirements, practitioners will likely want to know if the method assumptions are appropriate for the system, how difficult the method is to implement and how accurate the method is. This chapter has provided guidance on such issues and demonstrated that whilst SIPP can generate reasonable predictions with little computation time and effort for some systems, it is imprecise for some parameter values. Since the case studies have demonstrated that the technique occasionally generates requirements that considerably exceed the the targeted performance level, the headline message to managers determining staffing levels according to the SIPP methodology is to proceed with caution. If the assumptions of SIPP are not met or poor requirements are generated, several simple modifications that can be applied to the approach may improve its performance (see Table 6.2). Revisions such as Lag Avg, SIPP Mix, SIPP Max or Lag Max can also be used; but are unlikely to improve the accuracy of the standard SIPP predictions if the system already possesses the components for which are expected to allow the technique to perform well. When deciding the most appropriate revision, consideration should be given to particular system characteristics. Whilst Lag SIPP is commonly found to improve SIPP performance due to it's incorporation of the time-lag that commonly exists between peak arrival and peak congestion in service systems, it fails to significantly improve the results investigated in the case study (unless the primary goal is to minimise overpredictions).

Whilst this chapter has outlined many of the fundamental issues relating to the evaluation of time-dependent queueing systems, it leaves several issues unresolved that will be explored in the remainder of this thesis. Firstly, the techniques are

extended to cope with priority queueing systems in Chapter 7, in order to provide minimum period requirements that will ultimately be used to inform the development of optimal shift patterns and translated into actual work schedules in Chapter 9. It is undoubtedly essential to schedule employees in an optimal fashion at WAST where it is of upmost importance to have an adequate response team to rapidly respond to any life-threatening emergency (supporting the argument to consider the weighted RMSE_τ measure when evaluating the models). Chapter 7 explores the development of meta-heuristic approaches to obtain staffing rosters that match the staffing requirements generated by the techniques discussed throughout this chapter to optimise resource allocation within WAST. This is achieved using a two-step approach where shifts are first scheduled in an optimal fashion and employees are subsequently assigned to shifts.

Chapter 7

Computing service levels in $M(t)/M/s(t)/NPRP$ systems

7.1 Introductory remarks

This chapter demonstrates how the approximate and numerical time-dependent queueing theory techniques discussed in Chapter 6 may be extended to model a multi-server priority system with a time-varying arrival rate, which has only been investigated in a restricted capacity to date. Whilst the characteristics of a single server priority system have been well understood since the 1950's (Cobham, 1954), its multi-server counterpart is far less tractable for reasons described in Chapter 5. Most papers analysing multi-server priority queues are based on two priority classes with a Poisson arrival process and exponential service times; and the focus is generally on how the mean response times vary amongst the priority classes compared to those in single server systems under steady-state conditions. This research responds to the need for methods to evaluate further performance measures in time-dependent priority systems, and in particular the waiting time distribution. Expressions are developed that provide the probability of an excessive wait for each class of customer, and particular devotion is awarded to the examine the change in this probability, as servers enter and leave the system at shift boundaries.

This chapter considers a time-dependent dual-class priority service system with s servers and an unrestricted waiting line, as illustrated in Figure 7.1. Customers are triaged instantaneously when they arrive and processed according to the head-of-the-line priority rule (see Chapter 5.4), so HP items join the front of the a queue if there is

no server available upon arrival, to be seen by the first available server. HP customers arrive according to a Poisson process with rate λ_H and LP customers arrive with rate λ_L; so the rate of customers arriving for service is $\lambda = \lambda_H + \lambda_L$. Service times are independently and exponentially distributed (not class-dependent) with mean time $\frac{1}{\mu}$. All servers have identical capabilities, operate under the exhaustive service discipline, and if multiple servers are available to process a job, each available server has equal an probability of taking on this job.

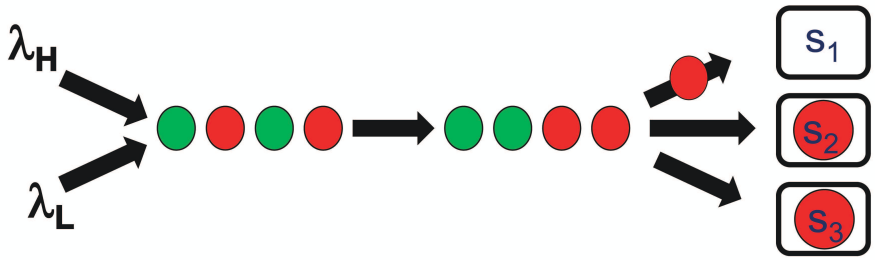

Figure 7.1: A schematic diagram of the priority queueing system

The research performed in this chapter is motivated by a problem facing WAST to set staffing levels for EAs in the Cardiff area (a subsection of the SE Region). EAs are required to serve both HP patients (those with Category A life-threatening conditions) and LP patients (those exhibiting Category B/C type injuries) and WAST is interested in determining the number of crews necessary to to ensure that 95% of HP and LP patients are reached within 14 minutes by an EA i.e. Targets 1 and 2 outlined in Chapter 1. The current literature only allows an approximate answer to this question, as the numerical techniques described in Chapter 6, which accurately track the behaviour time-dependent systems over time and allow evaluation of the probability of an excessive wait over shift boundaries, have seemingly not been extended to priority queueing systems. This chapter accordingly considers how the techniques can be adjusted to determine the probability of an excessive wait in generic time-dependent priority queueing systems in Figure 7.1, before applying the developed methodology to the particular problem facing WAST.

The analysis is structured in the following way. First, Section 7.2.1 considers appropriate extensions to the SIPP approximation technique and demonstrates how the probability of an excessive wait may be computed under steady-state conditions for two customer classes. Complementing this research, Section 7.2.2 presents extensions

to the Euler numerical method described in Chapter 6, through constructing balance equations to accurately track the movement of customers through the system and devoting close attention to the adjustments necessary to account for the behaviour of departing servers over shift boundaries. Since the presence of a time-dependent arrival rate makes numerical analysis of the system computationally expensive, methods to increase the efficiency of the Euler solver in finding the minimum quantity of staff necessary to keep the expected proportion of unacceptable waits below a given threshold are considered in Section 7.2.3. The techniques are ultimately applied to the particular WAST scenario described above in Section 7.3, which includes an evaluation of the methods in terms of their accuracy and efficiency. A short summary of the chapter concludes this analysis in Section 7.4.

7.2 Approximation and numerical methods

This section presents extensions that can be applied to the approximation and numerical methods considered in Chapter 6 to enable their employment within priority systems subject to time-dependent demand.

7.2.1 Approximation methodology

Since exact analysis of systems with time-varying demand is extremely difficult, approximation methods are commonly used to approximate the time-dependent behaviour. The SIPP methodology can easily be extended to approximate the time-varying behaviour of a dual-class priority queueing system by computing stationary measures in a set of stationary systems which are subsequently adjoined by the technique, as outlined in the following steps:

i. Segment the time period into distinct intervals

ii. Find the average arrival rate of HP and LP customers within each interval

iii. Assume the system reaches steady-state within each interval, so each interval may be modelled as a $M/M/s/NPRP/\infty/\infty$ system

iv. Use mathematical expressions to evaluate performance measures in each interval and use these to set staffing levels based on system quality

Hence by assuming that the behaviour of the system in consecutive intervals is statistically independent and that the system reaches steady state within each one (so

$\rho_H + \rho_L < s$); stationary measures may be used to approximate the system behaviour and recommend minimum staffing levels that ensure that given performance metrics are kept below acceptable thresholds. Whilst equivalent methodology has been considered to set staffing levels in Chen and Henderson (2001), the paper did not distinguish the application of the approach from its application within $M/M/s/FIFO/\infty/\infty$ systems. To avoid confusion with the application of SIPP in Chapter 6, the approach outlined in the steps above shall be referred to as '**Priority SIPP**' herein.

In cases where the performance measure of interest relates to overall system performance, in place of targets distinctly specified for HP and LP customers, mathematical expressions that evaluate stationary performance measures in simpler $M/M/s/FIFO/\infty/\infty$ systems are sufficient to evaluate the overriding steady-state performance of equivalent $M/M/s/NPRP/\infty/\infty$ systems. The total number of customers in the system may of course be computed from the state probability vector for a non-priority system (where $p_n, n = 0, 1, ...$ is given in equation (5.3)). Closed-form formulae to evaluate other system characteristics of $M/M/s/FIFO/\infty/\infty$ systems also hold and may be evaluated by substituting the value for λ with $\lambda_H + \lambda_L$. However, if a particular level of service quality is specified for either HP and/or LP customers, the expressions necessary to evaluate performance measures are generally more complex.

Expressions that provide the expected total quantity of HP and LP customers in priority service systems have been developed using various approaches (see Chapter 5). Knowledge of this basic summary measure aids evaluation of further performance measures, such as the probability that a customer experiences an excessively long wait in the queue. The expected number of customers of each class in the queue may be evaluated using generating functions, described for a head-of-the-line priority queueing system in equation (7.1). Further details regarding the derivation of the formulas are provided in Cohen (1956) and the notation used below closely follows that given within the paper.

Let $g(h, l)$ denote the probability that there are h HP customers and l LP customers in the queue. If all servers are busy, the probability generating function $F_y(d)$ of state $g(h, l)$ is therefore defined as $F_y(d) = \sum_{l=0}^{\infty} g(h, l)d^l$ and given by:

$$F_y(d) = \gamma_1^y \, \frac{(s-r)E_s(r)}{(s-r)+rE_s(r)} \, \frac{1-d}{1-\gamma_2 d} \tag{7.1}$$

where

$$\left.\begin{matrix}\gamma_2\\ \gamma_1\end{matrix}\right\} = \frac{1}{2}\left[1 + \frac{r_H}{s} + \frac{r_L}{s}(1-d) \pm \left[\left\{1 + \frac{r_H}{s} + \frac{r_L}{s}(1-d)\right\}^2 - 4\frac{r_H}{s}\right]^{1/2}\right],$$

$$N_s(r) = \sum_{i=0}^{s} \frac{r^i}{i!},$$

$$E_s(r) = \frac{r^s/s!}{N_s(r)},$$

$$r_H = \frac{\lambda_H}{\mu}, r_L = \frac{\lambda_L}{\mu} \text{ and } r = r_H + r_L$$

Using equation (7.1), Cohen (1956) shows that the probability that all servers are busy may be calculated as:

$$P(\text{All servers busy}) = \frac{sE_s(r)}{(s-r)+rE_s(r)} \tag{7.2}$$

and if not all servers are busy, the probability of the number of customers in the system may be calculated from:

$$p_n = \frac{(s-r)}{(s-r)+rE_s(r)} \times \frac{r^n/n!}{N_s(r)} \tag{7.3}$$

The average waits in the queue, $\overline{W_{qH}}$ and $\overline{W_{qL}}$, for HP and LP customers respectively, are given as:

$$\begin{aligned}\overline{W_{qH}} &= \frac{\mu}{s-r_H}P(\text{All servers busy})\\ \overline{W_{qL}} &= \frac{s\mu}{(s-r_H)(s-r)}P(\text{All servers busy})\end{aligned} \tag{7.4}$$

Equations (7.2 - 7.4) may all be embedded into the Priority SIPP methodology proposed above to approximate the performance of time-dependent priority systems. However, priority systems are often evaluated by more complex performance metrics than those such as the *average* waiting time, and closed-form expressions only exist for a limited set of performance standards. Thus whilst Priority SIPP may perform efficiently for some systems, the computational cost for its execution is higher for systems such as WAST, where quality is measured as a function of the waiting time distribution. The following section nevertheless proposes how Priority SIPP may be still employed in such systems. It outlines the previous research performed in this area and derives formulae that allows computation of the probability of an excessive wait for head-of-the-line priority systems.

7.2.1.1 Virtual waiting time distribution

For the case where the service rates are identical across classes of customers, Davis (1966) derived LSTs for the waiting time distribution of customers of different classes (see Chapter 5). Kella and Yechiali (1985) have since shown that the LST for the waiting time presented for HP customers may be inverted. Using W_{qH} to denote the time that a HP customer waits in the queue before commencing service, the probability this quantity is greater than the acceptable waiting time, x_H, may be simply expressed as:

$$P(W_{qH} > x_H) = P(\text{All servers busy})e^{-(s\mu-\lambda_H)x_H}. \tag{7.5}$$

Yet the equivalent inversion is analytically intractable for LP customers. Whilst Abate and Whitt (1995) show that this quantity may be determined using a numerical transform inversion, the calculations require considerable computational resources and are thus unsuitable to be embedded within SIPP, which seeks to provide a quick and efficient approximation.

Chen and Henderson (2001) suggest that simple inequalities may be used to obtain a bound on waiting time performance for LP customers, for example the probability that an LP customer waits less than time x_L in the queue, $P(W_{qL} > x_L)$, may be bounded using:

$$P(W_{qL} > x_L) \leq \min\left(\frac{\overline{W_{qL}}}{x_L}, \frac{\overline{W_{qL}}^2}{x_L^2}\right) \tag{7.6}$$

This bound provides a conservative estimate of waiting time in the queue; thus if implemented as part of a wider SIPP model to evaluate performance measures and recommend minimum staffing levels, it will provide staffing levels that will ensure the required performance target will be certainly met in $M/M/s/NPRP/\infty/\infty$ service systems, given the assumptions of SIPP are met. However whilst the staffing levels it recommends will always be sufficient, they may be higher than necessary. The research contained in the remainder of this section provides derivations of alternative formulations to compute the proportion of HP and LP customers waiting excessive times in the queue. By means of computing the expected number of customers of each class in the queue, and using equations (7.7) and (7.10), the risk of setting staffing schedules with unnecessarily high staff quantities which may be result from equation (7.6) is overcome.

The developed formulae are extensions of the basic expression given in equation (6.1), which provides the probability of an excessive wait for customers in $M/M/s$ systems. The key observations that enable the expression to be extended to priority systems

are that an arriving HP customer is effectively presented with a $M/M/s$ queue with $\lambda = \lambda_H$ (as its priority status allows it to advance in front of any LP customers currently waiting in the queue, so is unaffected by their presence); and an arriving LP customer sees a $M/M/s$ queue where the s servers must first serve all of the customers currently in the system, in addition to any HP customer arriving while the LP customer is awaiting service.

Hence the probability of an excessive wait for a HP customer may be computed by considering the number of HP calls in the system and the number of LP customers in service (i.e. it is not necessary to account for LP customers in the queue as HP customers are served before of these, so their presence has no effect on the system behaviour). Letting $\tilde{n}$ represent the cumulative number of LP customers in service and HP customers in the system (i.e. all customers in the system excluding LP customers in the queue), $p_{\tilde{n}}$ denote the probability that the system is in each state, and $W_{qH}^{\tilde{n}}$ denote the waiting time for HP customers that arrive to find $\tilde{n}$ people in the system ahead with s servers on duty; the probability that a HP customer waits longer than the acceptable time in the queue is given by:

$$P(W_{qH} > x_H) = \sum_{\tilde{n}=s}^{\infty} P(W_{qH}^{\tilde{n}} > x_H) p_{\tilde{n}} \tag{7.7}$$

$p_{\tilde{n}}$ may be evaluated using equation (7.3) for $\tilde{n} < s$, and the generating function in equation (7.1) for $\tilde{n} \geq s$; and $P(W_{qH}^{\tilde{n}} > x_H)$ can be evaluated for each $\tilde{n}$ using the formula presented in equation (6.9), adjusted for a stationary system i.e.:

$$P(W_{qH}^{\tilde{n}} > x_H) = \begin{cases} \sum_{b=0}^{\tilde{n}-s} \frac{a^b e^{-a}}{b!} & \text{if } \tilde{n} \geq s, \\ 0 & \text{if } \tilde{n} < s. \end{cases} \tag{7.8}$$

where $a = \mu s x_H$ since a HP customer will wait greater than the acceptable waiting time threshold x_H if there are $\tilde{n} - s$ or fewer departures over this interval.

Combining these results yields:

$$P(W_{qH} > x_H) = \begin{cases} \sum_{\tilde{n}=s}^{\infty} \sum_{b=0}^{\tilde{n}-s} \frac{a^b e^{-a}}{b!} p_{\tilde{n}} & \text{if } \tilde{n} \geq s, \\ 0 & \text{if } \tilde{n} < s. \end{cases} \tag{7.9}$$

A similar approach can be followed to compute the waiting tail probability for LP

customers, but for such customers it is necessary to consider the probability distribution of the number of HP arrivals during the acceptable waiting time threshold, in addition to the number of LP and HP customers in the system since any HP customers arriving within this interval are required to be served in advance of the arriving LP customer. Thus letting W_{qL}^n denote the waiting time for LP customers that arrive to find n people in the system ahead with s servers on duty; $P(W_{qL} > x_L)$ may be computed as:

$$P(W_{qL} > x_L) = \sum_{n=s}^{\infty} P(W_{qL}^n > x_L) p_n \tag{7.10}$$

where

$$P(W_{qL}^n > x_L) = \begin{cases} \sum\limits_{f=0}^{\infty} P(f \text{ HPs arrive in } x_L) \times \sum\limits_{b=0}^{n-s+f} \frac{a^b e^{-a}}{b!} & \text{if } n \geq s, \\ 0 & \text{if } n < s. \end{cases} \tag{7.11}$$

Here $a = \mu s x_L$ and since HP customers are assumed to arrive in a Poisson fashion, the probability of f HP arrivals in time x_L may be calculated as:

$$P(f \text{ HPs arrive in } x_L) = \frac{(\lambda_H x_L)^f e^{-(\lambda_H x_L)}}{f!} \tag{7.12}$$

Thus combining these results yields:

$$P(W_{qL} > x_L) = \begin{cases} \sum\limits_{n=s}^{\infty} \left(\sum\limits_{f=0}^{\infty} \frac{(\lambda_H x_L)^f e^{-(\lambda_H x_L)}}{f!} \sum\limits_{b=0}^{n-s+f} \frac{a^b e^{-a}}{b!} \right) p_n & \text{if } n \geq s, \\ 0 & \text{if } n < s. \end{cases} \tag{7.13}$$

Whilst the expressions presented for HP and LP customers in equations (7.9) and (7.13) are logical formulations of the computations required to evaluate the probability of excessive waits, the calculation for HP customers requires knowledge of the number of HP or LP customers in the queue, which in turn necessitates consideration of the generating function given in (7.1). Thus whilst it is useful to appreciate the formulation of the metric in this context, the closed-form version of the formula given in equation (7.5) requires less computational effort, and is accordingly employed in the Priority SIPP methodology herein.

Contrastingly, the equivalent LP formula given in equation (7.13) only requires knowledge of the probability vector denoting the total number of customers in the system (which may be evaluated using the steady-state probabilities for a stationary system provided in (5.3)) and the evaluation of equation (7.11). Thus the additional computational effort needed to evaluate the probability accurately is negligible compared to effort required to compute the bound proposed by Kella and Yechiali (1985) (see equation (7.6)).

In summary of the above analysis, balancing the need to achieve reasonably accurate analysis in a computationally efficient manner, the equations that shall be used within the Priority SIPP methodology to calculate the probability of an excessive wait for HP and LP customers (to ultimately find the minimum staffing levels necessary to achieve specific service targets) are equations (7.5) and (7.13), i.e.:

$$P(W_{qH} > x_H) = P(\text{All servers busy})e^{-(s\mu-\lambda_H)x_H}.$$

and

$$P(W_{qL} > x_L) = \begin{cases} \sum_{n=s}^{\infty}\left(\sum_{f=0}^{\infty} P(f \text{ HPs arrive in } x_L) \times \sum_{b=0}^{n-s+f} \frac{a^b e^{-a}}{b!}\right) p_n & \text{if } n \geq s, \\ 0 & \text{if } n < s. \end{cases}$$

Although inclusion of the revised LP formula means the SIPP approach is slightly less efficient, it still possesses the key benefits that it requires less computation than the numerical technique, and allows the system performance to be computed efficiently at any independent time point, in place of the requirement to continually track the system behaviour over time. Thus whilst the implementation of SIPP Priority requires more computational time for systems where quality is measured as a function of the waiting time distribution compared to measures which may be computed using a closed-form formulae, it may still be implemented in a reasonably efficient manner.

7.2.2 Numerical methodology

Following the argument outlined in Section 6.2.2 which presents the Euler method as a comprehensive numerical technique, this section considers the potential of the Euler approach to accurately evaluate system performance and provide a benchmark to

evaluate the Priority SIPP results. The Euler method has been investigated in several papers dealing with $M/M/s/NPRP/\infty/\infty$ systems (see Gail et al. (1988) and Wagner (1997)); and the following analysis extends the technique so is suitable for use in time-dependent priority systems i.e. $M(t)/M/s(t)/NPRP/\infty/\infty$ systems, where staff operate under the head-of-the-line service discipline. This research devotes particular attention to the behaviour of the system at shift boundaries and additionally derives formulae to calculate the excessive wait probabilities for both priority classes over time.

Priority queueing is difficult to analyse in a multi-server setting because customers of different classes may be in service at the same time; thus to know which customer will leave in a time interval, it is necessary to know the exact composition of the customers being served within the interval and those awaiting service in the queue. Hence the Markov chain representation of the multi-class, multi-server queue requires tracking the number of customers of each class through the system, meaning the Markov chain is infinite in k dimensions (where $k = 2$ in this research, is the number of priority classes). In the formulae presented to analyse the system below, i and j represent the number of HP and LP customers in service respectively; and h and l are used to denote the number of HP and LP customers in the queue respectively. Hence the inequalities $i, j, h, l \geq 0$ and $i + j \leq s(t)$ must hold at all times. In cases where it is relevant to track the total number of customers in the system, n is used to represent the cumulative total of HP and LP customers.

For computational efficiency, Chapter 5.3.2 stated that numerical analysis of a $M(t)/M/s(t)/FIFO/\infty/\infty$ system requires the approximation of the infinite capacity system with its finite equivalent, with a limit imposed on the number of customers considered in the system that is large enough to allow accurate analysis whilst ensuring that the dimension of the Markov chain is finite. The same approximation is clearly required for the numerical analysis of $M(t)/M/s(t)/NPRP/\infty/\infty$ systems, and following similar reasoning to the case presented by Izady (2010) for non-priority systems, this research recommends this upper limit be chosen such that $P_G(t) \leq 10^{-6}\ \forall t$.

In order to accurately track the movement of all customers through the system, it is necessary to compute the number of customers of types i, j, h and l in the system over time, represented by the quadruple $S = (i, j, h, l)$. Following the methodology presented by Gail et al. (1988), it is easily shown that the description of this state space quadruple $S = (i, j, h, l)$ may be reduced to:

- $S = (i, j)$ if at least one server is idle (as both h and l must both be null, since there will be no customers in the queue)
  - For example, in a service system staffed by 5 servers with 2 HP customers and 2 LP customers in service, one server would be idle and no customers would be in the queue. Thus the state space could be represented using the notation $S = (2, 2)$ in place of the extended version $S = (2, 2, 0, 0)$.
- $S = (i, h, l)$ if all servers are busy (since j may be derived from the description of the other parameter values)
  - For example, a service system staffed by 5 servers with 4 HP and 1 LP customer in service, and 2 LP customers in the queue, could be represented using the notation $S = (4, 0, 2)$ in place of the extended version $S = (4, 1, 0, 2)$ since element j may be calculated directly from knowledge of the total number of servers in the system and the number of HP customers in service i.e. $j = 5 - 4 = 1$.

As such, this convenient notation simplifies the state space description and increases the computational efficiency of the numerical solver.

Further to the methodology outlined by Gail et al. (1988), this research considers head-of-the-line priority systems where time-variable and stochastic demand is served by a time-varying number of servers (permitted to change at the beginning of each planning period), in order to maintain a consistent level of service throughout the period of service operation. The numerical solution of the balance equations for such a system, presented in the following two pages, are determined using the Euler method.

For the case where at least one server is idle (i.e. $i + j < s(t)$) the balance equations are:

$$\begin{aligned}
(\lambda(t) + (i+j)\mu)P(i,j) =& \lambda_H(t)P(i-1,j) + \lambda_L(t)P(i,j-1) && \text{for } 0 < i, 0 < j,\\
&+(i+1)\mu P(i+1,j) + (j+1)\mu P(i,j+1), && \text{and } i + j < s(t)\\
(\lambda(t) + i\mu)P(i,0) =& \lambda_H(t)P(i-1,0) + (i+1)\mu P(i+1,0)\\
&+\mu P(i,1), && \text{for } 0 < i < s(t)\\
(\lambda(t) + j\mu)P(0,j) =& \lambda_L(t)P(0,j-1) + (j+1)\mu P(0,j+1)\\
&+\mu P(1,j), && \text{for } 0 < j < s(t)\\
\lambda(t)P(0,0) =& \mu P(1,0) + \mu P(0,1), && \text{otherwise}
\end{aligned} \tag{7.14}$$

For states in which all servers are busy, and only LP customers are in service (i.e. $i = 0$), (7.14) becomes:

$$\begin{aligned}
(\lambda(t) + s(t)\mu)P(0,h,l) =& \lambda_H(t)P(0,h-1,l) + \lambda_L(t)P(0,h,l-1), && \text{for } 0 < h, 0 < l\\
(\lambda(t) + s(t)\mu)P(0,h,0) =& \lambda_H(t)P(0,h-1,0), && \text{for } 0 < h\\
(\lambda(t) + s(t)\mu)P(0,0,l) =& \lambda_L(t)P(0,0,l-1) + s(t)\mu P(0,0,l+1)\\
&+\mu P(1,0,l+1), && \text{for } 0 < l\\
(\lambda(t) + s(t)\mu)P(0,0,0) =& s(t)\mu P(0,0,1) + \mu P(1,0,1)\\
&+\lambda_L(t)P(0,s(t)-1), && \text{otherwise}
\end{aligned} \tag{7.15}$$

The next set of balance equations define the probabilities for scenarios where all servers are busy and at least one HP and LP customer is in service. To aid interpretation of the formulae, Figure 7.2 overleaf illustrates the possible system states which may result in $S = (i, 0, 0)$, when $i = 2$ and $s(t) = 3$.

If all servers are busy and at least one HP and LP customer is in service (i.e. $0 < i < s(t)$):

$$\begin{aligned}
(\lambda(t) + s(t)\mu)P(i,h,l) =& \lambda_H(t)P(i,h-1,l) + \lambda_L(t)P(i,h,l-1) \\
&+ i\mu P(i,h+1,l) \\
&+ (s(t)-i+1)\mu P(i-1,h+1,l), && \text{for } 0 < h, 0 < l \\
(\lambda(t) + s(t)\mu)P(i,h,0) =& \lambda_H(t)P(i,h-1,0) + i\mu P(i,h+1,0) \\
&+ (s(t)-i+1)\mu P(i-1,h+1,0), && \text{for } 0 < h \\
(\lambda(t) + s(t)\mu)P(i,0,l) =& \lambda_L(t)P(i,0,l-1) + i\mu P(i,1,l) + \\
&(s(t)-i+1)\mu P(i-1,1,l) \\
&+ (s(t)-i)\mu P(i,0,l+1) \\
&+ (i+1)\mu P(i+1,0,l+1), && \text{for } 0 < l \\
(\lambda(t) + s(t)\mu)P(i,0,0) =& i\mu P(i,1,0) + (s(t)-i+1)\mu P(i-1,1,0) \\
&+ (s(t)-i)\mu P(i,0,1) + (i+1)\mu P(i+1,0,1) \\
&+ \lambda_H(t)P(i-1,s(t)-i) \\
&+ \lambda_L(t)P(i,s(t)-i-1) \\
&[\text{See Figure 7.2 for diagrammatic illustration}], && \text{otherwise}
\end{aligned} \tag{7.16}$$

Finally, if all servers are busy, and only HP customers are in service (i.e. $i = s(t)$), the balance equations are:

$$\begin{aligned}
(\lambda(t) + s(t)\mu)P(s(t),h,l) =& \lambda_H(t)P(s(t),h-1,l) + \lambda_L(t)P(s(t),h,l-1) && \text{for } 0 < h \\
&+ s(t)\mu P(s(t),h+1,l) + \mu P(s(t)-1,h+1,l), && \text{and } 0 < l \\
(\lambda(t) + s(t)\mu)P(s(t),h,0) =& \lambda_H(t)P(s(t),h-1,0) + s(t)\mu P(s(t),h+1,0) \\
&+ \mu P(s(t)-1,h+1,0), && \text{for } 0 < h \\
(\lambda(t) + s(t)\mu)P(s(t),0,l) =& \lambda_L(t)P(s(t),0,l-1) + s\mu P(s(t),1,l) \\
&+ \mu P(s(t)-1,1,l), && \text{for } 0 < l \\
(\lambda(t) + s(t)\mu)P(s(t),0,0) =& s(t)\mu P(s(t),1,0) + \mu P(s(t)-1,1,0) \\
&+ \lambda_H P(s(t)-1,0), && \text{otherwise}
\end{aligned} \tag{7.17}$$

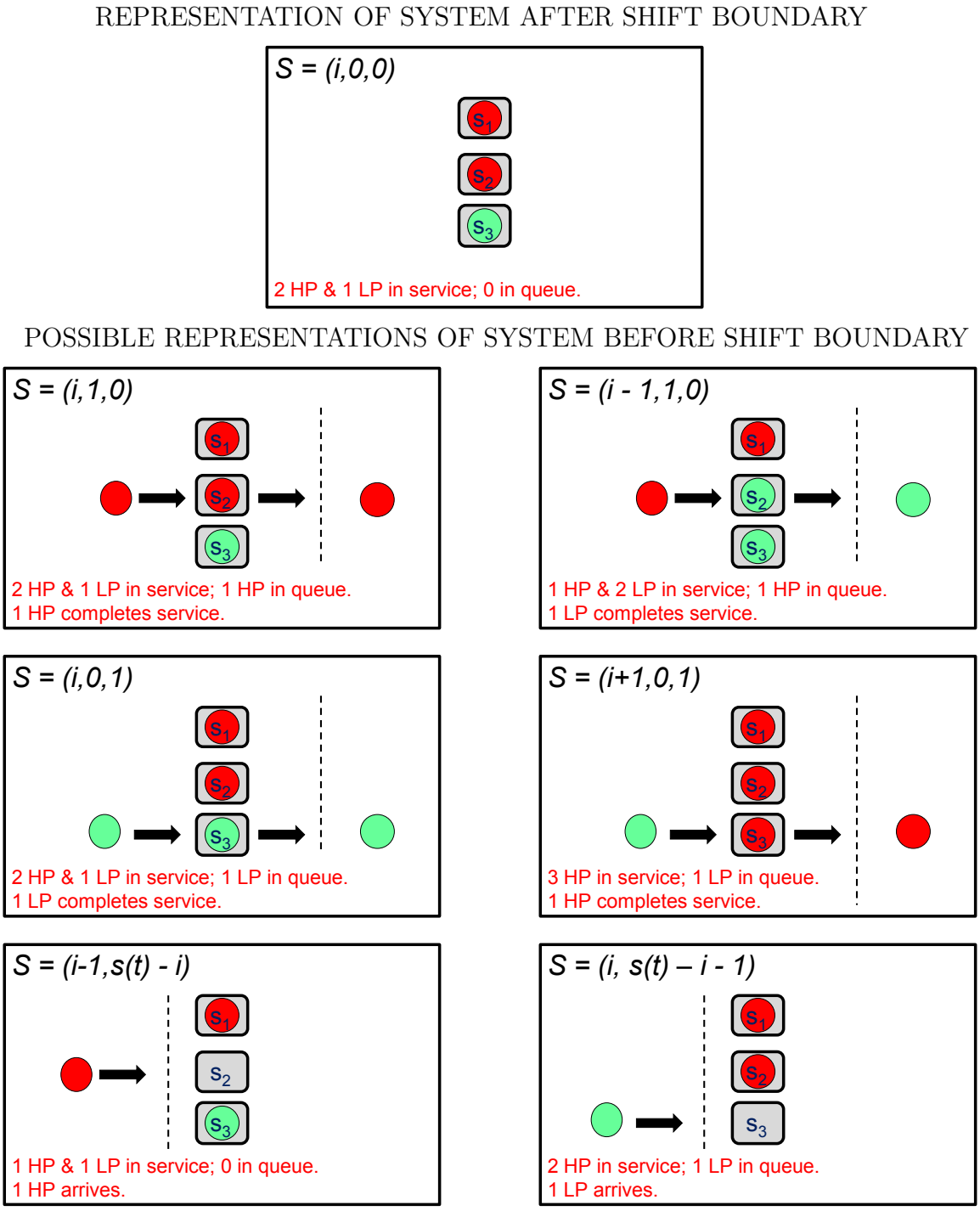

Figure 7.2: State spaces that may result in $S = (1, 0, 0)$ after shift boundary when $i = 2$ and $s(t) = 3$. (Red and green circles denote HP and LP customers respectively)

The probabilities of the various combinations of HP and LP customers in the system fluctuate over time according to the above equations, and may be tracked using a Euler solver. The complicating factor is how these probabilities evolve at the start of each planning period, where the arrival rates and number of servers on duty may change. The transformation of the probability vector depends on type of shift boundary (i.e. true or dummy), and further upon if the servers are busy when scheduled to leave in the case of the dummy shift boundary. Such scenarios have not been investigated for priority systems to date; thus the work below contributes to the research in this area through defining the mappings and transition probability matrices separately for each case. The notation used is a logical extension of the above to convert the standard probabilities to time-dependent likelihoods, such that if a new shift begins at epoch t,

$$P(i,h,l)(t) = P\{i \text{ HP in service, } h \text{ HP in queue and } l \text{ LP in queue at time } t\},$$
$$P(i,j)(t) = P\{i \text{ HP in service and } j \text{ LP in service, no customers in the queue and at least one server idle at time } t\},$$

Further, $P(i,h,l)(t)^- = \lim_{r\to t_z} P(i,h,l)(r)$ and $P(i,j)(t)^- = \lim_{r\to t_z} P(i,j)(r)$ are the probability vectors immediately before the shift boundary (i.e. at time points where $t = t_z$); and $s(t)^-$ and $s(t)^+$ represent the number of servers on duty for the shifts preceding and following the boundary respectively.

7.2.2.1 Mappings of state probability vector across shift boundaries

The precise transformations required to appropriately track the number of customers remaining in the system after true and dummy shift boundaries, where departing servers operate under the exhaustive discipline, are defined below. Figure 6.1, which was used to provide a visual representation the effect of applying each type of shift boundary on the number of customers in the queue for a non-priority time-dependent system in Chapter 6, is also useful to aid the comprehension of the analysis below since the same idea applies to time-dependent priority systems; only here all HP customers move ahead of LP customers in the queue to be seen by the first available server.

Case A: True shift boundary

At the end of a planning period bordered by a true shift boundary, all customers in service are ejected from the system under exhaustive discipline rules; thus the probability vector mappings are identical for all adjustments made to the number

of servers on servers on duty (i.e. independent of whether this number increases, decreases or remains the same). The new servers begin serving the customers in the queue at the immediate commencement of their shift, so all customers at the front of the queue move into service. Recalling that LP customers are only served when a server becomes free if there are no HP customers in the queue, then the mappings that define the instantaneous transitions of the probability vectors may be expressed as:

For $0 \leq h + l + s(t)^+ \leq G, i <= s(t)^+$(if $i = s(t)^+$ then $P(i,h,l)(t)$ is only defined for $h = 0$):

$$P(i,h,l)(t) = \sum_{u=0}^{s(t)^-} P(u, h+i, l+(s(t)^+ - i))(t)^- \tag{7.18}$$

And the transitions for the dual state vectors, defined for the case where $i + j < s(t)^+$ are:

$$\begin{aligned} &\text{For } i+j=0: \\ &\qquad P(i,j)(t) = \sum_{u=0}^{s(t)^-} P(u,0,0)(t)^- + \sum_{u=0}^{s(t)^- -1} \sum_{q=0}^{s(t)^- -1-q} P(u,q)(t)^- \\ &\text{For } 0 < i+j < s(t)^+: \\ &\qquad P(i,j)(t) = \sum_{u=0}^{s(t)^-} P(u,i,j)(t)^- \end{aligned} \tag{7.19}$$

Due to the way in which an artificial limit G is placed on the number of customers considered within the system to allow computational of the solution in reasonable time, there will be some cases where the revised probability vectors will be assigned zero values (for example, if $h + l + s(t)^- > G$ then $P(i,h,l)(t)^-$ will not be defined).

Case B: Dummy shift boundary

When dealing with a dummy shift boundary, the probability vector mappings depend on the nature of the change made to staffing levels over the boundary. In the case where the number of servers remains the same or is increased, the probability vectors require little or no modification, since the same set of staff are assumed to work both shifts (thus each server may continue working without disruption). The only potential modification that needs to be accounted for by a mapping is the movement of customers from the head of the queue into service that receive service from any additional employees who join the team at the shift boundary.

However, if the number of servers is reduced over the shift boundary, the behaviour of

the system is additionally dependent on the current occupation of the servers who are selected to leave. Thus the precise mappings necessary for each of the three scenarios are separately defined in cases B1 - B3 below.

Case B1: Number of servers remains the same

If the number of servers on duty over two consecutive shifts remains consistent, then the Markov process evolves as a continuous time Markov chain, as each server is available to work at all times across the shifts. Thus all probability vectors remain identical across the shift boundary, and may be defined as follows:

$$\begin{aligned} P(i,h,l)(t) &= P(i,h,l)(t)^- \qquad \text{for } 0 \le i \le s(t)^+, 0 \le h+l+s(t)^+ \le G \\ P(i,j)(t) &= P(i,j)(t)^- \qquad \text{for } 0 \le i+j < s(t)^+ \end{aligned} \tag{7.20}$$

Case B2: Number of servers is increased

For the case where more servers are supplied in the period following a dummy shift boundary, vector mappings are required to account for the fact that customers at the front of the queue move into service to be attended to by the additional servers who commence their duty at time t. Using $s_c = (s(t)^+ - s(t)^-)$ to represent the change in the number of servers, which in case B2 will always be positive; the instantaneous transitions may be defined for the triple state probability vector as:

$$\begin{aligned} &\text{For } i = s(t)^+, 0 \le h+l+s(t)^+ \le G: \\ &\qquad P(i,h,l)(t) = P(i-s_c, h+s_c, l)(t)^- \\ &\text{For } i < s(t)^+, 0 \le l+s(t)^+ \le G: \\ &\qquad P(i,0,l)(t) = \sum_{u=\max(0,i-s(t)^-)}^{\min(i,s_c)} P(i-u, u, l+s_c-u))(t)^- \\ &\text{For } s_c \le i < s(t)^+, 0 < h, 0 \le h+l+s(t)^+ \le G: \\ &\qquad P(i,h,l)(t) = P(i-s_c, h+s_c, l)(t)^- \end{aligned} \tag{7.21}$$

Concurrently the dual state space probability vectors remain identical, except for extra states which may arise if the number of customers in the queue is less than the quantity

of additional servers joining at the boundary. Thus for $0 < i + j < s(t)^+$:

$$\begin{aligned}
&\text{For } i + j < s(t)^- : \\
&\qquad P(i,j)(t) = P(i,j)(t)^- \\
&\text{For } i + j = s(t)^- : \\
&\qquad P(i,j)(t) = P(i,0,0)(t)^- \\
&\text{For } s(t)^- < i + j < s(t)^+ : \\
&\qquad P(i,j)(t) = \sum_{u=max(0,i-s(t)^-)}^{min(i,s_c,i+j-s(t)^-)} P(i-u,u,i+j-s(t)^- - u)(t)^-
\end{aligned} \tag{7.22}$$

Case B3: Number of servers is reduced

The transitions for the case where the number of servers is lower in the shift following the dummy shift boundary are generally more complex to define, since additional to the requirement in Section 6.2.2 which illustrates that it is necessary to specify the probability that a busy or idle server leaves; it is also necessary to determine the probability that a busy server selected to leave is serving a HP or LP customer at that epoch.

The number of customers ejected from the system can be seen to follow a series of hypergeometric distributions, similar to the distribution discussed in Section 6.2.2; only here after initially calculating the probabilities of various numbers of busy servers departing (equivalent to the total number of customers ejected) using a specific hypogeometric distribution, calculations are performed to compute the various compositions of HP and LP customers that could comprise this total quantity. Recalling that $s(t)^-, i$ and j denote the total number of servers on duty, number of HP customers in service and number of LP customers in service *before* the shift boundary respectively; and letting δn represent the total number of customers ejected from system and δs represent the total number of servers leaving at shift boundary; then the probability that δi HP customers are ejected from the system is defined for:

$$\begin{aligned}
&\max(0, i + j \ -s(t)^+) \leq \delta n \leq \min(\delta s, i + j) \\
&\text{and } \max(0, \delta n - j) \leq \delta i \leq \min(\delta n, i)
\end{aligned}$$

and given by:

$$\varphi(\delta n; \delta i; s(t)^-, i + j, i) = \frac{\binom{i+j}{\delta n}\binom{s(t)^- - i - j}{\delta s - \delta n}}{\binom{s(t)^-}{\delta s}} \times \frac{\binom{i}{\delta i}\binom{j}{\delta n - \delta i}}{\binom{i+j}{\delta n}} \tag{7.23}$$

Equation 7.23 can be used to compute the dual state probability state vectors. Considering the different ways in which each of these states may arise, it directly follows that the transitions are given by:

$$\begin{aligned}&\text{For } i+j < s(t)^+ : \\ &P(i,j) = \sum_{\delta n=0}^{\delta s} \sum_{\delta i=0}^{\delta n} \varphi(\delta n; \delta i; s(t)^-, i+j, i+\delta i) P(i+\delta i, j+(\delta n - \delta i))^- \end{aligned} \tag{7.24}$$

For all cases where $i + j \geq s(t)^+$, probabilities are derived by considering the triple state vectors, as defined below.

The triple state probability vectors $P(i,h,l)$ are defined at the dummy shift boundary for cases where all servers are busy. The probability that δi HP customers are ejected from the system is somewhat simpler to define for this scenario, since it is certain that all departing servers will each eject a customer from the system, so it is only necessary to take into account the probability that those ejected are HP or LP customers. If there are no idle servers, it can be easily shown that the probability that δi HP customers are ejected from the system follows a hypergeometric distribution. Following the notation given in equation (6.4), this is given by:

$$\theta(\delta i; \delta s, s(t)^-, i+j) = \frac{\binom{i+j}{\delta i}\binom{s(t)^- - i - j}{\delta s - \delta i}}{\binom{s(t)^-}{\delta s}} \tag{7.25}$$

One may observe that the number of HP and LP customers in the queue remain identical over the dummy boundary: since staff numbers decrease, there are no additional servers available to accept new customers at the commencement of the new shift. Thus the only parameter value to experience a transition in the triple state vector over the shift boundary is i (representative of the number of HP customers in service). Section 6.2.2 demonstrated that if servers depart at the shift boundary, the non-priority probability vector experiences an instantaneous transition according to $P(t) = P(t)^- B(t)$. The same methodology is applied here to model the instantaneous transitions, giving:

$$\begin{aligned} &\text{For } 0 < h+l, 0 \leq s(t)^+ + h + l \leq G: \\ &P(i,h,l)(t) = \sum_{u=0}^{s(t)^+} P(u,h,l)(t)^- B(t) \end{aligned} \tag{7.26}$$

where transition matrix $B(t)$ has the following non-zero entries:

$$b_{n,n-\delta i} = \theta(\delta i; \delta s, s(t)^-, n) \quad \begin{cases} \text{for } n = 0, 1, ...s(t)^- - 1 \text{ and} \\ \max(0, n - s(t)^+) \leq \delta i \leq \min(\delta s, n) \end{cases} \tag{7.27}$$

Equation (7.24) is valid for $i + j < s(t)^+$. Yet it also gives the probability for the boundary state for the case when all idle servers leave the system, leaving no customers in the queue, but all remaining servers busy. Thus the triple state probability vector defining the case where there are no customers in the queue additionally needs to take into account this event, so the transition may be defined by:

$$\begin{aligned} &\text{For } i \leq s(t)^+: \\ &P(i,0,0) = \sum_{\delta n=0}^{\delta s} \sum_{\delta i=0}^{\delta n} \varphi(\delta n; \delta i; s(t)^-, i+j, i+\delta i) P(i+\delta i, j+(\delta n - \delta i))^- \\ &\qquad + \sum_{u=0}^{s(t)^+} P(u,0,0)(t)^- B(t) \end{aligned} \tag{7.28}$$

7.2.2.2 Virtual waiting time distribution

Whilst Green et al. (2007) devote much attention to the computation of time-dependent waiting time probabilities in $M(t)/M/s(t)$ queueing systems, recognising the wide variety of service systems which are evaluated based on functions of this measure (usually waiting tail probabilities); equivalent mathematical expressions have not been derived in the literature to allow such meaningful analysis of priority systems despite their widespread applicability in industry. Gurvich et al. (2010) recently investigated the minimum number of servers required to limit the fraction of abandoning customers by means of considering the steady-state fraction of abandonments as a random variable; but the majority of previous papers investigating waiting times in a priority systems have based their analysis on the probability of delay or the mean waiting time experienced by customers (see Kao and Wilson (1999), Gail et al. (1992)).

The derivation of the waiting tail formulae presented below contributes to the literature through proposing expressions that allow the probability of an excessive wait to be accurately evaluated in time-dependent head-of-the-line priority systems. The derivations closely follow those presented in Section 6.2.2.2 which extended the standard formulae to deal with staffing changes over shift boundaries - only here the adjustments are applied to the distinct expressions which evaluate the probability of an excessive wait for HP and LP customers at specific time instances t, presented in equations (7.9) and (7.13). Whilst these equations are slightly more complex than the equations for systems with a single class of customer, the adjustments that need to be applied to take account of changes over shift boundaries are effectively identical. Cases A and B1 - B3 below define the expressions required to evaluate the probability of an excessive wait if the number of servers changes over a true or a dummy shift boundary. As defined in Section 6.2.2.2, $a = \mu s(t)^{-}\Delta t + \mu s(t)^{+}(x - \Delta t)$ is the mean departure rate over $[t, t + x]$ (where x may be adjusted to x_H or x_L accordingly); and $s(t)^{-}$ and $s(t)^{+}$ denote the number of servers on duty before and after the shift boundary respectively. For a visual representation of the specific events accounted for by the expressions, the reader is directed to Figure 6.1 which illustrated the case for a single customer class in Chapter 6. Due to the equivalence in the adjustments for systems with one and two priority classes, the same diagram may be used to aid comprehension of the extensions to the waiting time formulae for HP and LP customers described below.

Case A: True shift boundary

At a true shift boundary where servers operate under the exhaustive discipline, all customers in service are ejected from the system. Hence following the formulation of the expression derived for the non-priority case, since all servers leave the system and are replaced by an entirely new set, one may observe that only one standard adjustment is needed to account for all possible changes in staffing levels (i.e. an increase, decrease or equal levels); giving:

$$\begin{aligned} P(W_{qH}(t) > x_H) &= \sum_{\tilde{n}=s(t)^{-}+s(t)^{+}}^{\infty} P(W_{qH}^{\tilde{n}}(t) > x_H) p_{\tilde{n}}(t), \text{ where} \\ P(W_{qH}^{\tilde{n}}(t) > x_H) &= \sum_{b=0}^{\tilde{n}-s(t)^{-}-s(t)^{+}} \frac{a^b e^{-a}}{b!} \quad \text{if } \tilde{n} \geq s(t)^{-} + s(t)^{+} \end{aligned} \tag{7.29}$$

And

$$P(W_qL(t) > x_L) = \sum_{n=s(t)^-+s(t)^+}^{\infty} P(W_q^nL > x_L)\, p_n(t), \text{ where}$$
$$P(W_q^nL > x_L) = \sum_{f=0}^{\infty} P(f \text{ HPs arrive in } x_L) \sum_{b=0}^{n-s(t)^--s(t)^++f} \frac{a^be^{-a}}{b!} \text{ if } n \geq s(t)^- + s(t)^+ \tag{7.30}$$

Case B: Dummy shift boundary

Case B1: Number of servers remains the same

If the number of servers remains unchanged over the shift boundary, the same formula may be used to calculate the probability of an excessive wait as if no boundary was imposed, since the servers are unaffected by the occurrence of the shift boundary and continue working as normal. Note that as $s(t)^- = s(t)^+ = s$, they may be used interchangeably within the expression:

$$P(W_{qH}(t) > x_H) = \sum_{\tilde{n}=s(t)^+}^{\infty} P(W_{qH}^{\tilde{n}}(t) > x_H)p_{\tilde{n}}(t), \text{ where}$$
$$P(W_{qH}^{\tilde{n}}(t) > x_H) = \sum_{b=0}^{\tilde{n}-s(t)^+} \frac{a^be^{-a}}{b!} \quad \text{if } \tilde{n} \geq s(t)^+ \tag{7.31}$$

And

$$P(W_qL(t) > x_L) = \sum_{n=s(t)^+}^{\infty} P(W_q^nL > x_L)\, p_n(t), \text{ where}$$
$$P(W_q^nL > x_L) = \sum_{f=0}^{\infty} P(f \text{ HPs arrive in } x_L) \sum_{b=0}^{n-s(t)^++f} \frac{a^be^{-a}}{b!} \text{ if } n \geq s(t)^+ \tag{7.32}$$

Case B2: Number of servers is increased

If the number of servers increases at time $t + \Delta t$, so $s(t)^+ > s(t)^-$ the waiting tail probabilities $P(W_q^nH(t) > x_H)$ and $P(W_q^nL(t) > x_L)$ may be calculated as:

$$P(W_{qH}(t) > x_H) = \sum_{\tilde{n}=s(t)^+}^{\infty} P(W_{qH}^{\tilde{n}}(t) > x_H)p_{\tilde{n}}(t), \text{ where}$$
$$P(W_{qH}^{\tilde{n}}(t) > x_H) = \sum_{b=0}^{\tilde{n}-s(t)^+} \frac{a^be^{-a}}{b!} \quad \text{if } \tilde{n} \geq s(t)^+ \tag{7.33}$$

And

$$\begin{aligned} P(W_qL(t) > x_L) &= \sum_{n=s(t)^+}^{\infty} P(W_q^nL > x_L)\, p_n(t), \text{ where} \\ P(W_q^nL > x_L) &= \sum_{f=0}^{\infty} P(f \text{ HPs arrive in } x_L) \sum_{b=0}^{n-s(t)^+ +f} \tfrac{a^b e^{-a}}{b!} \text{ if } n \geq s(t)^+ \end{aligned} \tag{7.34}$$

Case B3: Number of servers is reduced

Finally considering the case where the number of servers decreases at time $t + \Delta t$, i.e. $s(t)^+ < s(t)^-$, the probability of an excessive wait may be computed as:

$$\begin{aligned} P(W_{qH}(t) > x_H) &= \sum_{\tilde{n}=s(t)^-}^{\infty} P(W_{qH}^{\tilde{n}}(t) > x_H) p_{\tilde{n}}(t), \text{ where} \\ P(W_{qH}^{\tilde{n}}(t) > x_H) &= \sum_{b=0}^{\tilde{n}-s(t)^-} \tfrac{a^b e^{-a}}{b!} \quad \text{if } \tilde{n} \geq s(t)^- \end{aligned} \tag{7.35}$$

And

$$\begin{aligned} P(W_qL(t) > x_L) &= \sum_{n=s(t)^-}^{\infty} P(W_q^nL > x_L)\, p_n(t), \text{ where} \\ P(W_q^nL > x_L) &= \sum_{f=0}^{\infty} P(f \text{ HPs arrive in } x_L) \sum_{b=0}^{n-s(t)^- +f} \tfrac{a^b e^{-a}}{b!} \text{ if } n \geq s(t)^- \end{aligned} \tag{7.36}$$

7.2.3 Hybrid methodology

Analysis of time dependent systems is often performed for the purpose of developing low-cost employee schedules that guarantee that the service level is always at or above a specified minimum level. Accurate recommendations of such quantities necessitates iterative analysis of system performance with varying staffing levels to determine the minimum quantity of staff, s, that allow the target to be achieved; and Figure 7.3 demonstrates that the approximate methodology is capable of recommending staffing levels that are close to these requirements, but often not identical. In recognition of the fact that the approximation methodology is sometimes unreliable, whilst the numeric technique allows accurate analysis at the system at the expense of computation time (which can be especially long in priority systems), this research proposes a hybrid approach which considers using the staffing levels generated from approximate methodologies as initial staffing levels for each period, to be accurately analysed with

numerical techniques. Under the assumption that the approximated levels are fairly close to the correct quantities; by directing the methodology to consider staffing levels just below those suggested by approximate methods (if numerical analysis finds that the approximate predictions are sufficient) or just above those suggested (if they are insufficient) it is hypothesised that considerable time savings could be made, especially if the quantities of staff significantly vary in different periods or if the demand continually rises in several consecutive periods, meaning a large number of staff quantities would otherwise need to be iterated through using the numeric approach before a sufficient number could be found to allow the required service level to be achieved. Empirical analysis supports this approach in the case study included in Section 7.3.3.

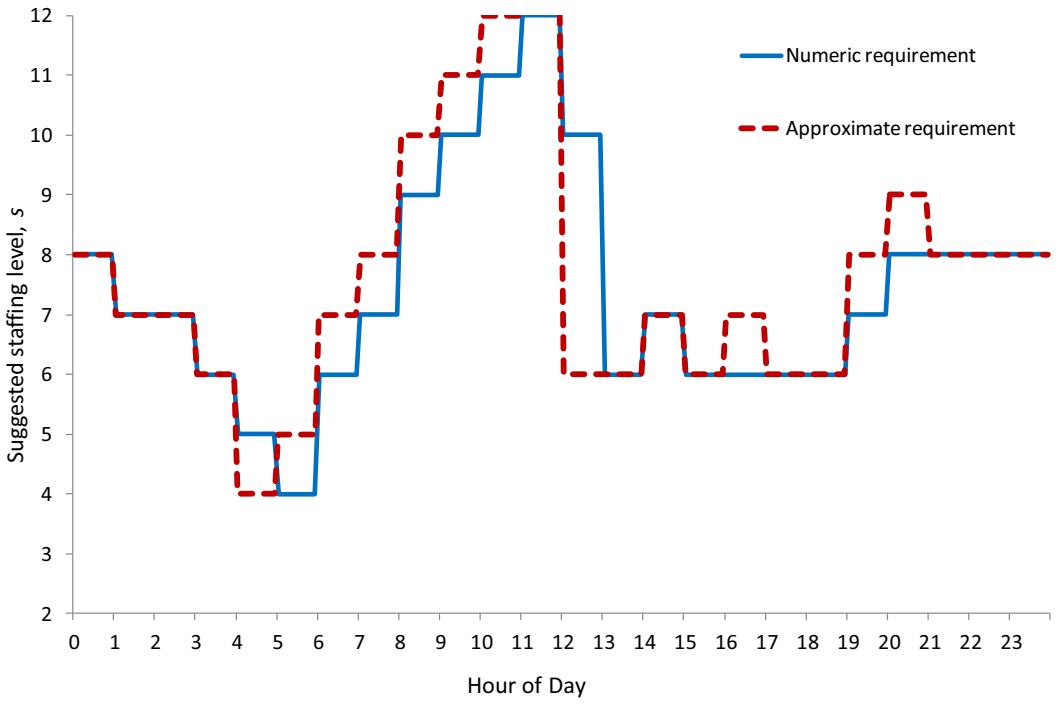

Figure 7.3: Approximate and numerical staffing requirements for EAs in the Cardiff region, 1st July 2009

7.3 Application to WAST data (Cardiff area)

This section investigates the potential of the aforementioned methods in terms of both accuracy and efficiency, to compute patients' virtual waiting times in the WAST priority system, with the objective of setting minimum staffing levels to limit the

proportion of patients waiting longer than targeted times. As discussed in Section 1.4, WAST responds to emergencies of a life-threatening nature (Category A calls) with precedence over all other injuries and thus typifies a head-of-the-line priority queueing system, meaning that ambulances may only attend to a Category B/C incident if there are no Category A incidents logged in the system awaiting a response. However, once an ambulance has been assigned to attend a Category B/C call, it cannot be re-routed to attend one of a more serious nature until it completes its service with the current patient. Hence if a Category A call is reported when all ambulances are busy, this incident will be placed on the waiting list with precedence above all other Category B/C calls within the system.

The response time target for Category B/C calls varies throughout Wales, and is set for different regions according to their population density. Thus in order to impose a consistent target in the analysis, when evaluating various methods to approximate state probabilities for the queueing system, the data employed in this case study is confined to demand arising within the Cardiff area only (a subsection of the SE Region) for which a consistent target is imposed that 95% of Category B and C calls should be responded to by an EA within 14 minutes. The same 14 minute target also applies for EA responses to Category A emergencies in the region, concurrent with the target that a RRV should arrive at the scene within 8 minutes. There are situations in which the first response to a Category A call is an EA rather than a RRV - in reality, such a call would not require an additional EA to attend within the 14 minutes, but to simplify the situation for this analysis, it is assumed that exactly one EA is required to attend all Category A, B and C incidents. One may model the queue for EAs as a $M(t)/M/s(t)/NPRP/\infty/\infty$ queue i.e. a queue with non-preemtive priority, a time-varying number of servers (practicing the exhaustive discipline) and two classes of customers, each with their own exponential time-varying arrival rate. The class of each patient is determined by the calltaker who classifies their degree of injury according to the AMPDS as described in Chapter 6. This analysis refers to those classed as suffering from Category A conditions as HP customers, and those with Category B or C symptoms as LP customers. The precise problem, data inputs and constraints that are used to determine the optimum staffing levels are as follows:

- This investigation is concerned with determining the minimum number of crews required to man EAs, to ensure that 95% of all emergency calls arising within the Cardiff area are reached within 14 minutes for July and December 2009 (i.e.

to attain the EA requirements specified by Target 1 and Target 2 presented in Chapter 1).

- The queueing model is based on the assumption that exactly one EA is required to attend each incident reported to WAST. Whilst every EA is required to be staffed by two ambulance officers, it is assumed that the two officers are paired for the purpose of constructing coverage requirements; thus the investigations design minimum staffing levels for each pair of officers (referred to as 'crew'). If a crew is assisting a patient when their shift is due to end, they must follow exhaustive service discipline rules and first complete their service before finishing duty.

- Category A calls are classed as HP; whilst Category B and C calls are considered LP. HP calls have head-of-the-line priority over LP calls, i.e. if a EA is unavailable when an emergency is reported, then that emergency is placed in the virtual queue for assistance, irrespective of its urgency; and queued LP calls are only dealt with after all queued HP calls have been allocated a responder.

- The expected number of HP and LP emergencies requiring EA assistance for each period of each day in the scheduling horizon are obtained from SSA forecasts (see Chapters 3 and 4) based on historic demand data, known ,from 2005.

  - Following the guidance provided in Chapter 4, the first 20 components and a window length of 1,367 are selected to construct the forecasts, and to reduce the risk of understaffing, the predicted counts are uplifted by 10%.
  - Individual counts are predicted for the total number of unique incidents requiring emergency assistance for each shift, pre-defined by WAST as 6am-12pm, 12pm-7pm and 7pm-6am (see Chapter 2.4).
  - SSA is initially applied to the the time series recording all EMS calls, and in order to estimate the proportion calls within each shift that are HP or LP, the technique is subsequently re-applied to the time series recording the proportion of HP calls for the period of known historic demand (with 6 components and a window length of 1,367). The forecasted series specifying the expected number of HP and LP calls in each shift is ultimately obtained by applying the projected proportions to the projected counts.
  - Finally the expected counts per shift are converted to expected counts per hour by analysing the typical distribution of calls for each shift for each weekday throughout 2008 (to gain a relatively recent representation of the

system), and applying the average determined proportions to each shift projection. Distinct proportions are specified for each weekday to improve the forecast quality, since a two-way ANOVA reveals that the expected hourly demands are not equal for all weekdays ($p < 0.05$).

- All parameter values employed in the analysis are evaluated from 2008 data alone to capture the most recent representation of the service. Since the travel times are found to be significantly different for HP and LP responses; the acceptable waiting times set in the program are adjusted to be:

  $x_H = 14 - 8.27$ (average travel time) $= 5.73$ minutes for HP incidents; and

  $x_L = 14 - 9.21$ (average travel time) $= 4.79$ minutes for LP incidents.

  Hence although the response time targets are identical for HP and LP calls; the acceptable times that a patient should wait before an EA is mobilised are different.

- Similar analysis provides an average service time of $\mu = 54.55$ minutes which is applied to all call categories in the analysis.

- Minimum crew requirements are sought for one-hour planning periods throughout the scheduling horizon. Within each period, requests for assistance are assumed to arrive according to a homogeneous Poisson process with the forecasted mean rate for that hour, and all servers are assumed to have independent exponentially distributed service times, with the same mean length (independent of the category of incident they are responding to).

Sections 7.3.1 and 7.3.2 describe the implementation of the numerical and approximation methods that have been developed throughout this chapter, to determine the probability of an excessive wait for each category of call (i.e. a wait greater than the response time targets), and ultimately recommend the minimum number of EA crews required to ensure that this probability is no more than 95%. Similarly to Chapter 6, the analysis assumes that the minimum service level must hold for *every* time point in the scheduling horizon (rather than being considered as an aggregate service level that is to be achieved on average across all hourly periods in the scheduling horizon), to ensure a consistent quality of service is provided.

7.3.1 Numerical requirements

This section demonstrates how the extensions that have been proposed for the numerical method throughout this chapter (i.e. to map the probability state vector over shift boundaries, and accurately evaluate the probability of an excessive wait over time), may be applied to real-life data and generate minimum staffing requirements. Modelling the system described above as an MDCTMC, the Euler method may be readily implemented to analyse system performance. Discussions surrounding the importance of incorporating the appropriate type of shift boundary in the analysis have previously been provided in Chapter 6.2.2.2. Since the ultimate aim of the analysis is to construct hourly coverage requirements for the purpose of informing the shift scheduling model, a dummy boundary with exhaustive service discipline is applied to every period boundary, similarly to the approach described in Chapter 6.3.1.

For computational efficiency, the infinite capacity $M(t)/M/s(t)$ system is approximated by a finite equivalent, with a cap of $G = 40$ imposed on the number of patients considered in the system at any specific time instance; and following the discussion surrounding the selection of appropriate calculation intervals presented in Chapter 5.3.2, calculation periods are selected as $\delta_c = 0.04$ hours (i.e. 2.4 minutes) which is a common divisor of the length of the planning periods $\delta_{pp} = 1$ hour. To ensure that dynamic steady state is reached (see Heyman and Whitt (1984) for a definition and conditions for achieving dynamic steady state), the service quality is also computed for a one-day warm-up period, using the forecasted demand data for the first day in the scheduling horizon, before meaningful analysis of the system is performed.

The computation time required to implement Euler in a priority time-dependent service system is however considerably longer that that required for one with a single class of customer. For example, it takes around 120 minutes to produce hourly requirements a 3 month scheduling horizon on a 3GHz machine with 2.96GB RAM, compared to 6 minutes for a system without priorities. Approximation techniques can generate estimates in a far more efficient manner; but are only suitable if they can produce results within a reasonable accuracy. The staffing requirements generated by Euler are accordingly used as a benchmark to evaluate the potential of SIPP and its variants to approximate minimum EA crew requirements in the Cardiff area in Section 7.3.2.

7.3.2 Approximate requirements

This section evaluates the potential of the Priority SIPP, Priority Lag Avg and Priority SIPP Mix methodologies to construct hourly requirements for the minimum number of EAs required within the Cardiff region for first 28 days (i.e. 672 hourly periods) of July and December, consistent with the periods selected in Chapter 6. The Priority Lag Avg and Priority SIPP Mix approaches are direct extensions of the standard Lag Avg and SIPP Mix approaches described in Chapter 6: Priority Lag Avg estimates the required staffing level based on the average arrival rates predicted for the relevant period shifted back by L units (estimated as the average service time) in an attempt to incorporate a suitable estimation of the lag; and Priority SIPP Mix uses the average planning period arrival rates in all periods where the overall arrival rate is strictly increasing, and the maximum arrival rates otherwise (calculated as $1.2 \times$ average rate, based on preliminary investigations), to avoid the problem of understaffing. Thus the Priority SIPP methodology remains consistent in all cases and the only adjustments are those made to the arrival rate function prior to the application of the technique.

For the scheduling horizons investigated, the average expected hourly demands are 4.2 and 4.6 for July and December respectively; and the average minimum number of EAs recommended are 7.5 and 8.0 with an average service time of 54.6 minutes. This in turn means there is a relatively low server utilisation rate, which is favourable for SIPP. Supplementary analysis however uncovers that there are come periods with relatively high relative amplitudes which is potentially problematical.

Table 7.1 displays the average relative error associated with the staffing requirements generated for each hour of each day by the approximate methods when compared with the Euler requirements, where periods with a RMSE greater than or equal to 1 are highlighted with bold text. The results show that the Priority SIPP approach performs well for the scheduling horizons investigated. This corresponds with the conclusions drawn in Green et al. (2001) that standard SIPP should perform well in systems with low relative amplitudes and low presented loads. It appears logical that the technique requires the same conditions as its non-priority counterpart to generate accurate results, since the steps followed by the methods are analogous: the sole differentiating factor is the computation required to evaluate the performance measure. As this research has uncovered a suitable method to allow the probability of an excessive wait to be calculated in a time-dependent priority system, it has enabled

the SIPP technique to be aptly applied.

Table 7.1: Pri SIPP, Pri Lag Avg and Pri SIPP Mix accuracy

		RMSE		
Hour	$\lambda_H + \lambda_L$	Pri SIPP July/Dec	Pri Lag Avg July/Dec	Pri SIPP Mix July/Dec
0	4.8 / 5.1	0.37 / 0.62	0.83 / 0.72	**1.29** / **1.35**
1	4.5 / 4.8	0.42 / 0.38	0.68 / 0.80	0.82 / 0.65
2	3.8 / 4.0	0.33 / 0.33	0.78 / **1.00**	0.89 / **1.00**
3	3.1 / 3.3	0.46 / 0.65	0.96 / **1.41**	0.85 / **1.25**
4	2.2 / 2.3	0.53 / **1.05**	**1.21** / 0.73	0.63 / 0.87
5	1.9 / 2.0	0.53 / 0.60	0.76 / 0.57	0.76 / 0.73
6	3.2 / 3.5	0.76 / 0.94	**1.13** / **1.13**	0.76 / 0.94
7	3.8 / 4.2	0.85 / 0.73	0.38 / 0.68	0.93 / 0.82
8	5.4 / 5.9	**1.00** / **1.10**	**1.25** / **1.16**	**1.00** / **1.10**
9	6.9 / 7.4	0.98 / **1.05**	0.85 / 0.78	0.98 / **1.05**
10	7.7 / 8.3	0.93 / 0.82	0.38 / 0.71	0.93 / 0.82
11	7.5 / 8.1	0.63 / 0.65	0.73 / **1.27**	**1.68** / **1.74**
12	2.9 / 3.2	**2.43** / **2.34**	**3.07** / **3.16**	**1.77** / **1.52**
13	3.1 / 3.5	0.65 / 0.38	0.38 / 0.27	0.65 / 0.38
14	3.1 / 3.5	0.00 / 0.50	0.53 / 0.19	0.76 / 0.89
15	3.0 / 3.4	0.38 / 0.42	0.76 / 0.46	**1.00** / 0.78
16	3.1 / 3.4	0.46 / 0.42	0.57 / 0.42	0.96 / 0.85
17	3.0 / 3.3	0.60 / 0.19	0.73 / 0.46	**1.04** / 0.80
18	3.2 / 3.5	0.00 / 0.00	0.65 / 0.38	0.53 / 0.53
19	4.9 / 5.3	1.00 / 0.87	**1.13** / **1.20**	**1.00** / 0.87
20	4.9 / 5.3	0.53 / 0.65	0.53 / 0.93	0.76 / **1.00**
21	5.3 / 5.7	0.57 / 0.42	0.80 / 0.42	**1.25**/ 0.82
22	5.1 / 5.4	0.50 / 0.38	0.76 / 0.91	0.98 / **1.30**
23	5.0 / 5.3	0.19 / 0.33	0.57 / 0.73	0.68 / 0.73
MEAN	4.2 / 4.6	0.78 / 0.80	**1.00** / **1.03**	**1.00** / **1.00**

The standard Priority SIPP approach provides the best results overall, with identical results to the Euler method in $\frac{404}{672} = 60\%$ of cases for July and $\frac{397}{672} = 59\%$ for December. Whilst these results are less accurate than those found for the non-priority case presented in Table 6.1, the results can be seen to be more consistent across the two investigated months. Similarly to case study presented in Chapter 6, the main problem with the SIPP approach is that it overstaffs a number of periods in this case study. For example, SIPP overestimates $\frac{227}{672} = 34\%$ of the hourly periods for July (although never by more than a single crew), and underestimates $\frac{41}{672} = 6\%$ of the hourly periods (18 of which are underestimated by a single crew, 11 by two crews, 10

by three crews and 2 by four crews). Thus although Priority SIPP overestimates the requirement frequently, the quantity that it overestimates by is never prodigious.

Table 7.1 reveals that periods which SIPP fails to produce reliable requirements for are predominantly the 08:00-09:00 and 12:00-13:00 periods. The main problem in using SIPP to construct requirements for the 08:00-09:00 period is that it fails to recognise the link with previous periods and account for the fact that it takes time for the queue to build up to a level great enough to employ additional crew. Contrastingly, the error associated with the 12:00-13:00 period is mainly attributable to underpredictions, since the demand exhibited in this period is considerably lower than the previous period. Whilst the Euler methodology recognises that more crews are required from 12:00-13:00 to deal with the backlog of demand remaining in the system from previous periods, by assuming the requirements can be generated independently for each period, SIPP estimates that far fewer crews are required.

The results for the adjusted SIPP approaches are inferior: this is as expected for the case study, since Lag Avg and SIPP Mix are unreliable when relative amplitude is high (as is the case for a large portion of the data). SIPP Mix additionally commonly understaffs when the arrival rate is decreasing, and thus attributes even greater errors in these periods; although it does reduce the error associated with overstaffing from 12:00-13:00. It appears logical that in order to generate accurate results, each of the Priority SIPP extensions require the same conditions to hold as their non-priority equivalent, i.e. Priority Lag Avg and Priority SIPP Mix are expected to be reliable if the RA is low (around 0.1 - 0.5) and planning periods are short (around 0.25 - 0.5 hours). Since neither of these conditions are strictly met in this case study, it is not surprising that they fail to improve the results of the standard Priority SIPP approach, which already performs quite well, and understaffs just 6% of periods in July.

In recognition of the varying impacts arising from overpredictions and underpredictions in differing service industries, it can be more useful to consider a modified measure of the RMSE, RMSE_τ, defined in Chapter 6.3.2 which allows overstaffing and understaffing to be penalised accordingly, using weights that may be prudently selected by management. Figure 7.4 presents the overall RMSE_τ for various levels of τ for the month of July. It demonstrates that the standard Priority SIPP approach achieves around the same error, no matter what the value selected for τ, since although the method overestimates demand more frequently than it underestimates, the periods

that it understaffs are understaffed by larger quantities, meaning it is penalised for overstaffing and understaffing in similar quantities. On the other hand, Priority Lag Avg and Priority SIPP Mix both overstaff more frequently; and thus are evaluated more favourably for larger values of τ (when underpredictions are penalised more harshly). They however only perform better than Priority SIPP if overpredictions are virtually ignored in the error calculation. Priority Lag Avg and Priority SIPP Mix achieve less error than the standard approach for values of $\tau \geq 0.95$ and $\tau \geq 0.79$ respectively.

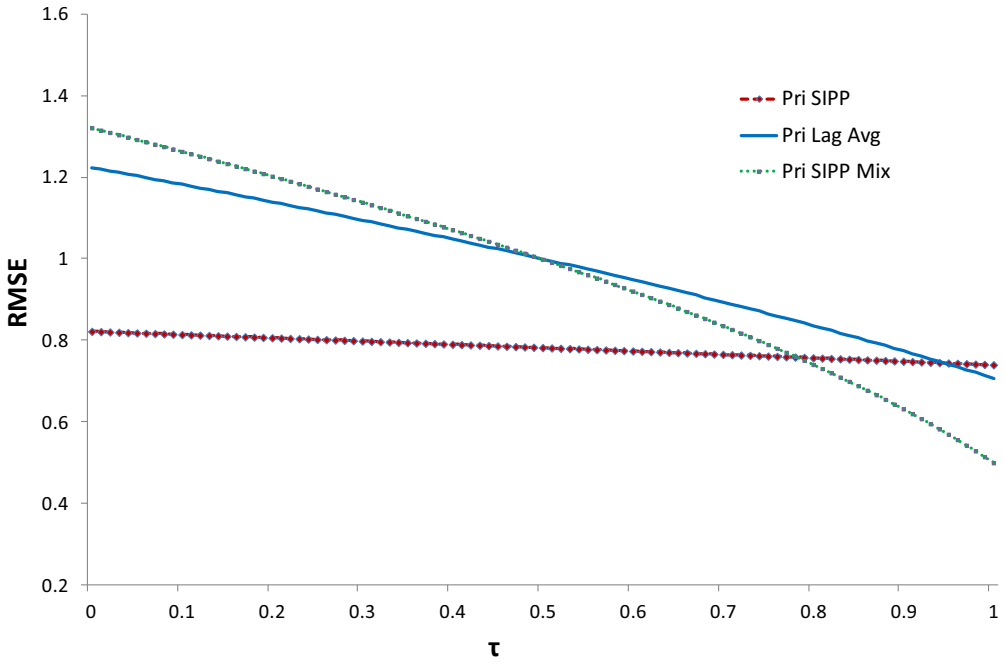

Figure 7.4: Weighted RMSE for various levels of $\tau \in [0, 1]_{\mathbb{R}}$

Since SIPP more commonly overstaffs than understaffs, the potential revisions suggested in the literature (such as Lag Avg and Lag SIPP) are unsuccessful in improving the predictions. It appears that the revisions proposed in Chapter 6.3.2 to adjust the arrival rate prior to the application of the approximation technique are more suitable, and Table 7.2 illustrates their potential to improve the quality of the predictions. The adjustments applied to the arrival rates are of an identical format to those proposed in Chapter 6.3.2, but some of the parameter values used in the techniques are adjusted to reflect the different targeted performance standards applied in this specific investigation.

The revisions that are investigated are:

i. Producing a modified arrival rate for each period where the original rate differs by more than 20% to the rate expected in the preceding period. Two variants of the arrival rate are considered: one resulting from taking an average of the current and preceding period, and another calculated as a weighted average (25%/75% in favour of the current period). This revision aims to incorporate the effect of previous arrivals in each period, and avoid the problem of over/understaffing where the approximation methods fail to recognise that it takes time for the queue to increase/decrease significantly.

ii. Given the staffing levels as output by the original SIPP model, scaling the arrival rate within each period where the expected proportion of patients seen within the acceptable time is considerably greater than the minimum proportion required, before re-running the model to provide revised requirements. For the data employed in this case study, the arrival rates are scaled by a factor of 0.9 for periods where over 97.5% of patients are seen within the acceptable time following preliminary investigations. Given the discussions surrounding Figure 6.5 in Chapter 6, it appears logical to only investigate if the reduced demand rate can be satisfied with one less paramedic if there is little congestion in the pervious period. However, adding this constraint to the model produces marginally inferior results for the test data.

iii. Given the staffing levels as output by the original SIPP model, directly reducing the number of staff recommended to be deployed by one member in periods where the expected proportion of patients seen within the acceptable time is considerably greater than the minimum proportion required. Again, for the case study data it appears sensible reduce the requirement by one paramedic if over 97.5% of patients are expected to be reached within the acceptable waiting time. Consideration is also given to the expected congestion in the system in the previous period, but find that this extra consideration fails to improve the model performance for the test data.

The results of the above revisions are summarised in Table 7.2. The 'Pri SIPP' column serves as a reminder of the RMSEs achieved by the standard SIPP approach, and the remaining columns present the revised RMSEs for the variants of Priority SIPP described above. The results show that the adjustment applied to the arrival rate as

described in Method (i) is highly successful in improving the accuracy of the Priority SIPP requirements, but Methods (ii) and (iii) yield similar results. Table 7.2 reveals that Method (i) considerably improves the accuracy of the requirements produced for both months investigated, which ever weighting measure is selected (-whilst the results for July are slightly better using the average rate from the current and previous period, the results for December are slightly better when a weighted average (25%/75% in favour of the current period) is computed). Although the same method only generated results of roughly equivalent accuracy as the SIPP approximations for the non-priority case presented in Table 6.2; this is likely because the necessary characteristics for SIPP to perform well were present in the system investigated.

The findings reported in Table 7.2 support the use of Method (i) as a practical technique to improve the approximate approach, and directly demonstrates how improved requirements can be generated in systems where the assumptions of Priority SIPP are not met. Due to the methodology followed by the technique, it is equally capable of improving Priority SIPP staffing requirements in systems where overstaffing or understaffing are shortfalls of the standard approach, if the error is attributable to the failure of the technique to recognise the impact of staffing levels and arrival rates in previous periods. Since the methodology is only applied to periods in which the arrival rate differs considerably to that in the preceding period, meaning the earlier arrival rate will likely impact on the service level in the current period (as the system will take longer to reach steady state), it is expected to perform better than the Lag Avg approach which is applied to all periods, regardless in the change in system behaviour. Since the particular version of the technique that computes an average arrival rate between two consecutive periods is essentially modified version of Lag Avg approach, and is demonstrated to perform well, it shall be promoted as a perspicacious variant of the SIPP approach, referred to as Modified Lag Avg (or Priority Modified Lag Avg if it is applied within priority systems). The technique is expected to be more robust to higher RA than the standard SIPP/Priority SIPP techniques, due to its capacity to prudently adjust the arrival rate in consecutive periods with widely differing arrival rates, and also more robust in systems with longer service rates, as it considers the effect of the time-lag that exists between arrival and service times. The technique is also expected to offer improved predictions if staffing requirements are desired for moderately longer planning periods than 0.25 or 0.5 hours (required by SIPP/Priority SIPP), as it accounts for the effect of arrivals in earlier periods in its calculations, and is shown in this case study to offer improved predictions for planning periods of 1 hour

Table 7.2: Priority SIPP Revised Reliability (RMSEs)

Hour of day	Pri SIPP	If 20% change in λ		SIPP re-run		Modified SIPP	
		$\lambda' = 0.5\lambda(t-1) + 0.5\lambda(t)$	$\lambda' = 0.25\lambda(t-1) + 0.75\lambda(t)$	If $P(W_q(p) > x) > 0.975$, $\lambda' = 0.9\lambda$	If $P(W_q(p) > x) > 0.965$ and $P(W_q(p-1) > x) > 0.965$, $\lambda' = 0.9\lambda$	If $P(W_q(p) > x) > 0.975$, EAs' = EAs $-$ 1	If $P(W_q(p) > x) > 0.965$ and $P(W_q(p-1) > x) > 0.975$, EAs' = EAs $-$ 1
0	0.37 / 0.62	0.53 / 0.62	0.53 / 0.62	0.53 / 0.00	0.42 / 0.00	0.00 / 0.56	0.62 / 0.56
1	0.42 / 0.38	0.63 / 0.38	0.46 / 0.38	0.46 / 0.19	0.38 / 0.19	0.42 / 0.60	0.57 / 1.07
2	0.33 / 0.33	0.42 / 0.33	0.33 / 0.33	0.33 / 0.50	0.53 / 0.42	0.50 / 0.71	0.65 / 0.57
3	0.46 / 0.65	0.46 / 0.76	0.27 / 0.71	0.27 / 0.68	0.38 / 0.73	0.78 / 0.65	0.60 / 0.65
4	0.53 / 1.05	0.53 / 0.50	0.00 / 0.78	0.00 / 0.76	1.35 / 0.65	0.91 / 1.35	0.82 / 1.21
5	0.53 / 0.60	0.65 / 0.50	0.53 / 0.42	0.53 / 0.53	0.50 / 0.76	0.65 / 0.68	0.91 / 0.76
6	0.76 / 0.94	0.68 / 0.38	0.19 / 0.19	0.19 / 0.38	0.53 / 0.27	0.00 / 0.38	0.78 / 0.71
7	0.85 / 0.73	0.53 / 0.38	0.65 / 0.60	0.65 / 0.53	0.57 / 0.53	0.60 / 0.53	0.65 / 0.57
8	1.00 / 1.10	0.53 / 0.19	0.53 / 0.87	0.53 / 0.76	0.93 / 0.53	0.53 / 0.93	0.46 / 0.94
9	0.98 / 1.05	0.38 / 0.19	0.65 / 0.73	0.65 / 0.73	0.89 / 0.68	0.65 / 0.71	0.57 / 0.94
10	0.93 / 0.82	0.76 / 0.63	0.93 / 0.63	0.93 / 0.76	0.42 / 0.71	0.76 / 0.33	0.78 / 0.63
11	0.63 / 0.65	0.63 / 0.65	0.63 / 0.65	0.63 / 0.68	0.38 / 0.63	0.68 / 0.33	0.42 / 0.53
12	2.43 / 2.34	1.20 / 1.07	1.22 / 1.07	1.22 / 2.76	2.88 / 2.62	2.76 / 3.05	2.94 / 2.83
13	0.65 / 0.38	0.65 / 0.38	0.65 / 0.38	0.65 / 0.27	0.27 / 0.53	0.38 / 0.27	0.76 / 0.78
14	0.00 / 0.50	0.00 / 0.50	0.00 / 0.50	0.00 / 0.63	0.65 / 0.53	0.65 / 0.68	0.65 / 0.68
15	0.38 / 0.42	0.38 / 0.42	0.38 / 0.42	0.38 / 0.53	0.5 / 0.38	0.53 / 0.57	0.53 / 0.91
16	0.46 / 0.42	0.27 / 0.19	0.27 / 0.19	0.27 / 0.60	0.33 / 0.46	0.71 / 0.50	0.46 / 0.53
17	0.60 / 0.19	0.60 / 0.19	0.60 / 0.19	0.60 / 0.50	0.42 / 0.50	0.50 / 0.60	0.57 / 0.68
18	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.19 / 0.38	0.65 / 0.53	0.42 / 0.57
19	1.00 / 0.87	0.00 / 0.38	0.53 / 0.42	0.53 / 1.00	0.85 / 0.85	0.78 / 0.71	0.76 / 0.76
20	0.53 / 0.65	0.53 / 0.65	0.53 / 0.65	0.53 / 0.38	0.00 / 0.46	0.38 / 0.38	0.53 / 0.63
21	0.57 / 0.42	0.65 / 0.19	0.57 / 0.42	0.57 / 0.00	0.50 / 0.27	0.00 / 0.60	0.57 / 0.63
22	0.50 / 0.38	0.42 / 0.19	0.50 / 0.19	0.50 / 0.60	0.33 / 0.71	0.68 / 0.60	0.60 / 0.60
23	0.19 / 0.33	0.19 / 0.33	0.19 / 0.33	0.19 / 0.68	0.50 / 0.82	0.82 / 0.68	0.63 / 0.63
MEAN:	0.78 / 0.80	0.55 / 0.48	0.54 / 0.55	0.54 / 0.79	0.82 / 0.77	0.82 / 0.88	0.86 / 0.93

durations. However, the technique should not be used in systems with planning periods spanning several hours, or which exhibit low presented loads, since such systems are more likely to reach steady state within each planning period, so the modifications considered by the technique would be unsuitable to improve the accuracy of predictions. The real benefit of approximate techniques lies in their ability to generate estimated requirements at a rapid rate. For example, whilst the Euler method requires around 120 minutes to generate hourly staffing requirements for a 3 month forecasting horizon on a 3GHz machine with 2.96GB RAM; Priority SIPP can offer an approximate solution in around 10 minutes, and the additional time required by the variants of the Priority SIPP technique is only that required to obtain adjusted arrival rate functions. When deciding if it is appropriate to use a numerical or approximate technique, the client should therefore importantly consider this factor, in conjunction with accuracy considerations.

In situations where accuracy is of upmost importance, the numerical method should always be selected, since the approximate methods will always be susceptible to a certain degree of error. In an attempt to reduce the computation time required by the Euler method to generate accurate staffing requirements, Section 7.3.3 investigates the potential of hybrid approaches to increase its computational efficiency.

7.3.3 Hybrid Approach

This section examines the potential of various hybrid approaches to increase the computational efficiency of the Euler method to produce minimum staffing requirements for hourly periods, that ensure emergencies are responded to within the targeted times outlined in Targets 1 and 2 (see Chapter 1). When applied in its standard format, the Euler method begins by considering the performance that would be achieved in each period if two staff were to be employed. It computes the relevant performance measure at computation intervals of length δ_c across the period, and if the measure is below the required level at any given point, the method returns to the start of that shift to re-start the calculations with one additional member of staff. The method continues to iterate through staffing levels (incremented in integer steps) for each period in this way, until a sufficient quantity is found to satisfy the response time targets. Three methods are considered to increase the efficiency of finding the minimum required staffing levels, and discussions surrounding why they are/are not appropriate:

i. Using the staffing levels generated from the Priority SIPP approach (or one of

its variants) as initial staffing levels for each period, to be accurately analysed with numeric methodology. If numerical analysis finds that the Priority SIPP predictions are sufficient, the Euler method can be used to test if the performance measure could in fact be achieved with a lower quantity, by decrementing the Priority SIPP suggested staffing levels for each period in integer steps and recomputing the performance measure for all calculation intervals throughout the period, until a staffing level is reached that violates the waiting time targets. If the converse is true and the initial quantities are found to be insufficient, the minimum required staffing levels can be obtained by incrementing the initial quantity in integer steps until a sufficient number is found.

- Applying this methodology to generate staffing requirements for the month of July 2009 actually requires greater computation time (45 minutes on a 3GHz machine with 2.96GB RAM) than the standard methodology (40 minutes). Closer inspection finds that this is primarily due to periods in which Priority SIPP accurately predicts or overpredicts the staffing requirement. If Priority SIPP generates the correct requirement, the methodological steps outlined above means that the Euler method must first be employed to compute the performance measure at all calculation intervals throughout the period with the suggested level to find that it is sufficient. It must then re-perform these calculations with one fewer staff member until the calculation interval is reached where the staffing quantity fails to achieve the targeted level. If Priority SIPP however overpredicted the staffing level, no such interval will exist, so at least part of the period must further be investigated with two fewer staff members until an infeasible calculation interval is found. This can take considerably longer than iterating through all staffing levels using the standard Euler methodology, since the low staffing levels will often be found to be insufficient for a calculation interval near the start of the period, so very few computations may be required for such levels. The shortfall of this method motivates the investigation of the proposed Method (ii) below.

ii. Using the same approach as described above, but with the initial staffing levels considered in each period being one below those suggested by Priority SIPP (or one of its variants).

- Applying this methodology to generate staffing requirements for the month of July reduces the computation time from 40 to 34 minutes on a 3GHz

machine with 2.96GB RAM. Whilst this is a small saving for this dataset, a 15% reduction in time could become quite considerable if requirements are requested for longer periods. The approach is particularly successful for the July dataset, as Priority SIPP frequently overstaffs. However, due to the requirement of the methodology to consider lower staffing levels even when the prediction is correct, it is also found to require fewer computations in many periods where Priority SIPP correctly predicts the minimum required staffing level. Even if Priority SIPP understaffs a period, the methodology increases the computational efficiency of the Euler method since it no longer needs to consider the excessively low staffing quantities that are below those predicted by Priority SIPP. The sole extra time required is that necessary to calculate the initial approximate requirements, but this is relatively small in comparison (around 3 minutes for the month of July).

iii. Using the staffing levels suggested generated from the standard SIPP approach (or one of its variants) with an aggregate arrival rate $\lambda = \lambda_H + \lambda_L$ and average allowable waiting time $x = \frac{x_H + x_L}{2}$ as initial staffing levels for each period, to be accurately analysed with the numeric methodology. The true minimum required staffing level can subsequently be confirmed by directing the Euler methodology to consider staffing levels just below those suggested by SIPP (if numerical analysis finds that the SIPP predictions are sufficient) until the waiting time targets are violated, or to consider quantities just above those suggested (if they are insufficient), until a sufficient number is found.

- Although standard SIPP requires less computation time than Priority SIPP (a few seconds for a one month scheduling horizon, compared to a few minutes), it is not always suitable for generating approximate staffing requirements for priority service systems, since the two priority classes cannot be treated as a single class. Even if the service requirements are the same for each class, the requirement that it must be achieved for *both* classes can require more servers than if it were only needed to be achieved on average for the two customer classes. Although the approximated staffing level can be reasonably close to those suggested by Priority SIPP if the waiting time targets are reasonably similar for each class of customer (as is the case with this investigation), the methodology marginally increases the computation time by around one minute, and if the targets are very different for each customer class the methodology would be less suitable again.

Figure 7.5 graphs the initial requirements suggested by each of the methods discussed above. It illustrates that the initial requirements suggested by Method (ii) (i.e. those one below the requirement estimated by Priority SIPP), match or fall just below the Euler requirements, providing motivation for this approach to be used as an efficient hybrid method.

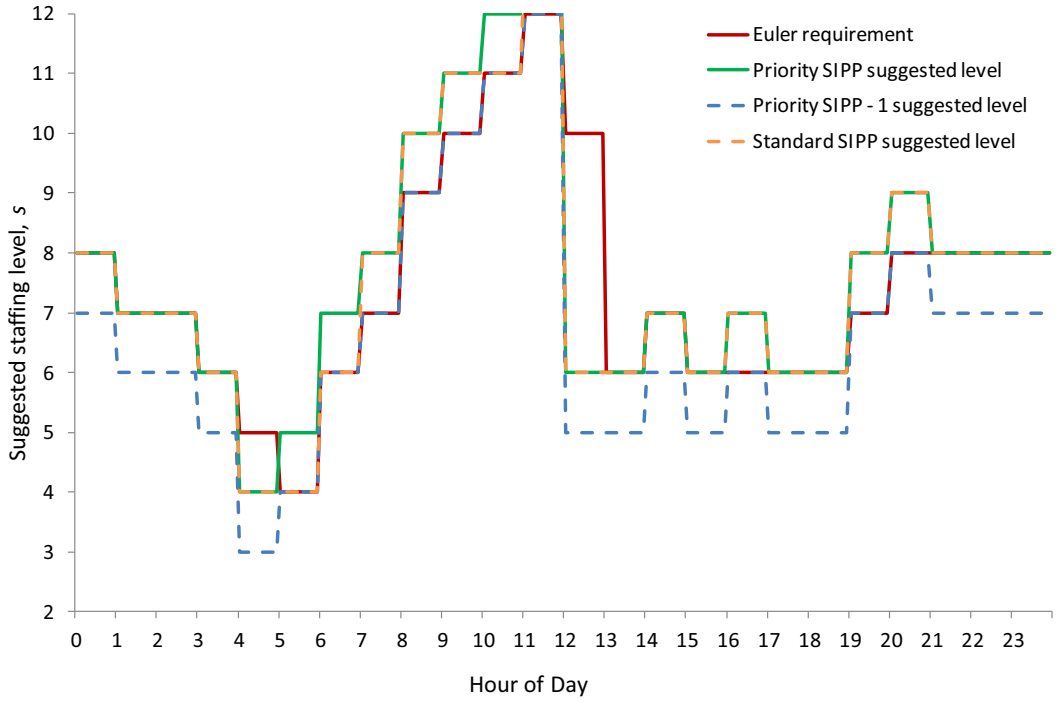

Figure 7.5: Initial staffing levels suggested by approximate approaches

It is observable that the approach would allow even greater time savings if it were applied to periods of longer durations, especially in epochs with increasing arrival rates, since the standard Euler approach could require the computations to be performed at several calculation intervals until one was found that failed to satisfy the performance target, for each incremented staffing level. A different approach to consider in increasing the efficiency of the Euler method could also involve increasing/decreasing the staffing level in a step function from requirement produced for the previous period, depending on if there is an increase/decrease in the average arrival rate. However, supplementary analysis finds this to be less efficient for the July data (in the order of a few minutes) than the proposed hybrid methodology, and it is also recognised as less appropriate for staffing scheduling horizons with greatly fluctuating demands from period to period. Therefore, this research recommends the hybrid approach outlined

in Method (ii) that uses staffing levels one below those suggested by the Priority SIPP methodology as initial staffing levels to be considered for each period, as the most appropriate technique to increase the efficiency of the Euler method.

7.4 Summary

This chapter has extended standard techniques commonly used in the literature to evaluate performance measures in time-dependent systems, so they may be applied to multi-server priority systems subject to time-dependent demand. It has demonstrated that with small revisions, both the approximation and numerical approach may be duly adapted to evaluate more complex priority systems, which are widespread throughout industry and commonly evaluated against a set of minimum performance standards. To date, staffing requirements have only been generated for multi-server priority systems with time-varying arrival rates using approximation methods, and whilst this study has furthered the work by Chen and Henderson (2001) in this area through providing greater insights into the Priority SIPP approach and proposing variations which increase the accuracy of its predictions; its real contribution to the literature lies in the extensions it proposes to the numerical approach. This chapter has explained how mappings may be derived to accurately track the probability state vector detailing the composition of customers in the system over shift boundaries, and proposed formulae to accurately evaluate the probability of an excessive wait for two customer classes. To the best of our knowledge, this study represents the first time that such analysis has been applied to a time-dependent priority system.

Section 7.2.1 noted that SIPP could be extended and efficiently applied to priority systems that are evaluated according to performance metrics for which closed-form formulas exist (such as the average expected waiting time for each priority of customer) by means of the methodology outlined by Priority SIPP. However, it is more laborious to apply the technique to systems where the performance measure of interest is a function of the waiting time of customers in the queue. In such systems, this research has confirmed the method proposed by Chen and Henderson (2001) to compute the probability of an excessive wait for HP customers using the inverted LST as an efficient technique, and developed formulae to predict the probability of an excessive wait for LP customers more accurately than the bounds presented in the same paper. Whilst greater computation time is required to compute the performance measure for two customer classes, meaning the approximation methodology is not as efficient as

for the non-priority case, Priority SIPP still has the benefit that it can be applied independently to any period of the day/week so no lengthy tracking of the system is required over time to recommend a minimum staffing levels for a particular shift, and also performs far more efficiently than the corresponding numerical method. In scenarios where the assumptions of Priority SIPP (i.e. low presented loads, relatively low RA) are broken, a method known as Priority Modified Lag Avg has been proposed to offer improved predictions. Whilst the method is not expected to improve the standard Priority SIPP performance in systems if the planning periods are excessively long or the presented loads are low, it is expected to be more robust in systems with higher loads and higher RA.

With regards to the numerical technique, this chapter investigated the Euler methodology, consistent with the work presented in Chapter 6. The notation first derived by Gail et al. (1988) was used to represent the composition of customers in the system, along with the extended balance equations to track the movement of customers through the priority system. The research has extended the work of Ingolfsson (2002) who defined the instantaneous transitions necessary to apply to track the movement of customers across shift boundaries in $M(t)/M/s(t)/FIFO/\infty/\infty$ systems, to track the composition of customers present in priority service systems over shift boundaries where servers operate under the exhaustive discipline. Not least has the approach been extended to define the transitions for a head-of-the-line priority system, but transitions have also been defined for both dummy and true shift boundaries. Furthermore, this research has devoted particular attention to further approaches followed by Green et al. (2007) and Ingolfsson (2002) to derive distinct waiting time formulas for HP and LP customers and to evaluate the probability of an excessive wait for both dummy and true shift boundaries.

Section 7.3 demonstrated how a subsection of WAST could be accurately modelled as a time-dependent priority system using the techniques developed throughout the chapter. In particular, the case study evaluated how close the staffing levels predicted by the Priority SIPP technique compared to those generated by the Euler solver, and illustrated that whilst the predictions were reasonably accurate, they were erroneous in certain periods, and could be improved by a method proposed as Priority Modified Lag Avg. The main conclusion that can be drawn from this work is that when computation speed is important, the approximate method is preferable; but the numerical method should always be used if accuracy is important. In situations where the

numerical method is selected, this research has offered an advanced hybrid approach which enables the accurate Euler requirements to be produced with improved efficiency.

All the models presented in this chapter have been logical extensions of the existing literature; and are both simple and practical to implement in a spreadsheet model that could be adopted by WAST. The workforce capacity planning tool developed in conjunction with this thesis and discussed in Chapter 10 directly demonstrates how both the Priority SIPP and Euler techniques can be incorporated into a practical tool to construct minimum staffing requirements, that can ultimately be used to inform the production of optimised staff schedules and rosters. The tool includes options to construct staffing requirements using either the approximate or numerical approach; so for each specific investigation, the client may choose if the approximate method is sufficient or if greater accuracy is required. If accuracy is of upmost importance, the investigations contained in this chapter have provided evidence to support the use of the hybrid approach over the standard numerical methodology, since the outputs are of equivalent accuracy - simply generated at different speeds. In future work, it would be useful to add the option to allow requirements to be generated using one of the variants of SIPP to the tool; however if this option were to be provided to an organisation, it would need to be accompanied by clear guidelines specifying the appropriate conditions under which each variant should be performed.

Chapter 8

Literature review (part 3): Scheduling and rostering

8.1 Introductory remarks

The term 'labour scheduling and rostering' collectively combines all techniques involved in the process of constructing work timetables for staff. Since many organisations (call centres, airlines, the NHS, etc) desire staffing levels that consistently match time-varying demands in order to remain competitive and meet minimum performance standards, it is imperative that staff are scheduled effectively. Whilst models that promote effective staffing and employee scheduling have a long history in the literature, there has recently been a major effort to develop the theory and application of the models to reduce the ever-increasing labour costs associated with the accelerating growth and development of the service sector (Atlason et al. (2008), Izady (2010) and Petrovik and Berghe (2012)). The models are primarily developed for systems subject to various constraints and stochastic conditions, and are concerned with two central issues (a) scheduling shifts optimally; and (b) assigning employees to shifts. These essentially mirror steps (iii) and (iv) of the typical process to schedule employees (Buffa et al. (1976) (see Chapter 1)) listed as:

i. Forecast demand

ii. Convert demand forecasts into staffing requirements

iii. Schedule shifts optimally

iv. Rostering: assign employees to shifts

Accordingly, most approaches to schedule employees assume that period-by-period staffing requirements are known or estimated (from step (ii) or otherwise), and focus on determining the lowest cost schedule that provides the required number of employees in each period.

This review explains the major research directions and modelling techniques investigated in the literature to schedule and roster employees in a range of service industries; highlighting the benefits and limitations in the methods. For a wider appreciation of the full scope of the scheduling literature, the reader is referred to Ernst et al. (2004) which provides a comprehensive summary of around 700 references in the area of personnel scheduling and rostering, classifying the methods according to the type of problem addressed, the application areas covered and the techniques used. The authors report that integer programming and constructive heuristic approaches are by far the most popular methods used to solve the staffing problem (each implemented in around 20% of the papers).

This chapter provides an overview of the common approaches, focussing on problems solved via integer programming, which is later considered to solve the WAST scheduling problem in Chapter 9 and incorporated in the workforce capacity planning tool discussed in Chapter 10. In particular, investigations that have developed schedules with the specific goal of achieving a minimum level of customer service at all times are described, since the techniques used to schedule WAST employees must meet comparable performance standards. Before beginning the review, the primary terminology used in the scheduling literature and this thesis is defined in Table 8.1:

Table 8.1: Glossary of terms used in scheduling and rostering

Scheduling models	
Labour staffing	Attempts to optimise the size and composition of the workforce for a medium planning horizon (e.g. a month / year)
Labour scheduling	Labour staffing for a shorter horizon (e.g. a day / week)
Shift scheduling	Constructing a work schedule for a given day
Tour scheduling	Constructing a work schedule for a given week
Constraints	
Hard constraints	Restrictions that must be satisfied at all costs
Soft constraints	Restrictions that are desirable but which may need to be violated in order to generate a solution
Working Time Directives (WTD)	The contract that defines allowable working hours for WAST personnel

Time intervals	
Period	The basic time interval for which planning typically occurs (e.g. 15 minutes, 30 minutes, 1 hour, etc)
Shift	A set of consecutive periods that represents a work schedule for an employee for a specific day
Shift length	The total quantity of time covered by the shift
Tour	A set of shifts that represents a work schedule for an employee for a week
Performance measures	
Service level	The percentage of customers served within a specific waiting time limit
Aggregate threshold level	The percentage of customers arriving within the duration of the entire planning horizon that management wishes to be served within the specified waiting time limit
Minimum acceptable service level	The lowest level of service that is acceptable in any planning period
Desired staffing level	The minimum quantity of staff required to allow the aggregate threshold level of service to be achieved for each period
Demand inputs	
Flexible demand	Future demand that is not known in advance, but may be predicted using forecasting techniques
Task based demand	Demand that is obtained from lists of individual tasks to be performed, which are usually defined in terms of a starting time and a duration, or a time window within which the task must be completed
Shift demand	Demand that is obtained directly from a specification of staffing requirements for different shifts

The main sets are defined as:

- I: the set of planning intervals
- T: the set of allowable tours

The optimisation models require the constants:

- c_t: the cost of assigning an employee to work tour t
- r_i: the desired staffing level in interval i
- b_i^{β}: the limit on the bounded shortage of employees in interval i
- b_i^{Π}: the limit on the bounded surplus of employees in interval i

- c^{α}: the cost of the unbounded shortage of employees, per employee-interval of shortage
- c^{β}: the cost of the bounded shortage of employees, per employee-interval of shortage
- c^{Θ}: the cost of the unbounded surplus of employees, per employee-interval of surplus
- c^{Π}: the cost of the bounded surplus of employees, per employee-interval of surplus
- $a_{ti} = \begin{cases} 1, & \text{if interval } i \text{ is included in tour (shift) } t \\ 0, & \text{otherwise} \end{cases}$

And the variables:

- x_t: the number of employees working tour t
- α_i: the unbounded shortage of employees in interval i
- β_i: the bounded shortage of employees in interval i
- Θ_i: the unbounded surplus of employees in interval i
- Π_i: the bounded surplus of employees in interval i

If both bounded and unbounded surpluses/shortages of employees are defined, then lower costs may be imposed for quantities falling within the pre-specified bounds.

8.2 Scheduling review

Shift scheduling involves selecting a sequence of shifts to be worked from a potentially large pool of candidates, and assigning a set number of employees to work each shift in order to satisfy the staffing requirements for each period. The classic formulation of the tour and shift labour scheduling problem was first addressed by Dantzig (1954) who approached the problem using linear integer programming (IP). The formulation of Dantzig's Labour Scheduling Model (DLSM) is given by:

Minimise,

$$Z = \sum_{t \in T} c_t x_t \tag{8.1}$$

Subject to constraints

$$\begin{gathered} \sum_{t \in T} a_{ti} x_t \geq r_i \text{ for } i \in I, \\ x_t \geq 0 \text{ and integer for } t \in T. \end{gathered} \tag{8.2}$$

Thus DLSM attempts to minimise the total labour cost (8.1) by allocating shifts subject to the constraint (8.2) that sufficient employees are present in all periods. In traditional scheduling models, the desired staffing levels for each period are set as the minimum number of staff who are able to provide the desired service level - (recall from Section 8.1 that the percentage of customers served within a set number of minutes after their arrival, constitutes the service level of interest in this study). In Chapters 6 and 7, formulations of SIPP and Euler methodologies were demonstrated that could be used to provide period-by-period staffing functions in line with this goal.

The basic form of the DLSM model may be extensively adapted to allow various constraints to be adhered to, such as restrictions imposed by management policy, working time directives or those that specify minimum service standards; and most current models used to solve the labour staffing or scheduling problem today are based on adapting formulations of this model. The adaption to Keith's Labour Scheduling Model (KLSM) (see Keith (1979)) has become a popular alternative that allows staff shortages and surpluses in certain periods. KLSM is formulated as:

Minimise,

$$Z = \sum_{t \in T} c_t x_t + \sum_{i \in I} \left(c^{\alpha} \alpha_i + c^{\beta} \beta_i + c^{\Theta} \Theta_i + c^{\Pi} \Pi_i \right), \tag{8.3}$$

Subject to constraints

$$\begin{gathered} \sum_{t \in T} a_{ti} x_t + \alpha_i + \beta_i - \Theta_i - \Pi_i = r_i \text{ for } i \in I, \\ \beta_i \leq b_i^{\beta} \text{ for } i \in I, \\ \Pi_i \leq b_i^{\Pi} \text{ for } i \in I, \\ x_t, \alpha_i, \beta_i, \Theta_i, \Pi_i \geq 0 \text{ and integer for } t \in T. \end{gathered} \tag{8.4}$$

KLSM allows staff shortages and surpluses to occur in all periods (subject to penalisation costs), and so the objective function aims to minimise the weighted costs

of staff wages and deviations from the desirable employee requirements. Lower costs may be imposed for shortages/surpluses falling within pre-specified bounds, meaning KLSM solutions often recommended staff numbers that deviate from the original requirements, but not by quantities so large that they exceed the two bounds.

The unique characteristics of service industries have lead to the investigations of numerous IP models in the literature, and several adaptions of the DLSM and KLSM have been proposed to account for particular settings relevant to various organisations. These include methods to account for periodic workforce requirements (Lin et al., 2000; Mason et al., 1998) part-time staff considerations (Burns and Carter, 1985), consecutive days off scheduling (Alfares, 1997), various break definitions (Aykin, 1996), complex working time directives (Beaumont, 1997a) and non-homogenous workforces (Lin et al., 2000). In recognition of the longer convergence time required by the models to solve problems with large constraint sets, research has been invested in developing more sophisticated methods to solve the underlying IPs by reducing the number of variables (Aykin, 1996; Brusco, 1998) and using cutting planes (Brusco, 1998; Jarrah et al., 1994).

Further revisions have also been applied to the objective function, such as those that allow staff to preference longer shift durations, achieve an even yearly workload distribution (see Beaumont, 1997b) and that consider the cost of customer inconvenience due to understaffing (Bailey, 1985). It is however not always possible to assign a monetary value to the cost of customer inconvenience; and in such situations management often aim to deliver a pre-specified service level as inexpensively as possible (e.g. through specifying target response times for WAST vehicles in the AOF targets (see Chapter 1.4)). The remainder of the review on the shift scheduling problem focusses on problems which schedule a flexible workforce in line with this goal, since the ultimate aim of the thesis is to roster WAST shifts that allow a consistent service quality to be provided, whilst minimising labour costs.

Although extensive studies have generated schedules based on formulations of DLSM and KLSM, these models have limitations as they often provide a level of service greater than the preferred level (since they are designed to ensure that every *period* delivers at least the aggregate level of service, rather than every *shift*). Also, if the period requirement is reduced to satisfy a *minimum* rather than *aggregate* level of service to overcome this problem, there is nothing to ensure that the aggregate level

is delivered for the entire shift.

Thompson (1993) further notes that managers are also often willing to reduce the staffing levels in some periods if this allows the overall labour scheduled to be considerably reduced. Such a reduction is usually performed manually and with little regard to the effect on service that the staff reduction may cause. Moreover, as DLSM and KLSM deliver staffing functions based on minimum staffing requirements for each period, approximations involved in previous steps to generate these requirements can sometimes lead to infeasible or suboptimal schedules. The problem arises as the staffing algorithms generally set the number of servers in each staffing interval independently of other periods, but in reality the service quality in each period is related to the staffing in previous and subsequent periods. To overcome the problem, iterative approaches have been proposed in the literature that integrate steps (ii) and (iii) of the typical process to schedule employees (see Section 8.1)).

The iterative approaches all switch between a schedule generator and a schedule evaluator, but differ in their approaches to obtain these components, and also in the way new constraints are added. The schedule generator searches for good shift schedules using exact or heuristic algorithms and the schedule evaluator estimates the service quality of the proposed schedules. Each time the service quality is evaluated, constraints are added to the schedule generator for periods where the service quality is below the specified targets. The iterative process continues until the service quality is met for all staffing intervals.

The Controlled Labour Scheduling Models (CLSM) described in Thompson (1993) avoid the limitations of the basic models through allowing both the number of employees needed in each period and the labour schedule, to be determined simultaneously. They also allow for the identification of the pareto frontier between service cost and service level. The first of the two revised models, the CLSM-CST, investigates the development of a staffing function that provides a pre-specified aggregate level of service to be provided as inexpensively as possible. The model includes one constraint that ensures that the minimum number of staff are provided for each period, and another to ensure that the aggregate level of service is provided over the duration of the operating period. As such, the constraint set enables the model to avoid the aforementioned limitations of the DLSM and KLSM. The second model, the CLSM-SRV, addresses the same problem but for situations where the workforce size or labour cost

is fixed. In place of minimising labour costs, the objective is formulated to maximise the proportion of customers served within the designated waiting time limit, whilst constraints are added to ensure that the minimum number of staff are provided for each period and that labour costs/labour size are below threshold levels specified by management. Both the CLSM-CST and CLSM-SRV offer advantages over the basic DLSM formulation and deliver the minimum acceptable level of service in all planning periods.

The investigation performed in Atlason et al. (2008) employs a simulation-based analytic center cutting-plane method. The study approaches the problems of determining minimum staffing levels while computing the best assignments to cover these staffing levels simultaneously; alongside queueing methods that compute the required staffing levels first and the shift assignments afterwards (using the staffing levels as input). In a different approach, Ingolfsson et al. (2010) uses an IP for the schedule generator, a randomization method for the schedule evaluation and adds constraints assuming the service quality is an exponential function of the number of servers. Whilst the method does not guarantee optimality (as do no others discussed above), it provides a good feasible solution with a lower bound on the minimum cost.

A more practical approach to the personnel scheduling problem is presented in Isken and Hancock (1991). Their approach successfully provides near-optimum solutions to large optimisation problems by combining a range of mathematical programming and heuristic techniques: the authors first relax the integer constraints in the scheduling model, secondly solve the resulting linear program, subsequently use a simple rounding heuristic to find a feasible solution to the integer program, and finally improve on the feasible solution found using a simulated annealing algorithm. The proposed model aims to provide a practical approach to difficult scheduling problems and is designed to complement current commercially available hospital staff scheduling systems.

Alternative approaches including staffing algorithms based on non-stationary infinite-server queues have also been investigated in the literature. These algorithms use the square root staffing formula (as discussed in Chapter 5.3) with an offered load equal to the mean number of busy servers in an associated time-dependent infinite-server system. Jennings et al. (1996) discover that this approach provides accurate results when generating staffing requirements for systems in which the delay probability (the probability a customer must wait before entering service) is the performance

measure of interest. Izady and Worthington (2012) further combine time-dependent infinite-server networks, the square-root staffing law, and use simulation to set staffing requirements in accident and emergency departments, modelled as time-dependent queueing networks.

Whilst the task of shift scheduling has been extensively studied in the literature, the particular application of the techniques to schedule ambulance crew shifts has surprisingly only recently been given more attention with three papers (Li and Kozan, 2009; Erdogan et al., 2010; Hari et al., 2011) which follow on from earlier studies by Trudeau et al. (1989) and Ernst et al. (1999).

The first study to simulate all operations of the ambulance service is presented in Trudeau et al. (1989); it investigates crew scheduling to satisfy the demand forecasts segmented into three-hour time blocks, combined with the investigation of optimal location of waiting points and a simulation model to account for the actual dispatch. Shifts are scheduled with the goal of minimising costs subject to a set of service-level constraints. The findings of this study are complemented by the descriptions in Ernst et al. (1999) of various network algorithms used to develop rosters for crew members, such as the shortest path algorithm and alternative network algorithm for cyclic rosters.

Li and Kozan (2009) investigate a two-stage approach to develop schedules with the particular goal of minimising the total number of shifts worked (essentially personnel costs) over a four-week planning horizon. In the first stage they develop a deterministic IP model to select shift start times and the requirement for ambulances in each shift, followed by an allocation model to assign ambulance officers to set shifts. An ambulance location model is also presented in Erdogan et al. (2010), complemented by two alternative IP models (with slightly different objective functions), both aiming to maximise expected ambulance coverage with probabilistic response times. Recently, Hari et al. (2011) also investigated a two-stage integrated approach for ambulance deployment and crew shift scheduling. For the sub-problem of shift scheduling, an IP model is formulated to determine optimal crew schedules, using the solution obtained from the first stage as input. The researchers ultimately develop optimal weekly schedules for ambulance crews from a pool of potential candidate shifts starting at various times of day and ranging from 10, 12 and 14 hour durations; and formulate an objective function, weighted by shift lengths, to minimise the number of shifts

scheduled subject to providing sufficient crew. The study discovers that solutions with multiple shift lengths weighted according to their shift lengths generate minimum slack solutions. Computational investigations show that the model and the algorithm are consistent, fast and provide optimal or near optimal solutions.

Much of the work discussed above is concerned only with the problem of deciding how many staff should be employed and when they are scheduled to work. However after solving any of these problems, it remains to assign individual staff to specific shifts. Ernst et al. (2004) comment that it is not usually computationally practical to deal simultaneously with all the elements needed to construct a roster, though such an approach is desirable from the perspective of achieving optimal rosters. When rostering to flexible demand, second order effects may arise as a result of selecting a particular roster. For example, it is not usually possible to exactly match the quantity of staff on duty for every period with the number required when employees are employed to work shifts that comprise several consecutive periods. Consequently, there may be certain periods with considerably higher staff numbers than the minimum recommended levels. If the congestion in the system is significantly reduced in these periods, it could lead to lower queue levels in later time periods.

In many settings staff assignment is performed manually through a consultive process or based on seniority (Ernst et al., 2004; Silvestro and Silvestro, 2000). In situations where this is not the case, the rostering depends on several factors, such as the type of demand forecast and the extent to which days off scheduling, line of work construction and task assignment are integrated. In contrast to the scheduling of EMS shifts (which has been extensively studied in the literature), the rostering of ambulance officers to these shifts has been given notably less attention as neither Erdogan et al. (2010) or Hari et al. (2011) consider this topic in their investigations. The following overview of the rostering literature thus focusses on applications of the methodology within other service industries. Particular attention is devoted to the nurse rostering problem that shares many parallels with the task to roster WAST employees, and has contrastingly become an attractive area of research.

8.3 Rostering review

Personnel rostering involves assigning human resources to a sequence of duties spanning a longer-term period of time (typically ranging from a week to a month),

subject to a set of given constraints (Petrovik and Berghe, 2012). The problem is often complex since it can contain a number of conflicting objectives and constraints; it is in these situations that automated approaches hold significant potential for improving the quality of those timetables. Mathematical or heuristic approaches can be used to generate several potential solutions, report upon the quality of schedules and attempt to divide the work evenly among personnel. One of the most significant benefits of automating the personnel scheduling process is a very considerable time-saving for the administrative staff involved (Burke et al., 2004).

There are a number of different line of work models, including cyclic rosters, acyclic rosters or stint-based rosters. Cyclic rosters develop timetables in which all employees of the same task perform an identical sequence of work, but begin their first shift at different times; and are thus generally developed for systems subject to repeating demand. Acyclic rosters are generally more applicable to hospitals, EMS and call centers as they allow greater flexibility in the construction of the timetable, which in turn allows fluctuating demand levels to be met through selecting independent lines of work for individual employees. Stint based rosters are also popular in service industries which operate over 24 hours a day since they only allow certain shift sequences called stints (Ernst et al., 2004).

Most rosters are significantly constrained by a number of legal, managerial and staffing requirements which govern allowable working patterns for individual employees. For example, such rules may impose limits on the total number of hours to be worked by an individual over a week, the number of sequential night shifts to be worked or the number of consecutive hours off between two shifts. For a nurse rostering problem, typical shift types include early (E), day (D), late (L), night (N) and off (O) shifts, but only certain sequences of stints may be allowable. Stint based rosters are thus popular in this setting as they may be used to prevent nurses from working certain shift sequences (e.g. two night shifts in a row (NN)) or to ensure that staff are awarded at least two consecutive days of within a week (OO) (Burke et al., 2004). A distinction must also be made between *hard* and *soft* constraints.

Once shifts are defined, various approaches can be used to assign lines of work to staff: the primary methods are *shift assignment* and *roster assignment*. Shift assignment involves allocating individual shifts to employees and roster assignment deals with the allocation of completed lines of work to staff members. If the assignment takes

place as the line of work is constructed then it is common to include individual staff preferences as part of the process. A number of approaches have been considered to allow for such preferences that include bidding systems in airline crew rostering (Gamache et al., 1998) or methods to formulate preferences made by staff as hard or soft constraints in nurse rostering.

Heuristics and metaheuristics have often been employed in the literature to solve staff scheduling problems since they allow complex objectives to be investigated, and can often produce a good feasible solution (although not necessarily optimal) in a limited amount of running time. Popular versions applied to scheduling and rostering include simulated annealing (SA) using swap and interchange based neighbourhood search heuristics. Constraint programming (CP) however tends to produce better solutions for more highly constrained problems; and some studies have further improved the proficiency of the approach by combining the methodology with traditional OR techniques (Ernst et al., 2004). In order to solve extremely complex set covering problems two primary methods have been suggested: one generating a limited number of columns to generate a reasonable sized formulation (Graves et al., 1993), and another partially generating all possible columns using column generation approaches (Lavoie et al., 1988; Ernst et al., 2004).

The major rostering focus in health systems has been in nurse scheduling, centering on the importance of the nature of the work and the difficulty in constructing timetables that can deal with complex individual work preferences. Whilst mathematical programming approaches are dominant in many areas of rostering such as call centers (Thompson, 1997) and retail (Haase, 1999), a wide range of sophisticated heuristic approaches have additionally been developed to construct good rosters for nurses. Some papers have approached the problem using goal programming (which extends general linear programming methods to consider multiple objectives, see Ozkarahan and Bailey (1998)) with weighted coverage and shift satisfaction terms in the objective function and the constraints to enforce hard rules; whilst others have used iterative algorithms to generate cyclic rosters, attempting to distribute the workload fairly (Smith, 1976). Millar and Kiragu (1998) further advance the previous methods considered in the literature by applying mixed IP models and optimisation techniques. Alternatively, tabu search has recently received attention in the literature (Bilgin et al., 2012) and has been shown to obtain optimal and near optimal solutions to a wide variety of classical and practical scheduling problems (Glover, 1990). Tabu

search is a particular type of metaheuristic which uses a local search procedure to iteratively move from one potential solution to another solution in the neighbourhood, until the stopping criterion is met. The unique feature of tabu search is that the exploration of the search space is governed by a set of rules which prevent all recently considered solutions from being repeatedly considered by the algorithm. The conditions for constraining and freeing the search process can be adapted by varying the number of iterations of the algorithm for which the solutions are retained in the memory function.

The primary difference between nurse and EMS rostering relates to the issue of service standards that are specified for the EMS in the form of targeted response times. The key paper published specifically related to the rostering of ambulance officers is that by Ernst et al. (1999). The approach taken to roster officers over a one year period is to first consider the allowable patterns of D, N and O shifts, together with annual leave periods. The authors use a transition matrix to specify which stints may follow each other and a shortest path algorithm to construct rosters, with a multi-objective function considering aspects such as demand coverage and an fair spread of workload across employees. A similar problem is considered in Taylor and Huxley (1989) to schedule police officers so that undercover is minimised. Shift patterns are fixed, but may be selected to start at different times, and improved rosters are generated using a local search heuristic with tabu restrictions.

8.4 Summary

The rostering process represents the final stage for optimising the efficiency of a service system; this commonly follows a sequence of previous steps (such as demand modelling, the specification of the work to be performed and the construction of optimised shift schedules). Although the modules are generally executed in a step by step procedure, in some implementations several of the modules may be combined into a single procedure. As can be seen from this review, a large amount of work has already been performed surrounding rostering and personnel scheduling as optimised staff schedules are recognised to provide enormous benefits and a contented workforce, since good rosters can minimise overtime, maximise coverage, and maximise satisfaction of personal preferences.

Much of the published work in the area of personnel scheduling has been shown to focus on a combination of heuristic and mathematical approaches. A summary

of the DLSM, which has become one of the most widely used and varied method for scheduling, has been provided in Section 8.1. Several alternative formulations of the model explored in the literature that have allowed improved performance in various settings have also been overviewed; and iterative approaches have been described that avoid infeasible or suboptimal solutions that may otherwise result from approximations. The next chapter proceeds to develop specific formulation of this model to optimise the scheduling of shifts for ambulance officers in WAST through refining the objective function and formulating constraints relevant to the Trust. An important issue when constructing schedules is if the service level constraints should apply at all times or be aggregated over a longer period: the decision should be taken carefully for each setting and may have a significant impact on the staffing required.

In the second part of the review, it was demonstrated that whilst little research has been invested in the application of the rostering literature specifically to the EMS, a large number of papers have investigated different approaches to particular rostering problems in other industries or sets of similar problem instances. Until recently staff assignment has commonly been performed manually, and even with today's technology it is a formidable task to efficiently solve complex linear integer programs whose search space may grow exponentially with the number of variables and constraints. Thus whilst it is often impossible to find an optimal solution, several approaches including metaheuristics, column generation and constraint programming have been used to provide good solutions. In cases where optimisation using a single one of these techniques proves inefficient, studies have attempted to combine heuristics with traditional OR techniques to solve extremely complex set covering problems.

Chapter 9 proceeds to apply a selection of the techniques discussed throughout this review to real-life data, and as such investigates the final stage in improving the effectiveness and efficiency of WAST. Whilst there is a wealth of material on labour scheduling, the majority of the research to date has notably focussed on developing a fixed work schedule for an employee to be repeated on a cyclical basic and this thesis attempts to narrow the gap in the literature through investigating methods that incorporate time-dependent demand to provide more flexible solutions.

Chapter 9

Scheduling and rostering

9.1 Introductory remarks

The main objective of this chapter is to develop the analytical models discussed in Chapter 8, in order to analyse staffing issues at WAST and determine the matching of personnel resources to fluctuating demand requirements. The first part of this process involves investigating the potential of IP models and heuristic search techniques to optimise the shift schedule and determine the minimum number of staff required to provide sufficient coverage for a monthly (four week) period, as inexpensively as possible. Secondly, a simple rostering model is proposed that assigns a set of ambulance staff to appropriate shifts.

As explained in Chapter 1, calls to the ambulance service in Wales have tripled over the last 20 years which has placed a tight constraint on resources. Coupled with the typical urgency of patient requirements for EMS assistance, proper staffing is critical to providing quality care. Yet staff scheduling is no easy task given the uncertainty of demand and the magnitude of staffing options, and this chapter accordingly investigates analytical methods to aid the resource allocation process. Whilst this research devotes significant effort to the development of high quality shift schedules, and addresses the problem using similar approaches to those summarised in Chapter 8; less focus is given to the construction of an optimised roster since the investigation of the entire list of methods considered in Chapter 8.3 lies outside the scope of this study, and a simple heuristic is shown to be capable of generating a feasible and effective result that adheres to various legal and managerial requirements. It would however be useful for future studies to explore the potential of a selection of the methods discussed in Chapter 8.3 to further optimise EMS rosters.

To demonstrate the application of the methodology developed in this chapter, the techniques are applied to a subsection of WAST data, namely EA responses in the Cardiff region (identical to the dataset used in Chapter 7). Following the approach described for the previous case study, the two ambulance officers required to staff each EA are assumed to be paired; thus the scheduling investigations that follow design joint schedules for each pair of officers, referred to as 'crew'.

This chapter is structured as follows. Section 9.2 considers methods that may be employed to solve the shift scheduling problem when it is formulated in terms of an IP model; followed by Section 9.3 which constructs an IP to roster ambulance crews to shifts. In both sections, the models are formulated prior to being applied to real-life WAST data in case studies that investigate the potential of heuristic approaches to generate high quality solutions. In cases where it is possible to solve the IP optimally, the optimal solutions are used to evaluate the heuristic performances. Finally, Section 9.4 summarises the performance of all approaches discussed above and suggests future areas for research.

The previous chapter demonstrated that numerous studies have developed models to schedule shifts, including several that are appropriate for an EMS (Li and Kozan, 2009; Erdogan et al., 2010; Hari et al., 2011). Thus Section 9.2 delves immediately into the investigation to develop suitable comparative models to optimally schedule EA shifts in the Cardiff region. Conversely, papers investigating the rostering of ambulance crews to shifts are not as common in the literature (with the exception of recent papers by Li and Kozan (2009) and Hari et al. (2011)). Hence the methods developed in this research to assign crews to shifts are first overviewed and discussed in Section 9.3 prior to investigation of the case study.

However, it is not the intention of this section to consider the 'best' method to construct a roster (considered outside of the scope of this thesis), but more to develop a simple and practical algorithm that may be embedded within the workforce capacity planning tool to ultimately assign shifts to employees in a reasonable fashion.

9.2 WAST shift scheduling model

Practically all EMS agencies plan their crew shifts in advance, although shift lengths vary by agency (Hari et al., 2011). The Cardiff branch of WAST currently use 11 pre-

defined shifts (detailed in Appendix A.4) that last for a combination of 5, 9, 11 or 12 hours, and start at various times of the day. The number of crews assigned to each of the shift currently varies on a daily basis, but this assignment is repeated identically for each week throughout the year. This section formulates a suitable IP model to construct a flexible shift schedule (i.e. one that specifies a unique selection of shifts for each day) that aims to minimise the total labour hours required whilst providing sufficient coverage. After considering a practical heuristic solution approach to the problem, the model is also solved optimally, which permits criticism of the heuristic solutions.

9.2.1 The IP model

Using the minimum number of EA crews required for each period generated from the Priority SIPP methodology in Chapter 7.3 as input, this section outlines the formulation of an IP model to determine the number of crew to assign to each shift type, to minimise the total number of labour hours required (essentially cost). The approach can fundamentally be considered as an adaptation of the DLSM, which was outlined in Chapter 8. The model formulation is broken down into two stages: the model is firstly presented in a generic format in order to describe the notation which is minimally revised from that discussed in Chapter 8 (to facilitate the formulation of the constraints in the following step); and this is followed by a specific parametrisation that provides optimised schedules for WAST. Defining the sets and variables as:

- D: the set of days of the week
- P: the set of periods in a day
- S: the set of allowable shifts
- x_{sd}: the number of crews working shift s on day d
- r_{dp}: the desired crew requirement in period p of day d
- c_s: the cost of assigning a crew to work shift s
- $a_{sp} = \begin{cases} 1, & \text{if period } p \text{ is included in shift } s \\ 0, & \text{otherwise} \end{cases}$

Then the general scheduling model can be written as:

Minimise,

$$Z = \sum_{s \in S} \sum_{d \in D} x_{sd} \tag{9.1}$$

Subject to constraints

$$\sum_{s \in S} x_{sd} a_{sp} \geq r_{dp} \text{ for } p \in P,\ d \in D, \tag{9.2}$$

$$x_{sd} \geq 0 \text{ and integer for } s \in S,\ d \in D. \tag{9.3}$$

The objective function presented in equation (9.1) attempts to minimise the number of crews assigned to each shift by allocating shifts subject to the constraint (9.2) so that sufficient employees are present in all periods. However, equation (9.1) simply minimises the total number of shifts, which implies that all shifts are weighted equally. Observing that WAST shifts are comprised from a combination of durations, in order to minimise the true cost (i.e. number of labour hours required), a more effective staffing model can be achieved by weighting each of the shifts in equation (9.1), either directly by shift lengths:

$$Z = \sum_{s \in S} \sum_{d \in D} x_{sd} \sum_{p \in P} a_{sp} \tag{9.4}$$

or by pre-defined weights c_s, $s \in S$:

$$Z = \sum_{s \in S} \sum_{d \in D} x_{sd} c_s \tag{9.5}$$

The weights proposed in equation (9.5) provide a more versatile approach, since they can either be chosen to directly reflect the shift length, or some other quantity of interest relating to the selection of each particular shift.

For the specific problem of generating an optimised shift schedule for EA crews in the Cardiff region, it is important to note that although the staffing requirements obtained from Priority SIPP (to be entered as the right-hand sides of the key constraints in equation (9.7)) exhibit some seasonality, for reasons as indicated in Chapter 2; they still vary to some extent from week to week, and significantly vary throughout the year. For this reason, it is the researcher's choice to allow greater flexibility in the model when investigating the potential of the model to produce an optimised shift schedule for the first four weeks of July (mirroring the period investigated in

Chapters 6 and 7), through permitting different shifts to be scheduled every day rather than repeated on a cyclical basis as is commonly seen in the literature (Bard et al., 2003; Ernst et al., 2004; Hari et al., 2011). The specific scheduling model to optimise the tour of shifts selected for the first 28 days of July may accordingly be formulated as:

Minimise,

$$Z = \sum_{s=1}^{11} \sum_{d=1}^{28} x_{sd} c_s \tag{9.6}$$

Subject to constraints

$$\begin{gathered} \sum_{s=1}^{11} x_{sd} a_{sp} \geq r_{dp} \text{ for } p = 1, 2, \ldots, 24, \ d = 1, \ldots, 28, \\ x_{sd} \geq 0 \text{ and integer for } s = 1, 2, \ldots, 11, \ d = 1, 2, \ldots, 28. \end{gathered} \tag{9.7}$$

Note that $s = 1, \ldots, 11$ to represent the 11 potential shifts currently used by WAST.

Whilst the weights assigned to the shifts are designed to be flexible for each specific formulation of the problem; for the purpose of this investigation, weights are selected that reflect both the duration and preference of each shift, such that:

$$c_s = l_s \times p_s \tag{9.8}$$

where l_s represents the length (hours) of each shift ($\sum_{p=1}^{24} a_{sp}$) and

$$p_s = \begin{cases} 0.95, & \text{if shift } s \text{ operates for less than 9 hours} \\ 1, & \text{if shift } s \text{ operates for exactly 9 hours} \\ 1.05, & \text{if shift } s \text{ operates for more than 9 hours.} \end{cases}$$

Hence the weights are defined as a product of the shift length and a factor which favours longer shifts. For example, the first shift listed in Appendix A.4 is defined as 06:00-15:00 (9 hours duration) and so $c_1 = 9 \times 1 = 9$.

Using the weights as defined above, the model to schedule EA crews to respond to all emergency requests in the Cardiff region for July 2009 becomes:

Minimise,

$$Z = \sum_{s=1}^{11} \sum_{d=1}^{28} x_{sd} l_s p_s \tag{9.9}$$

Subject to constraints

$$\begin{gathered} \sum_{s=1}^{11} x_{sd} a_{sp} \geq r_{dp}, \quad \forall\, p = 1, 2, \ldots, 24,\ d = 1, \ldots, 28, \\ x_{sd} \geq 0 \text{ and integer}, \quad \forall s = 1, 2, \ldots, 11,\ d = 1, 2, \ldots, 28. \end{gathered} \tag{9.10}$$

9.2.2 Solving the IP heuristically

This section considers heuristic search techniques that can be exploited to solve the shift scheduling problem defined above. Whilst commercially available IP solvers such as CPLEX or XPress-MP are often employed for this purpose, specialised software is not always available. In such cases, heuristics are often used to produce good quality solutions in a small amount of time; however they usually lack the ability to find an optimal solution (Abramson, 1991). This research is ultimately concerned with methods that may be integrated as part of the workforce capacity scheduling tool to be offered to WAST employees; thus whilst the ability of standard IP solvers to generate solutions for specific problem instances is acknowledged (and will later be investigated in Section 9.2.3), the research focus is centered on algorithms which may be embedded within the Excel software.

Local search is a general strategy for heuristic optimisation that involves iteratively applying small changes (moves) to a candidate solution, attempting to improve its quality by evaluating the effects of the move and then deciding whether to accept or reject the change. The criteria used for choosing a move and then deciding whether to accept or reject the move is the distinguishing feature between the various heuristic techniques. The random descent approach is the simplest type of local search: here a randomly chosen move is only accepted if it produces a solution at least as good as the current solution. SA is more sophisticated in that it may accept worsening moves with a certain probability. A comprehensive review of the theory and applications of this technique is contained in Laarhoven and Aarts (1987). The primary benefit of this methodology is that there is a lower probability of becoming stuck in a local optimum, as the probability of accepting a worsening move (given

in equation (9.11), where ν is the cooling parameter), is lowered during a run of the algorithm at a rate defined by the cooling schedule. The performance of SA is highly dependant on the choice of parameters such as the starting temperature, which may be adapted for each particular problem (Pirlot, 1996). The way in which the temperature is altered is critical to the success of the algorithm. Whilst decreasing the temperature more gradually allows worse solutions to be accepted with a higher probability for a greater number of iterations, it widens the exploration of the search space to a larger degree; giving an even greater probability of reaching the global optimum eventually through higher computational effort (Laarhoven and Aarts, 1987).

$$e^{-\frac{\Delta c}{\nu}}, \qquad \Delta c = \text{Solution Score}_{new} - \text{Solution Score}_{old} \tag{9.11}$$

The construction of both a random descent and SA algorithm to find an approximate solution to the shift scheduling problem is explained below. The algorithms are implemented in Excel VBA (the package which also implements the methodology of the previous sections) and allow a flexible formulation of the scheduling problem. For both algorithms, the initial feasible schedule is produced using a greedy algorithm: taking shifts in order of starting times as presented in Appendix A.4, the shift which first includes coverage of this period is scheduled as many times as required to provide sufficient coverage for this period. Shifts are continually assigned in this way until all periods have sufficient coverage.

In the algorithms, a swap operator is employed as the move operator. This operates by randomly selecting a shift $x_{s_1 d_1}$ ($s_1 \in S,\ d_1 \in D$) in the schedule such that $x_{s_1 d_1} > 0$, and reducing the number of staff scheduled to work shift $x_{s_1 d_1}$ by one. If all constraints remain satisfied after the removal of the shift from the schedule, then no further action is needed. Otherwise, if the reduction in the number of shifts scheduled causes $\sum_{s=1}^{12} x_{sd} a_{sp} < r_{dp}$ for some $p = 1, 2, \ldots, 24,\ d = 1, \ldots, 28$ (i.e. a violation of constraint (9.2)), then to maintain integrity of the solution, a corrective procedure is applied which randomly selects other shifts $x_{s_2 d_2}, x_{s_3 d_3}, \ldots$ that each resolve the violation of this constraint for at least one period, to add to the timetable until all periods regain sufficient coverage. It should be noted that due to the manner in which shifts are added and removed, there may be some iterations in which no new shifts are added, and others where several additional shifts may need to be assigned (if the

first shift added does not resolve the violations for all periods).

There are a vast number of different proposals in the literature regarding the stopping criterion of a heuristic algorithm (see Laarhoven and Aarts, 1987; Pirlot, 1996; Rayward-Smith et al., 1996). These include terminating the algorithm after a set number of iterations or if the expected improvement in cost, if the algorithm were to be continued, is small. The SA method may alternatively be terminated if the cooling parameter reduces to a particular quantity (Laarhoven and Aarts, 1987). For the purpose of this investigation, the neighbourhood operator is repeatedly applied for a set number of iterations. With regards to the selection of appropriate temperature values, Rayward-Smith et al. (1996) propose that the initial temperature be selected by choosing a very high temperature and cooling it rapidly until about 60% of worsening moves are being accepted: this forms the real starting temperature, found equal to 14 for the IP model presented in equation (9.9). Further experimentation finds that by applying a cooling parameter of 0.92 after every 500 iterations, it is ensured that enough iterations are performed so that the system stabilises at each temperature, whilst allowing its reduction to a small quantity in reasonable time.

Before the results of the heuristics are presented in Section 9.2.4, the problem of solving the problem optimally is considered.

9.2.3 Solving the IP optimally

This section presents the optimal results for the IP model presented in equation (9.9), solved using XPress-MP optimisation software (Dash Optimization Inc, 2004). As the software however requires all constraints to be coded formally (i.e. as a full IP model, in place of heuristics which may apply simple feasibility checks in conjunction with a well-coded objective function), the formulations of specific constraints first requires further consideration.

Recall that in the formulation of model (9.9), $s = 1, \ldots, 11$ to represent the 11 potential shifts currently used by WAST. However, as the scheduling process is concerned with assigning shift to 'days' considered as running from 6am-6am, then the 10pm-7am shift (which is the only shift to overlap the day boundary) must be formulated as comprising of two separate shifts (namely 10pm-6am (the 11th shift input) and 6am-7pm (the 12th shift input)). Subsequently, the model must instead specify $s = 1, 2, \ldots, 12$ with the

additional constraints:

$$x_{11,d} = x_{12,d+1}, \quad \forall\, d = 1, 2, \ldots, 27 \tag{9.12}$$

$$x_{12,1} = 0 \tag{9.13}$$

Constraint (9.12) ensures that any crew assigned to work the last shift on day d (10pm-6am) must also work for the first hour on day $d+1$ (6am-7am) because these two work stints technically form the same shift. Also, because the model schedules shifts for the 28 day period independent of the period immediately preceding and following this epoch, it is necessary to impose constraint (9.13) since the staffing function is considered to begin at 6am on day 1 (so there are assumed to be no staff previously working in the system, meaning the system is assumed empty for the last shift on day '0').

Hence the model to schedule EAs to respond to the all emergency requests in the Cardiff region for July 2009 becomes:

Minimise,

$$Z = \sum_{s=1}^{12} \sum_{d=1}^{28} x_{sd} l_s p_s \tag{9.14}$$

Subject to constraints

$$\begin{gathered} \sum_{s=1}^{12} x_{sd} a_{sp} \geq r_{dp}, \;\; \forall\, p = 1, 2, \ldots, 24, \;\; d = 1, \ldots, 28, \\ x_{sd} \geq 0 \text{ and integer}, \;\; \forall\, s = 1, 2, \ldots, 12, \;\; d = 1, 2, \ldots., 28, \\ x_{11,d} = x_{12,d+1}, \;\; \forall\, d = 1, 2, \ldots, 27 \\ x_{12,1} = 0. \end{gathered} \tag{9.15}$$

The results of model (9.14), solved using XPress-MP within the region of a few seconds on a 2002 operating system with 3GHz and 2.96GB of RAM, are displayed in Figure 9.1.

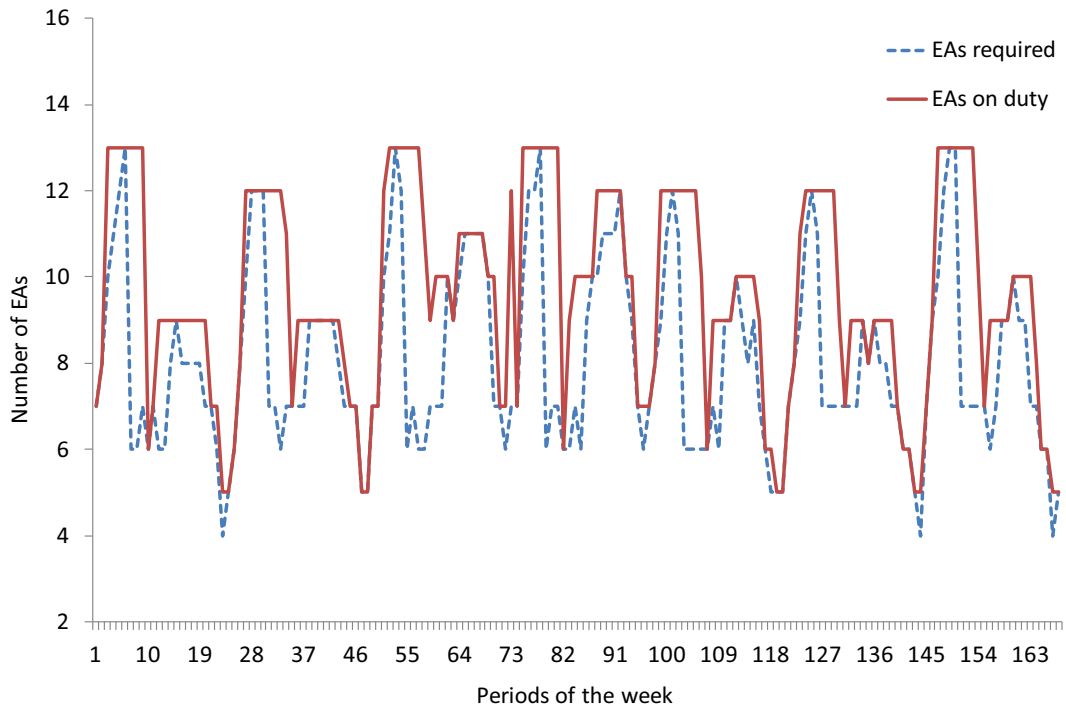

Figure 9.1: EA coverage for each period using optimised schedule from pool of original Cardiff shifts, July 2009. (For clarity, results are displayed for the first 168 periods (1 week) only)

The model suggests that staffing levels fairly accurately match the requirement for each period. However, due to the inconvenient overlapping times of some of the shifts currently used at Cardiff, Figure 9.1 demonstrates that there are some periods with an oversupply of resources. The main time periods subjected to this oversupply are highlighted as those between 12:00-14:00 in Figure 9.2.

In light of the results displayed in Figure 9.2, it can be noticed that a simple modification to the first shift entered in the candidate pool from 06:00-15:00 to 06:00-12:00 (see Appendix A.4), could allow a far lower optimal cost to be achieved. Figure 9.3 illustrates the improvement resulting from this change, followed by the optimised schedule of shifts in Table 9.1. Using the particular weights outlined in equation (9.8), the optimal achievable cost is reduced by 7% from 6,481 to 6,051.

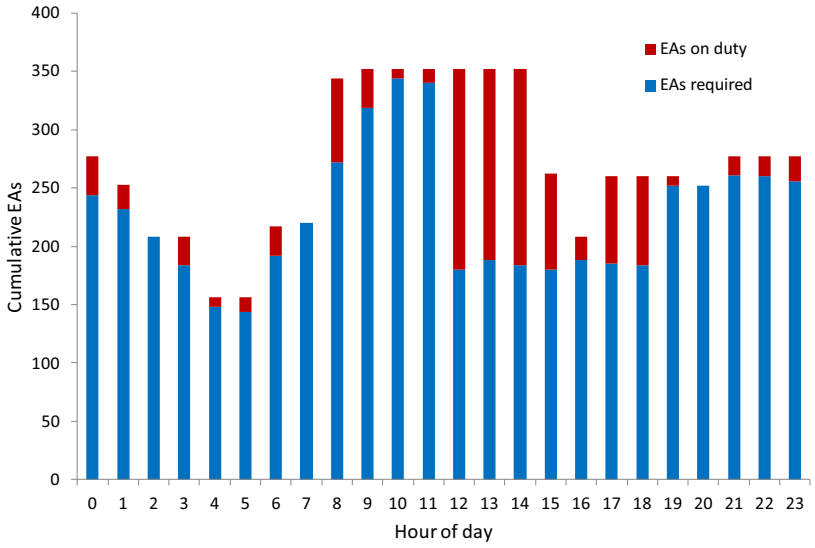

Figure 9.2: EA oversupply arising from scheduling original Cardiff shifts, July 2009. (The number of EAs supplied/required represent the cumulative number for all days in the first four weeks of July 2009)

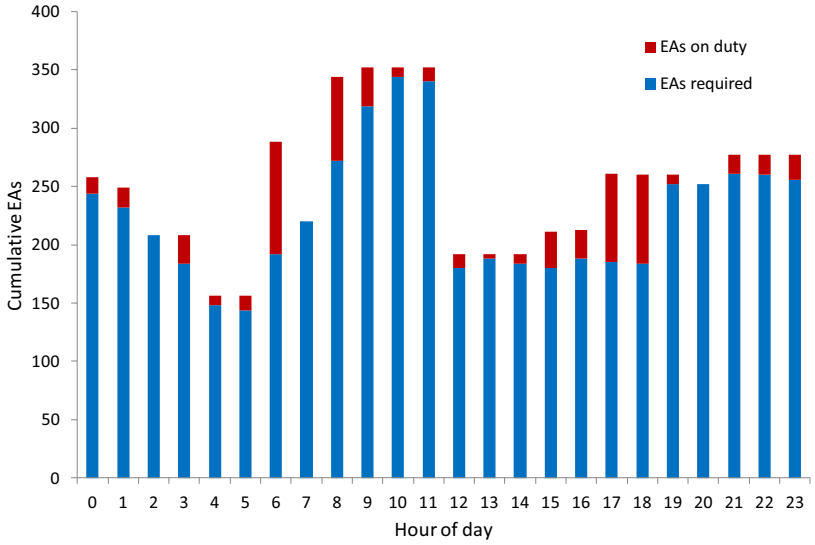

Figure 9.3: EA oversupply arising from scheduling revised Cardiff shifts, July 2009. (The number of EAs supplied/required represent the cumulative number for all days in the first four weeks of July 2009)

Table 9.1: Optimal shift assignments, July 2009

Shift Type	Total	Number of crew enrolled for each day of July																										
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	160	6	5	6	6	6	5	6	6	5	6	6	6	5	6	6	5	6	6	6	5	6	6	5	6	6	6	5	6
2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	59	1	3	1	1	2	3	3	2	3	1	1	2	3	3	2	3	1	1	2	3	3	2	3	1	1	2	3	3
4	124	5	4	5	6	4	3	4	5	4	5	6	4	3	4	5	4	5	6	4	3	4	5	4	5	6	4	3	4
5	8	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	1	0
6	19	1	1	0	0	1	2	2	1	1	0	0	0	2	0	1	1	0	0	0	0	2	1	1	0	0	0	0	2
7	9	0	0	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	1	2	0	0	0	0	0	1	2	0
8	52	2	2	3	3	1	1	1	2	2	3	3	1	1	1	2	2	3	3	1	1	1	2	2	3	3	1	1	1
9	172	6	6	6	7	7	5	6	6	6	6	7	7	5	6	6	6	6	7	7	5	6	6	6	6	7	7	5	6
10	25	0	0	2	2	1	1	1	0	0	2	2	0	1	1	0	0	2	2	0	1	1	0	0	2	2	0	1	1
11	131	5	5	5	5	4	4	4	5	5	5	5	5	4	4	5	5	5	5	5	4	4	5	5	5	5	5	4	4

It is of course possible to further reduce the cost by offering a different pool of shifts for selection. For example, the oversupply of resources in periods 06:00-07:00 and 17:00-19:00 in Figure 9.3 suggests that further reductions could be gained from adjusting the start and finish times associated with shift 4 and shift 9 which overlap these periods. However, Figure 9.4 demonstrates that the resulting reduction in the optimal cost for model (9.14) is marginal compared to the practical implications that are likely to be associated with changing additional shifts. Thus for the remainder of the investigations, the shifts input to the models are identical to those originally proposed for Cardiff, except for the single modification outlined above for shift 1.

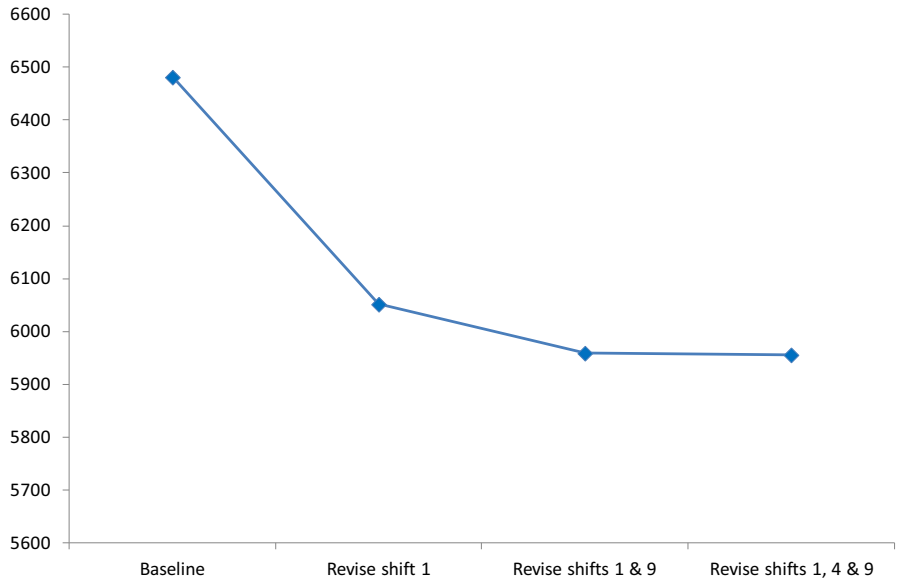

Figure 9.4: Total cost for various shift pools, July 2009. (Shift 9 is revised from 6:00-18:00 to 7:00-16:00 and shift 4 is revised from 8:00-17:00 to 9:00-17:00)

The final staffing schedule given in Table 9.1 suggests that the crew numbers employed for the 1st, 9th and 11th shift should remain fairly consistent on a daily basis; but the lowest labour cost is achieved by allowing the numbers employed for the 3rd, 4th, 6th, 8th and 10th to fluctuate to a greater extent depending on the daily demand. The 2nd, 5th and 7th shift are rarely selected by the model; but are nevertheless retained as they are considered 'acceptable' by WAST and their removal would increase the cost marginally, even if not by a significant quantity.

9.2.4 Evaluation of heuristic approaches

The ability of both a random descent and SA algorithm for finding an approximate solution to the shift scheduling problem as formulated in equation (9.9) with constraints (9.10) is explored below. The techniques are evaluated according to how well they schedule shifts using the data and revised shift patterns (determined at the conclusion of Section 9.2.3 and listed in Appendix A.4), and their performance is measured through their ability to reach the optimal cost achieved by the IP model.

Figure 9.5 reports the results of the two heuristic algorithms in constructing a desirable schedule for the July data. The improvement made to the solution over 80,000 iterations is depicted, which reqires around 2 minutes of computation time on a 2002 operating system with 3GHz and 2.96GB of RAM. Using the IP solver, the best cost achievable for this scenario was found to be 6,051 in Section 9.2.3; hence the heuristics may be evaluated by computing the rate at which the solutions converge to this optimal solution. It is clear that the SA method can be used to produce near optimal solutions if the algoirthm is run for a sufficient number of iterations, and Figure 9.5 demonstrates that the solution comes within 3.7% and 0.8% of the optimal solution after 20,000 iterations and 60,000 iterations respectively. The random descent method makes rapid improvements to the initial solution, but only marginal improvements after around 10,000 iterations until it reaches a local optimum. In summary, although the random descent may be used to quickly produce a reasonable quality solution, the SA technique provides the best overall schedule.

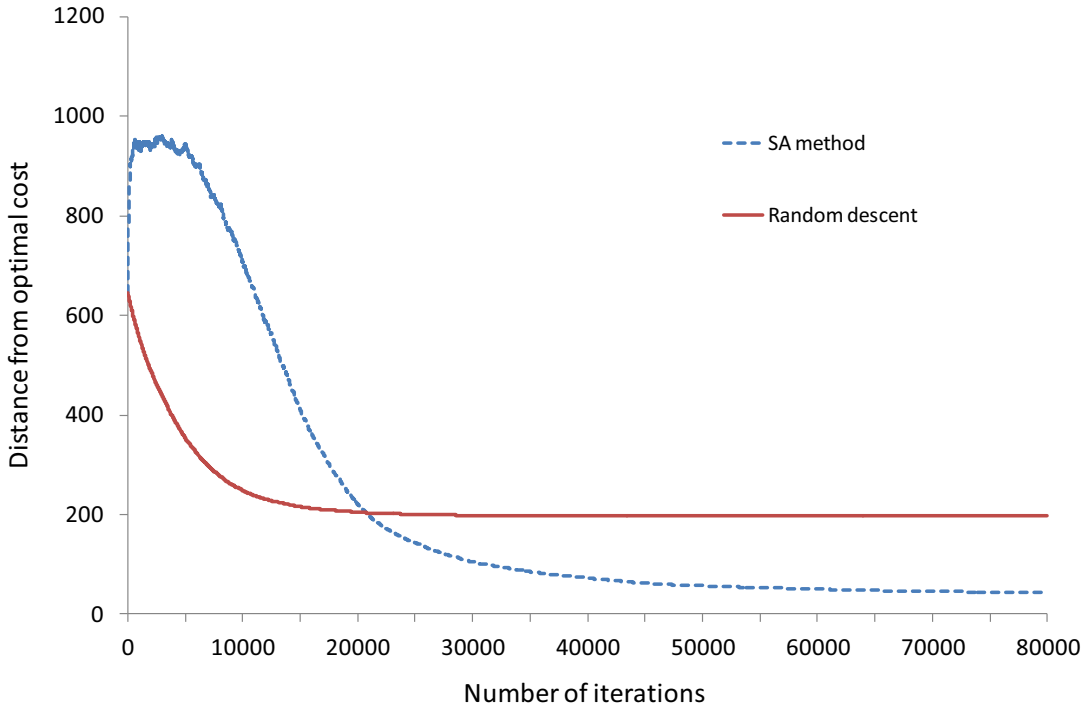

Figure 9.5: Rate at which heuristics converge to optimality (averaged over 50 runs)

Whilst IP solvers may be utilised to quickly generate solutions for small instances, as has been shown above, such software is not always available and is unsuitable to be be incorporated in the workforce capacity planning tool. Heuristic approaches also have the benefit that they may be applied to large problem instances, to problems with intricate objective functions that can be difficult to fomulate in the structure required by IP models, and can further cope with complex constraints. Moreover whilst IP solvers require constraints to be coded formally (so a different formulation is needed for every set of shifts, and additional consideration must be awarded to constraints that need to be added due to the definition of the working day such as constraint (9.12)), heuristics may be readily adapted to construct desirable schedules for organisations with different problem instances.

After solving any of the above shift scheduling models, it remains to assign individual schedules to specific personnel. This problem is addressed in Section 9.3.

9.3 WAST crew allocation model

Rostering ambulance officers is a highly constrained optimisation problem. As outlined in Chapter 8, when constructing a roster, healthcare institutions must take into account various legal, management and staff requirements. The imposed requirements are described by a large set of constraints which are usually divided into two categories: hard constraints (that must always be satisfied) and soft constraints (which are desirable to be met, but may be violated with a penalisation cost in certain circumstances). The set of constraints differs from Trust to Trust, and those relevant to WAST are discussed in the WTD which are outlined in the in the 'Agenda for Change' handbook (The NHS Staff Council, 2011).

Chapter 8 demonstrated that a large number of methods have been developed to solve rostering problems. IP approaches have been considered for simple cases such as assigning a small selection of shifts to a group of equally skilled employees over a limited time period (Li and Kozan, 2009), but exact methods are often not appropriate for larger and/or more complex problems. In general, heuristic or meta-heuristic searches are needed to solve real-world problems addressing requirements concerning shifts, work regulations, part time work, skill categories, legal constraints, personal requirements, etc. Rostering problems are often formulated as optimisation problems. Numerous objective functions have been formulated which generally attempt to address the subjective viewpoints, often including minimising the total number of workers, minimising the number of constraint violations, minimising overtime, maximising coverage and maximising satisfaction.

This research considers optimisation models that may be employed to roster EA crews in the Cardiff region in order to provide sufficient coverage, and to minimise labour costs. A heuristic approach to roster staff is developed and evaluated. It aims to produce a feasible and good quality timetable, although this is not guaranteed to be optimal. Since IP solvers are only capable of solving small instances, the rostering of crews is investigated for a limited one-week period, selected as the first week of July 2009. While the heuristic approach may be refined to consider numerous constraints (as they do not need to be strictly coded as formal entries to an IP model), the formulation of the model required by the IP solver means it is only capable of considering limited criteria; hence in order to evaluate the ability of the heuristics to produce good quality solutions, both the IP model and heuristics are initially developed for the limited set

of criteria outlined below as a subset of the rules imposed by the WTD. The list of constraints outlined by the WTD are given in terms of 'acceptable' working hours when averaged over a yearly period; hence the rules listed here are based upon a logical interpretation of the guidelines to clarify the components considered necessary for a feasible weekly timetable:

- *Coverage requirements:* The minimum number of crews that must be assigned to particular shifts (obtained from the outputs of the shift scheduling model in Section 9.2) or to particular periods (obtained from Priority SIPP methodology in Chapter 7) on specific days.
- *Max weekly work:* The standard hours of a full time worker is 37 $\frac{1}{2}$ hours. Each worker must not work more than 42 hours in the 7-day period.
- *Night work:* Each worker must work no more than 8 hours of night-time work (i.e. periods from midnight-5am inclusive) in the 7-day period.
- *Min rest between shifts:* Each worker must be allocated 11 hours of continuous rest between shifts.

A feasile schedule is one in which the coverage constraints are satisfied at all given times, and crews are allocated shifts that satisfy the WTD. A desirable schedule is one that further achieves a low cost (which may be defined in terms of working hours, crew size or the number of violations of soft constraints). Given the typical requirements for EAs in the Cardiff region, a typical problem requires the consideration of around 40-50 crews. It is common to develop timetables based on step-by-step procedure (Buffa et al., 1976), so the coverage requirements are taken as the outputs from the shift optimisation model provided in Section 9.2. Whilst this approach is considered desirable from the perspective that it provides good quality solutions to highly constrained problems in reasonable time, the separation of the tasks can result in the failure to find the global optimum; and a case study is presented to illustrate this issue in Section 9.3.3.

9.3.1 The IP model

The simplest version of the crew allocation IP model takes the coverage requirements generated from the optimised shift schedule reported in section 9.2 as input, and aims at reducing costs by rostering staff in such a way that the size of the workforce is

minimised. The following notation is needed in extension to that in Section 9.2.1:

Sets:

- J: the set of ambulance crew

The decision variable x is revised from a 2-dimensional to a 3-dimensional variable such that:

- $w_{jsd} = \begin{cases} 1, & \text{if crew } j \text{ is works shift } s \text{ on day } d \\ 0, & \text{otherwise} \end{cases}$

A dummy variable is needed to denote if a crew is employed at any point during the week, i.e.:

- $\text{crew}_j = \begin{cases} 1, & \text{if crew } j \text{ is assigned at least one shift} \\ 0, & \text{otherwise} \end{cases}$

The following model formulation investigates this approach to roster EA crews in the Cardiff region for the first week of July 2009. 50 staff are potentially offered to the model for selection, as preliminary investigations show that the quantity needed to satisfy the demand requirements should be far lower than this quantity.

Objective function

Adhering to the shift coverage requirements (as output from Section 9.2), the model directly minimises the total number of crew to effectively minimise cost. The objective function accordingly may be stated as:

Minimise,

$$Y = \sum_{j=1}^{50} \text{crew}_j \tag{9.16}$$

Constraints

The model must primarily ensure that enough crew are assigned to each shift to satisfy the coverage requirements:

$$\sum_{j=1}^{50} w_{jsd} = r_{sd}, \ \ \forall s \in 1, 2, \ldots, 12, \ d \in 1, 2, \ldots, 7 \tag{9.17}$$

Each crew also contributes a set of individual constraints (max hours, hours of night work etc) to the model, as each shift is created to fit a single crew. Recalling that $\sum_{p=1}^{24} a_{sp}$ represents length of each shift, equation (9.18) ensures that the maximum number of hours worked in the 7-day period does not exceed 42 hours:

$$\sum_{s=1}^{12}\left(\sum_{d=1}^{7} w_{jsd}\sum_{p=1}^{24} a_{sp}\right) \leq 42, \ \ \forall j \in 1,2,\ldots,50 \tag{9.18}$$

The total night hours worked by each crew should not exceed 8 hours per 7-day period. Recalling that night work includes periods from midnight-5am inclusive (i.e. periods 19-24 for a day considered to operate from 6am-6am), equation (9.19) ensures that this constraint is upheld:

$$\sum_{s=1}^{12}\left(\sum_{d=1}^{7} w_{jsd}\sum_{p=19}^{24} a_{sp}\right) \leq 8, \ \ \forall j \in 1,2,\ldots,50 \tag{9.19}$$

The more complex constraint to account for in the IP model is the rule specifying the minimum rest hours that must be awarded to employees between shifts. Due to the allowable shift types detailed in Appendix A.2, it is observable that a simple constraint to ensure that ambulance staff do not work more than one shift in a day (i.e. $\sum_{s=1}^{12} w_{jsd} \leq 1, \ \ \forall s \in 1,2,\ldots,12, \ d \in 1,2,\ldots,7$) would not appropriately deal with this constraint as certain shifts within the same 'day' may in fact be assigned to the same crew whilst still upholding the condition. In order to allow the particular assignment of such shifts to the same crew requires constraints in (9.20) to be added to the model $\forall\ j \in 1,2,\ldots,50, \ d \in 1,2,\ldots,7$:

$$\begin{aligned}\sum_{s=1}^{10} w_{j,s,d} &\leq 1\\ \sum_{s=2}^{11} w_{j,s,d} &\leq 1\end{aligned} \tag{9.20}$$

Additional tour types that violate the rest time constraint must be also be disallowed. The discovery of the precise tours that violate the constraint first requires the enumeration of all possible combinations of shifts, followed by a study to find those that require removal from the allowable pool. The addition of the constraints in equation (9.21) ensures that the rest break rules are upheld for this case study.

$\forall\ i \in 1, 2, \ldots, 50,\ d \in 1, 2, \ldots, 6$:

$$\begin{aligned}
\sum_{s=5}^{11} w_{j,s,d} + \sum_{s=1}^{2} w_{j,s,d+1} &\leq 1 \\
\sum_{s=6}^{11} w_{j,s,d} + \sum_{s=1}^{5} w_{j,s,d+1} &\leq 1 \\
\sum_{s=10}^{11} w_{j,d,s} + \sum_{s=1}^{8} w_{j,1,d+1} &\leq 1 \\
w_{j,d,11} + \sum_{s=1}^{9} w_{j,s,d+1} &\leq 1
\end{aligned} \tag{9.21}$$

Constraint (9.22) ensures that the same crew is assigned to the 11th shift on day d and 12th shift on day $d+1$ (essentially the same shift):

$$w_{j,11,d} = w_{j,12,d+1},\ \ \forall d = 1, 2, \ldots, 7 \tag{9.22}$$

Finally, the dummy variable to count the number of staff employed for at least once shift over the 7-day period may be constructed as:

$$\text{crew}_j \geq w_{jsd}\ \forall\ j \in 1, 2, \ldots, 50, s \in 1, 2, \ldots, 12, d \in 1, 2, \ldots, 7 \tag{9.23}$$

with the specifications:

$$\begin{gathered}
w_{jsd} \in 0, 1\ \ \forall j \in 1, 2, \ldots, 50,\ s \in 1, 2, \ldots, 12,\ d \in 1, 2, \ldots, 7; \\
\text{crew}_j \in 0, 1\ \ \forall j \in 1, 2, \ldots, 50.
\end{gathered} \tag{9.24}$$

The IP model can be adapted to produce revised schedules based on alternative shift assignments, coverage requirements and objectives by revising its formulation. Since individual crews are not able to work several tours of the pre-optimised shifts output from Section 9.2 due to the complex staffing constraints and WTD, it is observable that a better solution may be achieved if all allowable shift types are considered for selection in the formulation of *this* model, although it increases the model complexity. Hence by introducing the constraints attributed to each crew at the outset of this problem and listing the period-by-period coverage requirements in place of the shift requirements, it is possible to further reduce the size of the workforce.

Furthermore, recognising that methods employed to construct good quality rosters

should simultaneously seek to reduce the total size of the workforce and total labour hours; this research proposes a new objective function which considers the assignment of overtime hours as a violation of a soft constraint, and hence penalises any such assignment with an additional cost in the objective. The revised formulation presented in equation (9.25) aims to minimise a weighted function of the total crew size and the number of overtime hours assigned to staff members.

Minimise,

$$Y = 25 \sum_{j=1}^{50} \text{crew}_j + \sum_{j=1}^{50} \text{overtime}_j \tag{9.25}$$

where overtime_j represents the number of overtime hours assigned to each crew j ($j \in 1, 2, \ldots, 50$).

The WTD specify that a full time worker should work an average of 37 $\frac{1}{2}$ hours per week, but as the duration of all the shifts detailed in Appendix A.4 are integer multiples of hourly quantities, it is impossible to construct a 37 $\frac{1}{2}$ hour schedule for an independent week; therefore overtime is calculated as any quantity over 38 hours. Since the WTD specify that employees should not work more than 42 hours per week, each crew may be assigned between 0-4 hours of overtime. The coefficient in front of the crew_j variable may be adjusted to appropriately preference the goal of reducing the size of the workforce over the number of overtime hours assigned. The objective function given in equation (9.25) is built upon the assumption that overtime hours are paid at a rate 1.5 times higher than the standard wage; so as the standard working week consists of 38 hours, the weekly pay for 1 full member of staff should be equivalent to around 25 overtime hours.

9.3.2 Solving the IP heuristically

This section presents a practical approach to finding an approximate solution to the rostering problem by considering the potential of a SA algorithm to obtain near-optimal solutions to the IP model in reasonable time. Whilst other techniques overviewed in Chapter 8 such as column generation, CP or graph colouring may be employed to produce high quality rosters, these techniques have already received a great deal of attention in the literature and it is not the intention of this section to consider the 'best' method to construct a roster (considered outside of the scope of this thesis), but more to develop a simple algorithm that may be embedded within the workforce

capacity planning tool to ultimately assign shifts to employees in a reasonable fashion. The algorithm considered in this chapter offers a practical approach to solving the personnel scheduling problem, in that it may be executed to generate a good quality staffing function in reasonable time on a personal computer. In order to evaluate the technique against known optimal solutions, the heuristic is initially constructed for the problem described by the IP model in Section 9.3.1, adhering to an identical set of constraints. However since heuristics may use simple feasibility checks to determine whether certain constraints are upheld (rather than requiring them to be formally coded in IP formulations), these approaches are capable of considering additional WTD constraints than the limited set considered by the IP model, and an example of this is given in an extension of the model discussed at the conclusion of Section 9.3.4.

The heuristic approach presented below aims to find a set of meritorious solutions to the objective function defined in equation (9.25) which concerns the goal to simultaneously reduce cost in terms of the total workforce size and number of overtime hours required. In order to generate good quality solutions, two SA algorithms are combined (one considering total labour hours and another examining the overall cost); yet if only one variable is of interest in the objective (e.g. in cases where the composition of the workforce is pre-defined or employees work overtime at the same rate of pay) then individual subsections of the heuristic may be sufficient to provide good solutions. It is the decision of the researcher to outline the most complex heuristic in detail below, to be followed by discussions surrounding simplified versions for less complex objectives in the evaluation of the technique (see Section 9.4). The comprehensive model additionally permits the shifts allocated to crews to differ from those selected by the optimised shift schedule model, although the execution of the heuristic is aided by the provision of a reasonable selection of shifts to allocate initially (taken as the set of shifts generated by the SA model in section 9.2.2, when no improvement has been made to the cost for 1,500 iterations).

The initial feasible schedule is produced using a greedy algorithm as follows: taking shifts in chronological order, assign each one to the first crew (ordered in terms of desirability) that is able to work this shift without breaking any of the constraints imposed, until all shifts are assigned to crews.

The neighbourhood operator attempts to reduce the number of crews from the number output by the greedy algorithm using several processes (swapping of both the shifts

themselves and the assignment of shifts to employees) and permitting the changes to be accepted/rejected according to SA criterion at two points within each iteration. The approach is outlined in the following steps:

i. Select a random shift allocated to an employee $w_{j_1 d_1 s_1}$ and delete this shift from the timetable

ii. Test if the deletion of the shift violates the coverage constraint for any given period

- If the coverage constraint is met (so $\sum_{j=1}^{50} \sum_{s=1}^{12} w_{jsd} a_{sp} \geq r_{dp} \; \forall \; p = 1, 2, \ldots, 24, \; d = 1, \ldots, 7$) then no corrective procedure is necessary. Go to step (iii)
- If the coverage constraint is violated, then to maintain integrity of the solution, select a random shift to add to the schedule that resolves the violation of this constraint for at least one period, and assign this shift to the first crew in the list who is feasibly able to work this shift. Continue adding other shifts $x_{j_3 s_3 d_3}, x_{j_4 s_4 d_4}, \ldots$ in this fashion until the coverage constraints are re-satisfied

iii. **SA 1**: If the total labour hours are less than or equal to those required previously, accept the new solution as the updated current solution. If not, accept the new solution as the current solution with some probability

iv. When the set of shifts has been selected, the operator progresses to swap shifts between crews in an attempt to provide sufficient coverage with a smaller workforce. In order to widen the search space and rapidly reduce the number of crews employed, the neighbourhood operator considers the potential to re-assign *every* shift in the timetable in the following order:

- Crews are listed in ascending order of total weekly working hours. (The idea of first considering shifts assigned to the crew working the fewest number of hours is that the re-assignment of all of their work to other crew members generally requires the least effort. If all their work can be re-allocated to other crews already in the schedule, the size of the workforce may be immediately reduced)
- Within each crew, shifts are listed in descending order of shift durations. (The longest shifts are considered first, under the assumption that longer shifts are potentially harder to re-allocate)

- Alternative crews are investigated for their potential to take on each shift in order given at the outset of the investigation (so more 'desirable' staff are considered first)
- If an alternative crew is able to work the shift, it is re-allocated; otherwise the assignment remains unchanged and the next shift is considered
- After all shifts belonging to a particular crew have been considered, the remaining staff are re-ordered in terms of total weekly working hours (since this is likely to change with the re-assignment of shifts), before the remaining shifts are considered
- The process continues until all shifts have been considered for re-allocation

v. **SA 2**: If the current cost is lower than or equal to cost selected in step (iii), the new solution is accepted as the updated current solution. If not, it may be accepted with some probability

vi. The algorithm terminates after 2,000 iterations

The swapping function defined in step (iv) which swaps shifts between crews in an attempt to actively reduce the crew size, may additionally be applied to the initial solution before the main heuristic is executed. This single swap algorithm generally makes a vast improvement to the cost, and hence provides a good quality starting solution. The heuristic can then be applied to seek marginal improvements in an attempt to provide a better staffing solution. Yet to mirror the fact that a good solution is provided at the outset of the algorithm, the SA parameters must be adjusted accordingly to only accept a small proportion of worse solutions (to avoid nullifying the work of the initial swap function). In order to concurrently ensure that a large range of alternative feasible solutions are considered, an advantage is seen in allowing the cooling rate to be re-set to higher temperatures if no improvement is made to the cost in a set number of iterations, say after 200 if the algorithm is run for a total of 2,000 iterations.

Similarly to Section 9.2, consideration is now devoted to solve the problem optimally; before the results of the two approaches are displayed and compared in Section 9.3.4.

9.3.3 Solving the IP optimally

The result of the IP model formulated above to roster EA staff in the Cardiff region using the pre-optimised shifts output from Section 9.2 and the simple objective function given in equation (9.16), is presented in Table 9.2. As shown in the table, a weekly roster has been established using 39 ambulance crews, with each crew working an average of 38.56 hours (so a total of 3,008 labour hours are required to provide this coverage for the first week of July in total).

When the model is revised to allow the selection of all shifts given in the revised potential pool (see Appendix A.4) to satisfy the period-by-period coverage requirements, the IP solver finds that sufficient coverage may be provided with 38 crews, rather than the 39 required by the model based on the pre-optimised shift schedule. The benefit gained from the removal of a crew from the timetable is however compromised by an increase in the total labour hours assigned to staff, as the solution with 38 staff requires a total of 3,066 labour hours (2% more than previously).

However, standard IP models often fail to provide optimal solutions to sizeable problems which include variables that penalise the violation of soft constraints, since their formulations are generally more complex. For example, the standard XPress-MP MIP model often runs out of memory before successfully finding the optimal solution as its default strategy is best-bound search. For the objective function given in (9.25), with period-by-period coverage requirements, the best solution found by XPress-MP is 1,002; yet this cost is not necessarily optimal. The LP relaxation of the problem generates a lower bound of 600 which is far below the best cost found for the full IP model.

Restrictions of IP solvers

Whilst the above study demonstrates that IPs may be employed to rapidly find optimal solutions to small problem instances with relatively simple objectives, they are generally not utilized to optimise crew rosters due to several shortfalls as outlined below:

- IPs lack the power to find optimal solutions for rostering problems with more complex objectives
- The models cannot be flexibly adjusted to consider new shift types as a unique formulation is generally required for different problem instances
- Complex constraints cannot always be formulated in the format required by the

Table 9.2: An example schedule for Cardiff EA staff, first week of July 2009

Crew	Shifts assigned to each crew for each day of July						
	1st	2nd	3rd	4th	5th	6th	7th
1	9	-	3	4	11	12	9
2	1, 11	12	11	12	4	5	-
3	4	1	11	12	4	-	8
4	-	1	5	4	1,11	11,12	12
5	-	3	8	-	6	8	-
6	2	-	1,11	11,12	12	-	3
7	-	1	1,11	11,12	12	1	4
8	1	-	4	10	9	-	4
9	11	11,12	12	4	3	-	4
10	9	-	1	11	12	4	9
11	1	-	9	10	-	4	6
12	-	8	-	4	9	9	-
13	1	4	-	1	-	10	9
14	11	12	4	9	-	-	9
15	6	-	10	9	-	1	6
16	-	11	12	3	1	3	1,11
17	-	4	-	4	4	1,11	11,12
18	-	9	9	-	9	-	3
19	-	8	9	-	4	9	-
20	4	11	12	8	-	3	1
21	-	9	-	1	9	9	9
22	3	9	-	9	8	-	-
23	11	12	1	1,11	12	3	4
24	4	-	9	-	1	-	11
25	8	11	12	-	1	6	-
26	1,11	12	1,11	12	1	6	-
27	1	-	4	-	10	9	-
28	9	11	12	1	3	9	-
29	-	1	4	1,11	12	1	1,11
30	9	9	9	9	-	-	1
31	-	1	10	-	1	4	9
32	8	6	-	8	-	-	1
33	9	-	8	-	9	-	3
34	4	3	1	9	-	-	10
35	9	9	-	1	11	12	-
36	4	-	8	8	-	1	-
37	-	9	9	9	9	-	1
38	-	4	4	9	9	11	12
39	4	3	-	4	11	11,12	12

models. For example, the discovery of all tours disallowed by WTD may initially require complete enumeration of all possible shift combinations to eliminate those that violate certain constraints. Whilst this is not unreasonable for some problems (e.g. nurse rostering which commonly involves rostering nurses to a combination of four shifts (E/D/L/N), or to find the tours that violate the 11 hour rest hour constraint above), it is impractical to enumerate all possible tours for the data employed in this case study to allow the addition of the WTD constraint which specifies that employees should be awarded 35 consecutive hours of rest in a weekly period.

Although the investigation required to outline all the infeasible tours arising from the 35 hour constraint is intractable, it may be noted that the constraint is upheld for the majority of employees in the example timetable created in Table 9.2 anyhow, as the additional constraints specifying the maximum working hours (constraint 9.18) and minimum rest periods between shifts (constraints 9.20-9.21) only allow a limited number of tours which allocate more than 4 shifts to workers (as at least one of these must be less than 5 hours due to constraint 9.18). Further rules to then specify allowable days-on and days-off patterns could further be formulated to avoid the limited selection of such tours that still violate the 35 hour constraint; but the task becomes intractable if there are too many complex constraints to consider. In such cases, another technique known as the 'pattern approach' is sometimes considered which involves outlining a number of feasible tours to allocate to staff (in place of unique shifts); but this is again not so practical for this case study, due to large number of shift types offered in the allowable pool.

9.3.4 Evaluation of heuristic approaches

This section presents the results of the heuristic to produce a staff roster based on reducing the cost in objective function (9.25) subject to constraints (9.17-9.24). In light of the above IP restrictions, together with the need for specific software to efficiently find optimal solutions to the rostering problem, this research promotes the use of the proposed heuristic to construct desirable rosters. The heuristic benefits from being flexible in approach; so features such as shift times, the number of allowable shifts, etc can all be flexibly adjusted and entered into the model without requiring its re-formulation. Moreover it has the ability to consider constraints that are difficult to formulate in the format required by IP models, and may be ultimately embedded as

part of an Excel workforce capacity tool in line with the final goal of this thesis.

Figure 9.6 shows the average path taken by the algorithm on its journey to convergence over 50 runs, highlighting there are several points at which worsening moves are accepted. The parameter values used for the initial temperatures and cooling rate in the SA schedule presented in the chart are based upon the findings of preliminary investigations. In the final model, the initial temperatures are set as 20 and 7 for the labour hours and cost functions respectively, and are lowered by multiplying the current temperature by corresponding factors of 0.9 and 0.9995 after each iteration of the algorithm. However, if no improvement is made to the the cost function in 200 iterations, the temperature is raised (capped at a lower level each time the procedure is invoked) to allow a greater exploration of feasible solutions.

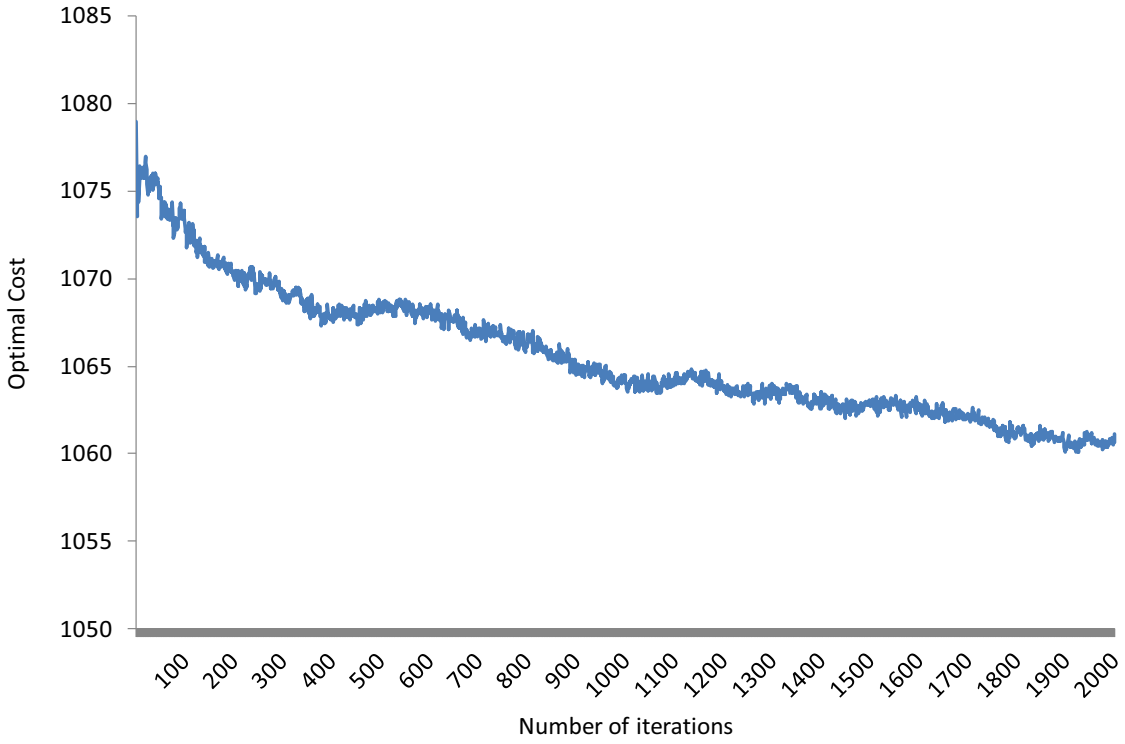

Figure 9.6: Rate at which heuristics converge to optimality (averaged over 50 runs)

Figure 9.6 shows that the optimal cost selected by the model after 2,000 runs is 1,061, although average *best* solution found for each run is actually far lower than this (equal to 1,054, SD=8.9), since the best solution is not necessarily the cost selected at the termination point of the SA algorithm. The vast majority of the improvement to the solution quality actually occurs with the first application of step (iv) to the initial

solution (which reduces the cost from around 1,250 to 1,080); and Figure 9.6 shows that the main algorithm proceeds to marginally reduce this cost by investigating alternative feasible solutions. The average optimal cost value obtained of 1,054 is actually within 5% of the best solution found by the IP model (recall that this was found as 1,002 in Section 9.3.1). In terms of computation time, the execution of 2,000 iterations of the heuristic method on a 2002 operating system with 3GHz and 2.96GB of RAM requires in the region of 30 minutes; and whilst XPress-MP continally seeks better solutions until it runs out of memory (after around 12 hours), it finds solutions of a similar quality to the heuristic in 1-2 minutes.

The real benefit of the heuristic lies in its ability to produce good quality rosters for complex objective functions and to consider complex constraints. The spreadsheet model is also flexible to consider alternative shifts and produce rosters for different durations based on varying constraints. Shifts may be changed/added to the potential pool; and parameter values for constraints such as maximum working hours, maximum night hours allowed, coverage requirements, etc, may be all directly adjusted by the user. These changes are possible due to the precise coding used to execute the algorithm which allows these new conditions to be automatically accounted for when constructing a roster, so no new formulation is needed (other than to update the *objective function* itself or add new *types* of constraints). Whilst the WTD constraint which specifies that employees should be awarded 35 consecutive hours of rest in a weekly period proved intractable to include in the IP model, it can be easily coded and embedded in the heuristic model to produce appropriate schedules. When this constraint is added to the heuristic model, the average best solution found over 50 runs of the model is 1,060: illustrating that it is indeed possible to produce a desirable staffing function with a very similar cost if staff are chosen in an effective way.

When the heuristic is applied to the same instance investigated for the IP model with the objective function (9.16) to simply minimise the size of the workforce adhering to constraints (9.17-9.24), it generates a roster requiring an average of 40 crews (sd = 0.64) over 50 runs. This is a significant improvement upon the initial solution generated by the greedy algorithm which uses around 46-48 crews. The execution of the swap function (prior to the execution of the main algorithm) subsequently reduces this quantity to around 43 crews, and the remainder of the improvement is made using the SA heuristics. Recall that the IP model found that the optimal timetable required 38 crews. Hence the case study demonstrates that the heuristic is capable of finding a

good quality solution, but not necessarily optimal.

Further investigations discover that the algorithm performance improves with simpler shift structures. Moreover, if set shifts are entered to the model (for example, the optimised shifts selected from the shift scheduling model), and the problem is simply to allocate the shifts to crews, then only steps (iv) - (vi) of the main algorithm are required. For example, for the simple cost function to reduce the size of the workforce given in equation (9.16) using the set of optimised shifts output from the scheduling model in Section 9.2.3, the objective cost is reduced from around 48 to 40 crews (compared to the optimal value of 39) in a single application of the swap function outlined in step (iv) alone. Thus if the objective is simply to find a good quality solution, the heuristic approach may be more appropriate for reasons such as it allows complex objective functions and constraints to be considered; can be flexibly adjusted; and is practical to be embedded into a spreadsheet model to produce good quality solutions to a large scale optimisation problem in reasonable time on a personal computer. Due a certain quantity of unavoidable uncertainty associated with future demand levels, it also seems sufficient for approximate solutions (not necessarily optimal) to be generated at this stage, and fine-tuned when short-term decisions are made that allow extra personnel to be added to the schedule based on updated information.

9.4 Summary

This chapter has analysed the problem of scheduling ambulance crews to shifts in order to reduce personnel costs whilst providing sufficient period-by-period coverage requirements as determined from the queueing models developed in earlier chapters. Whilst the majority of previous papers have analysed the issues of optimising shifts schedules and assigning crews to shifts as two separate problems (or ignored the latter problem completely), this chapter has acknowledged the benefit in constructing the shift schedule in conjunction with the allocation of shifts to specific crews, due to complex WTD constraints which prevent crews from working particular tours of the optimised shift schedule.

Prior to the consideration of allocating specific shifts to employees, the shift scheduling problem was considered in its own right. Both exact and heuristic approaches were considered to offer solutions to the problem; and the advantages and limitations were discussed for each. Whilst an IP solver was demonstrated to offer an optimal solution

to the case study, the SA heuristic was shown to produce a solution that came close to the optimal with the advantage that it may be embedded within a workforce capacity planning tool without requiring specialised software, and may be flexibly adjusted by the user to allow the consideration of different shift types.

Both an exact and heuristic method were also developed and compared to solve the rostering problem, and evaluated for their potentials to offer a good quality solutions. Whilst common IP solver packages such as XPress-MP were demonstrated to offer optimal solutions for simple cases, complex constraints and objectives were explained to be sometimes intractable to formulate, and the IP solvers were found to run out of memory on hard instances (since their default strategy is best-bound search). For such cases, a practical heuristic approach was proposed to provide good quality solutions, but not necessarily optimal, in a reasonable amount of time on a personal computer.

The main challenge that has been highlighted in this chapter is the challenge to integrate the separate steps of the rostering process into a single problem. When the stages are merged, the problem can become NP-hard, but the heuristic method proposed in this chapter has offered a practical approach and illustrated how approximate solutions can be generated to difficult problem instances. Chapter 10 shows how the proposed heuristics can be further incorporated as part of a workforce capactiy planning tool to produce efficient schedules and rosters. The resulting rosters produced by the tool will not necessarily be optimal, but nevertheless enable managers to avoid extreme undesirable understaffing an overstaffing situations.

There are a number of directions for future studies. If time permitted, the logical areas to progress this work would be to investigate iterative approaches to the problem to the rostering problem, to study the solutions generated from alternative techniques discussed in the literature review, to investigate alternative models formed by varying the objective function and set of constraints considered, and to investigate the benefit that could be gained from rostering each employee individually rather than as a paired 'crew' unit.

Chapter 10

Workforce capacity planning tool

10.1 Introductory Remarks

This chapter illustrates how each of the three components (forecasting, scheduling and rostering) investigated in this thesis are brought together in the final product. It contains a description of the workforce capacity planning and scheduling tool that has been produced in conjunction with the research, which allows the automation of the processes necessary to effectively allocate resources at WAST. The tool incorporates functions that allow future demand to be forecast, period requirements to be set in accordance with the response time targets, shift schedules to be optimised and rosters to be formulated following the methodology that has been discussed and developed throughout this thesis. The model is embedded within Excel software, and the algorithms and methodology to optimise the resource allocation process are implemented in VBA. The Excel VBA package has primarily been selected for the reason that it combines a programing language with a simple interface that is widely available within many organisations.

The workforce capacity planning tool is currently populated with historic data relating to requests for WAST EMS assistance arising in the SE region of Wales between April 2005-June 2009 and it is set up to optimise ambulance crews for EAs for the first week of July 2009, mirroring the test period investigated in several of the case studies contained within this thesis. The assumptions required by the model (such as the assumption that one paired EA crew is required to attend every incident reported, and the precise waiting time targets) stand as outlined in Chapters 1 and 2. The tool is nevertheless designed to cater to flexible requests so the default parameter values that have been programed in the tool can be adjusted by the user. For example,

different time periods can be investigated, various staff regulations and shifts can be considered, and different targets can be applied.

Further details regarding the individual functions offered by the tool are provided in Section 10.2. After outlining information regarding the general implementation of the tool, the section details the full range of adjustable parameters (see Section 10.2.1), and outlines the structure of the VBA programs (see Section 10.2.2). Whilst the tool is designed to be user-friendly and simple to operate, it is recommended that the section first be read by the user to gain familiarity with the tool and appreciate the range of functions that it offers. Section 10.3 ultimately concludes the chapter with a summary of the benefits of the tool and intentions for its development to allow independent resource optimisation within WAST.

The only function programed in the spreadsheet tool which requires an additional license to run is the SSA component, which allows forecasted demand values to be generated based on historic data. Since a licence needs to be purchased to execute the Dynamic Link Library (DLL) embedded in the code, the tool assumes by default that the SSA forecast has been computed externally and accordingly opens to display the sheet where the hourly demand forecasts can be input directly by the user, prior to the execution of the optimisation programs. However if the user possesses a licence, it is possible to compute the forecast within the tool itself by selecting to 'Compute Revised Demand Forecast Data Using SSA'. Once the demand forecasts have been generated, a range of optimisation programs are available (i.e. options to compute hourly period requirements, schedule shifts optimally, roster staff efficiently, or to execute a combination of these functions). Figure 10.1 displays a screenshot of the available functions, which can be accessed by selecting to view the main 'Scheduling Options' from any sheet within the spreadsheet tool.

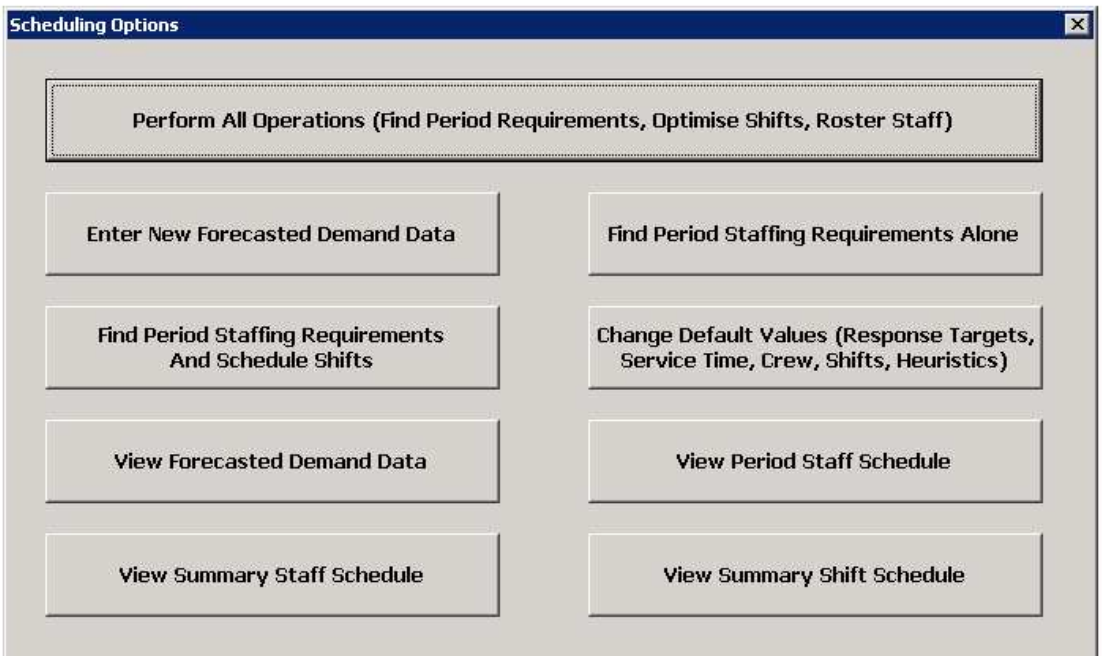

Figure 10.1: A screenshot of the main menu options

10.2 Functions offered by the planning tool

The spreadsheet tool is designed to be user-friendly and contains options to produce various staffing profiles (in terms of hourly period requirements, shift requirements or complete rosters). The options provided under the main menu form presented in Figure 10.1 are self-explanatory, and further details are given for each function below:

- **Perform all operations:** Once the period requirements have been generated, this option executes all the subsequent steps necessary to produce an efficient roster. It executes the VBA programs that convert the forecasted demand values into crew requirements using the priority queueing theory methodology developed in Chapter 7, and that produce a desirable roster using the shift scheduling and rostering heuristics developed in Chapter 9.

However, if only a subsection of the staffing functions are required, these can be selected using the options:

- **Find period staffing requirements alone:** This option executes the priority queueing theory methodology that converts the forecasted demand values into crew requirements for each shift.

- **Find period staffing requirements and schedule shifts:** This option runs the priority queueing theory methodology as above, and also produces an optimised shift schedule that details the number of crews that should be employed for particular shifts, based on the hourly requirements.

The particular version of the workforce capacity tool that has been supplied in conjunction with this thesis has been developed strictly for testing and exploratory purposes. It has been pre-populated with data relating to the forecasted demand for the first week of July 2009, and historic data from April 2005 (i.e. all known data before the forecasting period), with options to 'Restore Original Forecasted Demand Data' if the data is overwritten when different functions are viewed and explored. However, this option would not be included in a functional tool provided to an organisation: it is strictly present for test purposes. In general, the user may enter data for different forecasting periods and durations by selecting:

- **Enter New Forecasted Demand Data:** This option opens a new sheet where the user may manually enter the forecasted demand data (if this has been generated externally), or run the SSA algorithm within the tool itself to compute the forecasted values. If the option to 'Compute Revised Forecast Demand Forecasted Demand Data' is selected, the user should first select to 'Revise Historical Demand Data', secondly populate the spreadsheet with all known historic data (in the time periods required by the model) and finally select to 'Compute Demand Forecast Based on Historic Data And Export Forecast To Period Requirements Table', where further staffing operations may be performed. The SSA algorithm is set up to run with several default parameters, which are discussed in the parameter list in Section 10.2.1, and can be changed via the option to 'Change Default Parameter Values For SSA Model'. Observations for any number of days may be included in the historic demand time series or forecasted for future dates - the only restriction is that each day considered must be considered in its entirety (and not in fractions) due to the structure of some constraints in the staffing modules.

The remaining options provided in the form in Figure 10.1 allow the user to view the various outputs provided by the optimisation programs. When a program is executed, the output of the final stage is automatically be displayed, but it can be desirable to view other aspects of the staffing profile (i.e. to view how the period requirements fit in with the overall roster). Each of the viewing options displays the following output:

- **View Forecasted Demand Data:** This displays the forecasted demand data for which the staffing profiles have been/will be generated for. Within this sheet, there is also the option to revise these forecasts manually or by using SSA.
- **View Period Staff Schedule:** For each crew considered in the staffing heuristic, this table allocates a row populated with a 0/1 binary variable indicating whether the crew is off/on-duty for each period.
- **View Summary Staff Schedule:** For each crew considered in the staffing heuristic, this table allocates a column populated with a 0/1 binary variable indicating whether the crew is off/on-duty for each shift. Three summary measures are also provided for each crew: (i) the total number of hours for which they are employed, (ii) the number of night hours they are scheduled to work, and (iii) the number of overtime hours they are scheduled to work (i.e. any hours worked above standard 38 hours permitted per week).
- **View Summary Shift Schedule:** The table displayed in this sheet summarises the number of shifts of each type scheduled for each day included in the scheduling horizon.

By selecting any of the above options, the user may choose to view certain tables and consider how they relate to each other. Within each sheet, there are also options to execute further staffing functions based on the output of the preliminary programs (e.g. shift schedules can be developed around outputted period requirements; or a previously outputted shift schedule can be used to create an initial solution for a roster) so there is no need to re-start the rostering algorithm from scratch and re-calculate the period requirements based on the forecasted demand values if these have already been computed.

Since the user may select to only execute certain procedures, shift schedules and rosters will not always be developed around the period requirements. Hence in order to ensure that the outputs displayed in each sheet always coincide with each other, when the preliminary functions are performed, the outputs displayed on all other sheets (resulting from previous executions of the shift scheduling/rostering heuristics) are deleted.

The workforce capacity scheduling tool is programed in such a way that optimised staffing profiles may be generated for up to 8,000 hourly periods (around 4 months

of data, depending on the calculation interval chosen for the Euler methodology (see Chapters 5 - 7)). The only restrictions are that the working day is considered as running from 6am-6am, and due to the nature of the staffing constraints, days must be considered in their entirety (i.e. all 24 periods must be accounted for on every day entered into the model). It is however worth noting that the parameters are currently optimised for a scheduling horizon of 1 week, and that additional days entered into the model considerably increase the time required for it's execution. As an indication of the rough timings required to execute each of the staffing programs using the default parameter values programed in the model, Table 10.1 contains a summary of the approximate times required for the various functions when run on a 2002 operating system with 3.00GHz and 2.96GB of RAM:

Table 10.1: Run times required to execute programs for various forecasting horizons

Program	Forecasting horizon	
	1 week	3 months
Generate SSA demand forecast	3 mins	5 mins
Compute period requirements (Approximate)	0.3 mins	10 mins
Compute period requirements (Numerical)	10 mins	120 mins
Compute period requirements (Hybrid)	8 mins	100 mins
Produce optimised shift schedule	0.5 mins	7 mins
Produce optimised roster (2,000 iterations)	50 mins	180 mins
Produce optimised roster (200 iterations)	15 mins	120 mins

Thus as the scheduling horizon is increased the run times required to execute each of the programs considerably lengthen. Whilst the accuracy of the period requirements output for longer scheduling horizons should not be compromised, the quality of the shift schedule and roster are potentially poorer unless the parameter values are adjusted accordingly. (For example, the average cost achieved by 50 shift schedules developed heuristically using the default parameter values for a 1 week period were only 0.5% higher than the true optimal, where as the average cost achieved for schedules developed for 3 month periods were around 3% higher).

Table 10.1 also highlights that for all forecasting horizons, the programs to produce optimised rosters require considerably longer run times when 2,000 iterations are executed in place of 200. Whilst 2,000 iterations allow better quality rosters to be developed, the improvement is small in comparison to the additional time required. In

fact, in a test of 50 runs of the algorithm for a 1 week forecasting horizon, the average cost achieved for the roster was only 2% lower after 2,000 runs compared to 200. Due to the considerable time savings observed for lower number of runs, the heuristic in the test model is programed to run for a total of 200 iterations and the temperature is reset to a higher level if no improvement is made to the cost in 30 iterations (to simply illustrate the methodology followed by the approach). However, it is recommended that 2,000 runs should be used in a version of the tool provided to an organisation for practical purposes, with an increase in temperature if no improvement is made to the cost in 200 iterations.

The workforce capacity planning tool has been specifically developed to provide flexible scheduling options, and although it has been populated with several default parameter values, the majority of these can be adjusted by the user. The entire list of adjustable parameter values is replicated in Section 10.2.1 below. Once new values are input for these parameters, these hold until the default values are re-submitted. The user forms which contain the options to adjust the parameter values display the default values every time the form is viewed: thus the default values may be re-submitted very easily. There are however a few parameters which are adjusted directly in the spreadsheets rather than via user forms, and a hidden sheet included in the test version of the workforce capacity planning tool (entitled 'DefaultValues') so these default values may be conveniently reviewed and restored as desired.

10.2.1 List of adjustable parameters/variables

SSA Parameters

The SSA spreadsheet requires historical demand data to be input for three periods per day (namely Morning (6am-12pm), Afternoon (12pm-7pm) and Night (7pm-6am)) as are pre-defined by WAST and have been discussed in Chapter 2. Whilst the spreadsheet could equally be set up to consider hourly demands, the structure associated with the demands exerted in each of these three slots aids SSA to exploit the periodic patterns in the data and generate superior forecasts. The demand forecasts for each of these three slots are subsequently apportioned into hourly periods using tables populated with proportions which define the expected distribution of the demand across each of the hourly slots (based on historic demand). Different proportions are applied different days of the week as a two-way ANOVA reveals that the proportions are not all equal ($p < 0.05$). The default values used in these tables can be updated by selecting 'Change

Default Parameter Values For SSA Model' from the SSA sheet. This sheet also contains options to update several parameter values as follows:

- **Window Length:** This component represents the number of non-overlapping vertical windows the trajectory matrix X is subdivided into, defined as L in Chapter 4. For an accurate series reconstruction, this value should lie between $\frac{1}{3}$ and $\frac{1}{2}$ of the number of known historic data points and should be proportional to any known periodicity (e.g. a multiple of 7 to account for the weekly periodicity), but smaller values can be desirable for forecasting purposes.

- **Number of components used for reconstruction:** Discussed in Chapter 4, the number of principal components used to reconstruct the time series can be prudently selected by visually inspecting a plot of the logarithms of eigenvalues and noting the point at which the series plateaus. When the model is populated with 4-5 years of data, the default value entered for this parameter (20) is fairly robust for different forecasting horizons.

- **Number of days to forecast ahead:** The user may request a forecast to be generated for any forecasting horizon.

- **Uplift:** Given the responsibility of WAST to respond to potentially life-threatening emergencies, it is arguably favourable to have an oversupply rather than undersupply of resources; thus the option to 'uplift' the estimated demand forecasts is offered before the staffing profiles are generated. By default, the forecasted demand figures are inflated by 10%, but by making this parameter adjustable in the model, the ultimate decision lies with policy makers to choose how risky they wish to be.

Response Time Targets

The response time targets set by the government for WAST responses have been outlined in Chapter 1, and the average observed response and service times for EA responses within the SE Region have been given in Chapter 7. The default values that are programed in the tool can be revised by selecting 'View Scheduling Options' → 'Change Default Values' → 'Change Response Targets/Service Time', and are currently programed as follows:

Table 10.2: Default values for response targets and average service rate

Parameter	Default value
Acceptable wait for category A incidents (hours)	0.0955
Acceptable wait for category B/C incidents (hours)	0.0799
Target proportion for Category A incidents (hours)	0.95
Target proportion for Category B/C incidents (hours)	0.95
Average service rate (patients per hour)	1.0989

Parameters for Queueing Theory Model

The tool allows the period requirements to be generated using either of the approximate (Priority SIPP), numerical (Euler) or hybrid approaches as described in Chapters 6 and 7 of this thesis. Whilst the SIPP approach provides rough solutions rapidly, the numerical method is capable of producing accurate predictions at the expense of computation speed. The hybrid approach offers a method that increases the efficiency of the standard numerical Euler solver by suggesting initial staffing levels for each period. It shortens the computational time required to achieve an accurate staffing profile by around 15%, which can represent a significant quantity of time for larger problem instances and should thus always be selected in place of the standard numerical method. The standard method is however included in the test version of the tool supplied with this thesis, to complement the research discussed within it.

Instead of being presented in a user form which is accessed via the option to 'Change Default Values', the option to select an appropriate technique to generate the staffing requirements is presented to the user every time the staffing program is executed; along with the option to view the probability of an excessive wait for patients calculated at the last interval of each period with/without the dummy shift boundary transition applied. The standard probability (i.e. the measure computed prior to the application of the transition matrix) provides a more realistic view of the probability of an excessive wait for patients towards the end of each period, but there may be situations in which it is preferable to obtain the exact probability of an excessive wait at the commencement of the next period. Whilst various types of shift boundaries have been investigated in this thesis, a dummy boundary has been applied to all periods considered in the test model, since the requirements are generated for hourly periods (which may be ultimately be used to form any of the shifts presented in the potential pool).

If the numerical or hybrid approach is selected, the user may adjust the parameter values used in the methodology by selecting 'View Scheduling Options' → 'Change Default Values' → 'Change Parameters For Numerical Methodology'. The parameters, and default values, are presented in Table 10.3 below:

Table 10.3: Default values for numerical methodology

Parameter	Default value
Limit on the number of Cat As considered to arrive in Cat B/Cs waiting time	10
Limit on total number of emergencies considered in the system, G	40
Calculation interval, δc	0.04

Allowable Shifts

The default pool of potential shifts that may be scheduled have been provided in Appendix A.4 and this mirrors those currently used by WAST in the SE Region. The default pool can however be changed to include any number of shifts (lasting between 1-13 hours in duration) by selecting 'Change Default Values' → 'Change Allowable Shifts'. The start and end times may be adjusted for any of the shifts currently entered in the model, or entire shifts can be added/deleted from the potential pool directly. Only the start and end times need to be adjusted manually, since the number of shifts, shift durations, and number of 'night time' hours included in each shift is computed automatically within Excel. The only restrictions to the set of shifts allowed for selection is that they must ensure that every period of the day is covered by at least one shift, and one shift must begin at 6am to allow the demand for the first periods on the first day to be covered. The reason for this is that the model considers 'days' to run from 6am-6am; and as the scheduling period is considered independently from the preceding previous days, the tool assumes that no excess staff remain in the system from shifts starting on days prior to the scheduling period. The shifts scheduled for the last day of the model are accordingly allowed to be truncated, to ensure that no excess staffing is scheduled for following periods.

The shift scheduling heuristic additionally allows various preference weights to be assigned to shifts of different lengths that are included in the allowable pool. The weights, which may be adjusted by the user by selecting 'View Scheduling Options' →

'Change Default Values' → 'Change Allowable Shifts' → 'Adjust Cost', are recorded in Table 10.4 below:

Table 10.4: Default values for shift preference weights

Shift length	Preference weight
≤ 8 hours	1.05
9 hours	1
> 9 hours	0.95

Parameters for Shift Scheduling Heuristic

The shift scheduling heuristic is developed following the SA methodology as discussed in Chapter 9. The default values presented in Table 10.5 are those selected to produce an optimised schedule for WAST resources for the first week of July 2009 in Section 9.2.2, but they can be adjusted by the user if shift schedules are desired for different problem instances via the options to 'Change Default Values' → 'Change Parameters For Shift Schedule Heuristic'.

Table 10.5: Default parameter values for shift scheduling heuristic

Parameter	Default value
End heuristic if no improvement to cost seen in x iterations	1,500
Reduce temperature after every x iterations	250
Starting temperature	8
Cooling rate	0.86

Crew Constraints

Rostering ambulance officers is a highly constrained optimisation problem. As outlined in Chapter 8, when constructing a roster, healthcare institutions must take into account various legal, management and staff requirements. The set of constraints (and the default parameter values) presented in the table below that are considered in this spreadsheet tool are identical to those investigated to those in Chapter 9.3.2. Recall that 'night time' hours are those covering hourly periods from midnight-5am inclusive. However, the default values may be adjusted within acceptable bounds, if different regulations are to be considered via the options to 'Change Default Values'

→ 'Change Crew Constraints'.

Table 10.6: Default parameter values for crew constraints

Parameter	Default value	Allowable values
Available crew	100	0 - 300
Max work hours per week	42	0 - 60
Max night hours per week	8	0 - 40
Min rest hours between shifts	11	0 - 25
Continuous rest hours per week	35	0 - 72

Parameters for Rostering Heuristic

The rostering heuristic is developed using the SA methodology as discussed in Chapter 9. The default values presented in Table 10.7 are those selected to produce an optimised schedule for WAST resources for the first week of July 2009 in Section 9.3.2, but they can be adjusted by the user if rosters are desired for different problem instances via the options to 'Change Default Values' → 'Change Parameters For Rostering Heuristic'.

Table 10.7: Default parameter values for rostering heuristic

Parameter	Default value
End heuristic after x iterations	200
Starting temperature	20
Cooling rate	0.90
Increase temperature if no cost improvement after x iterations	30
Crew coefficient in objective function	25

10.2.2 VBA Code

This section provides a brief overview of the functions coded in Excel VBA and a description of what each of them achieves. Three modules are included in the VBA tool: namely 'Commands', 'SSABasic' and 'StaffingAllocation'. The macros compiled in each of these modules is discussed in turn below.

The 'Commands' module contains macros which enhance the user-friendliness of the tool. The macros primarily define which sheets are displayed when different functions

are selected, and outline the options to hide and display each of the menu forms. For testing purposes however, a further macro 'Testing' is also included at the end of the module that allows some of the hidden sheets within the tool to be viewed by the user.

The 'SSABasic' module contains the functions required to generate SSA forecasts of future demand levels. The historic demand is read in for three daily pre-defined slots (6am-12pm, 12pm-7pm, 7pm-6am), and the forecasts produced for each slot are subsequently apportioned into hourly demand estimates taking into account the day-of-week effects. Hence the forecasted time series is ultimately output at the hourly level to capture fine-scale dependence in addition to long-term structure.

The 'StaffingAllocation' module contains the programs which deal with staffing issues, including the programs to produce period-by-period requirements, optimise shift schedules and roster staff. The key macros are 'SIPP', 'Euler', 'OptimiseShifts', 'Swap1', 'Assignstaff', 'SwapStaff' and the remaining macros define which programs should be executed for specific options and how they link together. These macros are discussed in turn below:

- **SIPP:** This program executes the Priority SIPP methodology (outlined in Chapter 7) to provide approximate crew requirements for hourly periods, according to the specified waiting time targets. This program is also run to produce initial period requirements for the hybrid approach.

- **Euler:** This program executes the Euler methodology (outlined in Chapter 7) to provide approximate crew requirements for hourly periods, according to the specified waiting time targets. Before the period requirements are calculated, the data for the first day is initially used as a warm-up period, so the main program is operated from suitable dynamic steady state conditions. Dummy shift boundaries are applied to all periods considered in the model, as the Euler methodology is employed to produce hourly requirements that will later be used to create efficient shift schedules. The program outputs the probability of an excessive wait for each category of patient when the recommended minimum staffing profile is used, and also includes the option to view the probabilities over the shift boundary, when the transition matrix is applied to account for the effect of any departing servers at the end of each period. This program is also executed if the hybrid methodology is selected, and calculates period requirements using the recommendations obtained from the Priority SIPP methodology as initial

solutions.

- **OptimiseShifts:** This program contains the code required to produce an initial shift schedule, that is subsequently optimised using the 'Swap1' macro. It reads in all allowable shift types that may be included in the schedule, and selects particular shifts to produce a greedy feasible shift schedule.
- **Swap1:** This program executes the shift scheduling heuristic as described in Chapter 9.2.2 to produce an optimised shift schedule. It is programed to continue running until it fails to make an improvement in the cost in x iterations (1,500 by default).
- **AssignStaff:** This program assigns crews to the optimised shift schedule produced using the 'Swap1' algorithm in a greedy fashion.
- **SwapStaff:** The code contained in this module takes the initial solution provided by the 'AssignStaff' program and considers deleting/adding certain shifts to the schedule, in addition to swapping shifts between crews, using the heuristic presented in Chapter 9.3.2. A specific number of iterations can be specified by the user as suitable stopping criteria for this algorithm, and whilst the heuristic presented in Chapter 9.3.2 is run for 2,000 iterations, it is notable that a good quality solution is obtained in the first few iterations. Hence to illustrate the methodology followed by the heuristic in a quick run time, the test program has been set up to run for 200 iterations, although 2,000 are recommended for operational purposes.

Each of the macros discussed above additionally include several 'check' functions within them, to reduce the risk of attempting to produce solutions for incomplete or unsuitable scenarios input by the user. For example, the SSA macro produces an error messages if executed on a computer lacking the necessary software; the macros coded within the 'StaffingAllocation' module check that for every day included in scheduling horizon, demand data has been forecasted for all 24 periods; and if the user selects to change the shifts allowed in the potential pool for selection, the shift scheduling heuristic ensures that every period throughout the day is covered by at least one allowable shift.

10.3 Summary

This chapter has summarised the main features incorporated in the workforce capacity planning and scheduling tool that has been produced in conjunction with this thesis.

It has brought together the three main components investigated in the research and illustrated how they can be linked together in a single tool to offer efficient solutions to resource allocation problems.

While forecasts can of course be progressively updated, the inability to instantly summon crews implies that WAST has to schedule shifts and roster staff several weeks in advance; and this tool offers functions that generate potential solutions to complex problem instances. It is nevertheless important to remember that the version of the workforce planning tool supplied with this thesis has been developed strictly for illustrative purposes, and whilst it contains several features that would require fine-tuning before it was offered as an operational tool, it has been the intention of this chapter to outline its potential to produce optimised staffing profiles for a wide range of scenarios. Moreover, its user-friendly interface is intended to increase is suitability to be offered as an operational tool to organisations such as WAST, to inform staffing decisions and ultimately enable planners to optimise resources independently.

Chapter 11

Conclusions and future research

11.1 Introductory remarks

This final chapter concludes the thesis by drawing together the main results presented in the previous ten chapters, discussing the novel contributions offered and suggesting opportunities for future research. It is structured as follows: Section 11.2 summarises the key findings and conclusions in relation to the seven research questions posed in Chapter 1; Section 11.3 outlines the key contributions; and Section 11.4 discusses some of the limitations of this work, with suggestions for further research.

11.2 Key findings and conclusions

This research has demonstrated how OR techniques may be applied to promote effective and efficient management of EMS. This thesis has introduced a method for finding staffing requirements, while simultaneously selecting shifts that cover these requirements, that minimise costs and achieve pre-defined performance standards. The problem of staffing a multi-class multi-type call centre is recognised as notoriously hard even when demand rates are perfectly predictable (Gurvich et al., 2010); therefore this thesis has adopted a macro view utilising stochastic modelling techniques as and when necessary, to approach this task efficiently. The distinct techniques have ultimately been integrated into a workforce capacity planning tool, that is intended to provide organisations with the necessary tools to independently optimise their resources.

The research contained in this thesis has been motivated by problems facing WAST, who provided the real-life context for developing the operational models. WAST

receives emergency calls of varying urgency that are triaged by the call taker according to the AMPDS. Category A calls have head-of-the-line priority over Category B and C calls, i.e. if an emergency response vehicle is unavailable when an emergency is reported, then that emergency is placed in a virtual queue for assistance, irrespective of its urgency; and queued Category B and C calls are only dealt with after all queued Category A calls have been allocated a responder. Minimum acceptable response time targets to each category of call are specified by the Welsh Government, which stipulate minimum proportions of requests for assistance to be served within set time frames.

This research has addressed the staffing problem facing WAST, who are primarily interested in determining minimum coverage requirements, optimising shift schedules and constructing desirable rosters for EA and RRV crews, in order to satisfy the response time targets. The service level standards are in effect proxies for the underlying goals of saving lives and preventing suffering; but high costs are associated with the provision of staff and resources, so it is a problem of substantial economic and social interest to manage EMS systems efficiently. Due to the seriousness of the incidents WAST is required to deal with, rules specify that a paramedic serving a patient when he is scheduled to finish duty, must first complete the service currently in operation. Hence the techniques that have been developed throughout the thesis concern the exhaustive discipline (see Chapter 5). Whilst the methods have been primarily developed for WAST, they are nevertheless applicable to all time-dependent multi-class, multi-server systems operating under the head-of-the-line priority and exhaustive service disciplines (such as the police service, call centers that process two types of work and breakdown cover organisations, to mention just a few).

In developing the various models, this thesis attempted to satisfy the research objectives defined in Chapter 1 while answering seven research questions. Discussions surrounding how well the objectives have been addressed in this thesis are provided below.

The research was geared towards developing an integrated workforce capacity planning tool that:

(a) Incorporates time-series methods that adequately account for the stochastic nature of demand to produce accurate forecasts of future demand.

The tool forecasts demand for EMS assistance using SSA, since by taking

account of trend, periodic components and structureless noise, the technique is capable of simultaneously accounting for several factors believed to affect demand. The SSA technique has further been shown to offer forecasts of a superior quality to those generated by conventional time-series techniques, such as ARIMA and Holt-Winters in Chapter 4, and it is not constrained by parametric assumptions common-place in traditional forecasting methods.

(b) Provides both accurate and approximate evaluations of system performance over time.

The tool offers functions that allow approximate and accurate staffing requirements to be generated by appropriately analysing the system using Priority SIPP and Euler methodologies discussed in Chapters 5-7, which have been extended to time-dependent priority service systems in this thesis. Whilst the hybrid approach has been shown capable of generating the Euler outputs at a faster rate, both the hybrid and standard Euler approach are offered as viable approaches in the planning tool to compliment the research in Chapter 7.

(c) Permits a certain service quality to be met as inexpensively as possible by generating an efficient staffing function that accurately matches resources to fluctuating demand levels.

Akin to many service systems where the interests of the customer and server are not mutual (so it is difficult to assign a monetary value to the cost of waiting), WAST is evaluated according to the fraction of customers waiting 'unacceptable' times for service, as specified by governmental targets. Whilst the approximation method has previously been considered to set staffing levels based on such service standards in time-dependent priority systems (see Chen and Henderson (2001)), the numerical approach has not been developed for such purposes. In addition to proposing suitable formulae to calculate this metric to be embedded within numerical methodologies, with adjustments to account for the effect of departing/arriving servers over shift boundaries, this thesis has proposed an adjustment to the approximation approach which allows the performance measure to be computed for LP customers with greater accuracy. By incorporating the expressions in SIPP and Euler methodologies, the tool generates minimum staffing requirements that match fluctuating demand levels and satisfy the governmental response time targets.

(d) Generates an optimised shift schedule.

Considering the shift scheduling problem in terms of an IP model, the tool generates a schedule using a SA heuristic that has been shown capable of producing close to optimal solutions (see Chapter 9). The heuristic approach has the advantage that it may be embedded within a workforce capacity planning tool without requiring specialised software, and may be flexibly adjusted by the user to allow the consideration of different shift types.

(e) Assigns staff to shifts in an efficient manner, whilst adhering to governmental regulations and working time directives (WTD).

A heuristic is programmed within the tool that provides good quality rosters, but not necessarily optimal, in a reasonable amount of time on a personal computer. The case study included in Chapter 9 has demonstrated that the heuristic is capable of generating a feasible and effective solution that adheres to various legal and managerial requirements.

(f) Is user-friendly and practical; so it could be used to inform WAST's staffing decisions and readily adopted by planners to optimise resources independently.

The workforce capacity planning tool has been designed with a user-friendly interface, providing results based on calculations that are easily implemented in a spreadsheet setting (hidden from the user in VBA code). The tool is embedded within Excel for the reason that this computer package is widely available in many organisations and managers are generally familiar with the software. It is nevertheless important to remember that the version of the workforce planning tool supplied with this thesis has been developed primarily for illustrative purposes, and the discussions provided in Chapter 9 have outlined its potential to be fine-tuned to produce optimised staffing profiles for a wide range of scenarios. Specifically, this flexibility increases its desirability as a practical workforce planning tool.

In addition, the following research questions were posed:

(I) Is it possible to improve the accuracy of demand forecasts, by adequately accounting for seasonality in the data?

Whilst scheduling models rely heavily on accurate demand forecasts to perform well, the arrival rate is often not known with certainty, either because

it varies randomly over time or simply due to a lack of information. In either case, the uncertainty in the arrival rate has major implications for the validity of traditional performance measures and consequently on the quality of staffing decisions. Obviously inaccurate forecasts are very costly because they result in a mismatch of supply and demand (Matteson et al., 2011).

Current practice for forecasting call arrivals is often rudimentary. For instance, WAST currently estimates arrival rates for future periods based on peak call rates observed in the past; but Chen and Henderson (2001) have identified three sources of potential error in estimating the arrival rate in this fashion: (i) estimation error arising from taking the average of a finite number of random variables, (ii) failure to detect nonstationarities that could be present in the data, and (iii) the presence of a random arrival rate (which may be a function of external factors e.g. weather conditions). In particular, their research revealed that the presence of a random arrival rate can lead to overpredictions of service performance; therefore if one ignores randomness, the risk of underestimating the number of staff required to achieve a given performance level is increased.

By decomposing and reconstructing the time-series considering the trend, periodic components and structureless noise, SSA overcomes many of the above shortfalls. It captures the effect of nonstationarities, seasonalities, trends and random fluctuations in the forecasts, and allows the count-valued arrivals per hour to be directly modelled.

Motivation for the utilisation of SSA as a tool to accurately predict Welsh ambulance demand has been provided in Chapter 4, with empirical results demonstrating that it produces superior longer-term forecasts (especially helpful for EMS planning), and at least comparable short-term forecasts to well-established methods. The benefit of the SSA technique is however not only in its ability to forecast; but in its capability to recognise periodicities in the data and be flexible in approach. The incorporation of the technique into the workforce capacity planning tool, discussed in Chapter 9, illustrates how the SSA forecasts could ultimately be used to inform scheduling models and allow realistic simulation of the system.

(II) Can time-dependent queueing theory be extended to appropriately model system behaviour as servers enter and leave the system in differing fashions across shift

boundaries?

This thesis has demonstrated that the Euler methodology may be duly extended to track system behaviour in time-dependent priority systems, using dual and triple state vectors to represent the composition of customers within the system. The research contained in Chapters 6 and 7 has furthered the work of Ingolfsson (2002) (who defined the instantaneous transitions necessary to apply to track the movement of customers across shift boundaries in $M(t)/M/s(t)/FIFO$ systems), to track the composition of customers present in priority service systems over shift boundaries where servers operate under the exhaustive discipline. Not only has the approach been extended to define the transitions for a head-of-the-line priority system, but transitions have also been defined for various types of shift boundaries.

Ingolfsson (2002) showed that the incorporation of an exhaustive discipline has a considerable impact on performance levels; but the research contained in Chapter 6 has demonstrated that this boundary discipline is not adequate for the purpose of generating minimum period requirements that ultimately lead to the development of an optimised shift schedule. A case study provided justification for the creation of an alternative boundary (proposed as a 'dummy' shift boundary) that is able to take account of additional staff who join the workforce at the boundary, excess staff who leave, and staff who continue to work as normal both sides of the boundary; by demonstrating the widely differing performance levels arising in periods following the application of the two boundaries. The instantaneous transitions necessary to track the composition of customers in the system over the dummy and true period boundaries where staff operate under the exhaustive discipline have further been defined within the thesis.

(III) To what degree do staffing levels in one period affect another? Can guidelines be provided regarding situations under which it is appropriate to approximate time-dependent behaviour, how accurate the approximations are, and if steps can be taken to increase their accuracy?

In addition to the insights listed in (II) regarding the impact of various shift boundaries on service levels in subsequent periods, this thesis has evaluated the potential of various approximation methods used in the literature in de-

termining staffing levels to match the workload in service systems subject to time-varying demand. The main simplifying assumption in the approximate approach is that the staffing requirements for a particular period can be determined independently of the staffing in previous periods. The extent to which this assumption is valid determines whether it is reasonable to use it, and formal guidance accompanying case studies have been included in Chapters 6 and 7, illustrating that whilst the approach will always be subject to a certain degree of error, it can provide reasonable approximations at speed. However, in situations where the SIPP approach is justified, then it should be used because it is has a much lower computational cost than the numerical approach. In general, SIPP and Priority SIPP should provide reliable results in systems with short planning periods, high service rates, low presented loads and relatively low RA. In cases where the standard approach is unreliable, revisions have been suggested to improve the accuracy of its predictions, including Modified Lag Avg SIPP which has been proposed for the first time in this thesis, and is expected to be more robust in systems with higher loads and higher RA.

Greater insights regarding the impact of staffing levels in one period upon another have been provided in Chapters 6 and 7. In particular, Figure 6.5 has demonstrated that in periods where the arrival rate changes drastically, or where it continually increases/decreases over several consecutive periods, the failure to account for staffing levels in previous periods can considerably affect the staffing profile generated. Additionally, Figure 6.3 has shown the that the incorporation of the truly exhaustive shift boundary can greatly impact on customers expected waiting time in the system for a short duration after the boundary, since it suggests that greater numbers of customers are served by paramedics working past their scheduled end time, meaning lower staffing quantities are needed. However, the degree of these effects could possibly be considered as second order for WAST, since for the case study included in Chapter 6 around 7-10 crews were found sufficient for each hourly period, but the approximate and numerical requirements rarely differed by more than 1 crew. The decision over which performance evaluation method to use is therefore ultimately left to the client. To guide their decision, this thesis has attempted to outline the suitability of each approach with information detailing the degree of accuracy offered, when the assumptions of the method appropriate for the system and the compu-

tational time required to produce the staffing requirements.

(IV) Can time-dependent and approximate queueing theory techniques be extended to compute waiting-time probabilities in time-dependent multi-class, multi-server systems?

Building on non-stationary networks of finite server queues, this thesis has developed waiting time formulae to calculate the probability of excessive waits for both HP and LP customers over time. Whilst suitable formulae have previously been developed for the Priority SIPP approach (see Chen and Henderson (2001)), this research has proposed an alternative expression which allows the probability of an excessive wait to be calculated for an LP customer with a greater degree of accuracy.

The real contribution of this thesis however lies in the extensions of the waiting time formulae presented for the numerical method, since whilst the methodology required to limit the quantity of unacceptable waits in time-dependent queueing systems has been well studied in literature (Ingolfsson, 2002; Ingolfsson et al., 2007; Green and Soares, 2007); transient analysis of the probability of an excessive wait has not been investigated for priority systems, despite their high prevalence in industry. Not least has this thesis extended the waiting time formulas to enable their application within such systems, but it has further devoted particular attention to extend approaches followed by Green et al. (2007) and Ingolfsson (2002) to evaluate the probability of excessive waits for HP and LP customers over both dummy and true shift boundaries.

(V) Is it possible to increase the efficiency of numerical methods to accurately evaluate system performance?

For situations when the client requires particularly accurate analysis of system performance, this thesis has proposed a hybrid method that enables the accurate requirements generated by the numerical method to be produced at a quicker rate.

The hybrid approach produces staffing requirements by considering a function of the staffing levels output from approximate methodologies as initial staffing levels for each period. Under the assumption that the approximate levels are close to the numerical requirements; by considering staffing levels just below those initially suggested (if the approximate predictions are

found to be sufficient with numerical analysis) or just above those suggested (if they are insufficient), the case study included in Chapter 7 has shown that considerable time savings can be achieved.

(VI) Is it possible to develop suitable heuristics to optimise shift schedules and rosters that minimise cost and ensure satisfactory customer service?

Whilst this thesis has awarded less focus to the construction of optimised shifts and rosters (which itself could potentially be a whole research thesis in its own right), it has presented a practical approach to the problem through developing heuristics that produce good quality solutions, though not necessarily optimal, in reasonable computation time. Both exact and heuristic approaches have been considered; and whilst IP solvers have been demonstrated to offer optimal solution to the shift scheduling problem, common IP solver packages such as XPress-MP were found to run out of memory on hard rostering instances, and complex constraints were sometimes intractable. In light of these shortfalls, this thesis has provided justification for the use of SA heuristics to produce close to optimal solutions that can be embedded within the workforce capacity planning tool, without requiring specialised software, and flexibly adjusted by the user to consider different shifts.

The main challenge that has been highlighted in this research is the challenge to integrate the separate steps of the rostering process into a single problem. The investigations performed within Chapter 9 have shown that it is advantageous to consider solving the shift scheduling and rostering problem simultaneously, rather than as two separate problems which can result in sub-optimal rosters.

(VII) Can the individual forecasting, modelling and optimisation techniques be combined into a generic integrated workforce planning tool to optimise staffing schedules in stochastic environments that must consistently deliver a certain service quality?

The workforce capacity planning and scheduling tool developed in conjunction with this thesis has ultimately combined the individual steps required to produce an optimised shift schedule into a single integrated user-friendly workforce planning tool; allowing automation of the process to optimise

resources in time-dependent multi-class, multi-server service systems operating under the head-of-the-line priority and exhaustive service disciplines.

The tool contains numerous options which may be flexibly adjusted to model various scenarios, and is therefore applicable to a wide range of organisations that are interested in determining minimum staffing requirements, to ensure that a given fraction of customers are seen within pre-specified 'acceptable' time frames as inexpensively as possible. The models have been tested, and shown to perform well in case studies related to specific problems facing WAST, and are accordingly expected to promote efficient resource allocation within other $M(t)/M/M(t)/NPRP/\infty/\infty$ systems.

11.3 Novel contributions

In light of the research questions addressed above, the main contributions of this research may be summarised as follows.

i. The research has incorporated the forecasts generated by SSA (which is a powerful nonparametric technique, that appropriately deals with the stochastic nature of demand), as input to staffing models. Whilst many successful applications have been made of SSA, this study seemingly represents the first application of SSA to EMS data. Numerous studies have recently highlighted the need for more accurate forecasts (see Chen and Henderson (2001), Gurvich et al. (2010), Matteson et al. (2011)) due to their important role in operations, serving as a critical input for both resource acquisition and resource deployment decisions. This research has illustrated that SSA is capable of achieving this goal through accounting for seasonalities in the data, directly modelling the count-valued arrivals per hour and producing superior forecasts to well-established conventional methods.

   As demand for EMS assistance is rising in Wales, it is becoming ever more critical to ensure accurate demand forecasts are input to WAST scheduling models, as use of inaccurate parameter estimates in these models can result in poor resource allocation. Underpredictions can lead to understaffing and low performance, whilst overstaffing can involve unnecessary personnel costs. The workforce capacity planning and scheduling tool discussed in Chapter 10 has illustrated how the SSA technique may be ultimately embedded into a spreadsheet model, and directly used to inform the scheduling functions integrated within it.

ii. The research has proposed approximate methodologies for converting demand profiles to minimum period requirements in time-dependent priority systems, so that the fraction of customers waiting greater than acceptable times for service is limited to a threshold level. Whilst the approximation methodology has been considered for such purposes in Chen and Henderson (2001), this research has further developed the methodology, permitting the probability that a LP customer experiences an excessive wait to be evaluated to a greater degree of accuracy, resulting in the risk of overstaffing being minimised.

    Guidelines have additionally been provided which outline the characteristics required for the approximation methods to perform well, suggest adjustments that may be applied to improve the accuracy of the approximate approaches if some of the assumptions are broken (including a proposition of a new Modified Priority Lag Avg approach), provide insights regarding the computation time required, and also give an indication of the accuracy of the results generated.

iii. With regards to the numerical approach, this thesis has demonstrated that the incorporation of a truly exhaustive boundary at hourly intervals across the scheduling horizon is insufficient for the purpose of developing minimum staffing requirements that lead to the development of a shift schedule. This is primarily because only excess staff leave the system at such intervals in reality, in place of all staff on duty, as suggested at a truly exhaustive boundary. The research has accordingly defined a new type of shift boundary (a 'dummy' boundary), along with relevant mappings, to appropriately account for the behaviour of staff at the end of hourly periods (where some staff may leave the system, others may join the workforce and a base set continue to work as normal, to match the demand level changes as closely as possible). The transitions necessary to account for the effect of departing/additional staff at such intervals have been defined for both standard time-dependent and priority time-dependent systems, allowing for accurate numerical analysis of system behaviour at all times, including across shift boundaries.

iv. Whilst the methodology required to limit the quantity of unacceptable waits in time-dependent queueing systems has been well studied in literature (Ingolfsson, 2002; Ingolfsson et al., 2007; Green and Soares, 2007), transient analysis of the probability of an excessive wait has not been investigated for priority systems; despite their high prevalence in industry. Not only has this thesis extended the waiting time formulas to enable their application within such systems, but it has

further devoted particular attention to extend approaches taken by Green et al. (2007) and Ingolfsson (2002), to evaluate the probability of an excessive waits for HP and LP customers over both dummy and true shift boundaries.

v. In situations where the client requires accurate analysis of system performance, this thesis has proposed a hybrid method that enables the numerical requirements to be produced at a quicker rate.

vi. The research has developed practical heuristic algorithms that can be embedded in a spreadsheet tool, to produce feasible and desirable schedules and rosters.

vii. In the consideration of all the above functions this research has devoted particular attention to the development, solution and validation of sufficiently detailed stochastic models for time-dependent multi-server systems with varying service types, which can be ultimately employed to optimise resource allocation. Through integrating the steps involved in the rostering process into a single problem, the workforce capacity planning tool developed in conjunction with this thesis has essentially provided a macro view of multiple techniques required to optimise staffing profiles in complex systems.

11.4 Research limitations and directions for future research

Although this research has satisfied the objectives discussed in Section 11.2 and offers numerous benefits from previous studies in the field, the research has some limitations. This section demonstrates how some of these limitations may potentially be seen as areas for future research.

11.4.1 Errors associated with demand modelling techniques

Whilst the SSA forecasts allow several sources of errors arising from demand based on average demand rates experienced during specific periods in the past to be overcome, some sources of error still remain, since the arrival rates are random and not perfectly predictable. For instance, the staffing models assume that the HP and LP arrival rates follow inhomogeneous Poisson processes with mean arrival rates taken from SSA predictions for each hour, but demand within each hour is realistically far from stationary. Since the arrival rates are random, the analysis presented in this thesis essentially

produces requirements that ensure that the *expected* fraction of customers waiting longer than the targeted times are less than the threshold levels, where the expectation is taken with respect to the distribution of the arrival rates. Another possibility could be to consider that the constraint is met on some pre-specified fraction of the arrival rate values, to allow the standards to be violated on a small quantity of realistions.

It would also be interesting to investigate the quality of the forecasts for certain sub-categories of demand (such as falls, breathing problems or traffic accidents); and within different areas across Wales. The demand for assistance for certain sub-categories may possibly be linked to specific external factors, such as weather conditions, school holidays or the time of day; and to this end, it could be useful to consider the benefit of Multivariate SSA (MSSA). This considers the benefit of using causal time series, such as weather and climatic data, to improve forecasted ambulance demand.

Since the forecasts are ultimately used to inform staffing models, it would also be beneficial to provide some sort of estimate of the forecasting error, to be considered when making staffing decisions. For instance, whilst the forecasts are currently uplifted by 10% before being input to the scheduling models, in order to avoid the likelihood of understaffing, it could be more useful to allow this uplift to be some function of the confidence interval bounds.

11.4.2 Time-dependent and priority queueing theory

This research has illustrated the potential of numerical techniques to analyse time-dependent priority systems by considering the methodology followed by the Euler technique, and used variants of SIPP to illustrate how approximate techniques may be used to achieve the same goal. Whilst these have been sufficient to provide insights into the impact of staffing levels in one period upon service levels in following periods, several other methods have been developed for similar purposes in the literature. It would be useful to consider the potential of some of these other approximate (e.g. PSA, MOL) and numerical (e.g. randomisation, DTM) approaches to evaluate service quality in time-dependent priority systems. Furthermore, it would be beneficial to test the proposed Modified Priority Lag Avg approach on a number of test cases, with varying characteristics to confirm the conditions hypothesised in this thesis, under which it is expected to perform well.

When rostering to flexible demand, second order effects may also arise from the choice of shift schedules. For example, it is usually not possible to exactly match the staff on duty to a demand that varies on an hourly basis when using shifts that span several hours. As a result, there may be times when higher numbers of staff are on duty than the minimum required. This may lead to considerably lower queues in certain periods, which could create artificially lower demand in later time periods as it takes time for the queue to build back up to reach the steady-state level, and further cause the number of customers considered to have been ejected from the system at each hourly period over a dummy shift boundary to be somewhat artificial. In cases where the knock-on effect is considerable, its impact may be suitably reduced by applying a small number of iterations between the rostering modules and the (consequent) demand distribution (see Ingolfsson et al. (2010)). Since the iterative approach is not incorporated in the workforce capacity planning tool, it could suggest higher staffing quantities than necessary for some shifts. An additional simplifying assumption required for the system to operate along the principles of the exhaustive shift boundary is that there aren't any resource constraints over period boundaries (as discussed in Chapter 6.2.2). If this assumption is unrealistic, then an alternative queue discipline should be applied.

The approximate and numerical methodologies could potentially be extended to appropriately model scenarios where HP and LP customers require different service times, or where more than two categories of customers may be present. Moreover, the workforce capacity and scheduling tool could be improved to offer variants of the SIPP methodology to construct minimum period requirements. However this would also require additional guidance to be provided in conjunction with the tool, outlining the circumstances under which each approach is expected to provide the best results.

11.4.3 Improving the quality of the roster

The shift scheduling and rostering problems have been given less focus in the thesis, and have been approached using heuristics. These heuristics have been shown capable of generating good quality, feasible solutions; but these solutions are not necessarily optimal. The heuristics are simple and practical to implement within a spreadsheet model; however more appealing rosters could potentially be achieved if additional hard and soft constraints were added to the models (e.g. to take account of crew preferences), or if rosters were constructed for individual members with varying competency levels. The quality of the ultimate rosters could also be improved if more appropriate

methods, such as column generation (see Lavoie et al. (1988); Ernst et al. (2004)), were considered to solve the problem. The range of potential methods that may be investigated to roster staff efficiently is however so large that the study could be performed as a separate freestanding investigation.

11.4.4 WAST specifics

The workforce capacity scheduling tool that is provided in conjunction in this thesis has been primarily designed to illustrate the potential of the tool to optimise resources for various scenarios, but the representation of WAST provided by the model is a simplified version.

To simplify the analysis, whilst developing approximation and numerical techniques to deal with priority demand, the research has assumed that a singular average measure of service time μ may be applied to both customer classes. The preliminary analysis contained in Chapter 2 has however shown that the service time does differ between the two categories of patients at WAST. Following the argument presented in Ingolfsson et al. (2007) that the service rate typically changes more slowly than the arrival rate, this analysis has additionally assumed that the service rate is not dependent upon time, which may be unrealistic; as is the failure of the model to incorporate patient abandonments.

Another factor that could be taken into account to provide a more realistic representation of the system is the distribution of vehicles needed to attend each emergency. The case studies performed within this thesis have assumed that exactly one EA is required to attend all emergency calls, and a single RRV is additionally required to attend each Category A call; but the full model is more complex in reality. It would be interesting to collaborate further with WAST to explore more scenarios and refine the models.

The response time targets applied to WAST are regularly updated, and the latest set of National Ambulance Performance Standards are more focused on improved clinical outcomes for patients. Thus the research contained in this thesis is already slightly outdated. It has however been necessary to impose consistent targets throughout in order not to constantly recalculate scenarios. At the commencement of this research, hard targets were distinctly specified for three three main patient groups requesting

WAST assistance: Category A; Category B/C; and urgent requests. Since December 2011, Category B has however been removed and emergencies are now classified into two main types (for further details, see Appendix A.2): (i) Category A patients (including the most serious Category B calls and urgent requests from health care professionals) and (ii) Category C patients (including the less serious Category B calls, urgent and planned demand). Since the workforce capacity planning tool is already set up to deal with two categories of demand, if new data were to be provided, it could be instantly updated to produce revised staffing requirements in accordance with the current standards. If clinical outcomes were also recorded, it would be beneficial to perform greater exploratory analysis to obtain insights regarding the degree of association between efficient response times and positive patient outcomes, to compliment current work on survival functions for patients requesting emergency transportation (see Knight et al. (2012)).

In summary, the problems that have been solved within this thesis are specific instances of WAST's staffing problem, but because no assumptions have been made on the arrival rate prior to functions programmed within the workforce capacity planning tool, and the other parameters are all adjustable within the tool, it seems that the methods could be employed for a wide range of scenarios. Given that most countries adopt a similar system of ambulance deployment, the tool could in fact be picked up and populated with local data in principle by any EMS. Patient abandonments, new targets, and the consideration of distinct service rates for different vehicles/call categories are all relevant for EMS systems, and it would be interesting to investigate these issues in further work. Of course, there is always a balance to be struck on what occurs in reality and what can be modelled mathematically, but by accounting for both random and predictable sources of demand in the models, and developing modelling techniques to simultaneously deal with time-dependent and priority demand, the ultimate workforce planning tool produced as a result of this research is believed to outperform many of the existing models in the literature.

Facing ever increasing pressures to provide rapid responses, WAST is keen to develop new initiatives to overcome the wide range of challenges that are impeding their ability to meet the response time targets, and have been auspiciously keen to learn of the research findings resulting from the investigations within this thesis. The Trust have been enthusiastic to discuss issues for investigation as and when they have arisen through the investigation phase, and to recommend directions for research (as detailed

in Chapter 1.4). As a result of this consolidated relationship, it is anticipated that the models proposed in this thesis will soon begin to be incorporated alongside current practice as a pilot study. The Welsh Government have further expressed interest in this work and in particular the Head of Unscheduled Care (Roger Perks) and the Senior Emergency Care Policy and Performance Manager (Aled Brown) have stated their wish to oversee the implementation of the developed tools to support WAST going forward.

References

Abate, J. and Whitt, W. (1995). Numerical inversion of laplace transforms of probability distributions, *ORSA Journal on Computing* **7**: 36–43.

Abramson, D. (1991). Constructing school timetables using simulated annealing: sequential and parallel algorithms, *Management Science* **37**(1): 98–113.

Aldrich, C., Hisserich, J. and Lave, L. (1971). An analysis of the demand for emergency ambulance service in an urban area, *American Journal of Public Health* **61**: 1156–1169.

Alfares, H. (1997). An efficient two-stage algorithm for cyclic days-off scheduling, *Computers and Operations Research* **25**: 913–923.

Allen, M. and Smith, L. (1996). Monte Carlo SSA: Detecting irregular oscillations in the presence of coloured noise, *Journal of Climate* **9**: 3373–3404.

Alpert, J., Kosa, J., Haggerty, R., Robertson, L. and Heagarty, M. (1969). The types of families that use an emergency clinic, *Medical Care* **7**(2): 55–61.

Andrews, B. and Cunningham, S. (1995). L. L. Bean improves call-center forecasting, *Interfaces* **25**(6): 1–13.

Artalejo, J. and Lopez-Herrero, M. (2001). Analysis of the busy period for the $M/M/c$ queue: An algorithmic approach, *Journal of Applied Probability* **38**(1): 209–222.

Askin, Z., Armony, M. and Mehrota, V. (2007). The modern call center: A multi-disciplinary perspective on operations management research, *Production and Operations Management* **16**(6): 665–688.

Atlason, J., Epelman, M. and Henderson, S. (2008). Optimizing call center staffing using simulation and analytic center cutting plane methods, *Management Science* **54**(2): 295–309.

Aykin, T. (1996). Optimal shift scheduling with multiple break windows, *Management Science* **42**(4): 591–602.

Bailey, J. (1985). Integrated days off and shift personnel scheduling, *Computers and Industrial Engineering* **9**(4): 395–404.

Baker, J. and Fitzpatrick, K. (1986). Determination of an optimal forecast model for ambulance demand using goal programming, *Journal of Operational Research Society* **37**(11): 1047–1059.

Bard, J., Binici, C. and deSilva, A. (2003). Staff scheduling at the United States Postal Service, *Computers & Operations Research* **30**(30): 745–771.

Beaumont, N. (1997a). Scheduling staff using mixed integer programming, *European Journal of Operational Research* **98**(3): 473–484.

Beaumont, N. (1997b). Using mixed integer programming to design employee rosters, *Journal of the Operational Research Society* **48**: 585–590.

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

Bhat, U. (2008). *An Introduction to Queueing Theory: Modeling and Analysis in Applications*, Springer.

Bianci, L., Jarrett, J. and Hanumara, C. (1993). Forecasting incoming calls to telemarketing centers, *Journal of Business Forecasting* **12**(2): 3–11.

Bilgin, B., De Causmaecker, P., Rossie, B. and VandenBerghe, G. (2012). Local search neighbourhoods for dealing with a novel nurse rostering model, *Annals of Operations Research* **194**: 33–57.

Bondi, A. abd Buzen, J. (1984). The response times of priority classes under preemptive resume in $M/G/m$ queues, *Sigmetrics* (August): 195–1201.

Box, G. and Jenkins, G. (1970). *Time Series Analysis, Forecasting and Control*, Holden-Day, Incorporated, San Francisco.

Brewton, J. (1989). Teller staffing models, *Financial Managers' Statement* **11**(4): 22–24.

Brigandi, A., Dargon.D., Sheehan, M. and Spencer, T. (1994). AT & Ts call processing simulator (CAPS): Operational design for inbound call centers, *Interfaces* **24**(1): 6–28.

Brockwell, P. and Davis, R. (2002). *Introduction to Time Series and Forecasting, Second Edition*, Springer-Verlag, New York.

Broomhead, D., Jones, R., King, G. and Pike, E. (1987). *Singular system analysis with application to dynamical systems*, CRC Press, Bristol.

Broomhead, D. and King, G. (1986). Extracting qualitative dynamics from experimental data, *Physica D* **20**(2-3): 217–236.

Brotcorne, L., Laporte, G. and Semet, F. (2003). Ambulance location and relocation models, *European Journal of Operational Research* **147**: 451–463.

Brown, L., Lerner, E., Larmon, B., LeGassick, T. and Taigman, M. (2007). Are EMS call volume predictions based on demand pattern analysis accurate?, *Prehospital Emergency Care* **11**(2).

Brusco, M. (1998). Solving personnel tour scheduling problems using the dual all-integer cutting plane, *IIE Transactions on Operations Engineering* **30**(9).

Buffa, E., Cosgrove, M. and Luce, B. (1976). An integrated work shift scheduling system, *Decision Sciences* **7**: 620–630.

Burke, E., De Causmaecker, P., Berghe, G. and Van Landeghem, A. (2004). The state of the art of nurse rostering, *Journal of Scheduling* **7**: 441–499.

Burke, E. K., McCollum, B., Meisels, A., Petrovic, S. and Qu, R. (2007). A graph-based hyper heuristic for timetabling problems, *European Journal of Operational Research* **176**: 177–192.

Burke, E. and Newall, J. (2002). Enhancing timetable solutions with local search methods. Practice and theory of automated timetabling: selected papers from the 4th International Conference, *Springer Lecture Notes in Computer Science* **2740**: 195–206.

Burns, R. and Carter, M. (1985). Work force size and single shift schedules with variable demands, *Management Science* **31**(5).

Caiado, J. (2010). Performance of combined double seasonal univariate time series models for forecasting water demand, *Journal of Hydrologic Engineering* **15**(3).

Cambazard, H., Hebrard, E., O'Sullivan, B. and Papadopoulos, A. (2010). Local search and constraint programming for the post-enrolment-based course timetabling problem, *Annals of Operations Research* **177**: 1–25.

Channouf, N., L'Ecuyer, P., Ingolfsson, A. and Avramidis, A. (2007). The application of forecasting techniques to modelling Emergency Medical System calls in Calgary, Alberta, *Health Care Manage Science* **10**(1).

Chatfield, C. (2001). *Time Series Forecasting*, CRC Press, Florida.

Chen, B. and Henderson, S. (2001). Two issues in setting call centre staffing levels, *Annals of Operations Research* **108**: 175–192.

Cobham, A. (1954). Priority assignment in waiting line problems, *Operations Research* **2**: 70–76.

Cohen, J. (1956). Certain delay problems for a full available trunk group loaded by two traffic sources, *Communication News* **16**(3): 105–113.

Colebrook, J. (1978). Continuous plankton records - zooplankton and environment, northeast Atlantic and North Sea, 1948-1975, *Oceanologica Acta* **1**: 9–23.

Dantzig, G. (1954). A comment on Eddie's 'Traffic delay at toll booths', *Operations Research* **2**(3): 339–341.

Dash Optimization Inc (2004). *Xpress-Mosel User Guide, Release 1.4*.

Davis, L., Massey, W. and Whitt, W. (1995). Sensitivity to the service-time distribution in the nonstationary Erlang loss model, *Management Science* **41**(6): 1107–1116.

Davis, R. (1966). Waiting time distribution of a multi-server priority queueing system, *Operations Research* **14**: 133–136.

Defraeye, M. and Van Nieuwenhuyse, I. (2012). Controlling excessive waiting times in small service systems with time-varying demand: An extension of the ISA algorithm, *Decision Support Systems* **Forthcoming**.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman & Hall/CRC, New York - London.

Erdogan, G., Erkut, E., Ingolfsson, A. and Laporte, G. (2010). Scheduling ambulance crews for maximum coverage, *Journal of the Operational Research Society* **61**(4): 543–550.

Erlang, A. (1918). Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges, *Post Office Electrical Engineers' Journal* **10**: 189–197.

Ernst, A., Hourigan, P. abd Krishnamoorithy, M., Mills, G., Nott, H. and Sier, D. (1999). Rostering ambulance officers, *Proceeding of the 15th National Conference of the Australian Society for Operations Research* pp. 470–481.

Ernst, A., Jiang, H., Krishnamoorthy, M., Owens, B. and Sier, D. (2004). An annotated bibliography of personnel scheduling and rostering, *Annals of Operations Research* **127**: 21–144.

Feldman, Z., Mandelbaum, A., Massey, W. and Whitt, W. (2008). Staffing of time-varying queues to achieve time-stable performance, *Management Science* **54**(2): 324–338.

Fildes, R., Nikolopoulos, K., Crone, S. and Syntetos, A. (2008). Forecasting and operational research: A review, *Journal of the Operational Research Society* **59**(9): 1150–1172.

Fogiel, M. (1983). *The Numerical Analysis Problem Solver*, Research & Education Association.

Gail, H., Hantler, S. and Taylor, B. (1988). Analysis of a non-preemptive priority multiserver queue, *Advances in Applied Probability* **20**: 852–879.

Gail, H., Hantler, S. and Taylor, B. (1992). On a preemptive markovian queue with multiple servers and two priority classes, *Mahematics of Operations Research* **17**(2): 365–391.

Gamache, M., Soumis, F., Villeneuve, D., Desrosiers, J. and Gélinas, E. (1998). The preferencial bidding system at Air Canada, *Transportation Science* **32**: 246–255.

Gans, N., Koole, G. and Mandelbaum, A. (2003). Telephone call centers: Tutorial, review and research prospects, *Manufacturing and Service Operations Management* **5**: 79–141.

Gardner, E. (1985). Exponential smoothing: the state of the art, *Journal of Forecasting* **4**(1): 1–28.

Gelenbe, E. and Pujolle, G. (1998). *Introduction to Queueing Networks*, Wiley.

Ghil, M., Allen, M., Dettinger, M., Ide, K., Kondrashov, D., Mann, M., Robertson, A., Saunders, A., Tian, Y., Varadi, F. and Yiou, P. (2001). Advanced spectral methods for climatic time series, *Reviews of Geophysics* **40**(1): 3.1–3.41.

Gillard, J. (2010). Cadzow's basic algorith, alternating projections and singular spectrum analysis, *Statistics and Its Interface* **3**: 335–343.

Gillard, J. and Knight, V. (2012). Using singular spectrum analysis to obtain staffing level requirements in emergency units, *JORS* . Submitted.

Gillard, J., Knight, V., Vile, J. and Wilson, R. (2012). Staffing a mathematics support service. Submitted.

Glover, F. (1990). Tabu search: A tutorial, *Interfaces* **20**(4): 74–94.

Goldberg, J. (2004). Operations research models for the deployment of emergency service vehicles, *EMS Management Journal* **1**(1): 20–39.

Golyandina, N. (2010). On the choice of parameters in singular spectrum analysis and related subspace-based methods, *Statistics and Its Interface* **3**: 259–279.

Golyandina, N., Nekrutkin, V. and Zhigljavsky, A. (2001). *Analysis of Time Series Structure: SSA and related techniques*, Chapman & Hall/CRC, New York - London.

Grassmann, W. (1977). Transient solutions in Markovian queueing systems, *Computers & Operations Research* **4**: 47–53.

Graves, G., McBride, R., Gershkoff, I., Anderson, D. and Mahidhara, D. (1993). Flight crew scheduling, *Management Science* **39**(6): 736–745.

Green, L. and Kolesar, P. (1991). The pointwise stationary approximation for queues with nonstationary arrivals, *Management Science* **37**: 84–97.

Green, L. and Kolesar, P. (1997). The lagged PSA for estimating peak congestion in multiserver markovian queues with periodic arrival rates, *Management Science* **43**: 80–87.

Green, L. and Kolesar, P. (2004). Improving emergency responsiveness within management science, *Management Science* **50**(8): 1001–1014.

Green, L., Kolesar, P. and Soares, J. (2001). Improving the SIPP approach for staffing service systems that have cyclic demands, *Operations Research* **49**: 549–564.

Green, L., Kolesar, P. and Soares, J. (2003). An improved heuristic for staffing telephone call centers with limited operating hours, *Production and Operations Management* **12**(1): 46–61.

Green, L., Kolesar, P. and Svoronos, A. (1991). Some effects of nonstationary on multiserver markovian queueing systems, *Operations Research* **39**: 502–511.

Green, L., Kolesar, P. and Whitt, W. (2007). Coping with time-varying demand when setting staffing requirements for a service system, *Production and Operations Management* **16**: 13–39.

Green, L. and Soares, J. (2007). Computing time-dependent probabilities in $M(t)/M/s(t)$ queuing systems, *Manufacturing & Service Operations Management* **9**: 54–61.

Green, L., Soares, J., Giglio, J. and Green, R. (2006). Using queuing theory to increase effectiveness of emergency department provider staffing, *Academic Emergency Medicine* **13**: 61–69.

Gross, D. and Harris, C. (1998). *Fundamentals of Queueing Theory; 3rd edition*, Wiley.

Gurvich, I., Luedtke, J. and Tezcan, T. (2010). Staffing call-centers with uncertain demand forecasts: a chance-constrained optimisation approach, *Management Sciences* **56**(7): 1093–1115.

Haase, K. (1999). *Advanced Column Generation Techniques with Applications to Marketing, Retail and Logistics Management*, Ph.d. thesis, University of Kiel.

Hall, W. (1971). Management science approaches to the determination of urban ambulance requirements, *Socio-economic Planning Sciences* **5**(5): 491–499.

Haque, L. and Armstrong, M. (2007). A survey of the machine interference problem, *European Journal of Operational Research* **179**(2): 469–482.

Harchol-Balter, M., Osogami, T., Scheller-Wolf, A. and Wierman, A. (2005). Multi-server queueing systems with multiple priority classes, *Queueing Systems* **51**: 331–360.

Harding, C. (2011). Measuring the effectiveness of a maths support service, *Cardiff University Undergraduate Project* .

Hari, R., Saydam, C. and Sharer, E. abd Setzler, H. (2011). Ambulance deployment and shift scheduling: An integrated approach, *Journal of Service Science and Management* **4**: 66–78.

Hassani, H. (2007). Singular spectrum analysis: Methodology and comparison, *Journal of Data Science* **5**: 239–257.

Hassani, H., Heravi, S. and Zhigljavsky, A. (2009). Forecasting European industrial production with singular spectrum analysis, *International Journal of Forecasting* **25**(1): 103–118.

Hassani, H., Soofi, A. and Zhigljavsky, A. (2010). Predicting daily exchange rate with singular spectrum analysis, *Nonlinear Analysis: Real World Applications* **11**: 2023–2034.

Hawkes, T. and Savage, M. (2000). Measuring the mathematics problem, *London: Engineering Council* .

Hershey, J., Weiss, E. and Cohen, M. (1981). A stochastic service network model with application to hospital facilities, *Operations Research* **29**(1): 1–22.

Heyman, D. and Whitt, W. (1984). The asymptotic behaviour of queues with time-varying arrival rates, *Journal of Applied Probability* **21**: 143–156.

Holcomb, J. and Sharpe, N. (2007). Forecasting police calls during peak times for the city of Cleveland, *CS-BIGS* **1**(1): 47–53.

Holloran, T. and Byrne, J. (1986). United Airlines station manpower planning system, *Interfaces* **16**: 39–50.

Hyndman, R. and Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R, *Journal of Statistical Software* **27**(3).

Ingolfsson, A. (2002). Modelling the $M(t)/M/s(t)$ queue with an exhaustive discipline, *Thinking Beyond the Old 80/20 Rule. Call Center Magazine 15*, pp. 54–56.

Ingolfsson, A., Akhmetshina, E., Budge, S., Li, Y. and Wu, X. (2007). A survey and experimental comparison of service-level-approximation methods for nonstationary $M(t)/M/s(t)$ queueing systems with exhaustive discipline, *INFORMS Jounal on Computing* **19**(2): 201–214.

Ingolfsson, A., Campello, F., Wu.X. and Cabral, E. (2010). Combining integer programming and the randomisation method to schedule employees, *European Journal of Operational Research* **202**(1): 153–163.

Isken, M. and Hancock, W. (1991). A heuristic approach to nurse scheduling in hospital units with non-stationary, urgent demand, and a fixed staff size, *Journal of the Society for Health Systems* **2**(2): 24–41.

Izady, N. (2010). *On Queues with Time-varying Demand*, PhD thesis, Lancaster University Management School.

Izady, N. and Worthington, D. (2012). Setting staffing requirements for time dependent queueing networks: The case of Accident and Emergency departments, *European Journal of Operational Research* **219**(3): 531–540.

Jaiswal, N. (1962). Time-dependent solution of the head-of-the-line priority queue, *Journal of the Royal Statistical Society. Series B (Methodological)* **24**(1): 91–101.

Jarrah, A., Bard, J. and de Silva, A. (1994). Solving large-scale tour scheduling problems, *Management Science* **40**(9): 1124–1145.

Jennings, O., Mandelbaum, A., Massey, W. and Whitt, W. (1996). Server staffing to meet time-varying demand, *Management Science* **42**(10): 1383–1394.

Johnson, N., Kotz, S. and Kemp, A. (1993). *Univariate Discrete Distributions*, John Wiley & Sons, New York.

Jolliffe, I. (2008). *Principal Component Analysis, Second Edition*, Springer-Verlag, New York.

Kamentzky, R., Shuman, L. and Wolfe, H. (1982). Estimating need and demand for prehospital care, *Operations Research* **30**(6): 1148–1167.

Kao, E. and Narayanan, K. (1990). Computing steady-state probabilities of a nonpreemptive multiserver queue, *Journal on Computing* **2**(3): 211–218.

Kao, E. and Wilson, S. (1999). Analysis of nonpreemptive priority queues with multiple servers and two priority classes, *European Journal of Operational Research* **118**: 181–193.

Keith, E. (1979). Operator scheduling, *AIIE Transactions* **11**(1): 37–41.

Kella, O. and Yechiali, U. (1985). Waiting times in the non-preemptive priority $M/M/c$ queue, *Communications in Statistics: Stochastic Models* **1**: 257–262.

King, B. (1968). Estimating community requirements for the emergency care of highway accident victims, *American Journal of Public Health* **58**(8): 1422–1430.

King, B. and Sox, E. (1967). An Emergency Medical Service System - Analysis of workload, *Public Health Reports* **82**(11): 995–1008.

Knight, V. (2011). Personal webpage, *www.vincent-knight.com* (Last Accessed 12/08/11).
URL: *www.vincent-knight.com*

Knight, V., Harper, P. and Smith, L. (2012). Ambulance allocation for maximal survival with heterogeneous outcome measures, *OMEGA - The International Journal of Management Science* **40**(6): 918–926.

Kolesar, P., Rider, K., Crabill, T. and Walker, W. (1975). A queueing-linear approach to scheduling police patrol cars, *Operations Research* **23**: 1045–1062.

Koole, G. and Mandelbaum, A. (2002). Queueing models of call centers: An introduction, *Annals of Operations Research* **113**: 41–59.

Kvalseth, T. and Deems, J. (1979). Statistical models of the demand for Emergency Medical Services in an urban area, *American Journal of Public Health* **69**(3): 250–255.

Laarhoven, v. and Aarts, E. (1987). *Simulated Annealing: Theory and Applications*, D. Reidel Publishing Company, Kluwer Academic Publishers Group.

Lavenhar, M., Ratner, R. and Weinerman, E. (1968). Social class and medical care: Indices of nonurgency in use of hospital emergency services, *Medical Care* **6**(5): 368–380.

Lavoie, S., Minoux, M. and Odier, E. (1988). A new approach for crew pairing problems by column generation with an application to air transportation, *European Journal of Operational Research* **35**(1): 45–58.

Lewis, R. (2008). A survey of metaheuristic-based techniques for university timetabling problems, *OR Spectrum* **30**(1): 167–190.

Li, Y. and Kozan, E. (2009). Rostering ambulance services, *in* A. Casteli (ed.), *Industrial Engineering and Management Society, 14-16 December 2009, Kitakyushu International Conference Center, Kitakyushu, Japan*, pp. 14–16.

Lightfoot Solutions (2009). Time to make a difference: Transforming ambulance services in Wales. A modernisation plan for ambulance services and NHS Direct Wales, *Technical report*.
URL: *http://www.ambulance.wales.nhs.uk/assets/documents/c4cc0416-9fab-4dea-8753-247a9431c4c7633446359123733750.pdf, accessed 2 June 2010*

Lin, C., Lai, K. and Hung, S. (2000). Development of a workforce management system for a customer hotline service, *Computers and Operations Research* **27**: 987–1004.

Mason, A., Ryan, D. and Panton, D. (1998). Integerated simulation, heuristic and optimisation approaches to staff scheduling, *Operations Research* **46**: 161–175.

Massey, W. and Whitt, W. (1994). An analysis of the modified offered-load approximation for the nonstationary erlang loss model, *The Annals of Applied Probability* **4**(4): 1145–1160.

Massey, W. and Whitt, W. (1997). Peak congestion in multi-server service systems with slowly varying arrival rates, *Queueing Systems* **25**: 157–172.

Matteson, D., McLean, M., Woodard, D. and Henderson, S. (2011). Forecasting emergency medical service call arrival rates, *The Annals of Applied Statistics* **5**(2B).

McConnel, C. and Wilson, R. (1998). The demand for prehospital emergency services in an aging society, *Social Science and Medicine* **46**(8): 1027–1031.

Millar, H. and Kiragu, M. (1998). Cyclic and non-cyclic scheduling of 12 h shift nurses by network programming, *Journal of Operational Research* **104**(3): 582–591.

Miller, D. (1992). Steady-state algorithmic analysis of $M/M/c$ two-priority queues with heterogeneous severs, *Applied Probability - Computer science, The Interface* **2**: 207–222.

Morse, P. (1955). Stochastic properties of waiting lines, *Operations Research* **3**: 255–261.

Ngo, B. and Lee, H. (1990). Analysis of a preemptive priority $M/M/c$ model with two types of customers and restriction, *Electronic Letters* **26**: 1190–1192.

Nijdam, J. (1990). Forecasting telecommunications services using Box-Jenkins (ARIMA) models, *Telecommunication Journal of Australia* **40**(1): 31–37.

Nishida, T. (1992). Approximate analysis for heterogeneous multiprocessor systems with priority jobs, *Performance Evaluation* **15**: 77–88.

Office for National Statistics (2011). Statistical Bulletin: Index Of Services, January 2011.

Ozkarahan, I. and Bailey, J. (1998). Goal programming model subsystem of a flexible nurse scheduling support system, *IIE Transactions* **20**(3): 306–316.

Patterson, K., Hassani, H., Heravi, S. and Zhigljavsky, A. (2011). Multivariate singular spectrum analysis for forecasting revisions to real-time data, *Journal of Applied Statistics* **38**(10): 2183–2211.

Petrovik, S. and Berghe, G. (2012). A comparison of two approaches to nurse rostering problems, *Annals of Operations Research* **194**(1): 365–384.

Pirlot, M. (1996). General local search methods, *European Journal of Operational Research* **92**(3): 493– 511.

Pollaczek, F. (1934). "Uber das warteproblem", *Math Z* **38**: 492–537.

Quinn, P., Andrews, B. and Parsons, H. (1991). Allocating telecommunications resources at L. L. Bean, Inc., *Interfaces* **21**(1): 75–91.

Rayward-Smith, V., Osman, I., Reeves, C. and Smith, G. (1996). *Modern Heuristic Search Methods*, Jon Wiley & Sons.

Rodo, X., Pascual, M., Fuchs, G. and Faruque, A. (2002). ENSO and cholera: A nonstationary link related to climate change?, *PNAS* **99**(20): 12901–12906.

Samuels, P. C. and Patel, C. (2010). Scholarship in mathematics support services, *Journal of Learning Development in Higher Education* **2**.

Schimmelpfeng, K. and Helber, S. (2007). Application of a real-world university-course timetabling model solved by integer programming, *OR Spectrum* (4): 783–803.

Setzler, H., Park, S. and Saydam, C. (2005). Developing accurate forecasts for ambulance demand via artificial neural networks: a framework, 35th International Conference on Computers and Industrial Engineering, Istanbul, Turkey.

Setzler, H., Park, S. and Saydam, C. (2009). EMS call volume predictions: A comparative study, *Computers & Operations Research* **36**: 1843–1851.

Shen, H. and Huang, J. (2008). Interday forecasting and intraday updating of call center arrivals, *Manufacturing & Service Operations Management* **10**(3): 391–410.

Siler, K. (1975). Predicting demand for publicly dispatched ambulances in a metropolitan area, *Health Services Research* **10**(3): 254–263.

Silvestro, R. and Silvestro, C. (2000). An evaluation of nurse rostering practices in the National Health Service, *Journal of Advanced Nursing* **32**(3): 525–535.

Sleptchenko, A. (2005). An exact solution for the multi-class, multi-server queues with non-preemptive priorities, *Queueing Systems* **50**: 81–107.

Smith, H., Laporte, G. and Harper, P. (2009). Locational analysis: highlights of growth to maturity, *Journal of the Operational Research Society* **60**: S140–S148.

Smith, L. (1976). The application of an interactive algorithm to develop cyclical rotational schedules for nursing personnel, *INFOR* **14**(1): 57–70.

Stewart, W. (2009). *Probability, Markov Chains, Queues, and Simulation: The Mathematical Basis of Performance Modelling*, Princeton University Press.

Taylor, J. (2008). A comparison of univariate time series methods for forecasting intraday arrivals at a call center, *Management Science* **54**: 253–265.

Taylor, P. and Huxley, S. (1989). A break from tradition for the San Francisco police: Patrol officer scheduling using an optimization-based decision support system, *Interfaces* **19**(1): 4–24.

The NHS Staff Council (2011). *NHS Terms and Conditions of Service Handbook, Amendment Number 24, Pay Circular (Agenda for Change).*

Thomakos, D., Wang, T. and Wille, L. (2002). Modeling daily realized futures volatility with singular spectrum analysis, *Physica A: Statistical Mechanics and its Applications* **312**(3-4): 505–519.

Thompson, G. (1993). Accounting for the multi-period impact of service when determining employee requirements for labor scheduling, *Journal of Operations Management* **11**: 269–287.

Thompson, G. (1997). Labor staffing and scheduling models for controlling service levels, *Naval Research Logistics* **44**(8): 719–740.

Tomasek, O. (1972). Statistical forecasting of telephone time series, *Telecommunications Journal* **39**(12): 725–731.

Trudeau, P., Rousseau, J., Ferland, J. and Choquette, J. (1989). An operations research approach for the planning and operation of an ambulance service, *INFOR* **27**(1): 95–113.

Ulukus, M. (2011). *The M/M/s Queue*, Wiley Encyclopedia of Operations Research and Management Science.

Utley, M. and Worthington, D. (2011). Capacity Planning, *Handbook of Healthcare System Scheduling*, Springer, New York, pp. 11–30.

Vautard, R. and Ghil, M. (1989). Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series, *Physica D* **35**(3): 395–424.

Vile, J., Gillard, J., Harper, P. and Knight, V. (2012). Predicting ambulance demand using singular spectrum analysis, *Journal of the Operational Research Society* **63**(11): 1556–1565.

Wagner, D. (1997). Waiting time of a finite-capacity multi-server model with non-preemptive priorities, *European Journal of Operational Research* **102**: 227–241.

Wall, A. and Worthington, D. (2007). Time-dependent analysis of virtual waiting time behaviour in discrete time queues, *European Journal of Operational Research* **178**(2): 482–499.

Weare, B. and Nasstrom, J. (1982). Examples of extended empirical orthogonal functions, *Monthly Weather Review* **110**: 481–485.

Weinerman, E., Ratner, R., Robbins, A. and Lavenhar, M. (1966). Yales studies in ambulatory medical care, *American Journal of Public Health* **56**(7): 1037–1056.

Welsh Ambulance Services NHS Trust (2007). Time to make a difference: Transforming ambulance services in Wales. A modernisation plan for ambulance services and NHS Direct Wales, Final Report, *Technical report*.
URL: *http://www.ambulance.wales.nhs.uk/assets/documents/c4cc0416-9fab-4dea-8753-247a9431c4c7633446359123733750.pdf, accessed 2 June 2010*

Welsh Ambulance Services NHS Trust (2010). "Getting There", *Technical report*.
URL: *http://www.www.ambulance.wales.nhs.uk/assets/documents/ffcadc31-4fae-415b-9123-e32fbadf4a8363421259866578365 4.pdf, accessed 25 April 2011*

Welsh Government (2011). Ambulance Services in Wales: February 2011, *Technical Report SDR 59/2011*.

Welsh Government (2012a). Ambulance Services in Wales: June 2012, *Technical Report SDR 113/2012*.

Welsh Government (2012b). Ambulance Services in Wales: September 2012, *Technical Report SDR 187/2012*.

Yiou, P., Baert, E. and Loutre, M. (1996). Spectral analysis of climate data, *Surveys in Geophysics* **17**: 619–663.

Zhigljavsky, A. (2010). Singular spectrum analysis for time series: Introduction to this special issue, *Statistics and Its Interface* **3**: 255–258.

Appendices

Appendix A

Supporting documents

A.1 Annual response performance 2010/11

Table A.1: WAST performance statistics, by month, December 2010 - November 2011 (Source: StatsWales (*www.statswales.wales.gov.uk*), accessed 09/11/2012)

Standard[2]	A8	A14/18/21	B14/18/21	U15
Dec-10[3]	47.2	74.9	62.1	61.7
Jan-11	59.6	87.5	74.3	70.1
Feb-11	67.5	92.4	78.4	74.6
Mar-11	70.7	93.2	80.2	76.6
Apr-11	68.4	89.0	77.6	74.6
May-11	71.2	93.7	82.2	76.1
Jun-11	70.0	92.7	79.5	72.8
Jul-11	71.2	93.2	81.3	74.3
Aug-11	71.2	93.2	81.3	76.1
Sep-11	69.3	92.2	79.2	76.5
Oct-11	68.7	92.0	77.1	74.1
Nov-11	70.0	93.3	79.1	75.4
Average	**67.1**	**90.6**	**77.7**	**73.6**
Target	**65**	**95**	**95**	**95**

[2]The first character refers to the call category (A=Category A; B=Category B and U=Urgent) and the number refers to the number of minutes in which the percentage of calls are reported to have arrived within

[3]Severe weather conditions affected most of Wales in December 2010

A.2 Changes to National Ambulance Performance Standards

In March 2011, the Welsh Government published National Ambulance Performance Standards that are more focussed on improved clinical outcomes for patients. Only the most serious calls, classed as Category A (immediately life-threatening), are now guaranteed an emergency blue light response. All other calls receive an appropriate response, either face-to-face or telephone assessment, based on clinical need. In order to comply with the National Ambulance Performance Standards, the following changes to the ambulance service in Wales were introduced on 5th December 2011:

- Category B (serious but not immediately life-threatening calls) has been removed;
- Immediately life-threatening calls (where there is an imminent threat to life) continue to be identified as Category A calls but now include the most serious of the former Category B calls;
- Urgent and planned calls (serious but not life threatening and/or neither serious nor life threatening) are identified as Category C (urgent and planned) calls, but now also include the majority of the former Category B calls; and
- Calls to the ambulance service from health care professionals to order an ambulance for patients on an urgent basis for admission to hospital (previously called 'GP urgent patient journeys') are now included in the calls data. These calls are prioritised and classified as Category A or C in the same way as emergency 999 calls, although those classified as Category C have additional time bands/standards.

A.3 Proof: Calculation of δ_c

Section 6.2.2 recommends that δ_c be chosen such that $\delta_c = \frac{1}{2v}$. The following description outlines how v may be computed, as in Izady (2010).
Considering a continuous time Markov chain with finite state space $S = \{0, 1, ..., K\}$ and letting $E = \{e_1, e_2, ..., e_J\}$ denote the set of all types of events that may occur in this process; then for each $e_j \in E$, there corresponds two vectors: $r^j(t) = (r_0^j(t), ...r_K^j(t))$ (a transition rate vector corresponding to all types of events that may occur in the process) and $m^j = (m_0^j, ...m_K^j(t))$ (a target state vector). When the process is in state k at time t, $r_k^j(t)$ is the rate at which event e_j will occur and consequently put the system in state $k + m_k^j$ for $j = 1, 2, ..., J$. Thus v_k represents the total rate at which the process 'leaves' state k at time t, calculated as

$$v_k(t) = \sum_{j=1}^{J} r_k^j(t)$$

Letting $v(t) = (v_0(t), ...v_K(t))$ and $v = \sup_{0 \le t \le T}\{\max |v_k(t)|\}$, it follows that $\delta_c = \frac{1}{2v}$.

A.4 Cardiff EA shifts

Table A.2: Pool of potential shifts to be assigned to Cardiff EA crews

Shift number	Current (revised) shifts	Shift duration, hours
1	06:00 - 15:00 (06:00 - 12:00)	9 (6)
2	06:00 - 18:00	12
3	07:00 - 16:00	9
4	08:00 - 17:00	9
5	09:00 - 20:00	11
6	15:00 - 00:00	9
7	16:00 - 01:00	9
8	16:00 - 04:00	12
9	17:00 - 02:00	9
10	21:00 - 06:00	9
11	02:00 - 07:00	5

Appendix B

Publication: Predicting ambulance demand using singular spectrum analysis

B.1 Introductory remarks

This section contains a paper written in partnership with Professor Paul Harper, Dr Jonathan Gillard and Dr Vincent Knight that is to appear in JORS (Vile et al., 2012). It summarises the work discussed on demand forecasting in Chapter 4 of this thesis, demonstrating that the research is a topic of current interest and exemplifying how it contributes to the literature on advanced operational forecasting techniques.

Abstract

This paper demonstrates techniques to generate accurate predictions of demand exerted upon the Emergency Medical Services (EMS) using data provided by the Welsh Ambulance Service Trust (WAST). The aim is to explore new methods to produce accurate forecasts which can be subsequently embedded into current OR methodologies to optimise resource allocation of vehicles and staff, and allow rapid response to potentially life-threatening emergencies. Our analysis explores a relatively new nonparametric technique for time series analysis known as Singular Spectrum Analysis (SSA). We explain the theory of SSA and evaluate the performance of this approach by comparing the results with those produced by conventional time series methods. We show that in addition to being more flexible in approach, SSA produces superior longer-term forecasts (which are especially helpful for EMS planning), and comparable shorter-term forecasts to well established methods.

1 Introduction

The provision of an effective and efficient Emergency Medical Service (EMS) is a significant challenge in many developed nations. A particular difficulty for EMS planners is to allocate often limited resources whilst managing increasing demand for services, in a way to ensure high levels of geographical coverage and to improve key performance targets (Channouf et al., 2007; Setzler et al., 2009). To aid with the decision of the number of ambulances and paramedics to deploy, intensive OR has been conducted in the fields of optimal fleet size and vehicle deployment strategies; yet for these deployment schemes to be effective, the values used for forecasting future EMS demand must be accurate (Setzler et al., 2009).

Fildes et al. (2008) comment that from its foundation, OR has made many substantial contributions to forecasting as practitioners continue to recognise that the accuracy of predictions is important to their organisations; yet the authors note that major research opportunities still remain in forecasting, though they require a shift away from traditional statistical analysis. In this paper, we contribute to furthering such research by considering the ability of Singular Spectrum Analysis (SSA) to predict the arrivals of emergency incidents requiring EMS assistance. The forecasts can subsequently be embedded into a range of current OR methods such as queueing theory, simulation and optimisation, which will allow EMS managers to assign resources in a way that achieves a balance between service efficiency and service quality. By linking the effectiveness of a novel forecasting method to the organisational context in which it will be applied, we offer a unique contribution to forecasting through OR.

We motivate our investigations with a case study of ambulance demand in Wales. Akin to many developed nations, demand for EMS in Wales is increasing year on year and in particular is growing at a rate faster than the UK average. The country experiences a comparatively high number of emergency calls per head of population with above average proportions requiring patient transportation, impeding the Welsh Ambulance Service Trust's (WAST) ability to meet key performance targets set by the Welsh Assembly Government (Lightfoot Solutions, 2009). A recent review of the service found that WAST's performance relating to the target to respond to 65% of Category A[4] calls within 8 minutes significantly improved between 2007 and 2009; but there was little improvement in the Category B and Category C [5] 14/18/21 minute standards (i.e. to respond to 95% of Category B calls within 14, 18 or 21 minutes in urban, rural or sparsely populated areas respectively; and if the first response to a Category A call is not a fully equipped ambulance, to follow up with such an ambulance within the same time intervals). The report attributed this underperformance to insufficient staffing levels and a high reliance on overtime to fill core shifts, with estimated costs of employing additional staff to meet the performance targets surpassing £3,000,000 (Welsh Ambulance Services NHS Trust,

[4]Category A calls are immediately life-threatening calls

[5]Category B and C calls represent all other emergency calls

2007). Such extreme operational costs coupled with increased demand levels and public expectation have intensified the need for accurate forecasts to optimally deploy emergency response vehicles and personnel.

With increased demand and pressure for efficient EMS, many studies have investigated ways to improve the effectiveness of the service with much OR invested in the area of ambulance deployment to minimise response times and personnel scheduling (see Brotcorne et al., 2003; Goldberg, 2004; Li and Kozan, 2009). Whilst the models created each differ in complexity, they all require accurate predictions of demand to perform effectively; yet despite its fundamental importance less study has been directly invested in the forecasting aspect (as reported in Kamentzky et al. (1982) and Goldberg (2004)). The focus of this paper is how to effectively generate such forecasts.

We consider the ability of SSA and standard time series techniques to estimate and forecast the daily number of incidents reported to WAST and evaluate the prediction accuracy of the formulated models by inspecting the root mean squared errors (RMSEs) associated with the models. As the data analysed in this paper relates to national demand, we choose to predict daily demand levels. However, the SSA technique could be easily adjusted to predict demand levels on a shift-by-shift basis if these were required to make operational decisions for individual districts throughout Wales. SSA has been shown to be an effective method of time series analysis (see Broomhead and King, 1986; Broomhead et al., 1987; Vautard and Ghil, 1989; Yiou et al., 1996; Golyandina et al., 2001; Hassani, 2007) but this is the first time the technique has been applied to emergency demand. The advantage of SSA is that we do not need to fit a parametric model to the time series, but may apply the technique to any complex series with a potential structure (Hassani, 2007). Whilst traditional time series models require restrictive distributional and structural data assumptions, these assumptions are not required by SSA.

The organisation of this paper is as follows. Section 2 reviews previous research and Section 3 discusses the data used in the investigations with preliminary analyses. Section 4 explains the theory of the SSA technique, followed by a comparison of the model performance with ARIMA and Holt-Winters models in Section 5. The paper ends with conclusions and proposed directions for future research.

2 Previous research

OR investigations of EMS systems have developed considerably since the late 1960's when intensive research into the operations of the service was initiated (Setzler et al., 2005). Comprehensive reviews of OR models built for the deployment of emergency service vehicles are contained in Goldberg (2004) and Green and Kolesar (2004).

Whilst the most intensive research has been conducted in the fields of optimal fleet size and vehicle deployment strategies, EMS forecasting models have been

developed and are comparable to those designed for fire service and police services (see Holcomb and Sharpe, 2007). The initial models were very simplistic and had many shortcomings. The earliest used basic statistics to determine daily demand but failed to account for daily or weekly trend data, or other causal factors (Hall, 1971). Other early models were based on least squares regression, but were often performed on incomplete data sets with outdated socioeconomic and population covariates. Despite the data used in Aldrich et al. (1971) being subject to such errors, they successfully developed a model to predict total demand using 31 independent variables reflecting socio-demographic characteristics. Several similar investigations followed using standard regression techniques and variables to adjust for certain variations in demand. Sets of regression equations achieving high R^2 values were developed in Siler (1975) and Kvalseth and Deems (1979), whilst Kamentzky et al. (1982) successfully explained the variation in demand with only four independent variables, namely: area population, area employment rate, percentage of the population white and married, and housing units per area resident. More recent work using regression techniques has been performed by McConnel and Wilson (1998) who gave particular focus to the age distribution of the population.

A new collection of models were established at the end of the 1980's. Conventional time series models were successful in overcoming some of the shortfalls of regression techniques (such as multicollinearity, autocorrelation and the difficulty of selecting valuable covariates). Goal programing was used in Baker and Fitzpatrick (1986) to choose the optimal smoothing parameters in a Holt-Winter's exponential smoothing model (described in Section 5) to separately forecast the daily volume of emergency and non-emergency calls, whilst Channouf et al. (2007) recently developed and compared time series models to generate daily and hourly forecasts of EMS calls in Calgary, Alberta. The paper investigated the ability of autoregressive and ARIMA models to predict daily demand levels and concluded that regression models with residuals following the autoregressive process were able to forecast a few days into the future with a higher degree of accuracy than doubly seasonal ARIMA models.

The early demand prediction models were parametric in nature and required restrictive distributional and structural assumptions, such as stationarity of the data. Whilst these traditional methods proved useful for upper-level capacity planning and budgeting, recent advances in location analysis allowing ambulance deployment strategies to become more flexible and dynamic in nature, call for more responsive predictions of demand and model-free methods to predict call volumes. Some more recent methods developed to produce forecasts have proven to be successful, such as a feasible approach developed by Setzler et al. (2009) allowing forecasts to be developed on an hourly basis and for smaller areas through the consideration of Artificial Neural Networks as a viable alternative to standard causal forecasting. In this paper we investigate the model-free technique of SSA to predict demand levels which allows us to exploit the trend and seasonality patterns exhibited in the data. The problems inherent with the traditional methods are not present in SSA as it is able to expose important characteristics of the time series without requiring either a parametric

model, or assumptions concerning the signal or white noise (Golyandina et al., 2001).

SSA has been shown to be a powerful and effective nonparametric technique for time series analysis and forecasting in many diverse areas. From its origins associated with the papers Broomhead and King (1986) and Broomhead et al. (1987), SSA has been applied to many practical problems ranging from physics and meteorology to economics. Within the physical sciences, SSA has already become a standard tool in the analysis of climatic, meteorological and geophysical time series; see, for example, Yiou et al. (1996); Ghil et al. (2001) (climatology), Weare and Nasstrom (1982) (meteorology) and Colebrook (1978) (marine science). SSA has also recently been used to model of Cholera outbreaks in accordance with the El Niño cycle (Rodo et al., 2002). In the socio-economic sciences, SSA has been used to predict daily exchange rates and the volatility of the financial market (Thomakos et al., 2002; Hassani et al., 2010) . Whilst many successful applications have been made of SSA, to the best of our knowledge this study represents the first time the technique has been applied to emergency demand within an OR framework. Before we outline the SSA methodology, a brief description of the data will be given.

3 Data analysis

We analyse data provided by WAST from 1st April 2005 to 31st December 2009. The primary database contains information relating to the time and date of each incident reported to the service, the location of the incident, the assessed call priority, nature of call and data relating to the emergency vehicles sent in response. There are numerous cases in the dataset where an emergency call is logged but no ambulance is deployed e.g. deemed unnecessary due to a minor injury. In other cases several emergency response vehicles are dispatched. For the purpose of this paper, we consider daily demand to be the number of unique emergencies reported to the service which require the deployment of at least one emergency response vehicle.

An average of 1011 incidents (standard deviation 68.43) are reported to WAST each day, although the number reported fluctuates from 697 to 1485, as highlighted in Figure 1. Preliminary analysis of the data reveals daily, weekly and yearly periodicities; special-day effects; autocorrelations and a positive trend. Linear regression analysis applied to daily demand against time yields a significant slope coefficient of 0.045. All four high extreme values occur on January 1st, representing the repeating pattern of extreme demand for the service following annual New Year's Eve celebrations. The notable troughs occur on 21st March 2006, 31st October 2007 and 18th May 2009. There is no obvious reason for these low counts.

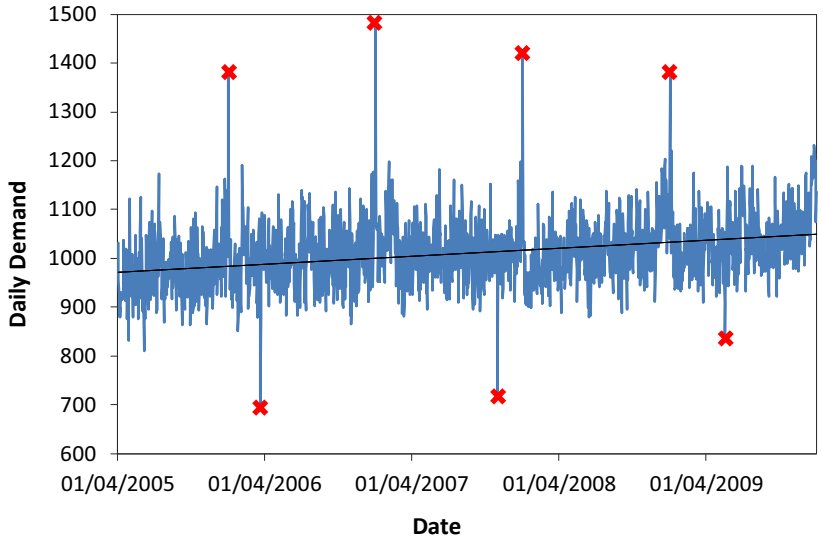

Figure 1: WAST daily demand (01/04/2005 - 31/12/2009)

Figure 2 displays the average number of daily calls requiring EMS assistance received by WAST each month over the same period. One can see a marked peak in demand for December in all years and a steady increase in demand levels over the five-year period.

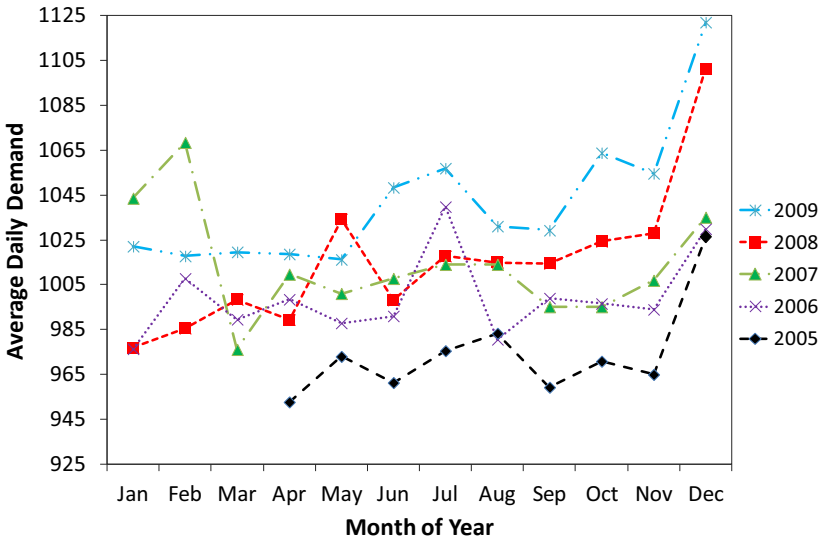

Figure 2: WAST average monthly demand (01/04/2005 - 31/12/2009)

Figure 3 displays box plots of daily demand volumes for each month of the year and day of the week. December is the busiest month with a median of 1063 incidents requiring WAST mobilisation a day. Higher demand is generally demonstrated during the winter months of November, February and October, although the lowest median demand occurs in January (984) despite the extreme peak each New Year's Day. Clear weekday effects are notable with larger volumes of incidents observed on Fridays and Saturdays. All such observations will become of key importance when designing schedules for ambulance crews.

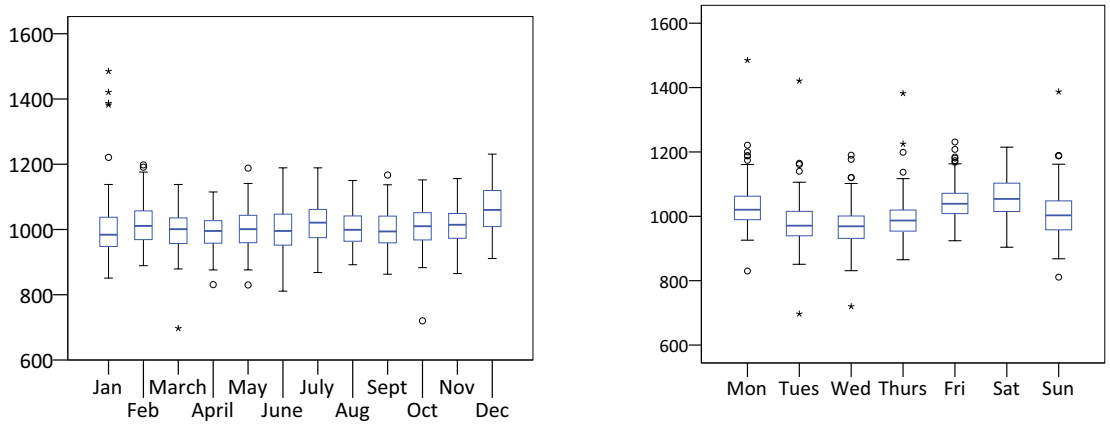

Figure 3: Box plots of demand volumes for each month and each day of week

In summary, we observe an increase in demand levels for EMS between 2005-2009, and note that demand is more volatile on weekends and in months associated with school holidays, such as July and December. Before applying SSA to the data, we will initially outline the theory. Further details concerning the theoretical underpinning of the method may be found in Golyandina et al. (2001).

4 SSA theory

SSA decomposes a time series into a sum of time series. Each component within this sum might be a trend component, periodic component, quasi-periodic component or noise. The main stages of SSA are as follows:

Stage 1: Decomposition
- Step 1: Embedding
- Step 2: Singular value decomposition (SVD)

Stage 2: Reconstruction
- Step 1: Grouping
- Step 2: Diagonal averaging

This Section will outline these stages for a real-valued nonzero time series with N observations $Y_N = (y_0, \ldots, y_{N-1})$.

4.1 Decomposition: Embedding

The first step of SSA is to map the given time series Y_N to a multidimensional series of L-lagged vectors $X_1, \ldots, X_K$. Here $X_i = (y_{i-1}, \ldots, y_{i+L-2})^T$ for $i = 1, \ldots, K$ where $K = N - L + 1$. The parameter L is known as the window length and is an integer such that $2 \leq L \leq N$. Usually L is selected so that it is proportional to the periodicity within Y_N and lies between $\frac{N}{3}$ and $\frac{N}{2}$. Some advice is given to the choice of L in Golyandina et al. (2001) and Hassani (2007).

The trajectory matrix X is formed:

$$X = [X_1, \ldots, X_K] = \begin{pmatrix} y_0 & y_1 & y_2 & \cdots & y_{K-1} \\ y_1 & y_2 & \iddots & \iddots & y_K \\ y_2 & \iddots & \iddots & \iddots & \vdots \\ \vdots & \iddots & \iddots & \iddots & \vdots \\ y_{L-1} & y_L & y_{L+1} & \cdots & y_{N-1} \end{pmatrix} \in \mathbb{R}^{L \times K}$$

X is a Hankel matrix as all elements along the anti-diagonals are identical.

4.2 Decomposition: Singular value decomposition (SVD)

Let $\lambda_1, \ldots, \lambda_L$ denote the eigenvalues of XX^T (ordered by magnitude such that $\lambda_1 \geq \ldots \geq \lambda_L \geq 0$) and $U_1, \ldots, U_L$ denote the orthogonal system of the eigenvectors of the matrix XX^T corresponding to $\lambda_1, \ldots, \lambda_L$.

If we denote $V_i = \frac{X_i^T U_i}{\sqrt{\lambda_i}}$ for $i = 1, \ldots, d$ then the SVD of the trajectory matrix can be written as

$$X = X_1 + \ldots + X_d \tag{B.1}$$

where $d = rank(X) = max(i : \lambda_i > 0)$ and $X_i = \sqrt{\lambda_i} U_i V_i^T$. The matrices $\{X_i, i = 1, \ldots, d\}$ have rank 1. The collection $(\sqrt{\lambda_i}, U_i, V_i)$ is called the i-th eigentriple of the matrix X.

4.3 Reconstruction: Grouping

Carefully selecting sets of the matrices within $\{X_i,\ i = 1, \ldots, d\}$ will give various trend or periodic components of Y_N. The grouping procedure partitions the set of indices $\{1, \ldots, d\}$ (obtained in expansion (B.1)) into m disjoint subsets $I_1, \ldots, I_m$.

Let $I = \{i_1, \ldots, i_p\}$. The resultant matrix X_I is defined as X_I as $X_I = X_{i_1} + \ldots + X_{i_p}$. This is computed for $I = I_1, \ldots, I_m$ and leads to the decomposition $X = X_{I_1} + \ldots + X_{I_m}$. For example, let $d = 10$ and $m = 2$. Then the set of indices is $\{1, \ldots, 10\}$. Let $I_1 = \{1\}$ and $I_2 = \{3, 4\}$. Then $X = X_{I_1} + X_{I_2}$ where $X_{I_1} = X_1$ and $X_{I_2} = X_3 + X_4$.

Auxiliary information may help us select particular components. For example, if it is known that there is a monthly periodicity within our time series, we may wish to identify the component(s) that reflect this. A plot of the singular values identifies the number of components to be taken (in a similar manner to principal component analysis, see (Jolliffe, 2008)). Explicit plateaux in the singular value spectra indicates pairs of components that are likely to be important. Pairwise scatter plots of components allow the visual identification of the components corresponding to harmonic elements of Y_N. Analysis of the periodograms from the original series, and of its components, will inform of the frequencies that need to be considered to reconstruct the time series. The "art" of SSA is in the selection of the subsets $I_1, \ldots, I_m$, and further details are provided in Golyandina et al. (2001). As more and more indices from $\{1, \ldots, d\}$ are selected, then more of the original signal is reconstructed. If too few indices are selected, then the reconstructed signal might not adequately explain the variation in Y_N (this might be sufficient to describe the overall trend of the series, however). If Y_N is a noisy time series, then taking too many indices would result in the noise forming part of the reconstructed signal.

4.4 Reconstruction: Diagonal averaging

Selecting $I_1, \ldots, I_m$ and computing $X = X_{I_1} + \ldots + X_{I_m}$ results in a matrix that is not of Hankel structure. In order to find the approximated time series, X must be transformed into a Hankel matrix. This may be done via diagonal averaging. Generally diagonal averaging can be described as follows.

If z_{ij} is an element within a matrix Z, the k-th term of the resulting time series is obtained by averaging z_{ij} over all i,j such that $i + j = k + 2$. This diagonal averaging operates on an $L \times K$ matrix Z ($L \leq K$) in the following way. For $i + j = s$ and $N = L + K - 1$ the element $\widetilde{z_{ij}}$ as a result of the diagonal averaging of Z is given by:

$$\widetilde{z_{ij}} = \begin{cases} \frac{1}{s-1} \sum_{l=1}^{s-1} z_{l,s-l} & 2 \leq s \leq L, \\ \frac{1}{L} \sum_{l=1}^{L} z_{l,s-l} & L+1 \leq s \leq K+1, \\ \frac{1}{K+L-s+1} \sum_{l=s-K}^{L} z_{l,s-l} & K+2 \leq s \leq K+L. \end{cases}$$

4.5 Forecasting

SSA uses linear recurrent formulae (LRF) in order to forecast future time series points. LRFs are extremely flexible; if a series is representable by a LRF then it may also be

represented as a product of exponentials, polynomials and harmonics (and vice versa). Technical details are provided in Golyandina et al. (2001). Y_N satisfies a LRF (of order q) if

$$y_{i+q} = \sum_{k=1}^{q} a_k y_{i+q-k} \qquad 1 \leq i \leq N - q + 1$$

The eigenvectors of XX^T provided in the SVD step yield the coefficients $a_1, \ldots, a_q$.

Confidence intervals for such forecasts can be obtained by bootstrapping (for further information see Efron and Tibshirani, 1993).

4.6 Measures of accuracy: root mean squared error

As a measure of prediction accuracy and to compare the goodness of fit of the models, we report the root mean squared errors (RMSE) for various models and for different forecast lags, defined in out case as:

$$\text{RMSE} = \sqrt{\frac{\sum_{n=1}^{N}(y_n - p_n)^2}{N}}$$

where y_n is the observed value, p_n is the predicted value and we have N fitted points in the time series. The RMSE is a commonly used forecast-accuracy metric in time series analysis to report how close forecasts or predictions are to the known data (Channouf et al., 2007; Matteson et al., 2011). Similar conclusions may be drawn from the results if the Mean Absolute Error (MAE) or Mean Absolute Percentage Error (MAPE) are used, but we choose to report the RMSE as in addition to overcoming the common problem encountered with the MAPE that the percentage error may become inflated if the actual value y_n in the denominator is relatively small compared to the forecast error; this performance measure also gives relatively higher weight to large errors, which are particularly undesirable for an EMS.

We will now apply this theory to the data described in Section 3, and compare the model fit against ARIMA and other standard time series models.

5 Model comparison

In this Section we compare the SSA technique with the well-established ARIMA and Holt-Winters forecasting methods based on the precision of the model fits as reflected by the RMSE. All the models are formulated using the first 51 months of data (1st April 2005 - 30th June 2009) and the forecasting error is measured using the data from the following six months. Figures 2 and 3 show that July and December are volatile months, and are thus expected to be harder to forecast, so we also individually report

the errors for these months positioned at each end of the six-month horizon. We show that whilst all methods produce reasonably accurate results for certain time periods, SSA is superior for the longer forecasting horizons. Sections 5.1 - 5.3 briefly outline the well-known forecasting algorithms of the conventional models, which are used to benchmark the forecasting accuracies.

5.1 SSA model formulation

As discussed in Section 4, the choice of the number of components to retain in the methodological process requires careful consideration. As a series of pure noise generally produces a slowly decreasing sequence of singular values, further guidance may be obtained by checking for breaks in the plot of logarithms of the eigenvalues. Explicit plateaux in the plot correspond to the components representative of clear periodicities within the data (because harmonic components with different frequencies produce two eigentriples with close singular values) and elbows in the chart demonstrate points at which retaining a larger number of components in the data explains little extra variation within the original series. When the data is truncated at various points of the post-sample period, we note an elbow in the chart commonly corresponding to the 14th eigenvalue, and thus choose to construct the SSA model using the first 14 components and a window length L = 581 (between $\frac{1}{3}$ and $\frac{1}{2}$ of the series). All of the results and figures in the following application are obtained by means of Caterpillar-SSA 3.30 software (available from www.gistatgroup.com).

Figure 4 shows the first two components generated from the SVD of the trajectory matrix obtained from the data. The majority of the variation (99.6%) is captured in the trend element accounted for by component 1, whilst the second plot reveals a periodic component in the data.

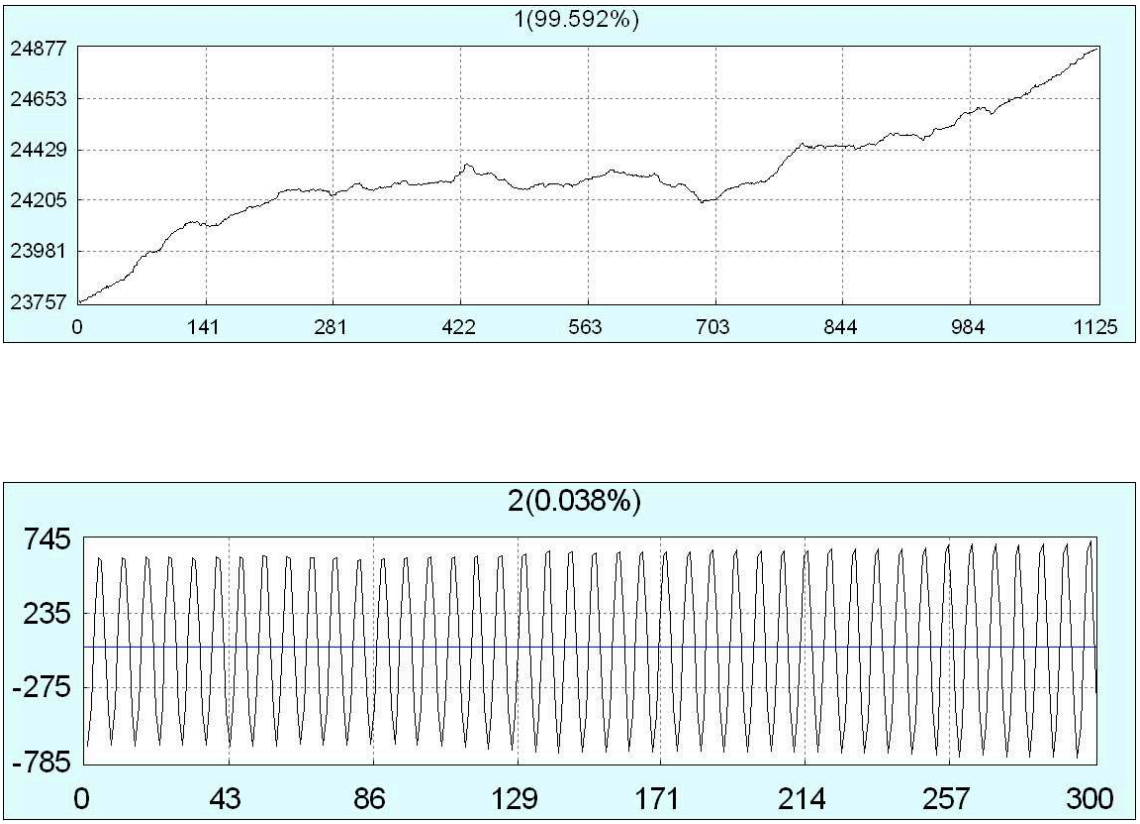

Figure 4: Principal components related to the first 2 eigentriples. [The second graph depicts up to the 300th lagged vector only, to allow the periodicity to be clearly visible].

Analysis of the pairwise scatter plots and their corresponding periodograms allows visual identification of the harmonic components. Figure 5 depicts an example scatter plot and periodogram corresponding to the second and third components. The 7-sided polygon and spike at $x = 7$ in the periodogram demonstrate that these components account for the 7 day (weekly) periodic effect. Other periodicities are accounted for by the remaining components; by retaining all significant components in the series reconstruction we account for the main periodicities in the data and build an accurate representation of demand.

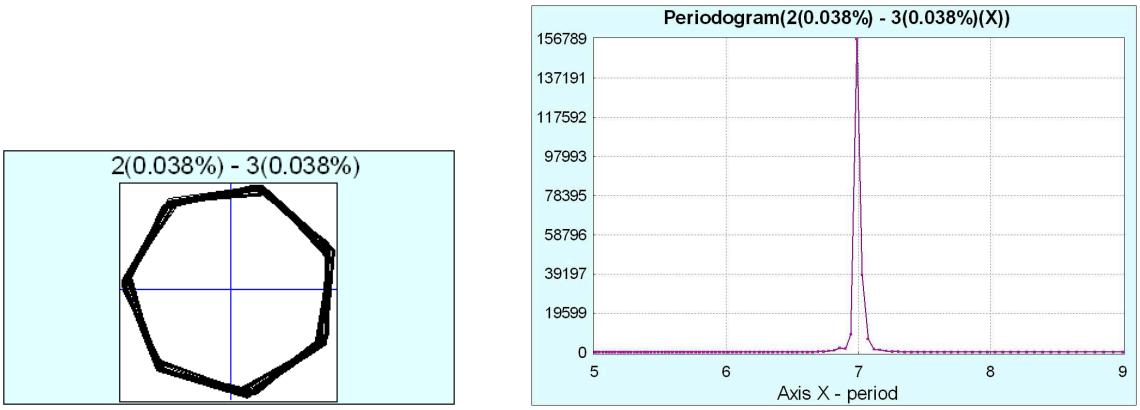

Figure 5: Scatter plot and periodogram of paired eigenvectors (2-3)

5.2 ARIMA model

ARIMA models, originally described by Box and Jenkins (1970), provide a class of models to approximate a time series after allowing the time series to be stationarised using transformations such as differencing and logging. These models account for temporal dependencies using autoregressive (AR) terms, which are lagged observations of the dependent variable and moving average (MA) terms, which are lagged error terms, as explanatory variables. Selection of the appropriate AR, MA and differencing terms to include in the model is usually considered subjective, but it does not have to be (Hyndman and Khandakar, 2008) as many attempts have been made to automate the ARIMA process. To select the optimal parameters, we use the 'forecast' package in R as described in Hyndman and Khandakar (2008). For all models built to forecast demand starting on the 1st day of the month between July-December 2009, the optimal model is ARIMA(1,0,1)(1,0,1).

5.3 Holt-Winters (HW) forecasting method

Exponential smoothing is a simple, but one of the most widely used, techniques for adaptive time series forecasting (Gardner, 1985). The model generates forecasts using a set of simple recursions and relies on a weighted average of historical data values, with the more recent values carrying more weight. Given the seasonality in the data, we consider the HW extension of the basic exponential smoothing model which includes additional terms to account for the linear trend and seasonality exhibited in the data (see Chatfield, 2001; Brockwell and Davis, 2002). For the purpose of this analysis, we observe that the HW additive model predicts the historic data more accurately, and thus select this version to forecast forward. The optimal model parameters are selected using the time series forecasting system within SAS, which includes a completely automatic forecasting model selection feature that selects the best fitting model for a time

series and reports diagnostic check results. When forecasting for each month post June 2009, we allow the model to re-parameterise at the start of each month, and note that the optimal values change slightly.

5.4 Results

In this Section, we evaluate the models outlined in Sections 5.1 – 5.3 in terms of their quality of fit and forecasting performances evaluated by the RMSE. For identification and estimation of the models we use the first 1552 daily counts (from 1st April 2005 - 30th June 2009) and reserve the last 6 months of daily counts from July - December 2009 to measure the forecasting performance of the models. We evaluate the accuracy of the forecasts by:

i. Performing a series of 1-through-28 day step-ahead forecasts beginning with the first 'unknown' observation on 1st July 2009;

ii. Updating the within-sample period by one observation and again performing a series of 1-through-28 day step-ahead forecasts;

iii. Repeating step 2 until less than 28 post-sample observations are available for evaluation purposes;

iv. Combining all forecasts to calculate the average RMSE for each forecasting method.

When calculating the average RMSE for each model, we report the measure individually for 7-day, 14-day, 21-day and 28-day forecasting horizons to provide a number of comparisons between the time series methods detailed in Sections 5.1 – 5.3. It is desirable for EMS planners to have an estimate of forecasts one month in advance so rosters may be finalised.

We first present the results when the models are run on a rolling basis over the entire post-sample six month period, starting with the first 'unknown' observation on 1st July 2009 and ending with the first 'unknown' observation on 4th December 2009 (to retain a 28-day post-sample period of known data for comparison purposes). We allow the models to re-parameterise at the start of each new month, but whilst the HW parameters vary slightly from month to month, the optimal ARIMA model (found using the 'forecast' package in R as described in Hyndman and Khandakar (2008)) remains consistently ARIMA(1,0,1)(1,0,1). For a fair comparison, we keep 14 components in the SSA model as the plot of logarithms of the eigenvalues suggests that this is a reasonable number of components to retain for all months. However, SSA does also provide flexibility for different components to be selected; hence the SSA model may be fine-tuned if necessary.

Table 1 reports the RMSE for each forecasting horizon, averaged over the 184 model runs between July and December. The first line of the table reports the retrospective

error, representing the closeness of fit of the model predictions with the initial true data used for the model construction (the period from 1st April 2005 - 30th June 2009). One may observe that the predicted values are very close to the data for all models considered.

Table 1: Comparison of model forecasts for daily demand (July - Dec 2009). Standard deviations are included in brackets.

Average RMSE	SSA	ARIMA	HW
Retrospective	6.19 (35.32)	6.11 (35.26)	6.37 (36.34)
7-day forecast	42.20 (12.92)	41.55 (13.69)	45.46 (15.79)
14-day forecast	42.86 (8.71)	44.06 (9.18)	47.47 (13.85)
21-day forecast	43.87 (7.14)	46.16 (9.50)	48.32 (12.74)
28-day forecast	45.46 (8.66)	48.75 (13.86)	51.14 (14.11)

The second part of Table 1 summarises the model forecasting performances. Values generated using the SSA technique generally follow the data more accurately than those predicted by the standard models, especially for the longer-term forecasts. The standard deviations show that the forecasts are additionally of consistent high-quality across all model runs.

Tables 2(a) and 2(b) display the segregated results when the rolling forecasts are computed for the first month (July) and last month (December) separately. These months have the highest demand levels [see Figure 3] and were noted by Channouf et al. (2007) as the most volatile months to forecast. Whilst the models were updated and re-run 31 times throughout July, lower numbers of runs were possible for December (e.g. the 28-day forecast could only be updated and re-run on 4 occasions as the true demand is only known until 31st December 2009). Both tables show that SSA often produces improved forecasts, especially for the longer forecasting horizons, but remains comparable at the least to other well-established methods for shorter periods. Further investigations have found that by selecting 200 components for December, far superior forecasts can be generated (this is an example of how SSA may be modified to produce even better forecasts for precise months). Yet we have chosen to display the results for the simpler 14-component model in this paper; whilst selecting a higher number of components can prove useful for forecasting the more volatile months, fewer components provide higher quality results over the greater part of the year.

Table 2: Comparison of model forecasts for daily demand (July & Dec 2009). Standard deviations are included in brackets.

(a) July 2009

Average RMSE	SSA	ARIMA	HW
7-day forecast	44.77 (11.03)	44.69 (13.20)	60.12 (15.87)
14-day forecast	44.25 (4.80)	48.96 (7.84)	63.52 (13.83)
21-day forecast	45.04 (3.31)	50.75 (4.63)	60.87 (12.20)
28-day forecast	45.76 (3.02)	50.74 (3.86)	62.50 (10.73)

(b) December 2009

Average RMSE	SSA	ARIMA	HW
7-day forecast	70.38 (35.52)	69.63 (38.26)	52.58 (29.80)
14-day forecast	70.96 (25.22)	86.32 (33.01)	63.20 (27.43)
21-day forecast	73.87 (10.87)	97.24 (13.00)	71.40 (6.98)
28-day forecast	80.85 (0.72)	105.47 (1.25)	90.19 (4.62)

An illustration of the 1-month SSA forecast beginning on 1st July is given in Figure 6. All 1552 daily counts up to 30th June 2009 are used in the estimation of the SSA model shown, although only the within-sample data for the month of June 2009 is displayed in the chart (to aid with clarity). In addition to the RMSE, visual inspection of Figure 6 shows that the HW method captures some element of the periodic nature of demand, but not the full variation of peaks and troughs throughout July. In contrast the SSA and ARIMA forecasts follow the true demand values reasonably closely, but of the two methods, the SSA forecast maintains the lowest RMSE when the rolling forecast is computed until December, as shown in Table 1. 95% confidence intervals for the SSA forecast are shown in Figure 7.

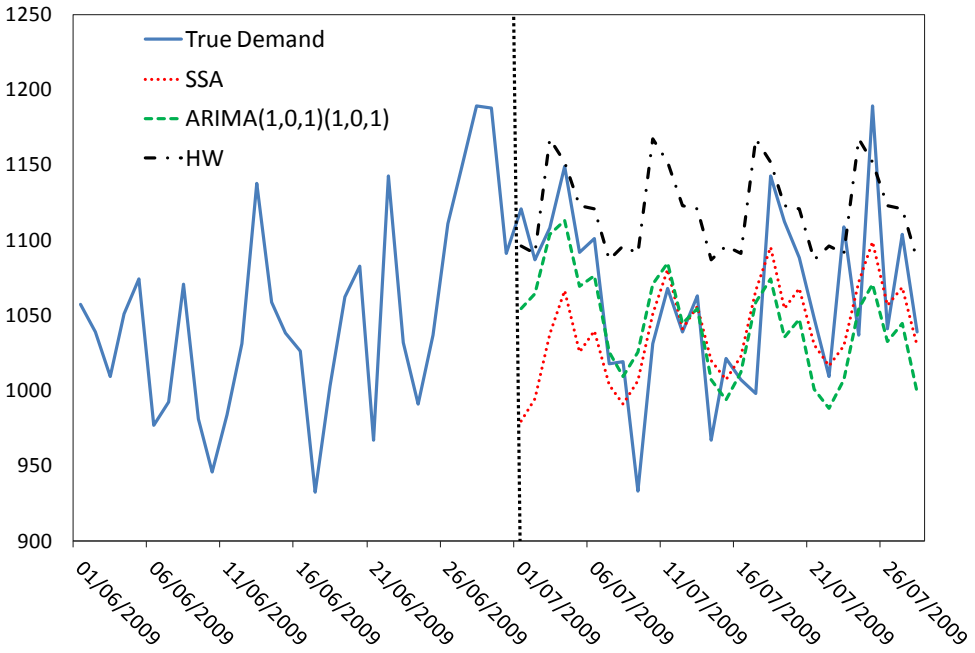

Figure 6: 28-day forecasts beginning on 1st July 2009

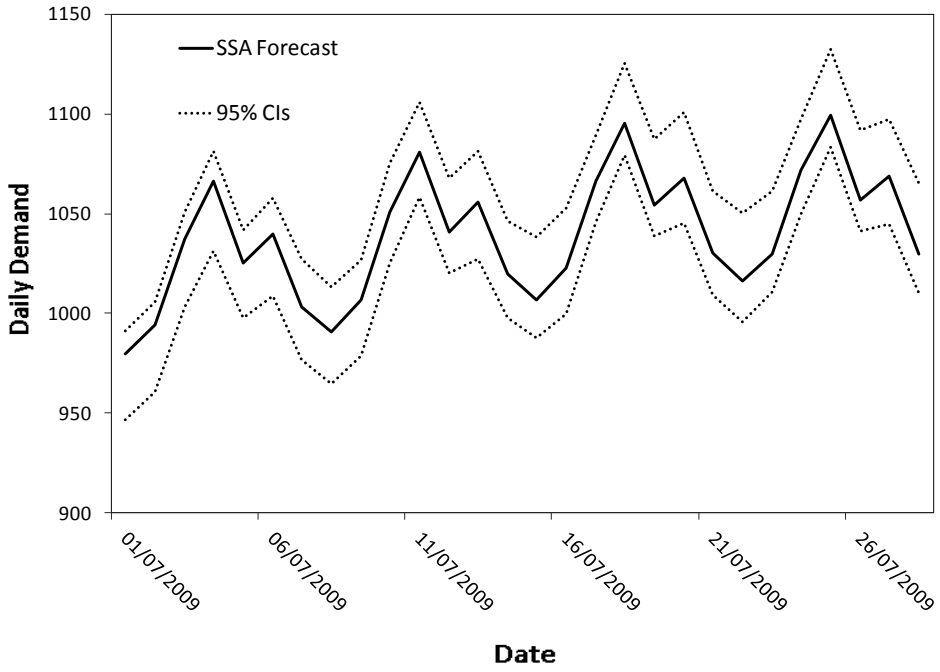

Figure 7: Confidence intervals for 28-day SSA forecast by the bootstrap method (1000 repetitions) beginning on 1st July 2009

For practical purposes, forecasts would additionally be updated as new demand levels are obtained and inputted in to the system, to finalise rostering and scheduling plans. Figure 8 illustrates the 7-step-ahead daily forecasts where the total demand levels for the current day are used to predict the demand levels for the same day one week ahead. In the same way, any "n-step" ahead forecast could be produced as required to allow WAST to update forecasts as and when required, leading to our decision to evaluate all 14-day, 21-day and 28-day forecasts in the preceding tables. One may observe the immeasurable value of updating forecasts, as the predicted SSA values forecasted one-week in advance in Figure 8 follow the SSA trend far more closely than those in the 28-day ahead forecast in Figure 6. However, the clear benefit that the SSA model provides more accurate predictions for longer forecasting horizons is a major advantage of the technique, as it is a costly operation to change staff rosters at the last minute.

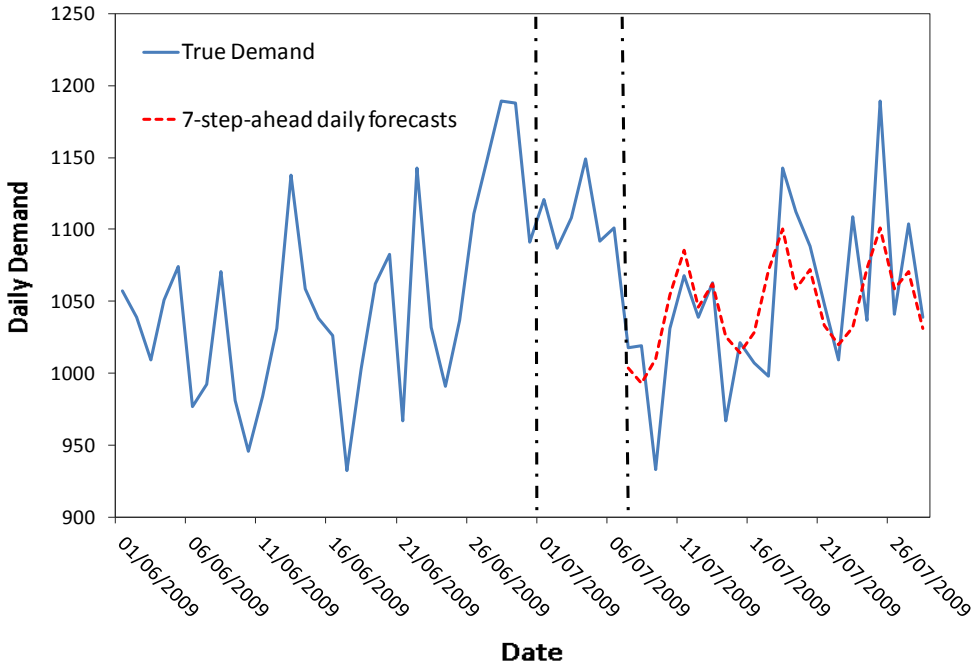

Figure 8: Original time series with 7-step-ahead daily SSA forecasts for July 2009, beginning with forecast when demand is known until 30th June

6 Conclusions

We have considered the ability of three time series analysis methods to model and forecast the number of incidents reported to WAST on a daily basis. Our analysis responds to the call by Fildes et al. (2008) to improve the accuracy of forecasts through considering novel methods, whilst providing motivation for the utilisation of SSA as a tool to accurately predict Welsh ambulance demand. The benefit of the SSA technique is not only in its ability to forecast as it produces superior, or in the least, comparable forecasts to other methods; but in its capability to recognise periodicities in the data and be flexible in approach, with the advantage that it may be easily implemented. SSA software may be readily obtained and provided to WAST employees to allow automatic revisions of the forecasts as new data becomes available, so the Trust can prepare more appropriately for future demand. Moreover, the advanced forecasts generated by SSA can be readily embedded into OR methodologies to determine the minimum number of ambulances to be deployed at any given time (Ingolfsson et al., 2007; Green et al., 2007). To be effective, these methods are heavily reliant on accurate predictions of demand, which we have developed in this paper. Further work may additionally include developing rosters based on these forecasts to promote proactive management of the service.

Ultimately, as well as investigating spatial distributions of demand using demographic variables, it might also be desirable to investigate other contributing factors such as links with weather conditions to further improve our predictions. In designing rosters for crew members, it will become necessary to model the data on a shift-by-shift basis, in line with current WAST practices (3 shifts per day) and for smaller areas, to further the research performed by Setzler et al. (2005) in this area. To maintain accurate SSA performance, we recommend forecasting on a shift, rather than hourly, basis to allow for a rostering algorithm to be more readily embedded in the forecasts, retain a large degree of seasonality in the data and reduce the number of 'zero' counts in the time series. If required, the shift forecasts could further be proportioned to calculate the expected number of incidents reported per hour.

Acknowledgements

This research is being funded by EPSRC grant EP/F033338/1 as part of the LANCS initiative.

The authors would like to thank the Welsh Ambulance Service Trust for the cooperation in providing the data and particularly Andrew Rees, Senior Information Analyst at the Health Informatics Department, for his helpful comments and advice.

Appendix C

Publication: Staffing a mathematics support service

C.1 Introductory remarks

The methods investigated within this thesis can be applied to various other scenarios, as indicated by the publication contained in this section. The paper has recently been submitted, and represents joint work completed in partnership with Dr Jonathan Gillard, Dr Vincent Knight and Dr Robert Wilson (Gillard et al., 2012). The submitted publication considers similar issues to those discussed in the time-dependent queueing theory and rostering sections of this thesis, but applies the techniques to the staffing of a mathematics support service. The scale of the problem is modest compared to the task to efficiently manage WAST, but the paper highlights that the same principles nevertheless apply. The research demonstrates that when the $M/M/c$ queue considered by the SIPP methodology in Chapter 6 is substituted by a finite source queue, the equivalent steady-state formula to compute the probability of an excessive wait still utilised to generate appropriate staffing requirements. The small-scale database compiled for the investigation of the maths support problem has also provided test bed for methods considered in this thesis; and whilst the exact rostering model considers an alternative objective, it emphasises that heuristic methods are capable of generating good quality solutions.

Abstract

We study the problem of staffing university mathematics support centers in which students drop in to the service (without appointment) for tutoring support. Our approach seeks to find the minimum sufficient number of tutors (with appropriate specialities) to present by hour and day to cover student demand with tolerable delays. We employ traditional operational research techniques to aid managers and administrators of mathematics support services to roster their services. The machine interference type queueing is adopted to model the number of student queries within a mathematics support session. We define and solve an appropriate integer program to roster the

number of tutors needed to run the service efficiently.

1 Introduction

Over the last decade the 'mathematics problem' (students lacking basic mathematical skills on entry into higher education), and proposed solutions of this problem have been debated in much detail Hawkes and Savage (2000). One approach that has been developed to help combat this issue has been the introduction of mathematics support services (MSSs) across higher education institutions. Such services in higher education can be traced back to 1990 and it is thought that they were launched even earlier in further education institutions Samuels and Patel (2010). The resources provided by MSS's often vary in nature but typically revolve around some form of drop-in service, whereby students call in to the service (without appointment) to discuss their mathematical query with a tutor.

The MSS at Cardiff University is open each weekday between 11:00 - 13:00. Many students go to the service immediately after the completion of their previous lecture, leading to large influxes of students arriving at 11:00 and 12:00, with fewer arrivals throughout the remaining period. Figure 3 shows the average number of students present at the MSS for every hour of the week between October 2010 to June 2011.

One goal of this paper is to obtain minimal staffing levels that ensure that no more than 10% of students wait longer than 15 minutes before being seen by a tutor. When employing queueing theoretical methods to estimate the optimum number of staff to ensure that a given performance metric is satisfied, we consider employing a member of staff for the two 1 hour periods throughout the two-hour session (i.e. 11:00-12:00 and 12:00-13:00). Over a given week, we thus require the staffing constraints for 10 time periods.

There is a vast quantity of literature that make use of queueing models to obtain staffing constraints for various types of service centres (for example: emergency departments Green et al. (2006) and police patrols Kolesar et al. (1975)). One novelty proposed in this paper is the use of a particular queueing model (a 'finite source queue') that is more often applied to the service of machinery than the service of individuals, to predict the amount of assistance students will require within a particular mathematics support drop-in session. Finally a mathematical program is defined; the solution of which rosters the available tutors. A heuristic is offered that gives an efficient approach to solving the rostering problem for larger problem instances.

To summarise the contribution of the work presented:

- A novel application of the 'finite source queue' is demonstrated.
- The MSS at Cardiff University is modelled using a particular queueing model to ensure sufficient levels of staffing.

- An integer program is developed that allows for an efficient rostering of a MSS, a heuristic to approximate the solution of this integer program is also discussed.
- The value of OR to aid in the provision of mathematics support services is demonstrated.

The structure of the paper is as follows. The presentation of the queueing model is given in Section 2 and the mathematical program and heuristic are presented in Section 3. Conclusions are given in Section 4.

2 MSS as a finite source queue

2.1 The mathematical model

Most of the literature dealing with setting staffing constraints assume the service under analysis to be an $M/M/c$ queue. A good review of approximation approaches that are relevant under this assumption is available in Green et al. (2007). For reasons that become apparent in Section 2.2 we choose to use a 'finite source queue' (also known as a 'machine interference model') to model the MSS at Cardiff University. This model is often used to represent situations where a finite number of machines run. Such machines run until breaking down, and consequently require repair, after which they are set to run again Haque and Armstrong (2007). The M machines break down with mean inter-breakdown rate λ per unit time (assumed to be negative exponentially distributed) and are repaired at one of c repair centres with mean repair rate μ per unit time (again assumed to have a negative exponential distribution). A diagrammatic representation is given in Figure 1.

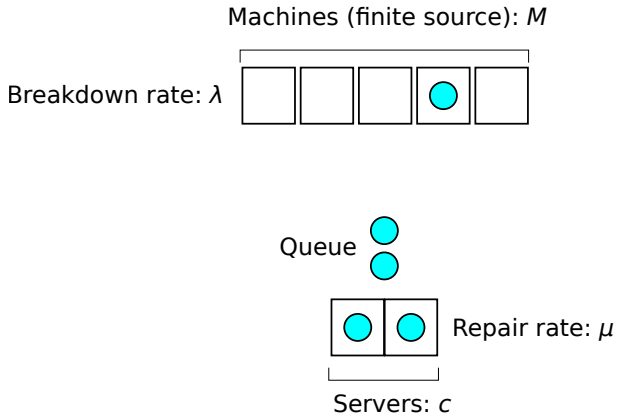

Figure 1: Diagrammatic representation of a finite source queue with $M = 5$ and $c = 2$.

Let p_n be the probability of having n of the M machines broken down at any point in time (in a queue or in service). The steady state probabilities are given by

Gross and Harris (1998):

$$p_n = \begin{cases} \binom{M}{n} \frac{\lambda}{\mu}^n p_0, & 1 \leq n < c \\ \binom{M}{n} \frac{n!}{c^{n-c} c!} p_0, & c \leq n \leq M \end{cases} \tag{C.1}$$

where p_0 can then be calculated by normalising the probabilities.

Of particular interest is the actual waiting time distribution $W_q(t)$: that is the proportion of individuals spending less than t time units waiting in the queue. This is given by:

$$W_q(t) = 1 - \sum_{n=c}^{M-1} q_n \sum_{i=0}^{n-c} \frac{c\mu t^i}{i!} e^{-c\mu t}, \tag{C.2}$$

where

$$q_n = \frac{(M-n)p_n}{L} \tag{C.3}$$

is the probability that an arrival finds the system in state n, and $L = \sum_{n=1}^{M} np_n$ is the average number of breakdowns in the system.

The next section describes the application of this finite source queueing model to Cardiff University's MSS.

2.2 Application to Cardiff University's MSS

The MSS at Cardiff University (and in many other institutions, see for example Samuels and Patel (2010)) encourages students to work in small groups and seek assistance from a tutor when required. As such the service can be modelled as a finite source queueing model where the students working in groups represent machines. A breakdown corresponds to a student requiring assistance from a tutor.

We can use (C.2) to obtain staffing levels that ensure a particular level of service. In particular we obtain a value of c that ensures $1 - W_q(.25) \leq 0.1$; the number of tutors required that ensures that the percentage of students waiting more than 15 minutes is less than 10%. Data has been collected that gives the breakdown rate λ as shown in Figure 2. This gives a breakdown rate of $\lambda = 2.19$ (i.e. students seek assistance about 2 times an hour). Furthermore a service rate of $\mu = 6$ is taken implying that a tutor will spend, on average, 10 minutes with a student.

Data collected at the MSS at Cardiff University over the academic year 2010/2011 Harding (2011) included the following variables:

i. The time when each student arrived

ii. Number of tutor consultations the student received during the session

iii. The time duration of each tutor consultation

iv. The length of time each student spent waiting for their tutor consultation to begin, from the moment they requested assistance.

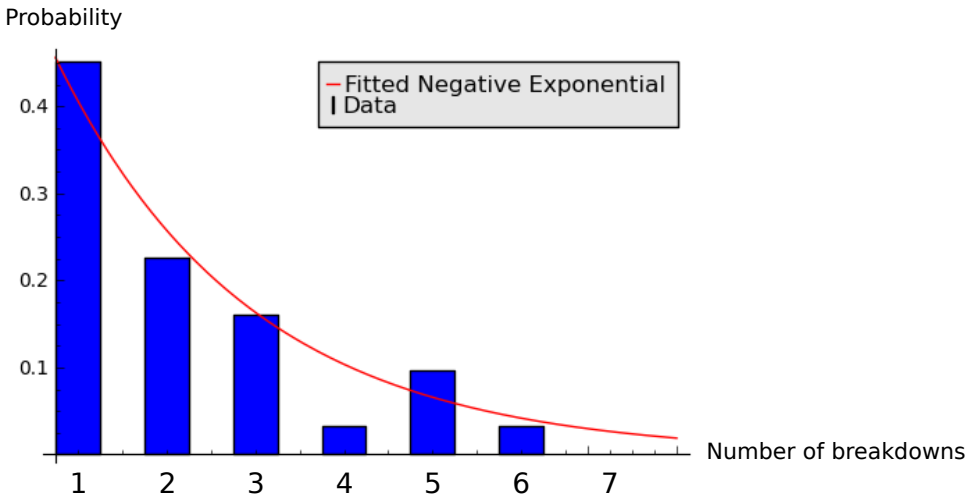

Figure 2: Negative exponential distribution with rate 2.19 fitted to the frequency of breakdowns per hour.

Figure 3 contains the mean number of students present at each session of Cardiff University's MSS, for the two semesters of the academic year, autumn and spring. The busiest semester was clearly the autumn semester, with a reduction in student attendees observed in the spring semester. In the autumn semester the busiest period was 11:00-12:00, with the following hour, 12:00-13:00 being much quieter. The busiest period was Thursday 11:00-12:00. In the spring semester, the demand is similar across all the time periods. There is some increase in attendance during the first hour of service 11:00-12:00, with the busiest period being Friday 11:00-12:00.

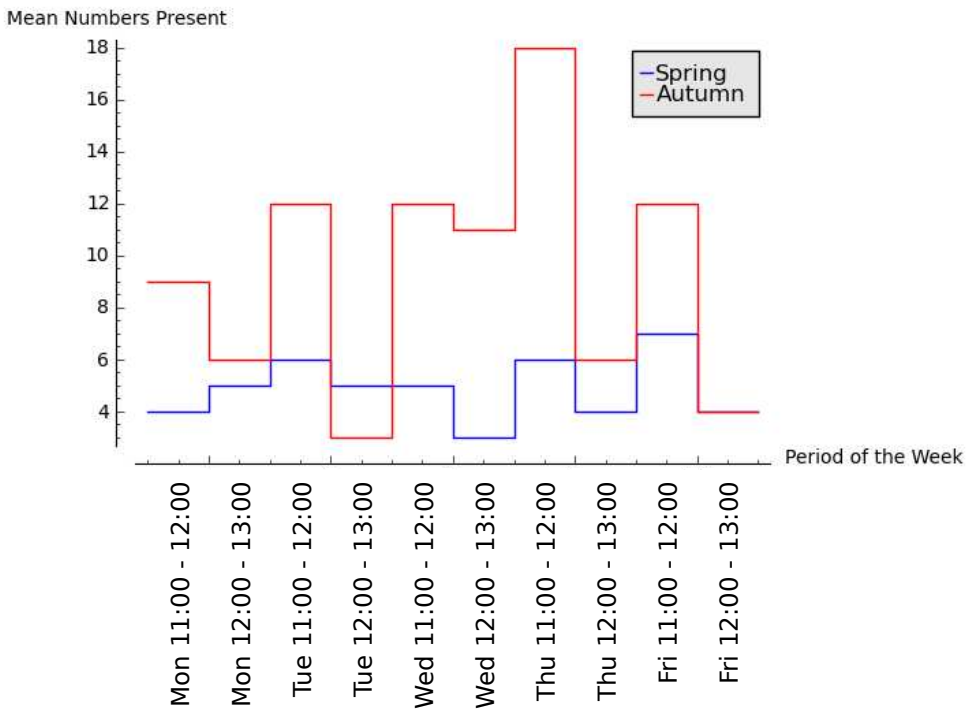

Figure 3: Mean number of students present by semester; autumn and spring.

2.3 Results

The staffing algorithm is implemented in Microsoft Excel (the package which also implements the rostering algorithm of Section 3.2 is available at Knight (2011)) and allows for an immediate calculation (using (C.2)) of the number of tutors required. Initial results are given in Figure 4. The amount of staff scheduled reflects:

i. The increased demand during 11:00-12:00 in the autumn semester

ii. More staff needed during the busiest session of the autumn semester (Thursday, 11:00-12:00)

iii. A constant number of staff needed during the spring semester, reflecting the demand viewed in Figure 3, with an increase in the number of staff during the busiest session of the spring semester (Friday, 11:00-12:00).

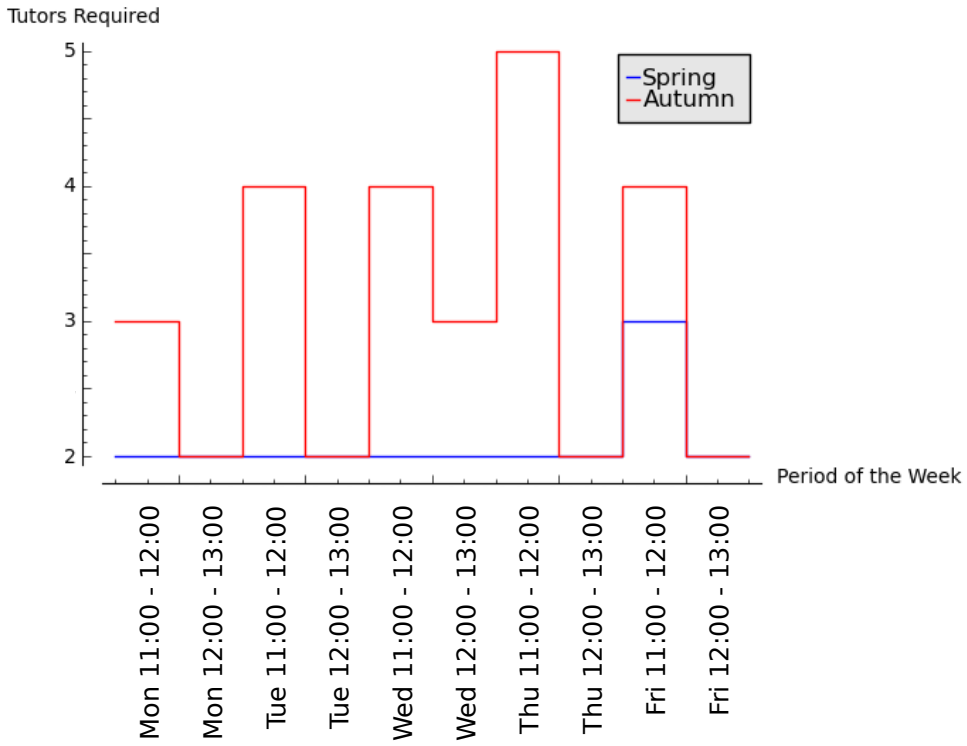

Figure 4: Number of tutors required by semester.

3 Optimisation of Staffing and Skill mix

3.1 Mathematical program

The task of assigning tutors to sessions is often the responsibility of the MSS coordinator, who attempts to find a suitable number of tutors to cover all shifts, that are additionally able to cover as many specialities (statistics, pure mathematics and applied mathematics) as possible so they may deal with all query types efficiently. However, to reduce the burden placed on this individual, we provide an integer program (IP) formulation of the problem which, in smaller cases, can be solved exactly using off-the-shelf software: in our case Xpress-MP. We utilise this method to solve Cardiff's MSS problem exactly, and also suggest a heuristic-based algorithm that may be used to produce a desirable timetable for larger problems.

The model assigns the correct number of tutors required to work for each shift as determined by the machine interference model, and incorporates a constraint for the maximum number of shifts that each tutor is willing to work during the week. Treated as an optimisation problem, the objective is to maximise a linear cost function to ensure that as many fields of mathematics (statistics, pure mathematics and applied mathematics are used as examples in this paper) are covered as possible within each shift.

Variables:

- m: number of tutors.
- n: number of shifts.
- α_i: the maximum number of shifts that can be worked by tutor $i \in [m]$.
- β_j: the number of tutors required for shift $j \in [n]$.
- k: the number of specialties.
- $C_i^{(k)} \in [0,1]_{\mathbb{R}}$ the efficiency of tutor i in speciality k.
- $D_i \in [0,1]_{\mathbb{R}}$ the desirability coefficient of tutor i.

We have the binary variable x_{ij}

$$x_{ij} = \begin{cases} 1, & \text{if tutor } i \text{ works in shift } j \\ 0, & \text{otherwise} \end{cases} \quad \text{for } i \in [m];\ j \in [n]$$

The desirability coefficient is a scaling factor which is multiplied by the subject scores to scale the values to account for unreliability, or any other factor, associated with each tutor. For example, if a particular tutor has asked that they are only scheduled to work if there is extreme need, we can give this tutor a low desirability score.

To achieve a roster that assigns tutors to sessions in a way that all fields are covered within each shift, the IP model representing our rostering model is given by the following:

Maximise,

$$\sum_{j=1}^{n} \sum_{l=1}^{k} \min \left(1, \sum_{i=1}^{m} x_{ij} C_i^{(l)} D_i \right) \tag{C.4}$$

Subject to constraints

$$\sum_{j=1}^{n} x_{ij} \leq \alpha_i \text{ for all } i \in [m] \tag{C.5}$$

$$\sum_{i=1}^{m} x_{ij} = \beta_j \text{ for all } j \in [n] \tag{C.6}$$

Here (C.5) is the worker constraint, representing the maximum number of shifts each tutor is allowed to work in a week, while (C.6) details the particular number of tutors are required per session.

Note that due to the formulation of the objective function, we will always have:

$$\sum_{j=1}^{n}\sum_{l=1}^{k}\min\left(1,\sum_{i=1}^{m}x_{ij}C_i^{(l)}D_i\right)\leqslant kn,$$

thus kn is an upper bound on the objective function.

Because we wish to maximise the number of specialities covered in each session, we constrain each of the weighted scores within the objective function (C.4) to be at most 1. This ensures that high scores may not arise from sessions where the tutors all have expertise in the *same* area of mathematics. The objective is then to maximise a cost function composed of k expressions for each session (each representing the degree to which a speciality of mathematics is covered within each session).

In our case, the timetable is composed of 10 one hour sessions a week, and is concerned with ensuring that three separate specialities of mathematics are covered within each shift; thus if we have a sufficient number of skilled tutors available, the maximum achievable cost is 30.

3.2 Finding a roster

This section considers ways for solving the MSS rostering problem defined above. Similar problems have been well studied in the literature since the 1960's with numerous heuristics, simulation approaches and graph colouring techniques suggested to solve variants of the timetabling problem in an effective way Burke and Newall (2002); Burke et al. (2007); Schimmelpfeng and Helber (2007); Lewis (2008); Cambazard et al. (2010). IP formulations have also been developed to make scheduling decisions Schimmelpfeng and Helber (2007), but in real-life applications where these problems are often too large to solve using exact methods, heuristic approaches have been shown to be successful Lewis (2008). Heuristics are acknowledged to produce good quality solutions for such problems in a short amount of time; however they often lack the ability to find an optimal solution Abramson (1991).

Cardiff University's MSS has 10 one hour sessions a week, with 8 tutors available, each with different specialities. The staff rostering problem with 10 one hour sessions and 8 tutors possesses a relatively small search space, with possibly several global optima; thus existing IP software such as Xpress-MP can be used to solve the problem exactly. However for larger instances (more sessions, and more tutors), we propose a simple heuristic (descent) algorithm for finding an approximate solution.

For the descent algorithm, the initial feasible schedule is produced using a greedy algorithm: taking the tutors in an arbitrary order, we assign as many time slots allowed by constraints (C.5) to the first tutor, and when this limit is reached, we consider the availability of the next tutor. Providing we have enough manpower to

allow a feasible solution, we continue to allocate all remaining shifts in this way.

Local search is a general strategy for optimisation that involves iteratively applying small changes (moves) to a candidate solution, attempting to improve it's quality. The random descent approach is the simplest type of local search; in each iteration a potential move is randomly selected and only accepted if it produces a solution which is at least as good as the current solution. In our algorithm, we employ a swap operator which randomly selects an element $x_{i_1 j_1}$ in the timetable such that $x_{i_1 j_1} = 1$ and a secondary element $x_{i_2 j_2}$ such that $x_{i_2 j_2} = 0$ $(1 \leqslant i_1, i_2 \leqslant m; 1 \leqslant j_1, j_2 \leqslant n)$ and swap the value of these elements. If $j_1 = j_2$ (i.e. both elements are in the same column) then no further action is needed, as we will maintain constraint (C.6) which ensures that the correct number of tutors are assigned to each session. Otherwise, to maintain integrity of the solution, we add a corrective procedure by randomly choosing two other elements $x_{i_3 j_2}$ and $x_{i_4 j_1}$, $(i_3 \neq i_1; i_4 \neq i_2; 1 \leqslant i_3, i_4 \leqslant m)$ and setting $x_{i_3 j_1} = 1$ and $x_{i_4 j_2} = 0$. In our descent method, the neighbourhood operator is repeatedly applied to produce new solutions for a set number of iterations, or until the upper bound kn is achieved for the objective function. For each solution, the new assignment of tutors to shifts is accepted if the move improves on (or maintains) the current cost.

3.3 Results

Using the IP solver, we found that the best possible cost of 30 *was* achievable for the autumn semester data; and the best cost achievable for the spring semester data was 29.52. We investigated the potential of the descent method to find good quality solutions for each of the MSS instances and found that it produced solutions that often came close to optimal, but not quite. In particular, the algorithm made rapid improvements to the initial solution, and marginal improvements after around 2000 iterations. Figure 5 demonstrates that the solution came within within 0.8% and 1.3% of the optimal solution after 2000 runs for the autumn and spring semester data respectively, and within 0.3% and 0.7% after 5000 runs. The algorithm provided better initial solutions for the autumn semester data (as more staff were required in each shift, it was an easier task to achieve the daily optimal score), and the capability of the algorithm to rapidly improve on the initial solution was indicated in particular for the spring semester data.

However, whereas heuristics are not necessary to roster Cardiff University's MSS, or for other instances that are small enough to optimally solve quickly and efficiently, they are appropriate for medium-sized instances.The descent algorithm would nevertheless take longer to find good quality solutions for large timetables; thus for such problems, we could consider using other heuristics such as simulated annealing.

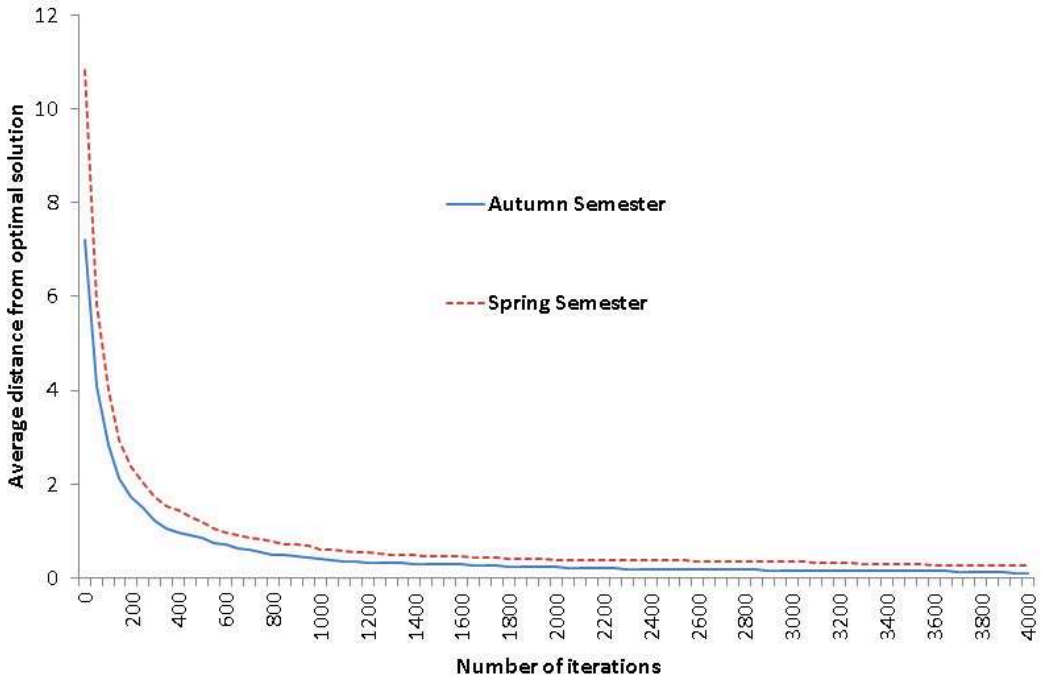

Figure 5: Rate at which descent method converges to optimality (averaged over 50 runs)

Tutor	Mon 11-12	Mon 12-1	Tue 11-12	Tue 12-1	Wed 11-12	Wed 12-1	Thu 11-12	Thu 12-1	Fri 11-12	Fri 12-1	Max no. of shifts	Stats score	Pure score	Applied score	Desirability coefficient
A					1		1				5	1	0.5	0.1	0.2
B							1		1		5	0.2	1	0.2	1
C			1		1	1	1		1		5	0.1	1	1	0.5
D	1	1		1				1		1	5	0.1	0.6	1	0.8
E	1		1						1		5	0.1	1	0.5	0.2
F		1		1		1		1		1	5	1	0.8	0.4	1
G	1		1		1		1		1		5	1	0.5	0.2	1
H			1		1	1	1				5	0.1	1	0.7	0.7
Tutors required	3	2	4	2	4	3	5	2	4	2					

Figure 6: Optimal timetable for given tutor profiles, autumn semester

The optimal roster for the autumn semester data displayed in Figure 6 appears visually logical. For example, on days where only two tutors are required, tutors D and F are consistently chosen as they are able to cover all areas of mathematics to a sufficient level amongst themselves, and are both highly desirable.

4 Conclusions

In this paper we have demonstrated that traditional operational research techniques are likely to be of much value to administrators of MSS's. Our approach in this paper

was to model the amount of student queries received during a session of the MSS at Cardiff University requiring tutor attention as breakdowns in a machine interference model. The results generated from this model were consequently input into a specially designed IP to ensure sufficient numbers of staff were scheduled, with a sufficient coverage of mathematics specialities to allow for most queries to be readily dealt with by the tutors.

The IP has been constructed with sufficient generality so that other MSS's, with different numbers of support sessions and staff available, each with different mathematical specialities, may roster their own staff. We suspect that many problem instances will be small enough to allow for an exact solution of this integer program to be found using available IP solvers. Nevertheless, for larger problem instances where IP solvers may take too long, we have demonstrated the potential use of a heuristic (descent approach).

Acknowledgements

The authors would like to acknowledge the time and valuable feedback offered by Dr Rhydian Lewis and Dr Jonathan Thompson during the writing of this manuscript.