Western Graduate&PostdoctoralStudies

Western University
Scholarship@Western

Electronic Thesis and Dissertation Repository

2-25-2016 12:00 AM

Probability Models for Health Care Operations with Application to Emergency Medicine

Azaz Bin Sharif, *The University of Western Ontario*

Supervisor: Prof. David A. Stanford, *The University of Western Ontario*
Joint Supervisor: Prof. Gregory S. Zaric, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Statistics and Actuarial Sciences
© Azaz Bin Sharif 2016

Follow this and additional works at: https://ir.lib.uwo.ca/etd

Part of the Statistics and Probability Commons

Recommended Citation

Sharif, Azaz Bin, "Probability Models for Health Care Operations with Application to Emergency Medicine" (2016). *Electronic Thesis and Dissertation Repository*. 3645.
https://ir.lib.uwo.ca/etd/3645

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlswadmin@uwo.ca.

Abstract

This thesis consists of four contributing chapters; two of which are inspired by practical problems related to emergency department (ED) operations management and the remaining two are motivated by the theoretical problem related to the time-dependent priority queue. Unlike classical priority queues, priorities in the time-dependent priority queue depends on the amount of time an arrival waits for service in addition to the priority class they belong.

The mismatch between the demand for ED services and the available resources have direct and indirect negative consequences. Moreover, ED physician pay in some jurisdictions reflects pay-for-performance contracts based on operational benchmarks. To assist in capacity planning and meeting these benchmarks, in chapter 4, I built a forecasting model to produce short-term forecasts of ED arrivals. In chapter 5, I empirically investigated the effect of workload on the productivity of ED services. Specifically, under discretionary work setting, different statistical models were fitted to identify the effect of workload and census on four measures of ED service processes, namely, number discharged, length of stay, service time, and waiting time.

The time-dependent priority model was first proposed by Kleinrock (1964), and, more recently, naming it accumulating priority queue (APQ), Stanford *et al.* (2014) derived the waiting time distributions for the various priority classes when the queue has a single server. In chapter 6, I derived expressions for the waiting time distributions for a multi-server APQ with Poisson arrivals for each class, and a common exponential service time distribution. In chapter 7, I worked with a KPI based service system where there are specific time targets by which each class of customers should commence their service and a compliance probability indicating the proportion of customers from that class meeting the target. Recognizing the fact that a customer who misses their KPI target is of greater, not lesser importance, I seek to minimize a weighted sum of the expected amount of excess waiting for each class. When minimizing the total expected excess, our numerical examples lead to an easily-implemented rule of thumb for the optimal priority accumulation rates, which can have an immediate impact on health care delivery.

Keywords: Time series analysis, regression models, emergency department, discretionary work, operations management, priority queue, time-dependent priority, non-preemptive priority, optimization, excess wait time

Co-Authorship Statement

I would like to acknowledge Prof. David A. Stanford, who contributed to all four papers (chapters 4, 5, 6, and 7) of this thesis by providing ideas, proposing some of the constructs used, helping to guide the analysis and providing suggestions and feedback for improving the texts.

I would like to acknowledge Prof. Gregory S. Zaric, who contributed to the first two papers (chapters 4 and 5) of this thesis by providing ideas, helping to guide the analysis and providing suggestions and feedback for improving the texts.

I would like to acknowledge Dr. Alim Pardhan, who contributed to the first two papers (chapters 4 and 5) by providing data, ideas, and suggestions for improving the text and analysis.

I would like to acknowledge Prof. Peter Taylor and Associate Prof. Ilze Zeidin, who contributed to the 3rd paper (chapter 6) by providing feedback and suggestions for improving the text and analysis. In particular, Professor Taylor proposed the bisection method for finding the optimal solution, and shared in the development of the particular measure of optimality along with Dr. Stanford and myself.

I would like to acknowledge Prof. Richard J. Caron, who contributed to the 4th paper (chapter 7) by providing ideas, analysis and suggestions for improving the text and analysis.

Acknowlegements

First, I would like to convey my heartiest gratitude to my supervisor, Prof. David A. Stanford without whose continuous guidance, support, and help my dissertation would not have been completed. Prof. Stanford was extremely encouraging, respectful, and approachable for any problems related to work or personal life.

Special thanks to Prof. Gregory S. Zaric for extending his hand to co-supervise me from the third year of my Ph.D studies until the end. Prof. Zaric not only introduced me with the applied operations management problems but also guided me to shape up my research ideas and problem solving techniques.

A special feeling of gratitude to my parents, siblings, relatives, in-laws, and above all friends who have supported me in every step of my life. I am also thankful to my fellow masters and Ph.D students in the department, who I had numerous discussions with and shared many thoughts.

I would also like to extend my thanks to all the professors and staffs of the department of statistical and actuarial sciences, Western University, whose doors were always open for me to receive any course, research, or administrative help.

Finally, I would like to thank my beloved wife Mariya, who have sacrificed a lot during my last few years of Ph.D. She invariably motivated me when I was frustrated, encouraged me when I was lost, and supported me whenever I needed it. Above all, Mariya recently presented me a beautiful gift, my cute little son Mayush, for whom I am doubly thankful to her.

Contents

Abstract **ii**

Co-Authorship Statement **iv**

Dedication **v**

Acknowlegements **v**

Table of Contents **vi**

List of Figures **xi**

List of Tables **xv**

1 Introduction **1**

2 Review of Statistical Methods **5**

- 2.1 Linear Models 5
- 2.2 Generalized Linear Models 6
- 2.3 Seasonal ARIMA Models 8
  - 2.3.1 Non-stationarity and Unit Root Tests 12
    - 2.3.1.1 Augmented Dickey-Fuller (ADF) Test 12
    - 2.3.1.2 Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test 13
  - 2.3.2 Determination of the Orders AR(p),MA(q),SAR(P),SMA(Q) 13
- 2.4 Harmonic Regression 15

    2.4.1 Determination of the Wavelengths Using a Periodogram 16
  2.5 Generalized Linear Autoregressive Moving Average Models 16
    2.5.1 Probability Integral Transforms (PIT) 18
  2.6 Forecast Accuracy Measures 19
  2.7 Cox-proportional Hazard Models 20

3 ED Process Flow and Data Description 22
  3.1 ED Process Flow 22
  3.2 Study Setting 24
    3.2.1 Missing Value Imputation 25
    3.2.2 Data Processing 28
  3.3 Clinical Setting and Data Description 30
    3.3.1 Clinical Setting 30
    3.3.2 Data Description and Definitions 34
      3.3.2.1 Dependent Variables 34
      3.3.2.2 Independent and Control Variables 35

4 Forecasting Emergency Department Arrivals 37
  4.1 Introduction 37
  4.2 Literature on Forecasting ED Arrivals 40
    4.2.1 Time Series Models 40
    4.2.2 Variants of Linear Models 42
    4.2.3 Rolling Horizon 45
    4.2.4 Contribution of this Chapter 45
  4.3 Statistical Methods 46
    4.3.1 Seasonal ARIMA models 46
    4.3.2 GLM with Calendar Variables as Predictors 47

4.3.3 Harmonic Regression Based on GLM and ARIMA 47
4.3.4 Generalized Linear Autoregressive Moving Average Models 48
4.3.5 Rolling Horizon Approach 49
4.4 Results and Interpretation 50
4.4.1 Descriptive Analysis 50
4.4.2 Univariate Analysis 55
4.4.3 Forecasting Outputs with Seasonal Time-Series Models 57
4.4.4 Forecasting Outputs for GLM with Calendar Variables 65
4.4.5 Forecasting Outputs for Harmonic Regression 70
4.4.6 Forecasting Outputs for the Generalized Linear Autoregressive Moving Average model 73
4.4.7 Rolling Horizon Approach to Produce Forecast 79
4.5 Discussion 91
4.6 Limitations 92
4.7 Conclusion and future research 93

5 Empirical Analysis of an Emergency Department Service Process 95
5.1 Introduction 95
5.2 Related Literature 97
5.2.1 Empirical Literature Related to ED 97
5.2.1.1 Emergency Medicine Literature 97
5.2.1.2 Operations Management Literature 98
5.2.2 Other Related Empirical Literature 99
5.3 Hypotheses Development 101
5.3.1 Expected Patient Discharge 101
5.3.2 Patient Length of Stay 102
5.3.3 Patient Waiting Time 103
5.3.4 Patient Service Time 103

5.4 Model Specifications 105
5.4.1 Model for Hypotheses 1 and 2 : Number Discharged & Census 106
5.4.2 Model for Hypotheses 3 and 4 : LOS & Census 107
5.4.3 Model for Hypothesis 5 and 6 : Waiting Time & Workload 108
5.4.4 Model for Hypotheses 7 and 8 : Service Time & Workload 108
5.5 Results and Interpretation 109
5.5.1 Descriptive analysis 109
5.5.2 Statistical Assessment of the Hypotheses 111
5.5.2.1 Hypotheses 1 and 2: Number Discharged and Census 111
5.5.2.2 Hypotheses 3 and 4: Length of Stay and Workload 114
5.5.2.3 Hypotheses 5 and 6: Waiting Time and Workload 115
5.5.2.4 Hypotheses 7 and 8: Service Time and Workload 118
5.6 Discussion 120
5.7 Limitations and Future Work 122

6 A multi-class multi-server accumulating priority queue 123

6.1 Introduction 123
6.2 Description of the Model 126
6.3 Waiting Time Distributions 127
6.4 Numerical Investigations 130
6.5 Conclusion and Future Research 141
6.6 Acknowledgments 142
6.7 Appendices 143

7 An Optimization Problem for Queues Operating under Waiting Targets 145

7.1 Introduction 145
7.2 Formulation of the Optimization Problem 149
7.2.1 Formulating the APQ Optimization Problem 150

7.3 Convexity in the Two-class APQ Optimization Problem 152
7.4 Relationship between the Transforms of the Waiting Time CDF and the Excess Wait 158
7.5 The Utility of Gaver-Stehfest Numerical Inversion for Evaluating Expected Excess 159
7.5.1 Necessary Formulae from the APQ Multi-server Model 161
7.6 Numerical Experiments 164
7.7 Conclusions and Future Research 171
7.8 Acknowledgements 173
7.9 Appendices 173

8 Conclusions **176**
8.1 Future Research Directions 177

Bibliography **179**

Curriculum Vitae **189**

List of Figures

3.1 A simplified emergency department process flow. 23

3.2 Imputation process for missing arrival time. 29

3.3 Staffing level for the emergency department under study in a particular day (24 hours period). 32

3.4 Emergency department patient flow process (this figure is not drawn to scale). . 33

4.1 Rolling horizon process flow. 49

4.2 Number of arrivals in an Emergency Department by block of the day, Jan 2012 to Dec 2013; R:Resuscitation, E:Emergent, U:Urgent, LU:Less-Urgent, NU:Non-Urgent. 51

4.3 Number of arrivals in an Emergency Department by day of the week, Jan 2012 to Dec 2013; R:Resuscitation, E:Emergent, U:Urgent, LU:Less-Urgent, NU:Non-Urgent. 51

4.4 Number of arrivals in an Emergency Department by month of the year, Jan 2012 to Dec 2013; R:Resuscitation, E:Emergent, U:Urgent, LU:Less-Urgent, NU:Non-Urgent. 51

4.5 Number of arrivals in an Emergency Department by month of the year, Jan 2012 to Dec 2013. 54

4.6 Number of arrivals in an Emergency Department by month of the year, Jan 2012 to Dec 2013. 54

4.7 Number of arrivals in an Emergency Department by month of the year, Jan 2012 to Dec 2013. 54

4.8 Time series plot for the first three months (year 2012) of high and low acuity block arrival data. 58
4.9 ACF of high and low acuity block arrivals (top row) and PACF for the high and low acuity block arrivals (bottom row), respectively. 60
4.10 ACF of the seasonally difference (order 8) high and low acuity block arrivals (top row) and PACF of the seasonally difference (order 8) high and low acuity block arrival (bottom row), respectively. 61
4.11 ACF of the residuals of the final model for high and low acuity block arrivals (top row) and PACF of the residuals of the final model for high and low acuity block arrivals (bottom row). 63
4.12 One month of preceding block arrival data with 56 blocks (7 days) ahead forecast with 95% confidence interval for high (top) and low (bottom) acuity arrivals. 64
4.13 Residual analysis of the fitted Poisson model for the high acuity arrivals. 69
4.14 Residual analysis of the fitted Poisson model for the low acuity arrivals. 69
4.15 Periodogram plots for the high and low acuity arrivals. 71
4.16 Residual analysis of the fitted GLARMA model for high acuity arrivals. 77
4.17 Residual analysis of the fitted GLARMA model for low acuity arrivals. 78
4.18 Rolling horizon forecasts for high acuity arrivals from months January to June, 2013. 82
4.19 Rolling horizon forecasts for high acuity arrivals from months July to December, 2013. 83
4.20 Rolling horizon forecasts for low acuity arrivals from months January to June, 2013. 84
4.21 Rolling horizon forecasts for low acuity arrivals from months July to December, 2013. 85
4.22 Density plot of residuals for the high acuity rolling horizon forecast 86
4.23 Density plot of residuals for the low acuity rolling horizon forecast. 87

5.1 Boxplots of length of stay, waiting time and service time, respectively, from left to right, with high and low acuity patients merged into a single class. 110

5.2 Boxplot of high and low acuity patients separately for length of stay, waiting time and service time, respectively, from left to right. 110

5.3 Fitted distributions for high acuity patient discharge where solid bars represents empirical distribution and the line bars with dot represents theoretical distribution. 112

5.4 Fitted distributions for low acuity patient discharge where solid bars represents empirical distribution and the line bars with dot represents theoretical distribution. 112

6.1 The waiting time distribution function for class-1, with arrival rates $\lambda_1 = 0.9$ and $\lambda_2 = 0.8$ and accumulation rates $b_1 = 1$ and $b_2 = 0, 0.5, 1$; simulation (solid line) and Gaver-Stehfest 8 point evaluation (dashed line). 133

6.2 The waiting time distribution function for class-2, with arrival rates $\lambda_1 = 0.9$ and $\lambda_2 = 0.8$ and accumulation rates $b_1 = 1$ and $b_2 = 0, 0.5, 1$; simulation (solid line) and Gaver-Stehfest 8 point evaluation (dashed line). 134

6.3 Permissible range of values of $0 < b_2 = b < 1$ to meet the class-1 and class-2 KPI where b_1 is set to 1 and arrival rates for both classes are considered equal, for one, two, three and five server cases, respectively (left to right, top to bottom). 136

6.4 Permissible range of values of $0 < b_3 < 0.9$ to meet the class-1, class-2 and class-3 KPI where $b_1 = 1$ and $b_2 = 0.9$ and arrival rates for all classes are considered equal. 139

6.5 Permissible range of values of $0 < b_3 < 0.4$ to meet the class-1, class-2 and class-3 KPI where $b_1 = 1$ and $b_2 = 0.4$ and arrival rates for all classes are considered equal. 139

6.6 Permissible range of values of $0 < b_3 < 0.9$ to meet the class-1, class-2 and class-3 KPI where $b_1 = 1$ and $b_2 = 0.9$ and arrival rates are considered unequal (37.5%,50%,12.5%, respectively). 140

6.7 Permissible range of values of $0 < b_3 < 0.4$ to meet the class-1, class-2 and class-3 KPI where $b_1 = 1$ and $b_2 = 0.4$ and arrival rates are considered unequal (37.5%,50%,12.5%, respectively). 140

7.1 Impact of change in Accumulation Rate 154

7.2 Two-class two-server APQ with equal arrival rates (80% occupancy) 164

7.3 Two-class two-server APQ with equal arrival rates (95% occupancy) 165

7.4 Two-class two-server APQ with 80% class-1 arrivals (95% occupancy) 168

7.5 Two-class two-server APQ with 80% class-2 arrivals (95% occupancy) 168

7.6 Two-class two-server APQ with equal arrivals (80% occupancy) where delay target for class-2 is 3 times that of class-1. 169

7.7 Two-class two-server APQ with equal arrivals (95% occupancy) where delay target for class-2 is 3 times that of class-1. 169

7.8 Three-class two-server APQ with equal arrivals (90% occupancy) where b_1 and b_2 are fixed at 1 and 1/2, respectively, while b_3 varies between 0 and 1/2, and the delay target for class-3 is 2 times that of class-2 and 4 times that of class-1. . 170

7.9 Three-class two-server APQ with equal arrivals (90% occupancy) where b_1 and b_3 are fixed at 1 and 1/4, respectively, while b_2 varies between 1/4 and 1, and the delay target for class-3 is 2 times that of class-2 and 4 times that of class-1. . 170

List of Tables

3.1 CTAS Key Performance Indicators 24

3.2 Definition of the variables on which individual time stamps were obtained. . . . 25

3.3 Definition of response variables for operational performance measures used in chapter 5. 35

3.4 Definitions of independent variables used in chapter 5. 36

3.5 Definitions of control variables used in chapter 4 and 5. 36

4.1 CTAS Key Performance Indicators 39

4.2 Summary table for the literature reviewed in this chapter (part 1). 43

4.3 Summary table for the literature reviewed in this chapter (part 2). 44

4.4 Descriptive statistics of the block, day, and month arrivals by CTAS categories. 52

4.5 Descriptive statistics of block, day, and month arrivals by Low (CTAS 3,4,5) and high (CTAS 1,2) categories. 53

4.6 Univariate analysis of the calendar (predictor) variables for the high and low acuity arrivals. 56

4.7 Test for stationarity time-series. 57

4.8 Performance of different seasonal ARIMA models for high and low acuity arrivals. 62

4.9 Parameter estimates of seasonal $ARIMA_{(2,0,1)(4,1,1)_8}$ model for high acuity arrivals. 62

4.10 Parameter estimates of seasonal $ARIMA_{(1,0,2)(4,1,3)_8}$ model for low arrivals. . . . 62

4.11 Performance comparison of GLM models. 65

4.12 Parameter estimates of Poisson and Negative Binomial model for high acuity arrivals. 66

4.13 Parameter estimates of Poisson and Negative Binomial model for low acuity arrivals. 67

4.14 Performance comparison of the GLM and ARIMA (including harmonic terms). 72

4.15 Parameter estimates of Poisson model with harmonic terms as predictors for high and low arrivals. 72

4.16 Parameter estimates of $ARIMA_{(1,0,2)}$ model with Fourier terms as predictors for high arrivals. 73

4.17 Parameter estimates of $ARIMA_{(1,0,2)}$ model with fourier terms as predictors for low arrivals. 73

4.18 Forecast accuracy measures for best fitted high and low acuity GLARMA models. 74

4.19 Parameter estimates of the full GLARMA model for high and low acuity arrivals. 75

4.20 Parameter estimates of the reduced GLARMA model for high and low acuity arrivals. 76

4.21 Accuracy measures of 12 months rolling horizon forecasts for high acuity arrivals. 80

4.22 Accuracy measures of 12 months rolling horizon forecasts for low acuity arrivals. 81

4.23 Point and interval forecast for the first 24 blocks of January based on past 12 months of fitted data. 89

4.24 Point and interval forecast for low acuity arrivals for the first 24 blocks of January based on past 12 months of fitted data. 90

5.1 Summary of the hypotheses and proposed models. 104

5.2 Definitions of response variables for operational performance measure. 105

5.3 Definitions of explanatory variables for operational performance measure. . . . 105

5.4 Definitions of control variables for operational performance measure. 106

5.5 Definitions of the indicies and parameters used in our models. 106

5.6 Model selection using Akaike Information Criterion for the high and low acuity patients, respectively. 113
5.7 Negative binomial regression output for high and low acuity patient discharges based on high and low acuity censuses, after controlling for calendar variables. . 113
5.8 Predicting patients LOS based on census at departure after controlling for calendar variables using Cox-Proportional Hazard models. 116
5.9 Cox Proportional Hazard model output for patient waiting time based on ED workload, after controlling for calendar variables. 117
5.10 Cox Proportional Hazard model output for patient service time based on ED workload, after controlling for calendar variables. 119

6.1 CTAS Key Performance Indicators 124
6.2 Key Performance Indicators for Hip and Knee Replacement Surgery, Canada . . 124
6.3 Maximum ρ and optimal b_2 137
6.4 Maximum ρ and optimal b_3 141

7.1 CTAS Key Performance Indicators 146
7.2 Coefficients for the Gaver-Stehfest Algorithm 173

Chapter 1

Introduction

Emergency departments (EDs) are an integrated part of modern healthcare systems where patients arrive with a variety of medical emergencies. Patients in an ED are usually prioritized based on their level of urgency or acuity and are served according to their relative positioning on the queue. Long wait times are common in EDs and which is associated not only with deteriorating health status but also with a decrease in level of satisfaction. Along with system inefficiencies, a mismatch between demand and supply for ED services is thought of as the primary cause of delayed care. Despite initiation of ’pay for performance’ from government funding and diligent efforts to increase efficiencies, emergency department waiting times and overcrowding continues to be an ongoing problem.

Emergency departments have to contend with patient arrivals with a diverse range of acuities, occurring at random points in time. They are also subject to time-varying patterns in demand that can be predicted to some degree, but which can also be subject to random surges. The service quality, such as, the wait times and the quality of care, are of growing concern not only to the patients but also to the stakeholders. There always remains political pressure to minimize ED wait times without sacrificing the quality of care. In this thesis I addressed several dimensions of the ED patient care process which will enable ED administrators to make informed decisions about efficient allocation of resources so as to minimize the wait times and to maximize the quality of care.

Various efforts to minimize wait times without distorting the quality of care have been emerged. For instance, empirical as well as analytical models have been used to investigate ED inefficiencies and wait times. I pursued both of these avenues in the course of this thesis. The objectives of this thesis are: (i) to accurately forecast future ED arrivals to meet patient demand, (ii) to empirically investigate the effect of ED workload on servers' productivity, (iii) to develop a queue analytic model where patients' priority depends not only on their level of acuity but also how long they have waited for the service, and (iv) to develop an optimization model to minimize total expected excess waiting time once a waiting time target is missed. In terms of novel contributions, the organization of this thesis is described below.

In chapter 4, I developed an appropriate forecasting model to accurately predict short-term ED arrivals. The demand for ED services are highly variable over the course of the day. Therefore, a uniform distribution of ED staffing levels throughout the day to serve patient demand leads to a mismatch between capacity and demand. In order to minimize wait times, resources should be distributed such that the gap between the demand and supply of ED services is minimized. I propose a combination of time series and regression models, namely, Generalized Linear Autoregressive Moving Average (GLARMA) models to predict future ED arrivals. I validate the appropriateness of the forecasting model using the "rolling horizon" approach borrowed from the operations management literature. An accurate short-term (3, 6, 9, 12, and 24 hours in advance) prediction of the ED patient arrivals will inform ED managers on how best to deploy service resources to meet anticipated demand.

In chapter 5, I looked at service processes. On the service side as well, there exists a substantial degree of variability in patient treatment time, with a variety of causes such as variability in patient conditions, server experience, dependence on the resources available, and other similar factors, based on how many patients are waiting (in the queue) for service. This phenomenon is known as queue dependent service; that is, the service process depends upon the length of the queue. Specifically, I investigated how the system workload, patient census, and congestion, affect certain ED service measures.

I then turn my attention to the analytical models for wait times. In chapter 6, I derived the waiting time distributions of different classes of patients under multi-class multi-server setting where patient priority accrues as a linear function of their wait times. Naming the model as Accumulating Priority Queue (APQ), Stanford *et al.* (2014) derived the waiting time distributions of different priority classes for the single server queue. Together with my co-authors, I have extended Stanford *et al.* (2014)'s model in the case of multi-server queues with a common exponential service time distribution. Under classical priority queue a lower priority patient continues to wait until all high priority services are completed, so that lower priority patients' waiting times are subject to considerable variability, but under APQ, the patient with highest accumulated priority goes into the service. The multi-server APQ model not only brings fairness to the priority based service selection system but also keeps waiting times for all classes of patients within a manageable threshold.

In chapter 7, I developed an optimization model to minimize total expected excess waiting time. There exist Key Performance Indicator (KPI) based service systems where a time standard is assigned to a certain class of patients along with a specified compliance probability of meeting that standard. From a KPI compliance viewpoint customers whose incurred waiting times already exceed their waiting time target are of greater concern than customers whose targets may yet be met. This may not be the view of a physician primarily concerned with clinical outcomes. I formulated an optimization model so as to minimize the frequency and amount of excess waiting, subject to the constraint that each class of patients meet their time standard with their specified compliance probability.

The foregoing novel contributions are complemented by chapter 2, which discussed relevant statistical models that I am going to use in this thesis. A detailed description of the process, data, and variables related to our study ED is outlined in chapter 3. Concluding remarks and future research directions are presented in chapter 8.

Initially my PhD studies involved the development of theoretical queueing models (chapters

6 and 7). Among them, chapter-6 has appeared in the journal of “Operations Research for Health Care” and an extension of chapter-7 is under revision. My empirical works, chapters 4 and 5, are in preparation for submission.

Chapter 2

Review of Statistical Methods

This chapter contains a brief description of the statistical methods that I have used throughout my thesis.

2.1 Linear Models

The simple linear model was developed to predict one continuous random variable (response variable, y) for a particular value of an explanatory variable (predictor variable, x). In the simple linear model, sample values of the response variable and the predictor variable are used to fit a best straight line equation such that the squared difference between the sampled observations and the straight line is minimized. The least square method is used to estimate the intercept and slope parameters of the fitted line and this minimizes the differences between the actual values and the fitted straight line. Mathematically, a simple linear model expresses the response variable as a linear function of the predictor variable plus an error term,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \ ; \ i = 1, 2, \ldots, n$$

where, β_0 and β_1 are the intercept and slope parameters. The usual assumption of a linear model is that the errors (ϵ_i's) are independent and identically distributed with zero mean ($E[\epsilon_i] = 0$)

and constant variance ($V[\epsilon_i] = \sigma^2$). An additional assumption of normality of the error terms is necessary to conduct parametric tests of the model parameters ($\epsilon_i \sim N(0, \sigma^2)$).

Often predicting a response variable requires more than one predictor variable and one can extend the simple linear model to multiple (general) linear models as follows,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \epsilon_i \ ; \ i = 1, 2, \ldots, n,$$

where, the responses $y_i, i = 1, \ldots, n$, are modeled by a linear function of a set of related predictor variables, $x_j, j = 1, \ldots, p$, plus an error term. General linear models constitute a very useful framework, however, their applicability is not appropriate in some situations: (i) if the response variable is not continuous (e.g., binary, count), and/or (ii) if the variance of the response variable depends on its mean.

2.2 Generalized Linear Models

Generalized linear models (GLM) are extensions of linear models, where the former can model both the normally and non-normally distributed random variables (responses) as a function of covariates (predictors, regressors), but the latter is only applicable to the normally distributed responses where the mean response is expressed as a linear function of the covariates. Generalized linear models overcome the aforementioned shortcomings where the response variable can take any values (e.g., continuous, binary, count) and the predictors are connected to the response via a function called the 'link function'. In GLM, the mean of the response variable, $E(y_i) = \mu_i$, is related to the linear predictor, $\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$, through a link function,

$$g(\mu_i) = \eta_i$$

In addition, the relationship between the mean and the variance of the response variable is defined by a variance function,

$$Var(y_i) = \phi Var(\mu_i),$$

where, ϕ is a constant known as a dispersion parameter. We can easily verify that the general linear model is a special case of GLM, where the link function $g(\mu_i) = \mu_i$, and the variance function $V(\mu_i) = 1$.

The link function provides a transformation to the mean response so that the transformed mean response is linearly related to the predictors. For example, in Poisson regression the predictors could be continuous variables that can take on values over the entire real line. However, the response is a count (the number of arrivals in an emergency department per hour, for example). Therefore, the response is constrained to take on whole numbers only with a possibility of zero counts. The link function in Poisson regression is the log function ($log(\mu)$). It can be seen that the log function transforms a variable constrained to natural numbers only to a variable that can take values over the entire real line. In this case, the link function makes the response compatible with the predictor variables and hence it is possible to make it a linear function of the predictors plus a random component. If λ is the average rate of arrival in an ED and $x_1, x_2, \ldots, x_k$ are predictors of ED arrivals, then a Poisson regression model can be written as,

$$log(\lambda) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \ ; \ k = 1, 2, \ldots, K$$

Generally, a response variable whose distribution belongs to the exponential family of distributions can be modeled under the GLM framework. (For instance, Normal, Binomial, Poisson, Negative Binomial, and Gamma distributions all belong to the exponential family of distributions.) An excellent introduction of GLM is documented in Dobson & Barnett (2011). A detailed description of the processes for parameter estimation and hypothesis testing are described in McCullagh & Nelder (1989). The implementation of GLMs in statistical software package R and their relevant interpretations are provided in Faraway (2005).

2.3 Seasonal ARIMA Models

Time series analysis is a statistical procedure that uses chronologically ordered past observations of a variable to predict its future values. Historical patterns (trend, cycle, seasonal variation, and irregular fluctuations) observed in a time series can be exploited to produce accurate forecasts. A nice introduction to the time series analysis and forecasting can be found in Brockwell & Davis (2002). One of the most popular and widely used time series models is the autoregressive integrated moving average (ARIMA) model. Box & Jenkins (1970) introduced the ARIMA model for non-stationary time series processes. A time series is called "stationary" if its statistical properties, such as the mean, variance, and autocorrelations, are constant in time; otherwise the time series is non-stationary. As a basis for visual inspection of stationarity, one might look at time series plots to see whether the mean or the variance of the time series changes over time. However, there are formal unit root tests (Ljung-Box test, Augmented Dickey-Fuller (ADF) test, Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test) to determine whether a time series is stationary or differencing is required to transform the time series into a stationary series (Makridakis *et al.* (2008)).

The ARIMA model combines an autoregressive (AR) process and a moving average (MA) process after "integrating" (I) the time series (i.e., differencing, logging, and/or deflating) to transform it into a stationary process. The integration applied to the original time series (if any) must be reversed to obtain a forecast for the original series. This process is automated in standard software used for time series analyses. If the time series under study appears to be stationary, then no integration is necessary and the model deduces to the ARMA model. Furthermore, if either the autoregressive or the moving average terms turn out to be insignificant, the model deduces to a moving average or an autoregressive model, respectively. Therefore, ARIMA is a general time series modeling framework where AR, MA, and ARMA models are special cases of the ARIMA model.

The aforementioned ARIMA model is non-seasonal. An extension of this is a seasonal

ARIMA model where another set of autoregressive, integration, and moving average parameters are incorporated to deal with the seasonality that may be present in a time series. Usually, a non-seasonal ARIMA model is symbolized as ARIMA(*p*,*d*,*q*) where "*p*" indicates the number of AR terms, "*d*" indicates the order of differencing, and "*q*" indicates the number of MA terms. A seasonal ARIMA model is symbolized as ARIMA(*p,d,q*)*(*P,D,Q*)$_s$, where the *p, d, q* indicates the model orders for the short-term components of the model (as described above) and *P, D, Q* indicate the model orders for the seasonal components of the model, where "*s*" stands for the seasonal cycle.

I will sequentially describe the building up process of a non-seasonal ARIMA model followed by its extension to the seasonal ARIMA model. Let y_t be a univariate time series observed at time points, $t = 1, 2, \ldots, T$.

In an AR process, the current observation y_t is modeled as a weighted sum of its past observations plus an error term. Therefore, an autoregressive model of order *p*, AR(*p*), is defined as,

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t \tag{2.1}$$

where, $\varepsilon_t \sim iidN(0, \sigma_t^2)$, and $\phi_1, \phi_2, \ldots, \phi_p$ are fixed coefficients. The above equation (2.1) is equivalent to a multiple linear regression of *y* on the preceding *p* values of *y*.

In an MA process, y_t is modeled as a weighted sum of past error terms. Therefore, a moving average model of order *q*, MA(*q*), is defined as,

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} \tag{2.2}$$

where again, $\varepsilon_t \sim iidN(0, \sigma_t^2)$, and $\theta_1, \theta_2, \ldots, \theta_q$ are fixed coefficients. Note that, in equation (2.2), y_t values are regressed on *q* previous error terms. Since error values are not known, one cannot estimate parameters using standard regression formulas. A few advanced methods, such as Durbin's method, the inverse covariance method, and the variance recursion method, implemented in statistical software (i.e., R), can be used to obtain MA parameter estimates.

By combining the two preceding models (2.1 and 2.2), a general non-seasonal autoregressive moving average model of order (*p, q*), ARMA*(p,q)* can be written as

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} \tag{2.3}$$

where, $\varepsilon_t \sim iidN(0, \sigma^2)$.

It is often convenient to use a backshift operator $B^h y_t = y_{t-h}$ to express the above equation concisely, i.e.,

$$\begin{aligned} & y_t = \phi_1 B y_t + \phi_2 B^2 y_t + \ldots + \phi_p B^p y_t + \varepsilon_t + \theta_1 B \varepsilon_t + \theta_2 B^2 \varepsilon_t + \ldots + \theta_q B^q \varepsilon_t \\ \Rightarrow \quad & (1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) y_t = (1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q) \varepsilon_t. \end{aligned}$$

Denoting $\phi_p(B) = (1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)$, and, $\theta_q(B) = (1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q)$, the above equation can be further simplified as

$$\phi_p(B) y_t = \theta_q(B) \varepsilon_t. \tag{2.4}$$

Now, if it appears that the time series under study is not stationary, but rather increasing approximately linearly with time, the first difference ($I(1) = y_t - y_{t-1} = (1-B)y_t$) may transform the series to stationary and one replace y_t by $(1-B)y_t$ in the above equation (2.3). However, if y_t has to be differenced d times to reach stationarity, one needs to replaces y_t by $\Delta^d y_t = y_t - y_{t-d} = (1-B)^d y_t$ and hence an $ARIMA(p, d, q)$ process can be written as

$$\phi_p(B)(1-B)^d y_t = \theta_q(B) \varepsilon_t \tag{2.5}$$

So far I have described the non-seasonal ARIMA model. However, extending the non-seasonal ARIMA to the seasonal ARIMA is trivial. One merely needs to incorporate seasonal AR, seasonal differencing, and seasonal MA terms in the model.

Seasonality may be defined as the periodic pattern of fluctuations in the time series values. Monthly average temperature in a particular area, for example, will exhibit a markedly periodic behavior every 12 months. As differencing between successive lags may alleviate non-stationarity due to possible trends in a time series, seasonal differences at successive periodic lags may reduce non-stationarity caused by seasonality in a time series. That is, the non-stationarity caused by the season change can be removed by taking seasonal differences. For the case of monthly average temperature, the difference of the values of the same months from consecutive years removes seasonality. If D is the degree of seasonal differencing used to reach stationarity (D can take only integer values 0, 1, 2, ...), then the stationary time series becomes $(1 - B^s)^D y_t$ and here s is the seasonal period.

Incorporating the P seasonal autoregressive terms and the Q seasonal moving average terms, the seasonal part of the process, which is identified by seasonal lags, can be written as,

$$\Phi_P(B^s)(1 - B^s)^D y_t = \Theta_Q(B^s)\varepsilon_t \tag{2.6}$$

where, $\Phi_P(B^s) = 1 - \Phi_{1,s}B^s - \Phi_{2,s}B^{2s} - \ldots - \Phi_{P,s}B^{Ps}$ is called the seasonal autoregressive operator of order P;, and $\Theta_Q(B^s) = 1 - \Theta_{1,s}B^s - \Theta_{2,s}B^{2s} - \ldots - \Theta_{Q,s}B^{Qs}$ is called the seasonal moving average operator of order Q.

By incorporating the non-seasonal and seasonal parts of the ARIMA model, one obtain a seasonal ARIMA model, denoted by $ARIMA_{(p,d,q),(P,D,Q)_s}$ as,

$$\phi_p(B)\Phi_P(B^s)(1 - B)^d(1 - B^s)^D y_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t \tag{2.7}$$

For example, if one fixes the parameters at, $p = 1, d = 0, q = 1, P = 0, D = 1, Q = 0, s = 12$, then the above model can be written as,

$$\begin{aligned} \phi_1(B)(1-B^{12})y_t &= \theta_1(B)\varepsilon_t \\ y_t - y_{t-12} &= \phi_1(y_{t-1} - y_{t-13}) + \varepsilon_t + \theta_1\varepsilon_{t-1} \end{aligned} \tag{2.8}$$

which is an $ARIMA_{(1,0,1)(0,1,0)_{12}}$ model where the first term on the left hand side in equation (2.7) is for the non-seasonal AR(1) process, and the second term is for the seasonal difference. The right hand side of equation (2.7) is for the non-seasonal MA(1) process.

A detailed description of the parameter estimation processes along with their hypothesis testing methods for ARIMA models can be found in Brockwell & Davis (2009). Implemetations of the ARIMA models in the statistical software package R can be found in Shumway & Stoffer (2010).

2.3.1 Non-stationarity and Unit Root Tests

In addition to graphical inspection of the original time series, there are few statistical tests, called unit root test, that are used to determine non-stationarity in the observed time series. A detailed description of such tests can be found in Greene (2011).

2.3.1.1 Augmented Dickey-Fuller (ADF) Test

An ADF Test is an extension of Dickey-Fuller test which accommodates serial correlation. The test is constructed based on the following autoregressive model,

$$y_t = \mu + \beta t + \gamma y_{t-1} + \gamma_1 \Delta y_{t-1} + \ldots + \gamma_p \Delta y_{t-p} + \epsilon_t$$

The null hypothesis to be tested here is $H_0 : \gamma = 1$, and the corresponding test statistics is a conventional t ratio (t-test like statistic),

$$DF_\gamma = \frac{\hat{\gamma} - 1}{\text{Estimated Std. Error}(\hat{\gamma})}$$

2.3.1.2 Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test

An alternative to the ADF Test is the KPSS test which is developed to test for non-stationarity against the null hypothesis that the series is stationary, i.e., $H_0 : \gamma = 0$. The KPSS test is formulated based on the following model:

$$y_t = \alpha + \beta t + \gamma \sum_{i=1}^{t} Z_i + \epsilon_t \; ; \; t = 1, 2, \ldots, T$$

where the vector Z_t is an i.i.d. stationary series with mean zero and variance one.

Estimates for α and β parameters can be obtained by ordinary least square estimates under the null hypothesis. Thus the residuals can be expressed as

$$e_t = \hat{\alpha} + \hat{\beta} t \; ; \; t = 1, 2, \ldots, T$$

Letting the sequence of partial sums $E_t = \sum_{i=1}^{t} e_i$, the KPSS statistic can be defined as,

$$\text{KPSS} = \frac{\sum_{i=1}^{t} e_i}{T^2 \hat{\sigma^2}}$$

where,

$$\hat{\sigma^2} = \frac{\sum_{t=1}^{T} e_t^2}{T} + 2 \sum_{j=1}^{L} (1 - \frac{j}{L+1}) r_j, \text{ and } r_j = \frac{\sum_{s=j+1}^{T} e_s e_{s-j}}{T}$$

Critical values for the KPSS statistic are estimated by simulation.

2.3.2 Determination of the Orders AR(p),MA(q),SAR(P),SMA(Q)

The plot of autocorrelation functions (ACF) and the plot of partial autocorrelation functions (PACF) aid us in determining the possible values of *p, q, P,* and *Q*. The autocorrelation function (ACF) plot shows the correlation of the series with itself at different lags. Therefore, the autocorrelation of y_t at lag k is the correlation between y_t and y_{t-k},

$$\begin{aligned}\rho_k &= \frac{\text{Autocovariance of } y_t \text{ at lag } k}{\text{Variance of } y_t} \\ &= \frac{Cov(y_t, y_{t-k})}{Var(y_t)} \\ &= \frac{E[(y_t-\mu)(y_{t-k}-\mu)]}{E[(y_t-\mu)]^2}.\end{aligned}$$

One can estimate the population autocorrelation function by the sample autocorrelation function at lag k, defined as, $r_k = \widehat{\gamma}(k)/\widehat{\gamma}(0)$, where

$$\widehat{\gamma}(k) = \frac{1}{n}\sum_{t=1}^{n-k}(y_t-\bar{y_n})(y_{t-k}-\bar{y_n}) \text{ and } \widehat{\gamma}(0) = \widehat{\sigma^2} = \sum_{t=1}^{n}(y_t-\bar{y_n})^2$$

are the sample auto-covariance function at lag k, and sample variance estimator, respectively. The highest autocorrelation of the stationary time series determines the moving average term, q to be considered in the model.

A partial autocorrelation is a conditional correlation where the correlation between two variables is calculated under the assumption that the known values of another set of variables are already accounted for. Suppose, one has three observations from a time series, Y_t, Y_{t-1} and Y_{t-2}. Then the partial autocorrelation between Y_t and Y_{t-2} is determined after taking into account how both Y_t and Y_{t-2} are related to Y_{t-1}, i.e., the correlation between Y_t and Y_{t-2} that is not predicted by Y_{t-1}. Mathematically, it can be written as,

$$\frac{\text{Covariance}(Y_t, Y_{t-2}|Y_{t-1})}{\sqrt{\text{Variance}(Y_t|Y_{t-1})\text{Variance}(Y_{t-2}|Y_{t-1})}}$$

In general, the partial autocorrelation between Y_t and Y_{t-k} is the conditional correlation between Y_t and Y_{t-k} where the condition is on the set of observations that comes between the time points t and $t-k$. This correlation can be obtained by correlating the residuals from two different regressions: (i) regression of Y_t on $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-k+1}$, and (ii) regression of Y_{t-k} on $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-k+1}$. Therefore, the partial autocorrelation is the autocorrelation which remains at lag k after the effects of shorter lags $(1, 2, \ldots, k-1)$ have been removed by regression. Notationally, $PACF(k) = Corr(y_t, y_{t-k} | y_{t-1}, \ldots, y_{t-k+1})$. If $k = 1$, then $PACF(1) = ACF(1)$. The highest partial autocorrelation of the stationary time series determines the autoregressive term, p to be considered in the model.

In order to obtain the orders for the Seasonal Autoregressive and the Seasonal Moving Average models, one has to inspect ACF and PACF plots after seasonal differences (if any) at multiple times of the seasonal period (s). If there are positive spikes in ACF at lag $s, 2s, 3s, \ldots,$ and a single positive spike in PACF at lag s, then SAR=1. Alternatively, if there is a negative spike in ACF at lag s, and negative spikes in PACF at lags $s, 2s, 3s, \ldots,$ then SMA=1.

2.4 Harmonic Regression

Harmonic analysis explains the variation in a time-series based on frequency rather than time. A time-series can be transformed into the frequency domain from the time domain using Fourier transformation. For a long memory time-series where periodic phenomena are existent, signals are more concentrated on the frequency domain than on time domain. This type of analysis is known as harmonic analysis, spectral analysis, or Fourier analysis.

In harmonic analysis, a stationary time series with infinite duration and evenly spaced observation is expressed as a function of sinusoidal periodic signals and a Gaussian (white) noise (see Babu (2012)). For a time-series Y_t, a common model formulation for a finite Fourier series is,

$$Y_t = \mu + a\ cos(2\pi\omega t) + b\ sin(2\pi\omega t) + e_t$$

where μ is the mean of the series, $\omega = 1/T$ is the fraction of the complete cycle completed in a single time period, period T is the number of time periods required to complete a single cycle of the cosine function, and e_t is the random error term.

Under harmonic regression a time series can be expressed as a combination of cosine (sine) waves with differing periods and a number of covariates (as in GLM). Harmonic regression can also be used under an ARIMA modeling framework, where the random error term e_t has a ARMA like structure. A detailed description of the harmonic regression can be found in Chatfield (2013).

2.4.1 Determination of the Wavelengths Using a Periodogram

An important part is to determine the appropriate set of frequencies to fit in the harmonic regression model. A periodogram is used to detect the dominant periods of a time series. One can determine those frequencies (wavelengths) by inspecting the periodogram. Let us partition the variability of the time-series into different components at frequencies $\frac{2\pi}{T}, \frac{4\pi}{T}, \ldots, \pi$. The component at frequency, $\omega_k = \frac{2\pi k}{T}$, is named as the kth harmonic. An equivalent form of the kth harmonic for $k \neq \frac{T}{2}$ can be written as,

$$a_k\ cos(\omega_k t) + b_k\ sin(\omega_k t) = R_k\ cos(\omega_k t + \phi_k)$$

where, $R_k = \sqrt{a_k + b_k}$, and $\phi_k = \tan^{-1}\left(\frac{-b_k}{a_k}\right)$.

The periodogram of a time-series data is the plot of $\frac{TR_K^2}{4\pi}$ against ω. The periodogram graphs a measure of the relative importance of possible frequency values that might explain the oscillation pattern of the observed data.

2.5 Generalized Linear Autoregressive Moving Average Models

Generalized linear autoregressive moving average (GLARMA) models have been developed to model discrete valued time series responses. GLARMA models overcome the shortcomings of both GLMs, which are capable of modeling non-normal data but cannot easily incorporate time series structures (unless autocorrelation is introduced to the model through a latent process), and time series models, which are suitable for time series data but not designed to handle non-normal responses. GLARMA models are a subclass of generalized state space models, (Davis *et al.* (1999)) which have two formulations: parameter driven and observation driven.

I will focus our attention on the observation driven GLARMA models for modeling time-dependent count data for two important reasons: (i) observation driven GLARMA models are easy to implement on long time series data with numerous predictors, (ii) they are readily available in statistical software packages (i.e., "glarma" R-package). Parameter driven GLARMA models, on the other hand, require high dimensional integrals to be evaluated and are not easy to interpret.

An excellent introduction of the "glarma" R-package for observation driven time series regression of counts is documented in Dunsmuir & Scott (2015). Let us denote, Y_t as the random variable for our time series of counts at time t, where $t = 1, 2, \ldots, T$. Associated with those count responses, there is a K-dimensional vector of covariates x_t. Letting, $\mathcal{F}_t = \{Y_s : s < t, x_s : s \leq t\}$ be the past information on the response series and past and present information on the covariates, the conditional distribution $Y_t \mid \mathcal{F}_t \sim Poisson(\lambda_t)$.

Defining $e_t = \frac{y_t - \lambda_t}{\sqrt{\lambda_t}}$, I have,

$$E(e_t|\mathcal{F}_{t-1}) = \frac{E(y_t|\mathcal{F}_{t-1}) - \lambda_t}{\sqrt{\lambda_t}} = 0$$

Therefore, by definition $\{e_t\}$ is a martingale difference sequence relative to $\{y_t\}$ (Brockwell & Davis (2009)). Now let $\{z_t, t \geq 0\}$ be an ARMA(p, q)-process as defined in (2.3) with noise given by the martingale difference sequence $\{e_t\}$,

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_p z_{t-p} + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_q e_{t-q} \tag{2.9}$$

One property of an ARMA(p, q)-process is that $\{z_t\}$ are causal and there exists a sequence

$\{\gamma_j, j = 1, 2, \ldots\}$ with $\sum_{j=0}^{\infty} |\gamma_j| < \infty$ such that

$$z_t = \sum_{j=0}^{\infty} \gamma_j e_{t-j} = e_t + \sum_{j=1}^{\infty} \gamma_j e_{t-j} = e_t + z_t^*. \tag{2.10}$$

the term z_t^*, can be regarded as a predictor for z_t given the past $\mathcal{F}_{t-1}$ since $E(e_t) = 0$. Therefore, it is reasonable to define the general Poisson-GLARMA(*p, q*) model as

$$log(\lambda_t) = x_t^T \beta + z_t^* = x_t^T \beta + \sum_{j=1}^{\infty} \gamma_j e_{t-j}. \tag{2.11}$$

The conditional distribution of y_t is specified via the transformed function of past observations $y_{t-1}, y_{t-2}, \ldots$ as $e_{t-j} = \frac{y_{t-j} - \lambda_t}{\sqrt{\lambda_t}}$. It is reasonable to set $e_t = 0$ and $z_t^* = 0$ for $t \leq 0$ since there are no observations for $t \leq 0$. Therefore, z_t^* as a predictor for $z_t | \mathcal{F}_{t-1}$ can be computed recursively using the ARMA recursion,

$$\begin{aligned} z_t^* &= \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_p z_{t-p} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_q e_{t-q} \\ &= \phi_1 (z_{t-1}^* + e_{t-1}) + \ldots + \phi_p (z_{t-p}^* + e_{t-p}) + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_q e_{t-q} \end{aligned}$$

Detailed descriptions of the maximum likelihood method for estimating parameters and their respective standard errors are described in Davis *et al.* (2003).

2.5.1 Probability Integral Transforms (PIT)

PIT provides a basis for testing whether a set of observations can reasonably be modeled as arising from a specified distribution. Specifically, PIT relates the results that the data values as being a random variable modelled from any given continuous distribution can be converted to random variables having a uniform distribution (Dodge (2006)). A number of authors have suggested the use of PIT to examine the validity of the assumed distribution in the GLARMA model (for example, Dunsmuir & Scott (2015)). In addition to residual plots and Q-Q plots, I also make use of PIT to check the validity of our proposed model. Czado *et al.* (2009) derived a PIT form which applies to discrete distributions based on the following formulae:

$$F^{(t)}(u|y_t) = \begin{cases} 0, & \text{if, } u \leq F(y_t-1), \\ \frac{u-F(y_t-1)}{F(y_t)-F(y_t-1)}, & \text{if, } F(y_t-1) \leq u \leq F(y_t), \\ 1, & \text{if, } u > F(y_t). \end{cases}$$

In the foregoing expression $F(y_t)$ and $F(y_t - 1)$ are the upper and lower conditional predictive probabilities, respectively. PIT is thus defined as

$$\bar{F}(u) = \frac{1}{T-1} \sum_{t=2}^{T} F^{(t)}(u|y_t)$$

A histogram of the PIT is drawn with a horizontal line at height 1, representing the density function of the uniform distribution at [0,1].

2.6 Forecast Accuracy Measures

Three different forecast accuracy measures are often used: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Those can be calculated as,

$$\begin{aligned} RMSE &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}, && (2.12) \\ MAE &= \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|, && (2.13) \\ \text{and, } MAPE &= (1/n)\sum_{i=1}^{n}\left(\frac{|y_i - \widehat{y_i}|}{|y_i|} \times 100\right). && (2.14) \end{aligned}$$

Additionally for model selection purposes, I calculated the second order information criterion (AIC_c) which is the Akaike information criterion (AIC) with a correction for finite sample sizes which pose greater penalty for extra parameters.

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1}. \tag{2.15}$$

In equation (2.15), n is the sample size, k is the number of parameters, and AIC is defined as the log-likelihood term penalized by the number of parameters, i.e., $AIC = -2LL + 2k$. Increasing number of parameter leads to a larger AIC value whereas the large value of the likelihood function makes the AIC value smaller. The AIC_c makes an additional correction for the number of parameters.

2.7 Cox-proportional Hazard Models

A Cox-Proportional Hazard (Cox-PH) model is a statistical technique developed by Cox (1972) to model the relationship between the time to an event (survival time) and several explanatory variables. One advantage of the Cox-PH model over parametric survival models is that no assumption regarding the survival time distribution is necessary. Under Cox-PH modeling framework, the log of the hazard function is linearly related to the covariates,

$$\begin{aligned} log(h(t; x_1, x_2, \ldots, x_p)) &= \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p & (2.16) \\ \implies h(t; x_1, x_2, \ldots, x_p) &= h_0(t)\ \exp(\beta_1 x_1)\ \ldots\ \exp(\beta_p x_p) & (2.17) \end{aligned}$$

where, $h_0(t) = exp(\beta_0)$, is the baseline hazard function, i.e., the value of the hazard function when all the covariates are zero. It is observed from equation (2.17) that the covariates have a multiplicative effect on the hazard function. Suppose two individuals have first covariate values x_{11} and x_{12} with all other covariate values remaining same. Then the ratio of the two hazard functions, which is the hazard ratio, can be written as

$$\frac{h(t, X_1)}{h(t, X_2)} = \frac{\exp(\beta_1 x_{11})}{\exp(\beta_1 x_{12})} = \exp\{\beta_1 (x_{11} - x_{12})\} \tag{2.18}$$

The term, $\exp\{\beta_1(x_{11} - x_{12})\}$ is the hazard ratio comparing x_{11} to x_{12}. If $\beta_1 = 0$, the hazard ratio becomes 1 ($e^0 = 1$), i.e., there is no effect of covariates on survival. A hazard ratio of 2 indicates individual 1 is twice as likely to die as individual 2 at any time instant.

From equation (2.18), I have, $h(t, X_1) \propto h(t, X_2)$, i.e., hazard functions are proportional to each other and are independent of time. This is why the Cox model is called a proportional hazard model. The parameters in a Cox-PH model are estimated by maximizing the partial likelihood function. Interested readers can find detailed estimation procedure in Kalbfleisch & Prentice (2011).

Chapter 3

ED Process Flow and Data Description

3.1 ED Process Flow

A schematic diagram of our ED process flow is presented in Figure 3.1. There are two modes of arrivals in our ED: walk-in and through ambulance or other emergency transportation. Patient arrivals by ambulance usually go through initial assessment immediately and a charge nurse decides their level of acuity. On the other hand, walk-in arrivals initially see a business clerk or charge nurse who swipes their health card and records their chief complaint. Patients then wait for a triage nurse who will come to assess their conditions.

Patients' priority levels are determined by a triage nurse after a general assessment of the chief complaint and overall condition. In Canada, the Canadian Triage and Acuity Scale (CTAS (2005)) is a guideline for more objective classification of severity where patients are classified into five priority classes (see Table 3.1). The highest priority class is assigned with a score of 1 and the score increases as the priority level decreases with a score of 5 for the lowest priority class. Each class is associated with a specified time target by which a patient must be

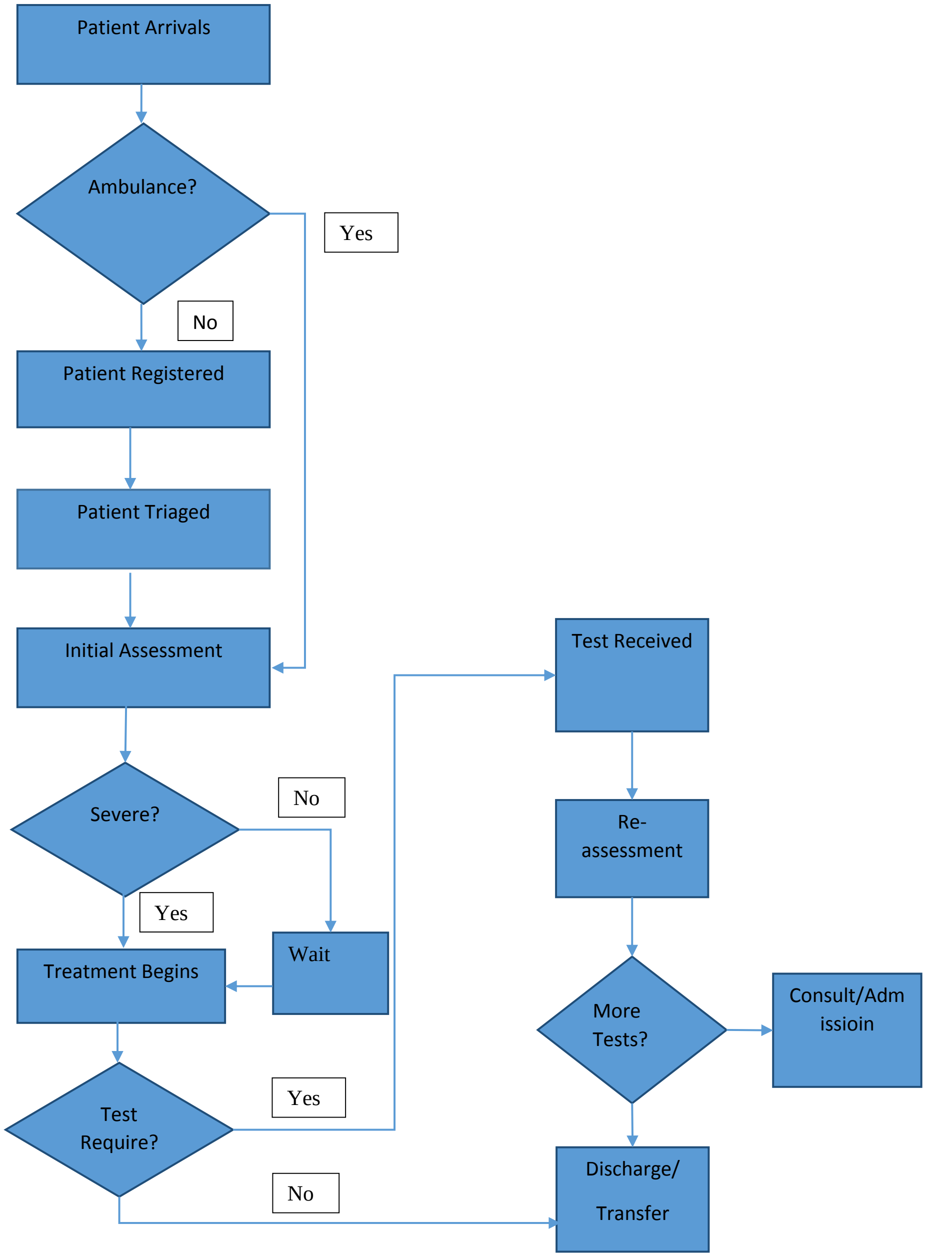


Figure 3.1: A simplified emergency department process flow.

seen in a certain percentage of the time. Once the triage process is complete, patients are then seen by a unit clerk who continues with the registration process by confirming all demographic information. Patients then wait in the waiting room for their turn to receive the first physician assessment.

Table 3.1: CTAS Key Performance Indicators

Category	Classification	Access	Performance Level
1	Resuscitation	Immediate	98%
2	Emergency	15 minute	95%
3	Urgent	30 minute	90%
4	Less urgent	60 minute	85%
5	Not urgent	120 minute	80%

According to the level of acuity, patients are then taken into a room for their initial assessment by the attending physician, a resident, or, clinical clerk (third or fourth year medical student). Residents or clinical clerks are able to initiate investigations and treatment of patients independent of their clinical supervisors, within the limit of their level of expertise and experience. Immediate treatment begins for the patients who are considered critically ill while other patients continue to wait until their turn comes. Patients from whom no further investigation is required and whose treatment plan is already assigned, are then discharged from the ED. All other patients who went through some sort of diagnostic procedures will wait for their test results to arrive. Patients are then reassessed by the attending physician who decides whether to discharge them, refer them to other facilities, or ask for further investigations. Repeated reassessments will continue unless the patient is finally discharged from the ED.

3.2 Study Setting

Two years worth of adult ED visit data was collected from an Ontario hospital. Patients who were registered and triaged between January 1, 2012 to December 31, 2013, were included in the study. Ethics approval for this study was obtained from this hospital as well. Discrete time stamps for operations that an individual patient encounters throughout the ED process flow (Figure 3.1) have been extracted from the ED database. Among other variables, age, triage levels, modes of arrival, discharge destinations, were also obtained.

Since the number of arrivals as well as service effort varies significantly from one triage level to another, I have stratified our analysis by triage level. The studied ED adheres to the Canadian Triage and Acuity Scale (CTAS) which stratify patients into five acuity level based on the severity of illness (Table 3.1). I aggregated CTAS categories 1 and 2 and named it the high acuity patient class, and CTAS categories 3, 4, and 5 and named it the low acuity patient class.

3.2.1 Missing Value Imputation

Our objective here is to achieve complete information about the arrival times of all the patients under study. In addition for later use, I need to have complete observations on physician initial assessments (PIA) and discharge times as well. The dataset consists of individual time stamps of several service activities for every single patient who flows through the ED service process. Definition and description of the short-form variable names of our dataset that were used to obtain values for arrivals, PIA, and discharges are provided in Table 3.2.

Table 3.2: Definition of the variables on which individual time stamps were obtained.

Variable Names	Definition
Received	Time at which charge nurse swipes patient's health card.
Triaged	Time at which patient was triaged.
TBS	Time at which the patient is expected to be seen.
SBMD	Time at which patient is seen by an MD.
DTD	Time at which decision about patient's discharge was made.
DC	Time at which patient was actually discharged from the ED.
Depart	Time at which chart is actually signed out.
TransWard	Time at which patient was transfered to another ward.
ConsAdm	Time at which consultant decides to admit the patient.
LWBS	Time at which patient was observed left without been seen by an MD.
LABS	Time at which patient was observed left after been seen by an MD.

Imputation is the process of substituting suitable values in places where there are missing values. Missing values are of three types: (i) missing completely at random (no pattern), (ii) missing at random (can be predicted from other values of the variable), and, (iii) missing not at random (depends upon unobserved variables). We assumed missing data occurs at random. In fact the missing data are unlikely to be random, but more likely to relate to the busy periods. However, this is not likely to lead to major problems due to the level of missing data involved in most cases.

I predicted the missing data from the observed values of the same variable after controlling for block of day, day of week, month of year, and triage category. There are several methods available for missing value imputation. I used single mean imputation to impute missing values which is the simplest among all imputation methods. Mean imputation uses the mean of the observed values as a substitute for missing values after matching for similar values of other variables.

Missing values are evident in the ED data collection process. This is due to the fact that providing services to a patient takes precedence over recoding data in an ED. A physician will start the treatment process for a critical patient before record the instant of service time initiation. In addition, while one observation is missing for one patient, it may not be so for others. Therefore, missing values can well be scattered across the dataset although one anticipates a greater preponderance of missing values during congested periods than during calm ones. During analysis, the major problem for having missing values in the dataset is that any standard statistical software disregards individual patients with any of their time stamps missing. Consequently, discarding a large portion of the data may introduce bias and reduces sample size. Since discarding missing values leads to loss of information, replacing missing observations through imputation improves the power of the study.

The algorithms used to obtain complete observations for arrival, PIA, and discharge times using observed and imputed values of the variables described in Table (3.2) are provided below,

- **Arrivals:** I have considered the patients' Received time as their arrival time to our ED.

There were 42 missing values for Received time stamp. The time stamp that follows Received time stamp is the TBS time stamp. After being received, patients are usually triaged followed by registration. When patients are being registered, they are assigned a TBS time stamp. I used TBS time stamp to impute received time stamp values for 35 patients because 35 of the 42 missing received time stamps have TBS time stamps (3.2).

Since TBS time stamps occur a certain time after the Received time stamp, in order to retrieve respective Received time stamps, I subtracted the average of the differences between the TBS and Received time stamps from the TBS time stamps for which complete observations were available. In our study ED, patient volume varied by the day of the week and by triage category. Therefore, further refinement was done by computing the averages of the differences between the TBS and Received time stamps after matching by the day of the week and triage category.

The remaining 7 missing Received time stamps had only corresponding Depart time stamps available, and I imputed those missing observations by subtracting the average of the time differences between the Depart and Received time stamps from Depart time stamps by day of the week and triage category. A schematic diagram of the algorithm described above to obtain arrival times is provided in Figure 3.2.

- **Physician First Assessment (PIA):** The Seen by MD (SBMD) time stamp was regarded as a PIA time. However, there are 10,411 missing SBMD time stamp. The closest time stamp is TBS and I approximated 9244 of the missing SBMD time from the corresponding TBS time. Since SBMD time stamps occur after TBS time stamps the majority of the time, I added the average of the differences between TBS and SBMD time stamps with TBS time to approximate PIA time. Additionally, as patient volume varied by the day of the week and by triage category, I computed average of the differences after matching by the day of the week and triage category. The similar procedure was applied to impute the remaining 1167 missing SBMD time stamps form DC (272), Depart (619), and DTD

(276), respectively.

- **Discharges:** Patients' are considered discharged from an ED when they are released home or admitted to other facilities of the hospital. Therefore, observed DC and TransWard time stamps were considered as discharge times. Among 7660 missing time stamps, 7286 and 374 were imputed from Depart and DTD times, respectively. The 724 of the remaining 768 missing time stamps have ConsAdm time stamps. I imputed discharge times by adding the average of the differences between the ConsAdm and DC times with DC times after matching by the day of the week and triage category. A similar procedure was applied to impute the remaining 44 missing discharge time stamps form LWBS (12), LABS (3), and Received (29) times, respectively.

3.2.2 Data Processing

Individual time stamps were aggregated to obtain arrival counts for hours, days, and months. Hourly data were further aggregated to 3-hourly blocks, where a day consists of 8 such blocks starting from midnight. In this study, our primary unit of analysis is the number of patient arrivals in such a 3-hour block. Two main advantages of using a 3-hour block are; (i) one can avoid majority of the zero counts which are dominant in the hourly ED arrival data, (ii) a frequency of every 3 hours coordinates better with shift changes that occur at (7 a.m. , 10 a.m., 3 p.m., 6 p.m., and 11 p.m.) than one that is less frequent (like every six hours, for instance).

Temporal variables such as hour-of day, day-of-week, month-of-year, and weekday/holiday were created using the hourly arrival data. Every 8 blocks were categorized to obtain day-of-week results whereas 240 blocks were accumulated to obtain month-of-year. Weekday/holiday is a binary variable representing 1 if the block belongs to a week day and 0 otherwise.

Before fitting forecasting models in chapter 4, I have divided the data set into two parts. The former is called training dataset which includes the first 22 months of data and will be used to fit the forecasting models. The latter is the test data set which consists of the remaining last

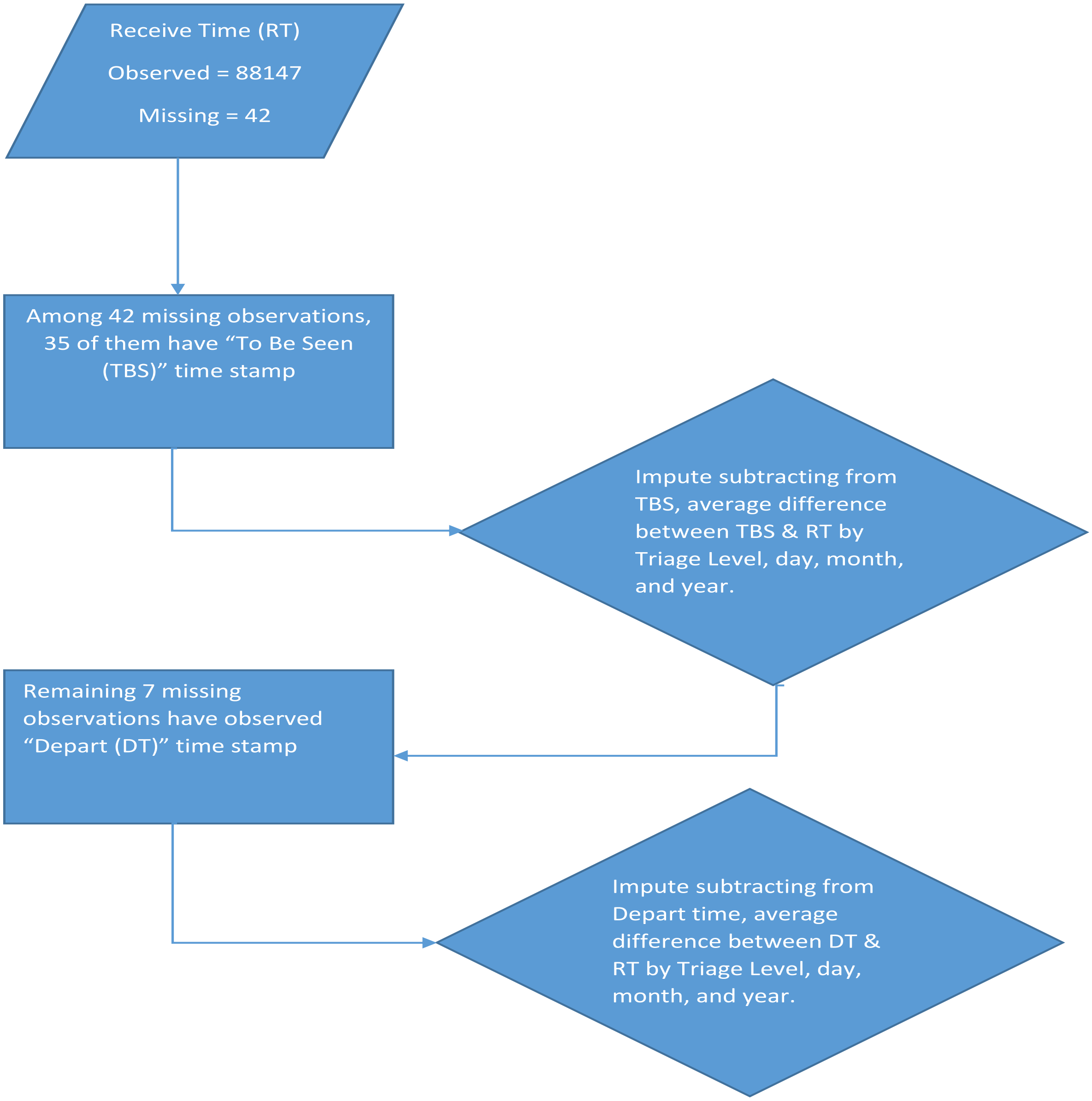

Receive Time (RT)
Observed = 88147
Missing = 42
Among 42 missing observations, 35 of them have “To Be Seen (TBS)” time stamp
Impute subtracting from TBS, average difference between TBS & RT by Triage Level, day, month, and year.
Remaining 7 missing observations have observed “Depart (DT)” time stamp
Impute subtracting from Depart time, average difference between DT & RT by Triage Level, day, month, and year.

Figure 3.2: Imputation process for missing arrival time.

2 months of data and will be used to validate the forecasting models.

3.3 Clinical Setting and Data Description

In an emergency department, along with consultation with a physician, an arriving patient frequently receives other services (diagnostic procedures) if necessary. These inter-connected activities together determine how long a patient will remain in the emergency department. A Patient's service to an emergency department is considered complete once they are discharged from the emergency department are admitted to a ward/hospital.

3.3.1 Clinical Setting

Emergency departments in Canada are publicly funded so that patients may seek treatment for any sort of emergency care. Due to the mismatch in demand and capacity, ED often experiences long wait times and operational inefficiencies. ED staff are mostly salaried employees of the hospital, and therefore their financial benefits are not affected by the number of patients seen.

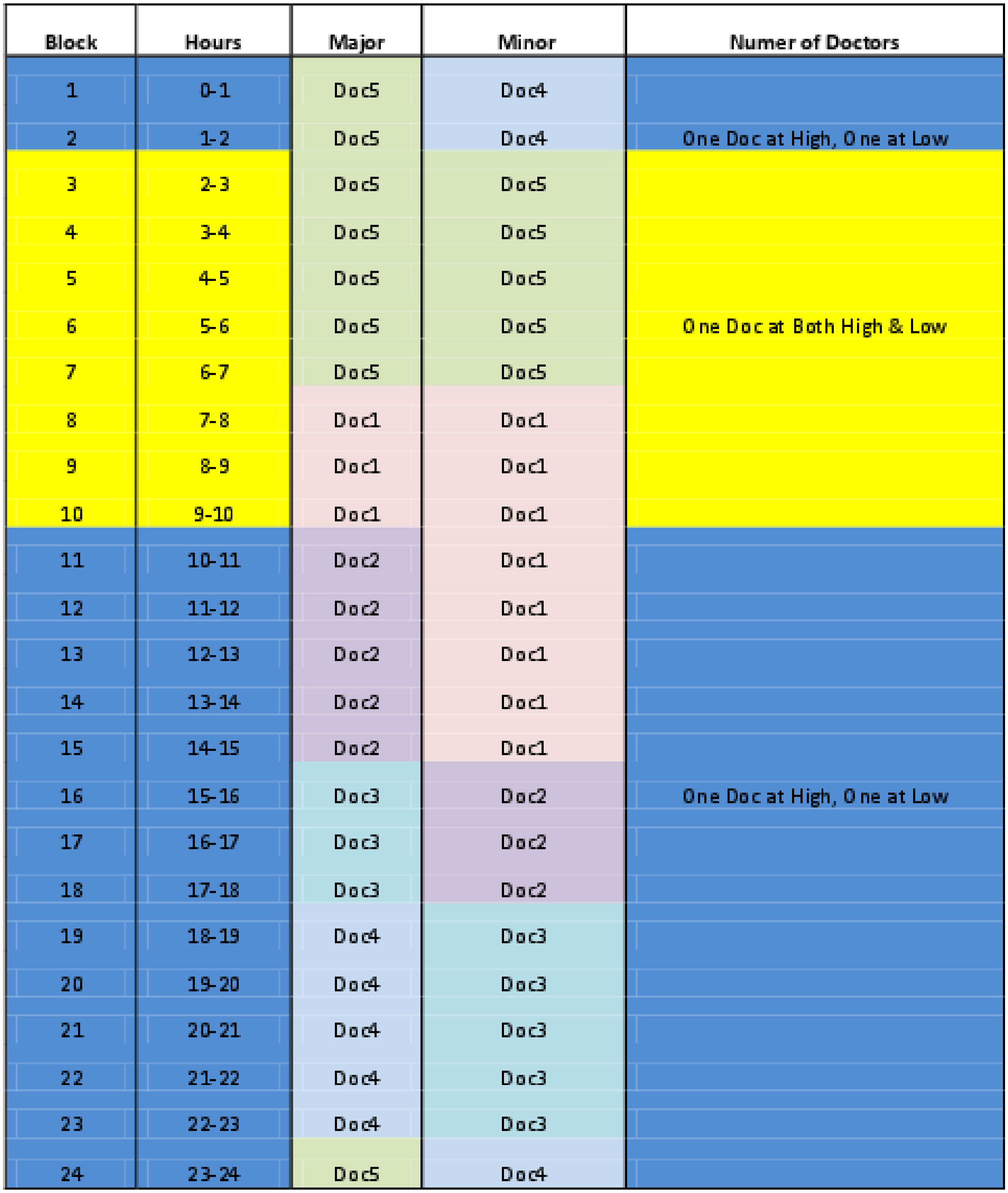
The data for our study have been collected from a large urban teaching hospital over a two year period from January, 2012 to December, 2013. An individual patient journey from arrival to the ED to exit is considered a single record where patients go through multiple procedures. There were approximately 44,000 emergency patient visits per year. Our ED contains 2 "red" areas for critical care; one for trauma care where there are 4 Beds and 2 nurses, and the other one for cardiac care where there are 4 beds, 2 nurses, and 2 hallway beds. There is a yellow area in our ED to provide intermediate care; Observation 1 where there are 7 beds, 2 nurses, and 2 hallway beds, Observation 2 where there are 8 beds and 2 nurses, and ambulance offload area where there are 4 beds, 1 nurse, and 2 hallway beds. The Rapid Assessment Zone (RAZ) is another area in our ED where there are 6 beds and 6 reclining chairs, 1 nurse at all times plus another from noon to midnight.

Among these three zones, high acuity patients are treated in the red and yellow zone areas whereas low acuity patients are treated on the RAZ area. Bed capacity for the high acuity patients totals to 31. Low acuity patients are treated in the 12 treatment beds of RAZ. However, depending on the variability in demand, actual treatment capacity may vary at any particular point in time.

The number of ED physicians at any time of the day varies between 1 to 2 in our study ED (Figure 3.3). Physician shift changes are predetermined and is stagged around different hours of the day (Figure 3.3). Demand for ED services usually slows down after midnight, and remains low until start of the morning work hours when demand again starts to peak up. Therefore, there is only one emergency doctor covering for both high and low acuity patients areas during late night and early morning hours (2 a.m. to 10 a.m.). At all other times, there are two dedicated emergency physicians, with one each serving the high and the low acuity patients, respectively.

Although physicians are dedicated to either high or low acuity patient services, respectively, their placement alters with the variability in demand for either or both of the high and low acuity patients. For instance, when there is a substantial need for high acuity patient services, both physicians work for high acuity patient services. The reverse could happen if there were no high acuity patient requiring intervention by its physician. Residents are present in our ED with no authority to prescribe treatment options for patients unless consulted with responsible attending ED physician.

In this hospital, physicians get paid at a flat rate for their shift of work, with no additional payment for extra work or overtime. Therefore, stretching out treatment times by providing additional services does not bring any incentives for the physicians (Song *et al.* (2015)). However, the usual practice for physicians is to work extra hours to finish the care process for their initiated cohort of patients before leaving their shift. Since physician shift changes are staggered throughout the day, there are always a physician to take on new patients and in that case, in order to finish their shift on time, a physician may stop taking new patients 2 hours prior to when their shift end.



Block	Hours	Major	Minor	Numer of Doctors
1	0-1	Doc5	Doc4	One Doc at High, One at Low
2	1-2	Doc5	Doc4	
3	2-3	Doc5	Doc5	One Doc at Both High & Low
4	3-4	Doc5	Doc5	
5	4-5	Doc5	Doc5	
6	5-6	Doc5	Doc5	
7	6-7	Doc5	Doc5	
8	7-8	Doc1	Doc1	
9	8-9	Doc1	Doc1	
10	9-10	Doc1	Doc1	
11	10-11	Doc2	Doc1	One Doc at High, One at Low
12	11-12	Doc2	Doc1	
13	12-13	Doc2	Doc1	
14	13-14	Doc2	Doc1	
15	14-15	Doc2	Doc1	
16	15-16	Doc3	Doc2	
17	16-17	Doc3	Doc2	
18	17-18	Doc3	Doc2	
19	18-19	Doc4	Doc3	
20	19-20	Doc4	Doc3	
21	20-21	Doc4	Doc3	
22	21-22	Doc4	Doc3	
23	22-23	Doc4	Doc3	
24	23-24	Doc5	Doc4	

Figure 3.3: Staffing level for the emergency department under study in a particular day (24 hours period).

Similar to other EDs in Canada, this ED follows a standard process flow. A simplified version of the processes flow for this ED is presented in Figure 3.4. Upon arrival, a patient's health card is swiped by a charge nurse along with documenting the chief reason for the ED visit. A triage nurse then assesses the patient's signs and symptoms and assign a priority score. This ED follows a commonly used 5 point priority scale, namely, the Canadian Triage and Acuity Scale (CTAS), where 1 denotes the highest level of severity of illness and 5 denotes the lowest. In this study, CTAS-1 (resuscitation) and CTAS-2 (emergent) patients are considered as high acuity patients, and CTAS-3 (urgent), CTAS-4 (less-urgent), and CTAS-5 (non-urgent) patients are considered as low acuity patients. Ancillary services, such as diagnostic testing, transport, laboratory, and pharmacy, are shared with all patients in the hospital.

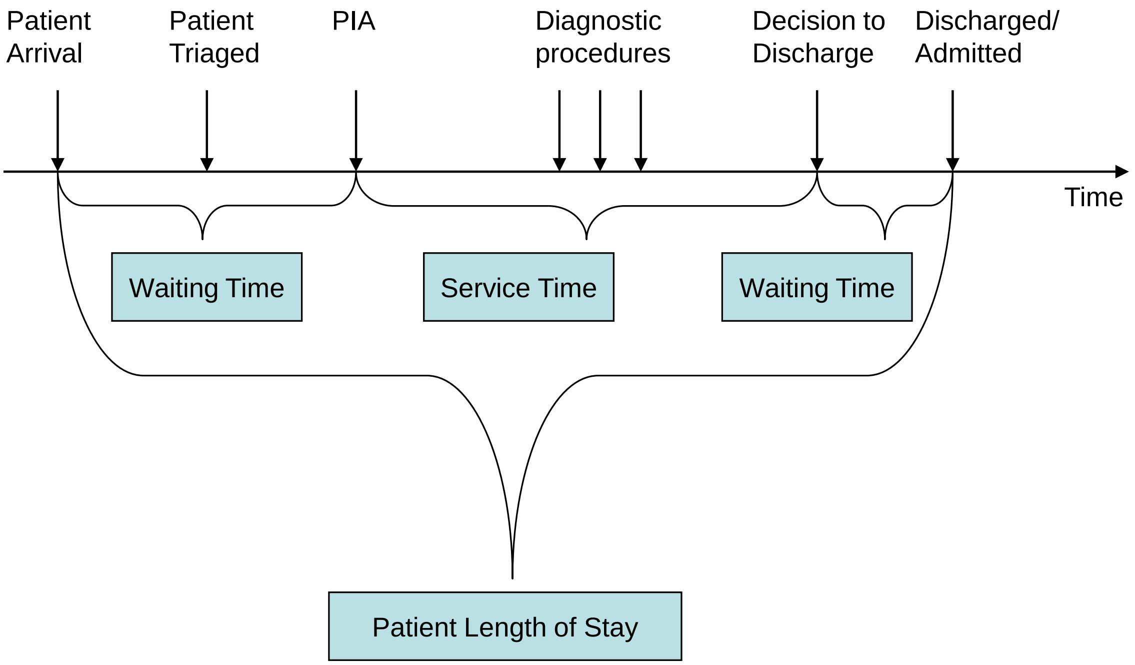


Figure 3.4: Emergency department patient flow process (this figure is not drawn to scale).

After being triaged, patients usually wait in a common waiting room to get the call for the physician assessment. It is usual practice for ED physician to work on multiple patients simultaneously. For example, while operation theater for a patient is in preparation, a physician may start working on a waiting patient in order to minimize her idle time. If physician assessment does not require any diagnostic testing or patient require hospital admission, a discharge or admission decision is made immediately. Otherwise, patients may have to go through different diagnostic procedures based on her sickness under study and then the previous step is repeated. Finally, once the discharge decision has been made, patients are then discharged home or admitted to the hospital for further care. Patients ED service can then be considered as complete.

3.3.2 Data Description and Definitions

Our data consists of 88,189 individual patients treated in an Ontario ED in 2 years between January 1, 2012 to December 31, 2013. Data were de-identified and ethics approval obtained from the pertinent authority. Times at which a patient received any service during her care process is documented as a time stamp, for example, received time, triaged time, first physician assessment, and so on. Patient level informations such as age, triage level, discharge destination, chief complaints, were also included in our dataset.

3.3.2.1 Dependent Variables

As described in Figure 3.4, individual time stamps have been used to calculate our key dependent variables which are going to used in chapter 5. A patient's ED LOS is defined as the time from a patient's registration at the ED to the time patient discharged/admitted from the ED. A patient's ED service time is defined as the time from physician's initial assessment to the ED to the time a patient discharged or admitted from the ED. Finally, a patient's ED waiting time is defined as the time from a patient's registration at the ED to the time physicians' start patient's initial assessment to the ED.

In order to conduct block-level analysis, a day was divided into 24 one hour blocks and individual data were aggregated to obtain block-level observations. The number of discharges per block is the sum of the individuals who have been discharged from that particular block.

Table 3.3: Definition of response variables for operational performance measures used in chapter 5.

Measure	Description and coding
DischHigh	Number of high acuity patients discharged from ED per block.
DischLow	Number of low acuity patients discharged from ED per block.
LOSHigh	High acuity patient length of stay (registration to discharge).
LOSLow	Low acuity patient length of stay (registration to discharge).
STHigh	High acuity patient service time (initial assessment to discharge).
STLow	Low acuity patient service time (initial assessment to discharge).
WTHigh	High acuity patient waiting time (registration to initial assessment).
WTLow	Low acuity patient waiting time (registration to initial assessment).

3.3.2.2 Independent and Control Variables

Similar to the calculation for the number of discharges, I can also calculate census per block, the number of arrivals per block, and the number served per block. The census in the ED at time $(t + 1)$ is obtained by using following formula,

$$\text{Census}(t + 1) = \text{Census}(t) + \text{Arrivals}(t, t + 1) - \text{Departures}(t, t + 1)$$

where $\text{Arrivals}(t, t + 1)$ and $\text{Departures}(t, t + 1)$ are the arrivals and departures between time t and $(t + 1)$, respectively. I have calculated the census for every block (hour) of a day for the duration of the study period. The formula above however requires us to know the census at the beginning of the study (at, $t = 0$).

Since our data starts at the midnight of January 1, 2012, we assumed the census is zero at time 0. However, I have considered dropping 24 blocks of observation from the beginning and the end of the study to deal with this truncation.

Since there exist natural variability in dependent variables depending on calendar variables such as, time of the day, day of the week, week of the month etc., I have used these variables as control variables in our models. I have incorporated control variables as dummy variables in our model (for n categories of a variable, I need $(n - 1)$ independent dummy variables).

Table 3.4: Definitions of independent variables used in chapter 5.

Measure	Description and coding
CenHigh	Number of high acuity patients in the system (counts).
CenLow	Number of low acuity patients in the system (counts).
CenHigh^2	Square of number of high acuity patients in the system (counts)
CenLow^2	Square of number of high acuity patients in the system (counts)
CenHighLag1	Number of high acuity patients in the system at previous block (counts)
CenLowLag1	Number of low acuity patients in the system at previous block (counts)
HighWorkload	Number of high acuity patients in the system divided by total number of beds for high acuity patients (ratio)
LowWorkload	Number of low acuity patients in the system divided by total number of beds for low acuity patients (ratio)
BusyHigh	Indicator variable takes 1 if high census is greater than the number of high beds
BusyLow	Indicator variable takes 1 if low census is greater than the number of low beds
Admitted	Binary variable indicating admitted to hospital beds or discharged home.
Age	Age of the patient (continuous variable).

Table 3.5: Definitions of control variables used in chapter 4 and 5.

Measures	Description and coding
BoD (1-24)	Block of day variables are defined using 23 dummy variables
DoW (Sun-Sat)	Day of week variables are defined using 6 dummy variables
MoY (Jan-Dec)	Month of year variables are defined using 11 dummy variables

Chapter 4

Forecasting Emergency Department Arrivals

4.1 Introduction

Emergency department (ED) overcrowding is prevalent in Canada (Bond *et al.* (2006)) as well as around the globe (Pines *et al.* (2011)). An ED becomes overcrowded when the demand for ED services exceeds the capacity of the ED, which in turn limits the ability of an ED to provide quality care in a reasonable amount of time (Rowe *et al.* (2006)). ED overcrowding has various negative consequences, such as increased wait times, decreased levels of service, poor patient outcomes, decreased physician productivity, patient suffering, and reduced patient and provider satisfaction (Derlet & Richards (2000)). Since across-the-board capacity expansion is not always a viable option due to budgetary constraints, an accurate short-term forecast of ED arrivals would help ED administrators to prepare in advance how to efficiently manage and allocate available resources so as to meet the anticipated demand and alleviate ED crowding. Of course, ED administrators do not have complete control over staffing. On one hand, ED employees are highly trained individuals who may not open to wait for works that may not come. On the other hand, hospitals may not afford to pay for keeping employees stand-by. However, ED administrators may still find a viable option to schedule staff accordingly to predicted demand or ask staff to stay few extra hours (before beginning of the shift or end of the shift) during the time of expected surge.

A considerable amount of research has been devoted towards predictions of ED arrivals, on a daily, weekly, monthly, and yearly forecast basis. However, very few studies have attempted to forecast ED arrivals for much shorter time into future and no particular method was shown to have consistent superiority over others. The objective of this chapter is to suggest appropriate models for short-run (3 hourly) ED arrival predictions. Forecasting short-term arrivals is important for three main reasons: (i) arrivals are not uniformly distributed throughout the day so that particular hours of the day have more arrivals than others, (ii) resources are mostly shared across the ED and the employees work in predetermined shifts within a day, and (iii) short term forecasts of a surge in load will enable an ED manager to know when to call in supplementary staff (doctors, nurses, etc.) and other resources.

In almost all jurisdictions, emergency patients are categorized into different priority classes based on their level of acuity. The Canadian Triage and Acuity Scale (CTAS) in Canada (CTAS (2005)) is the one where emergency patients are categorized into five priority classes and from the highest acuity to the lowest. They are arranged as; Resuscitation (1), Emergent (2), Urgent (3), Less-urgent (4), and Non-urgent (5). The priority of an arriving patient is determined based on expert clinical assessment. Associated with the wait time targets, a certain percentage of each acuity patient group should not exceed a particular wait time threshold (Table 4.1). Many EDs have introduced a separate "urgent care" units to serve lower acuity patients to meet their service level. However, variability in the demand for ED services impairs maximum utilization of the ED resources, and as a result certain performance benchmarks across the acuity categories often fail to be met.

Time series and regression models are two frequently used models in the literature for ED arrival predictions. Time series models assume prior arrivals are the most important predic-

Table 4.1: CTAS Key Performance Indicators

Category	Classification	Access	Performance Level
1	Resuscitation	Immediate	98%
2	Emergency	15 minute	95%
3	Urgent	30 minute	90%
4	Less urgent	60 minute	85%
5	Not urgent	120 minute	80%

tors of future arrivals, whereas the regression models assume variables other than the arrivals themselves are important predictors of future arrivals. Both models have advantages, so a combination of time series and regression models, namely, Generalized Linear Autoregressive Moving Average (GLARMA) models were used to predict future arrivals in this chapter. In addition, Harmonic regression models were used for comparison purposes, because such models were advocated to produce useful results for short-run prediction of time series data (Côté *et al.* (2013)). Separate forecasting models for high and low acuity ED patients were considered due to differences in arrival patterns, service mechanisms, and capacities (McCarthy *et al.* (2008), Sun *et al.* (2009), Au-Yeung *et al.* (2009), Chen *et al.* (2011)).

In order to validate the appropriateness of the forecasting model I adopted the "rolling horizon" approach from the operations management literature (eg., Sethi & Sorger (1991)). Starting from the initial segment of the original data the validation process moves one step forward by a fixed time interval and recalculates accuracy measures every time until the process reaches the end of the dataset.

The remainder of this chapter is assembled as follows. A detailed literature review is presented in Section 4.2. In Section 4.3, I present the materials and methods I employed in this chapter. Detailed results and interpretations are provided in Section 4.4. I discuss the results of our study in Section 4.5, along with the limitations in Section 4.6, conclusions and future research directions in Section 4.7.

4.2 Literature on Forecasting ED Arrivals

The problem of forecasting ED arrivals has been studied extensively in the literature. The majority of published articles regarding ED were focused on predicting ED arrivals, while others were interested in forecasting ED crowding (Hoot *et al.* (2008), Hoot *et al.* (2009)), ED bed occupancy (Schweigler *et al.* (2009)), or a combination of a few of other ED performance metrics (Hoot *et al.* (2008), Hoot *et al.* (2009)). Our focus in this chapter is to predict ED arrivals and the synopsis of the studies that have used forecasting methods to predict ED arrivals are provided in the tables 4.2 and 4.3. Different time horizons (daily to yearly) have been used in the literature to forecast ED arrivals. I am interested in very short-term forecasts, namely, several hours of a particular day.

Wargon *et al.* (2009) conducted a systematic review of the models that have been used to forecast the number of emergency department visits. They identified nine studies through a Medline search (date of access was 23 September, 2007), and observed that two types of models were predominant in those studies, namely, time series models and linear regression models. Therefore, I subdivided our review of literature based on these two modeling strategies because recent studies also tend to belong to one of these two modeling streams.

4.2.1 Time Series Models

An early attempt at using a time series model to predict ED arrivals was made by Milner (1988). Among time series models, the autoregressive integrated moving average (ARIMA) approach was the most widely used model in the literature for ED arrival forecast (Milner (1988), Tandberg & Qualls (1994), Milner (1997), Reis & Mandl (2003), Champion *et al.* (2007), Au-Yeung *et al.* (2009), Sun *et al.* (2009), Chan *et al.* (2011), Boyle *et al.* (2012)). A few other time series models such as, exponential smoothing (Tandberg & Qualls (1994), Champion *et al.* (2007), Boyle *et al.* (2010), Boyle *et al.* (2012), Bergs *et al.* (2014)), structural time series models (Au-Yeung *et al.* (2009)), multivariate time series models (Jones *et al.*

(2009)), and seasonal ARIMA models (Marcilio *et al.* (2013)) were also used for ED arrival forecasts.

Different time horizons were considered in the aforementioned studies such as yearly (Milner (1988), Boyle *et al.* (2012) Milner (1997)), monthly (Champion *et al.* (2007), Chan *et al.* (2011), Bergs *et al.* (2014), Boyle *et al.* (2012)), daily (Reis & Mandl (2003), Au-Yeung *et al.* (2009), Sun *et al.* (2009), Boyle *et al.* (2012), Marcilio *et al.* (2013)), and hourly (Tandberg & Qualls (1994), Jones *et al.* (2009), Boyle *et al.* (2012)). Yearly and monthly forecasts are important for strategic and tactical planning purposes, however, daily and hourly forecasts facilitate operational planning (Côté *et al.* (2013)).

Since different types of patients manifest variability in demand for services and availability of resources, separate forecasting models were proposed in the following studies: for overall and respiratory-related visits (Reis & Mandl (2003)), for trauma, non-trauma, and pediatric visits (Chan *et al.* (2011)), for three acuity categories P1, P2, and P3 (Sun *et al.* (2009)), and for walk-in and ambulance arrivals (Au-Yeung *et al.* (2009)). Therefore, I developed separate forecasting models for high (CTAS 1 and 2) and low (CTAS 3, 4, and 5) acuity patients because patients are treated in two different areas based on their level of acuity in our study ED.

I am only aware of two studies that applied time series models to forecast hourly ED arrivals. Tandberg & Qualls (1994) used time series models to forecast hourly ED arrivals using five different models (raw data, ARIMA, and 3 variants of a moving average model) based upon 2 years of hourly arrival data. The authors concluded that models based on arithmetic mean or seasonal indices with a single moving average term provides the most accurate forecast. Jones *et al.* (2009) on the other hand, advocated for the use of a multivariate time series model to simultaneously predict 8 different hourly measures including arrival counts, census counts, and diagnostic order counts. Using hourly data from 3 diverse hospitals for the year 2006, the authors suggested the use of multivariate time series models to accurately forecast all other measures, except the demand for diagnostic resources.

4.2.2 Variants of Linear Models

Earlier attempts to use linear regression models to forecast daily ED arrivals are considered by Holleman *et al.* (1996) and Rotstein *et al.* (1997). Using a linear regression model based on 34 months of walk-in clinic and ED daily arrival data, Holleman *et al.* (1996) argued that calendar and climate variables efficiently predict daily arrivals explaining 84% of the daily variability. Rotstein *et al.* (1997), on the other hand, proposed the analysis of covariance (ANCOVA) technique based on calendar variables to predict daily ED arrivals using 3 years of historical daily arrival data. According to Rotstein *et al.* (1997), 65% of the total variability in daily ED arrivals was explained by the calendar variables alone. A recent study by Ekström *et al.* (2015) reported a significant correlation between the number visits regional medical website between 6 P.M. and midnight and ED visits on the upcoming day ($r = 0.77, p < 0.001$). Ekström *et al.* (2015) claim that the web site visits along with calendar variables are capable of accurately predicting daily ED arrivals.

Among regression based approaches, McCarthy *et al.* (2008) selected Poisson regression to represent the count of hourly ED arrivals as a function of temporal, climatic, and patient factors. Advocating separate model fitting for high and low acuity patients, McCarthy *et al.* (2008) observed that there were more ED visits on Mondays, followed by weekends, relative to other weekdays, and climatic factors did not strongly influence patient arrivals to the ED. Marcilio *et al.* (2013) also argued that climate variables are not a good predictor of daily ED arrivals, finding rather that calendar variables alone are capable of detecting daily variability in ED volume. However, Marcilio *et al.* (2013) applied six different time series and regression models and showed that GLM and GEE manifested better forecast accuracy than SARIMA and

Table 4.2: Summary table for the literature reviewed in this chapter (part 1).

Authors (year)	**Country (data)**	**Forecast Horizon**	**Methods**	**Main findings**
Milner (1988)	UK (12 years)	Year	TS[1]	Accurate new attendance forecasts but not so for the returns. Forecasts at 1994 were no better than district projections made at 1984.
Tandberg & Qualls (1994)	USA (2 years)	Hour	TS[1]	Arithmetic mean model is better than ARIMA. No TS forecast explained > 1% variation in LOS or Acuity.
Holleman *et al.* (1996)	USA (34 months)	Day	LR[2]	Calendar & weather variables explained 84% variability in validation set. Decreased overstaffing by 44% but increased understaffing by 16%.
Rotstein *et al.* (1997)	Israel (3 years)	Day	LR[2]	Trend & time variables explained 65% of the variability. Trend variable was insignificant on 1 year validation data.
Milner (1997)	UK (12 years)	Year	TS[1]	Incorporating ARIMA improved forecast accuracy. Update forecast whenever new data available.
Reis & Mandl (2003)	USA (10 years)	Day	TS[1]	Respiratory-related and overall ED volume forecast with accuracy. ARIMA based alarm system may serve as a real-time outbreaks surveillance.
Champion *et al.* (2007)	Australia (6 years)	Month	TS[1]	Exponential smoothing and Box–Jenkins classes of models were adaptable. Accurate prediction facilitates planning of nursing rosters and staff allocation.
McCarthy *et al.* (2008)	USA (1 years)	Hour	PR[3]	Demand for ED services approximated well by Poisson model. Expected arrival rate is characterized by temporal,weather, diversion predictors.
Hoot *et al.* (2008)	USA (1 years)	Hour	DES[4]	Simulation forecasts of near-future ED crowding. Correlations between crowding forecasts and actual outcomes started high. The residual means were unbiased for all outcomes but the boarding time.
Au-Yeung *et al.* (2009)	UK (5 years)	Daily	STS[5]	Separate models for walk-in and ambulance arrivals. Walk-in and ambulance arrivals showed 7 day seasonality. Predictive power for walk-in arrivals (r=0.6205) were higher.
Jones *et al.* (2009)	USA (1 years)	Hourly	MTS[6]	Multivariate method provided more accurate forecasts. Descriptive analyses revealed little temporal interaction.

[1] Time-series
[2] Linear Regression
[3] Poisson Regression
[4] Discrete Event Simulation
[5] Structural Time-series
[6] Multivariate Time-series

Table 4.3: Summary table for the literature reviewed in this chapter (part 2).

Authors (year)	**Country (data)**	**Forecast Horizon**	**Methods**	**Main findings**
Sun *et al.* (2009)	Singapore (33 Months)	Daily	TS[1]	3 separate ARIMA models for P1, P2, and P3 patients. P1 attendance did not show weekly periodicity while P2 did. P3 attendance significantly correlated with day, month, holiday and $PSI > 50$.
Wargon *et al.* (2009)	Multiple	All	SR[3]	Medline search for forecast ED or walk-in patients. Mainly regression and time-series models were used. Meteorological data failed to improve the reliability.
Boyle *et al.* (2010)	Australia (5 years)	Daily	SM[4]	Presentations can be predicted with 90% accuracy. Highest accuracy observed on weekends and summer months. Variation in accuracy was highest on public holidays.
Chen *et al.* (2011)	Taiwan (57 months)	Monthly	TS[1]	Meteorological, clinical,& economic factors affect ED revenue and visits. Separate prediction models for non-trauma, trauma, and pediatric visits.
Boyle *et al.* (2012)	Australia (5 years)	Hourly Daily Monthly	LR[2] TS[1] SM	ED presentations were easier to forecast than admission. Forecasting performance roughly equivalent for sample sizes> 10. An automated dashboard generated as an admission prediction tool.
Marcilio *et al.* (2013)	Brazil (3 years)	Daily	GLM[5] GEE[6] SARIMA[7]	GLM, GEE showed better accuracy than SARIMA. Calendar variables alone can detect variability in ED volume. Climate variables are not predictive of daily presentations.
Côté *et al.* (2013)	USA (10 years)	Hourly,Daily Monthly,Yearly	LR[2]	A tutorial for ED directors to anticipate ED arrivals. Regression models was useful to variety of forecasting situation.
Ekström *et al.* (2015)	Sweden (15 months)	Daily	LR[2]	Internet data can be used to predict ED visits. Significant correlation between Web site visits and ED visits.
Kadri *et al.* (2014)	France (1 year)	Daily	TS[1]	Time series did not show trend or seasonality. univariate ARMA models were adequate.
Bergs *et al.* (2014)	Belgium (6 years)	Monthly	TS[1]	proposed an automated exponential smoothing approach. Forecast based on exponential smoothing was better.

[1] Time-series
[2] Linear Regression
[3] Systematic Review
[4] Smoothing
[5] Generalized Linear Models
[6] Generalized Estimating Equations
[7] Seasonal Autoregressive Integrated Moveing Average

three other time series models. Analogously, Boyle *et al.* (2012) also used both the regression and time series models to develop and validate forecasting models for hourly, daily, monthly, and yearly ED arrivals.

Finally, Côté *et al.* (2013) wrote a tutorial on how to use variants of regression-based forecasting models to predict hourly, daily, monthly, and yearly ED arrivals. The authors demonstrated that annual ED arrival data shows a long term growth and hence a trend regression analysis is appropriate for annual arrivals. However, hourly ED arrivals show a wavelike pattern and therefore Fourier regression is an appropriate method to predict hourly arrivals which effectively describes the wavelike pattern. For monthly and daily arrival forecasts, the authors used dummy variables for the months and days, respectively.

4.2.3 Rolling Horizon

The rolling horizon approach is widely used in the production and operations management literature. A systematic review of the use of rolling horizons in operations management problems is documented in Chand *et al.* (2002). Selective articles that applied rolling horizon principle to dynamic problems in economics, finance, marketing, and control engineering were also included in Chand *et al.* (2002). The theoretical framework of the rolling horizon process is described in Sethi & Sorger (1991). While exploring the accuracy of the time series methods, Makridakis *et al.* (1982) adopted the rolling horizon technique to choose the best method. The rolling horizon approach has also been applied to health care problems in the form of scheduling (Rohleder & Klassen (2002), Zhang & Vossen (2013), Addis *et al.* (2014)). I applied the rolling horizon approach to validate ED arrival forecasts.

4.2.4 Contribution of this Chapter

The closest paper to our work is that of McCarthy *et al.* (2008) who used Poisson regression to predict hourly ED arrival counts. Although McCarthy *et al.* (2008) tested for the dependence of current hourly arrivals on previous hourly arrivals, they did not consider any autoregressive or moving average terms in the model. Furthermore, hourly data consists of many zero counts (especially during night hours) which should be accounted for by some variant of a Poisson regression. Therefore, I have used the generalized linear autoregressive moving average (GLARMA) model to accurately forecast arrivals in 3, 6, 9, and 12 hours in future. The main contributions of this chapter are the following:

- I developed and validated an appropriate forecasting model (GLARMA) to predict hourly ED arrival counts based on calendar variables.
- I compared our proposed forecasting model with the existing ones; both time series models and regression based approach.
- I also compared our proposed forecasting model with the Harmonic regression model where the sine and cosine functions are used to capture the seasonal variability of the time series.
- I evaluated an ED forecasting method using a rolling horizon approach.

4.3 Statistical Methods

4.3.1 Seasonal ARIMA models

First we present the seasonal ARIMA model described in chapter 2, which has been applied to our block arrival data where the discrete time series is expressed as a function of non-seasonal and seasonal autoregressive terms, non-seasonal and seasonal moving average terms, and non-seasonal and seasonal differenced terms. The number of difference terms is determined after taking successive differences until the time series transforms to a stationary process, and the number of autoregressive and moving average terms are determined by using autocorrelation (ACF) and partial autocorrelation (PACF) functions. Once the time series is transformed to a stationary process, the following model is used to forecast future block ED arrivals,

$$\text{E}[\text{Arrival}_t] = \sum_{j=1}^{p} \phi_j \text{Arrival}_{t-j} + \sum_{j=1}^{P} \Phi_j \text{Arrival}_{t-sj} + \text{Error}_t + \sum_{j=1}^{q} \theta_j \text{Error}_{t-j} + \sum_{j=1}^{Q} \Theta_j \text{Error}_{t-sj}$$

where p, P, q, and Q represents the orders of the non-seasonal AR, seasonal AR, non-seasonal MA, and seasonal MA, respectively. The length of the season is denoted by s which is 8 in our case (corresponding to 8 3-hour blocks per day).

4.3.2 GLM with Calendar Variables as Predictors

Poisson and Negative Binomial regression models were applied to fit the original data. Unlike ARIMA models, block arrival counts are assumed to be distributed as Poisson or Negative Binomial, and the positive skewness of our data justifies their use. The Poisson regression model requires the variance to be proportional to their mean. However, Negative Binomial regression is preferred over Poisson regression when the observed variance is notably higher than the mean. While the Poisson is a one parameter model, the additional parameter in the Negative Binomial model allows the variance to be adjusted independently of the mean. For both of these models, the logarithm of the expected block arrivals is expressed as a linear function of the calendar variable:

$$\log(\text{Arrival}_i) = \alpha + B\ \text{BoD}_i + \Delta\ \text{DoW}_i + \Gamma\ \text{MoY}_i + \epsilon_i\ ;\ i = 1, 2, \ldots, n$$

where $B, \Delta,$ and Γ are vector of coefficients corresponding to the indicator variables for each of the block of day (BoD), the day of week (DoW), and the month of year (MoY) variables, respectively.

4.3.3 Harmonic Regression Based on GLM and ARIMA

A Harmonic Regression model was also considered, where the expected block arrival is expressed as a function of sine and cosine functions. Both the GLM and time series modeling framework can incorporate such terms as predictors in the model. The wavelike pattern of our time series data justifies the use of Fourier terms as predictors in our model. The GLM and time series representation of the Harmonic Regression can be expressed as follows,

$$\log(\text{Arrival}_t) = \alpha + B\,\text{BoD}_t + \Delta\,\text{DoW}_t + \Gamma\,\text{MoY}_t + \eta_1 \sin\left(\frac{2\pi t}{k}\right) + \eta_2 \cos\left(\frac{2\pi t}{k}\right) + \epsilon_i,$$

$$\begin{aligned} \text{E}[\text{Arrival}_t] = {} & \sum_{j=1}^{p} \phi_j \text{Arrival}_{t-j} + \sum_{j=1}^{P} \Phi_j \text{Arrival}_{t-sj} + \text{Error}_t + \sum_{j=1}^{q} \theta_j \text{Error}_{t-j} + \sum_{j=1}^{Q} \Theta_j \text{Error}_{t-sj} \\ & + \eta_1 \sin\left(\frac{2\pi t}{k}\right) + \eta_2 \cos\left(\frac{2\pi t}{k}\right) \end{aligned}$$

where t is the block number, 1 for the first block and 5848 for the last block (8 blocks/day * 731 days = 5848 blocks), and k is the value of the period required to complete one cycle of the time series.

4.3.4 Generalized Linear Autoregressive Moving Average Models

Generalized linear autoregressive moving average (GLARMA) models were developed to accommodate non-Gaussian (discrete valued) time series where successive responses are correlated. An important advantage of using GLARMA over either GLM or ARIMA is that inference on calendar variables are possible when accounting for the serial dependence among subsequent block arrivals. GLARMA models are easy to fit because the likelihood function is conditionally specified as a product of conditional distributions which belongs to exponential family (Dunsmuir & Scott (2015)). Parameter estimates were obtained through the method of Maximum Likelihood using Fisher scoring or Newton-Raphson iterations.

Generalized linear autoregressive moving average models combine the functionality of both the ARIMA and the GLM models under the state-space modeling framework. Under this framework, the logarithm of the expected block arrivals are expressed as a function of calendar variables as well as the autoregressive and moving average terms.

$$\log(\mathrm{E}[\mathrm{Arrival}_i]) = \alpha + B\,\mathrm{BoD}_i + \Delta\,\mathrm{DoW}_i + \Gamma\,\mathrm{MoY}_i + \sum_{i=1}^{p} \phi_i Z_{t-i} + \sum_{i=1}^{\tilde{q}} \theta_i e_{t-i}\ ;\ i = 1, 2, \ldots, n$$

where, $\tilde{q} = \max(p, q)$, and the serial dependence in the response process is introduced via Z_t which is a linear combination of past predictive residuals and which satisfies ARMA like recursions, $Z_t = \sum_{i=1}^{p} \phi_i (Z_{t-i} + e_{t-i}) + \sum_{i=1}^{q} \theta_i e_{t-i}$.

4.3.5 Rolling Horizon Approach

I applied a rolling horizon approach to validate our proposed forecasting models. The basis for the rolling horizon approach is to divide the forecast horizon into multiple periods and then to update and extend an existing plan in each period (e.g., Sethi *et al.* (2006)). The number of future periods for which the forecast is made can be termed as 'horizon' and these are the periods which 'roll over' once a forecast for that period is made (Sethi & Sorger (1991)). Our existing plan is to start the validation process with 13 months of data where the first 12 months of data are used for model fitting purposes (training data) and the remaining one month for the validation purposes (test data). The process is then moved forward by one month leaving the first month out and repeating the same procedure described above. A schematic representation of how the rolling horizon approach works is displayed in the following Figure 4.1.

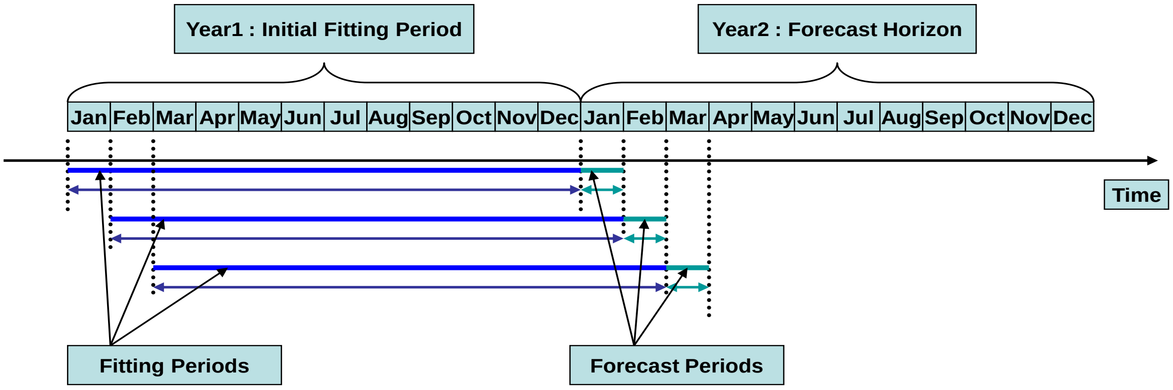


Figure 4.1: Rolling horizon process flow.

4.4 Results and Interpretation

4.4.1 Descriptive Analysis

The distribution of emergency patient arrivals for each of the five triage categories is presented in Figures 4.2 through 4.4 for blocks of the day, days of the week, and months of the year, respectively. Relevant summary statistics are documented in table 4.4. The three hour blocks are sequentially arranged from midnight to midnight. Thus it is not surprising that blocks 1, 2, and 3 have comparatively fewer arrivals than blocks 4, 5, 6, 7, and 8. This sort of daily cycle is well known in the ED literature, where arrivals continue to grow during the daytime but there are fewer arrivals overnight. The mean number of arrivals (8.83) per block was highest among Urgent patients, with maximum reaching to 27 patients, whereas Non-urgent and Resuscitation patients have comparatively lower mean arrivals per block, 0.32 and 0.63, respectively.

Irrespective of triage categories, Mondays have higher arrivals on average than any other day of the week. The number of arrivals on the weekend remains low compared to any weekdays. Among 120 average arrivals per day, approximately 70 of them are Urgent, 24 and 18 are Emergent and Less-Urgent, only 5 and 2 are Resuscitation and Non-Urgent. Arrivals vary around the months of the year without any specific pattern. The lowest number of arrivals for every triage category has been observed in the month of February. The decrease in numbers of patients in the month of February is at least in part because there are fewer days in February. July and August have higher arrivals.In descending order, Urgent, Emergent, Less-Urgent, Resuscitation, and Non-Urgent patient categories have on average, 2152, 735, 548, 153, and 78 arrivals per month.

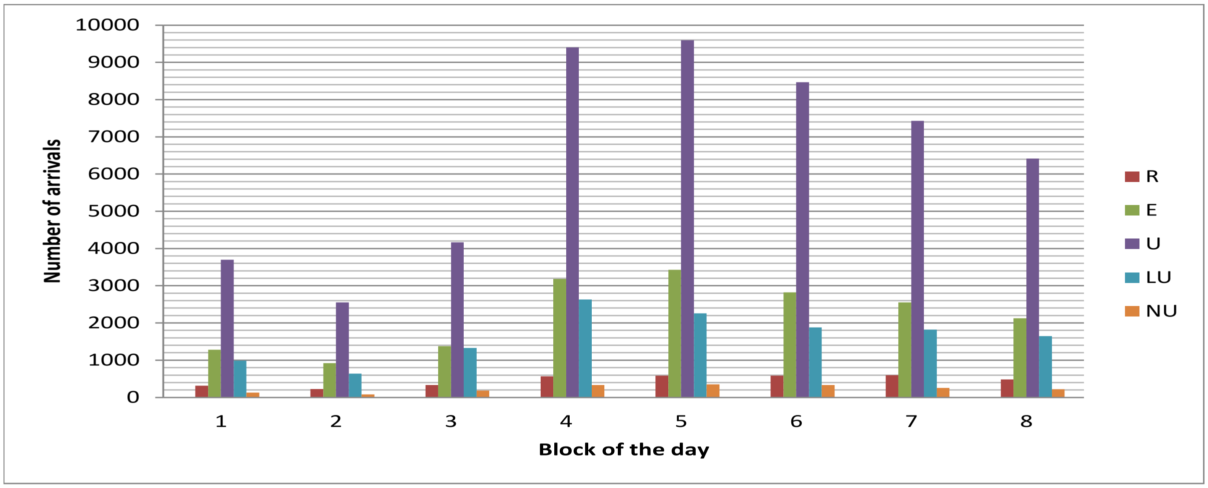


Figure 4.2: Number of arrivals in an Emergency Department by block of the day, Jan 2012 to Dec 2013; R:Resuscitation, E:Emergent, U:Urgent, LU:Less-Urgent, NU:Non-Urgent.

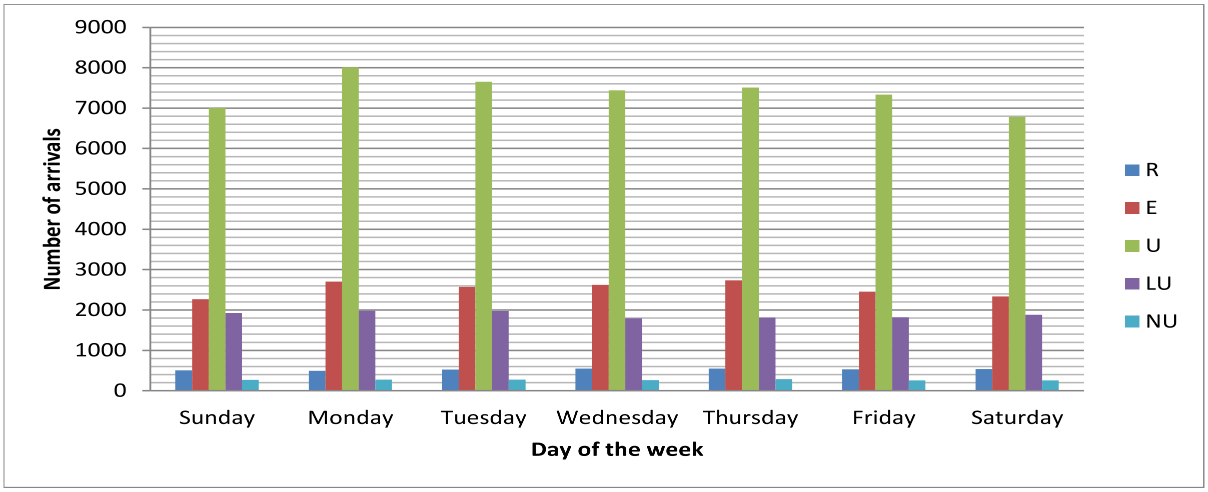


Figure 4.3: Number of arrivals in an Emergency Department by day of the week, Jan 2012 to Dec 2013; R:Resuscitation, E:Emergent, U:Urgent, LU:Less-Urgent, NU:Non-Urgent.

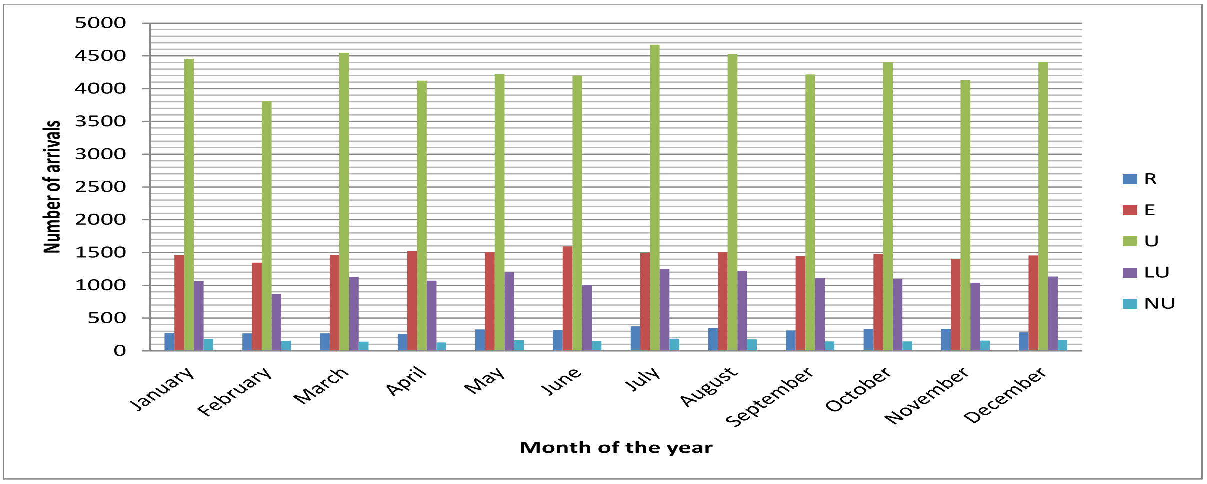


Figure 4.4: Number of arrivals in an Emergency Department by month of the year, Jan 2012 to Dec 2013; R:Resuscitation, E:Emergent, U:Urgent, LU:Less-Urgent, NU:Non-Urgent.

Table 4.4: Descriptive statistics of the block, day, and month arrivals by CTAS categories.

Horizon	**Triage**	**Mean**	**Median**	**Max**	**Min**	**Variance**	**StDev**
Block	Resuscitation	0.63	0.00	7.00	0.00	0.69	0.83
	Emergent	3.02	3.00	14.00	0.00	4.75	2.18
	Urgent	8.83	8.00	27.00	0.00	22.39	4.73
	Less-Urgent	2.25	2.00	12.00	0.00	3.80	1.95
	Non-Urgent	0.32	0.00	6.00	0.00	0.39	0.62
	Total	15.06	15.00	40.00	0.00	49.67	7.05
Day	Resuscitation	5.04	5.00	15.00	0.00	5.40	2.32
	Emergent	24.15	24.00	6.00	9.00	33.88	5.82
	Urgent	70.65	71.00	108.00	34.00	110.68	10.52
	Less-Urgent	18.02	17.00	44.00	4.00	40.05	6.33
	Non-Urgent	2.58	2.00	12.00	0.00	3.71	1.93
	Total	120.45	121.00	162.00	63.00	166.07	12.89
Month	Resuscitation	153.58	157.00	203.00	122.00	441.91	21.02
	Emergent	735.58	718.50	837.00	641.00	2660.95	51.58
	Urgent	2152.00	2165.00	2360.00	1791.00	17107.65	130.80
	Less-Urgent	548.79	556.50	680.00	366.00	6188.61	78.67
	Non-Urgent	78.71	79.00	100.00	48.00	189.35	13.76
	Total	3668.67	3674.00	4072.00	3058.00	41507.19	203.73

Since forecasting is done for high and low acuity patients separately, emergency patients have been classified by combining Resuscitation and Emergency patients, and as low by pooling Urgent, Less-Urgent, and Non-Urgent patients. Figure 4.7 displays arrivals by high and low acuity patients for blocks of the day, days of the week, and months of the year, from top to bottom. Related descriptive statistics are shown in table 4.5. Since Urgent arrivals are significantly higher than any other triage category, accumulated arrivals for low category are higher than that of high category. On average, there are approximately 11 arrivals per block for the low category as compared to 3 arrivals for the high category.

Among days of the week, Monday have slightly higher arrivals for both the high and low acuity cases. Approximately 91 and 29 low and high acuity arrivals correspond to roughly 120 daily arrivals. The maximum daily arrivals for low acuity patients reached to 130 whereas the minimum is at 43. Monthly arrivals fluctuate more for low acuity patients than for high acuity patients. low acuity arrivals are lowest in the month of February, as expected, and highest in the month of July. On average there are 2779 low category arrivals per month whereas there are 889 high category arrivals.

Table 4.5: Descriptive statistics of block, day, and month arrivals by Low (CTAS 3,4,5) and high (CTAS 1,2) categories.

Horizon	Triage	Mean	Median	Max	Min	Variance	StDev
Block	low	11.41	11.00	33.00	0.00	32.20	5.67
	high	3.65	3.00	16.00	0.00	5.80	2.41
	Total	15.06	15.00	40.00	0.00	49.67	7.05
Day	high	29.19	29.00	51.00	13.00	39.13	6.26
	low	91.25	91.00	130.00	43.00	134.61	11.60
	Total	120.44	121.00	162.00	63.00	166.07	12.89
Month	high	889.17	883.00	1016.00	770.00	3946.41	62.82
	low	2779.50	2793.50	3089.00	2216.00	36720.26	191.63
	Total	3668.67	3674.00	4072.00	3058.00	41507.19	203.73

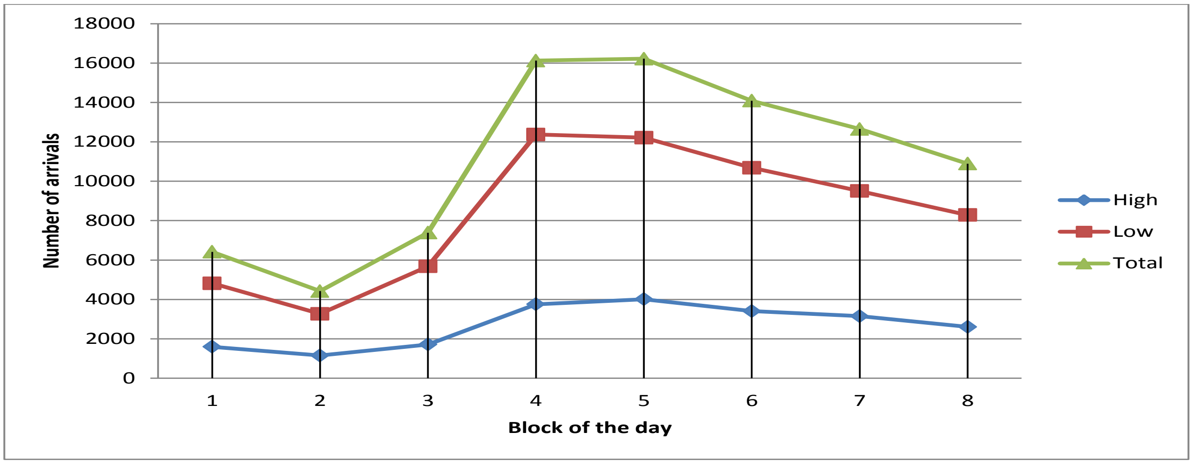


Figure 4.5: Number of arrivals in an Emergency Department by month of the year, Jan 2012 to Dec 2013.

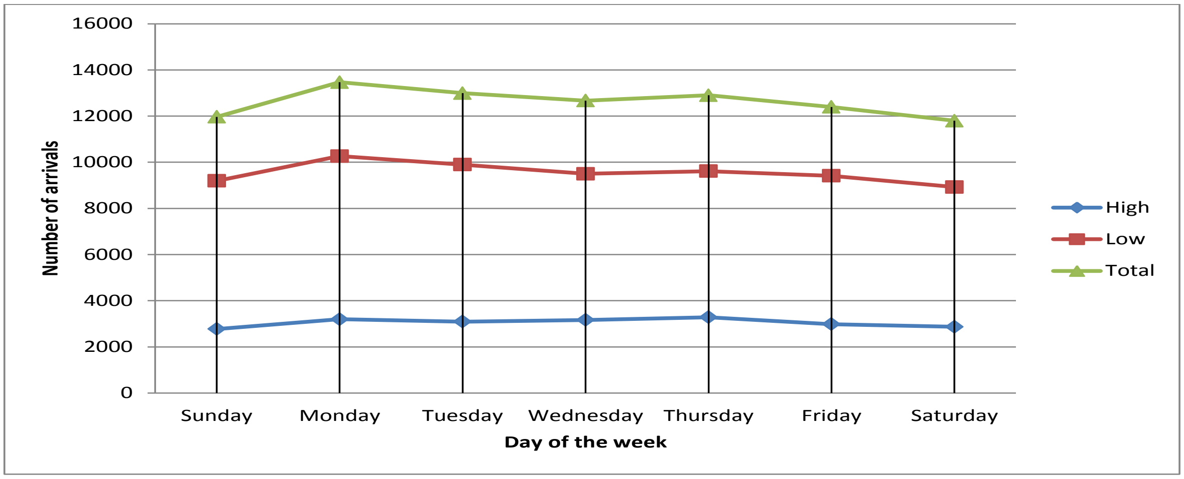


Figure 4.6: Number of arrivals in an Emergency Department by month of the year, Jan 2012 to Dec 2013.

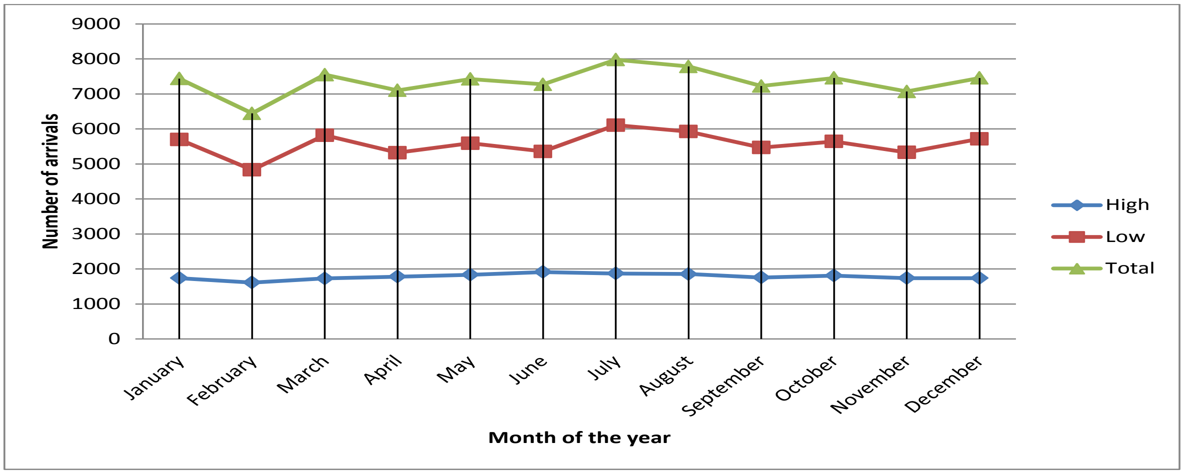


Figure 4.7: Number of arrivals in an Emergency Department by month of the year, Jan 2012 to Dec 2013.

4.4.2 Univariate Analysis

A complete univariate analysis was conducted on calendar variables as presented in table 4.6 for high and low acuity patients, respectively. I tested for the hypothesis that a given block mean arrivals is significantly different than the overall block mean. Similarly, for the day of week and for the month of year, I tested whether a particular day's and month's mean arrival are significantly different from the overall daily and monthly means, respectively. All the block means except block-8 are significantly different from the overall block mean for both the high and low acuity patient arrivals. Among days of the week, the mean arrivals for Saturday, Sunday, and Thursday were significantly different from the overall daily mean for high acuity patients, whereas Monday and Wednesday were moderately significant at 10% level of significance. It was apparent from visual inspection that Saturday and Sunday had lower mean patient arrivals than others and the statistical test confirmed this. Only Sunday and Monday showed a significant difference in mean arrivals for the case of low acuity patients. None of the monthly mean arrivals manifested a significant difference from the overall mean for the high acuity patient due to large variability caused by the small number of observations for monthly data. However for the low acuity patients, the month of July and November showed a significant difference in mean monthly arrivals as compared to the overall mean. The two years were not significantly different from each other.

Table 4.6: Univariate analysis of the calendar (predictor) variables for the high and low acuity arrivals.

	high			low		
Predictors	Mean	St Dev	p-value	Mean	St Dev	p-value
Block1	2.18	1.49	< 0.001	6.58	2.74	< 0.001
Block2	1.57	1.32	< 0.001	4.47	2.22	< 0.001
Block3	2.34	1.58	< 0.001	7.76	2.95	< 0.001
Block4	5.13	2.41	< 0.001	16.90	4.42	< 0.001
Block5	5.48	2.50	< 0.001	16.68	4.31	< 0.001
Block6	4.65	2.14	< 0.001	14.59	4.16	< 0.001
Block7	4.30	2.18	< 0.001	12.98	3.76	< 0.001
Block8	3.56	1.87	0.196	11.31	3.41	0.453
Mon	30.46	7.14	0.070	96.02	23.91	0.045
Tues	29.47	5.87	0.622	93.95	19.34	0.165
Wed	30.17	5.87	0.091	91.91	20.39	0.765
Thur	31.51	6.81	< 0.001	92.98	21.01	0.420
Fri	28.70	5.61	0.374	89.27	19.54	0.290
Sat	27.61	5.52	0.004	88.28	21.08	0.146
Sun	26.40	5.40	< 0.001	86.73	20.20	0.022
Jan	849.0	25.45	0.268	2799.5	26.16	0.475
Feb	806.0	50.91	0.260	2415.0	281.42	0.318
Mar	864.5	67.17	0.695	2910.0	253.14	0.599
Apr	890.0	107.48	0.993	2661.0	173.94	0.511
May	917.5	79.90	0.704	2795.5	82.73	0.830
Jun	957.0	48.08	0.295	2678.5	149.19	0.513
Jul	935.5	113.84	0.667	3053.5	3053.5	0.005
Aug	928.5	51.61	0.476	2962.5	112.42	0.260
Sep	878.0	72.12	0.862	2734.5	109.60	0.665
Oct	905.5	36.06	0.637	2821.5	26.16	0.264
Nov	869.5	23.33	0.444	2664.5	9.19	0.035
Dec	869.0	35.35	0.567	2858.0	52.32	0.280

4.4.3 Forecasting Outputs with Seasonal Time-Series Models

The fundamental requirement to apply Box-Jenkins ARIMA models to time series data is the assumption of stationarity. Therefore, the first step of a time series analysis is to check whether the data is stationary. A simple time series plot can aid in detecting stationarity in the data. Figure 4.8 shows time series plots for the first 3 months of high and low acuity patients block arrival data which reflects the behaviour of the whole 2 year of block data. It appears that both the high and low acuity patient arrival data satisfies the assumption of stationarity. Both the Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt–Shin (KPSS) (Greene (2011)) tests reject the hypothesis of non-stationarity.

Table 4.7: Test for stationarity time-series.

		high		low	
Tests	Null Hypothesis	p-value	Decision	p-value	Decision
ADF	Not Trend Stationary	< 0.01	Reject	< 0.01	Reject
KPSS	Not Trend Stationary	< 0.01	Reject	< 0.01	Reject

Figures 4.9 and 4.10 represents the ACF and PACF (defined in chapter 2) of the original and seasonal differenced high and low acuity block arrivals, respectively. The dotted lines in ACF and PACF plots provide the values beyond which the autocorrelations are significantly different from zero. Starting with a perfect correlation at lag 0, the ACF of the high and low acuity arrivals show that there is a positive correlation at lag 1 followed by negative correlations at lags 2, 3, 4, 5, and 6, subsequently followed by positive correlations at lags 7, 8, and 9. A similar pattern manifested in the lags 10 to 14, 15 to 17, 18 to 22, and so on. On the other hand, the PACF of the high and low acuity arrivals illustrate that the majority of the correlations in the data are accounted for by the first 8 lags, which covers a 24-hour period, and the correlations reduced significantly thereafter. This motivates us to take a seasonal difference of our time-series data at lag 8 which was also manifested in our descriptive analysis where ED arrivals have highest peaks at late afternoons and becomes virtually empty at early mornings (daily

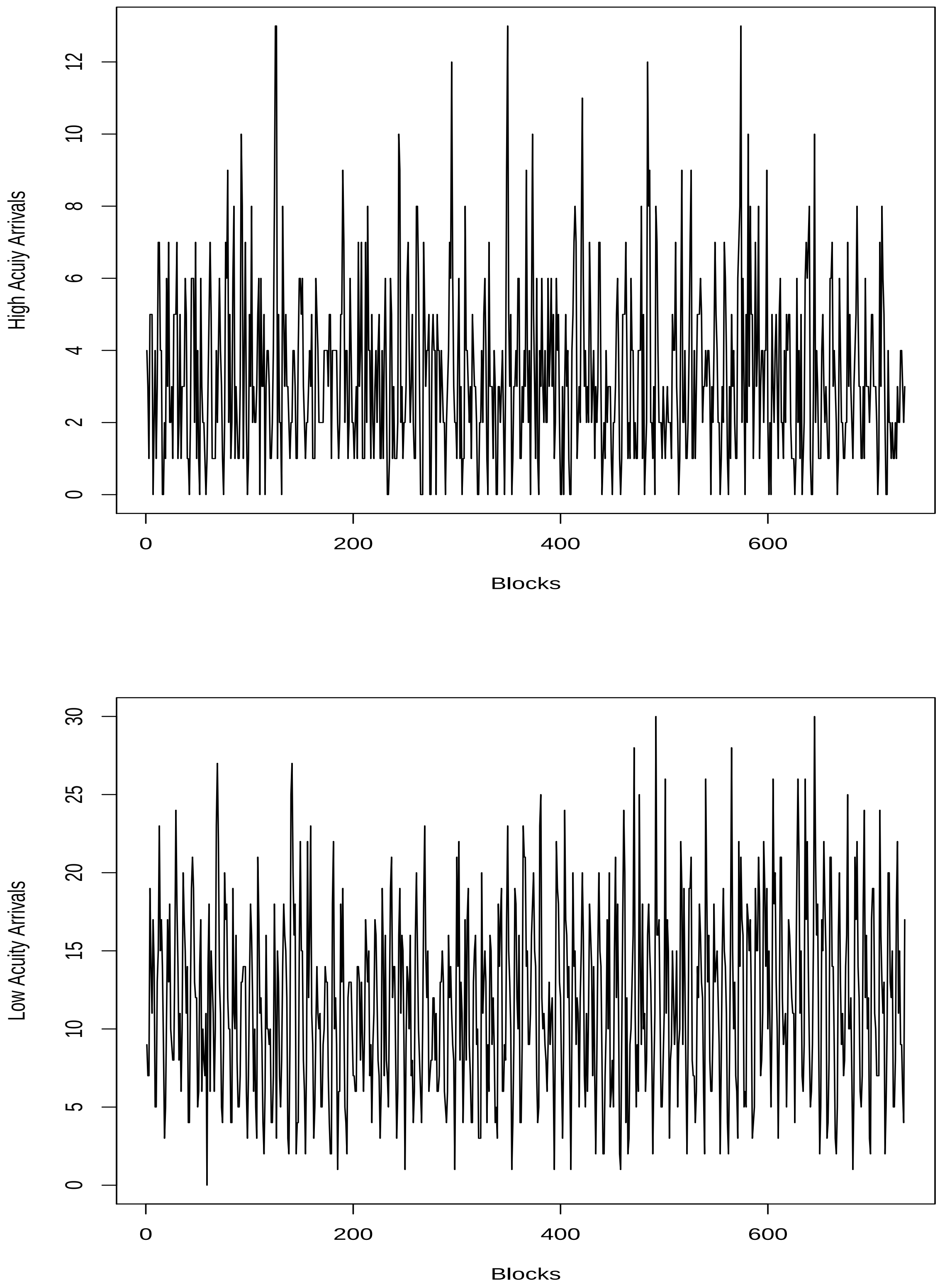


Figure 4.8: Time series plot for the first three months (year 2012) of high and low acuity block arrival data.

cycle consists of 8 3-hour blocks).

The ACF and PACF of the differenced data in Figure 4.10 now display only a few large spikes. Seasonal and non-seasonal autoregressive (AR) and moving average (MA) terms to be included in our initial Seasonal ARIMA model are identified by the ACF and PACF plots. The fact that there are only a few spikes at the beginning of the ACF and PACF plot suggests that zero or at most one or two nonseasonal AR and MA terms are required. A large spike at lag 8 of the ACF plot suggest inclusion of a seasonal MA term in the model, whereas 4 large spikes in the PACF plot indicates the necessity of 4 seasonal autoregressive terms in the model. Therefore, we start by fitting an $ARIMA_{(0,0,0),(4,1,1)_8}$ model followed by different combination of the non-seasonal and seasonal AR, MA terms.

Table 4.8 listed 13 SARIMA models that produced comparatively low AIC_c (defined in chapter 2) values. The smaller the values of AIC_c, the better the model. In addition to AIC_c for training data, I have also reported forecasting accuracy of the SARIMA models on the test data. For comparative purposes, starting with the initial SARIMA model, a variety of SARIMA models were considered, changing one parameter at a time (Table 4.8). Among all these models, the $ARIMA_{(2,0,1),(4,1,1)_8}$ and $ARIMA_{(1,0,2),(4,1,3)_8}$ are the preferred models for high and low acuity arrivals, respectively. As criteria, the Smallest AIC_c, RMSE, MAE, and, MAPE values were used to select the best model. Parameter estimates and their respective standard errors for $ARIMA_{(2,0,1),(4,1,1)_8}$ and $ARIMA_{(1,0,2),(4,1,3)_8}$ are provided in tables 4.9 and 4.10, respectively. ACF and PACF plots of the residuals in Figure 4.11 show that all but a few spikes exceed the upper and lower limit and hence provide a good fit. Out of sample forecast for future 7 days for 56 blocks along with 95% confidence band is presented in Figure 4.12 for the high and low acuity block arrivals.

The best fitted SARIMA model for the high acuity patient arrival is $ARIMA_{(2,0,1),(4,1,1)_8}$. This means that the data needs to be differenced seasonally once (at lag 8 for block arrivals) and the response series (the high acuity arrivals) depend on four seasonally lagged previous observations, one seasonally lagged error term, two previous observations, and one previous

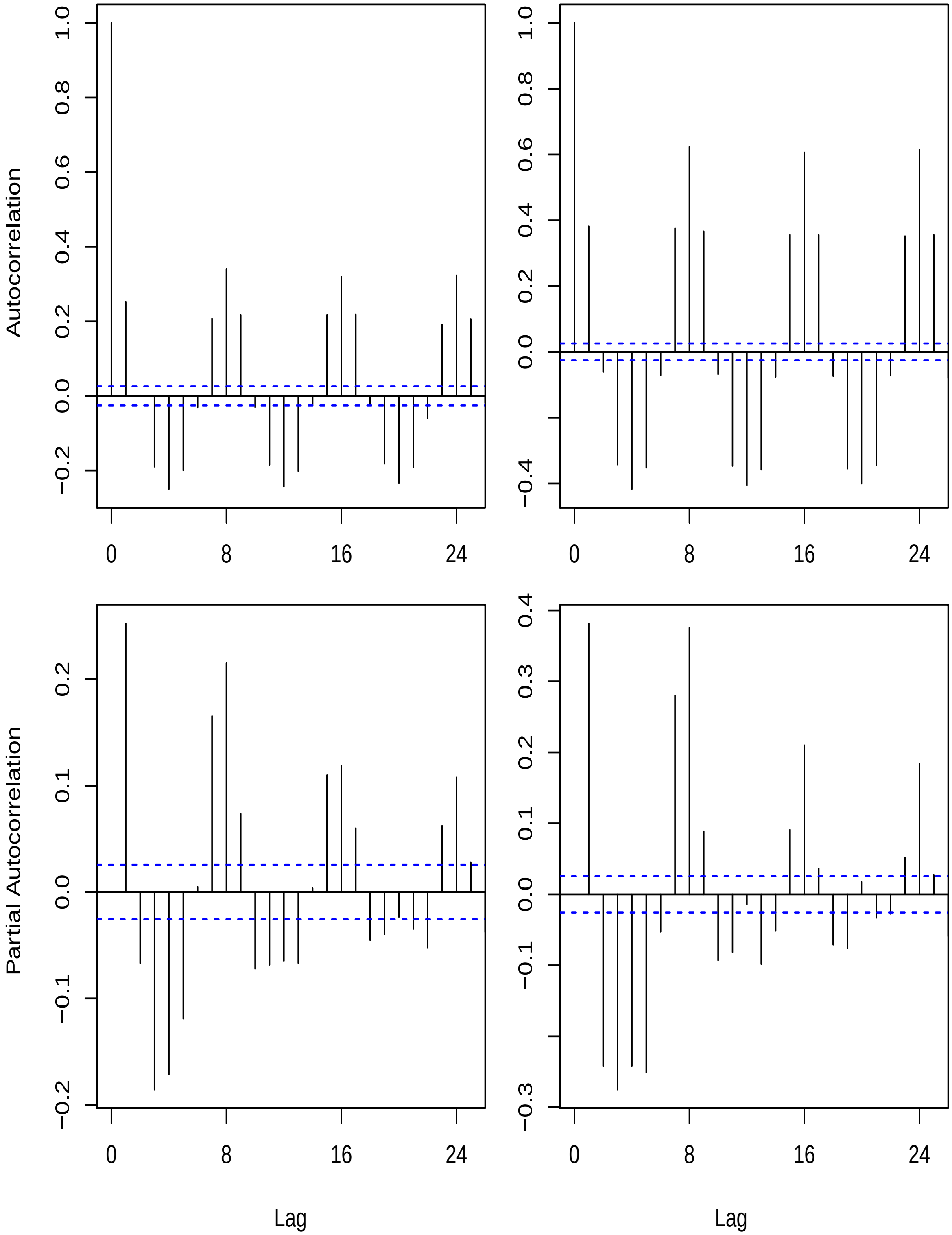


Figure 4.9: ACF of high and low acuity block arrivals (top row) and PACF for the high and low acuity block arrivals (bottom row), respectively.

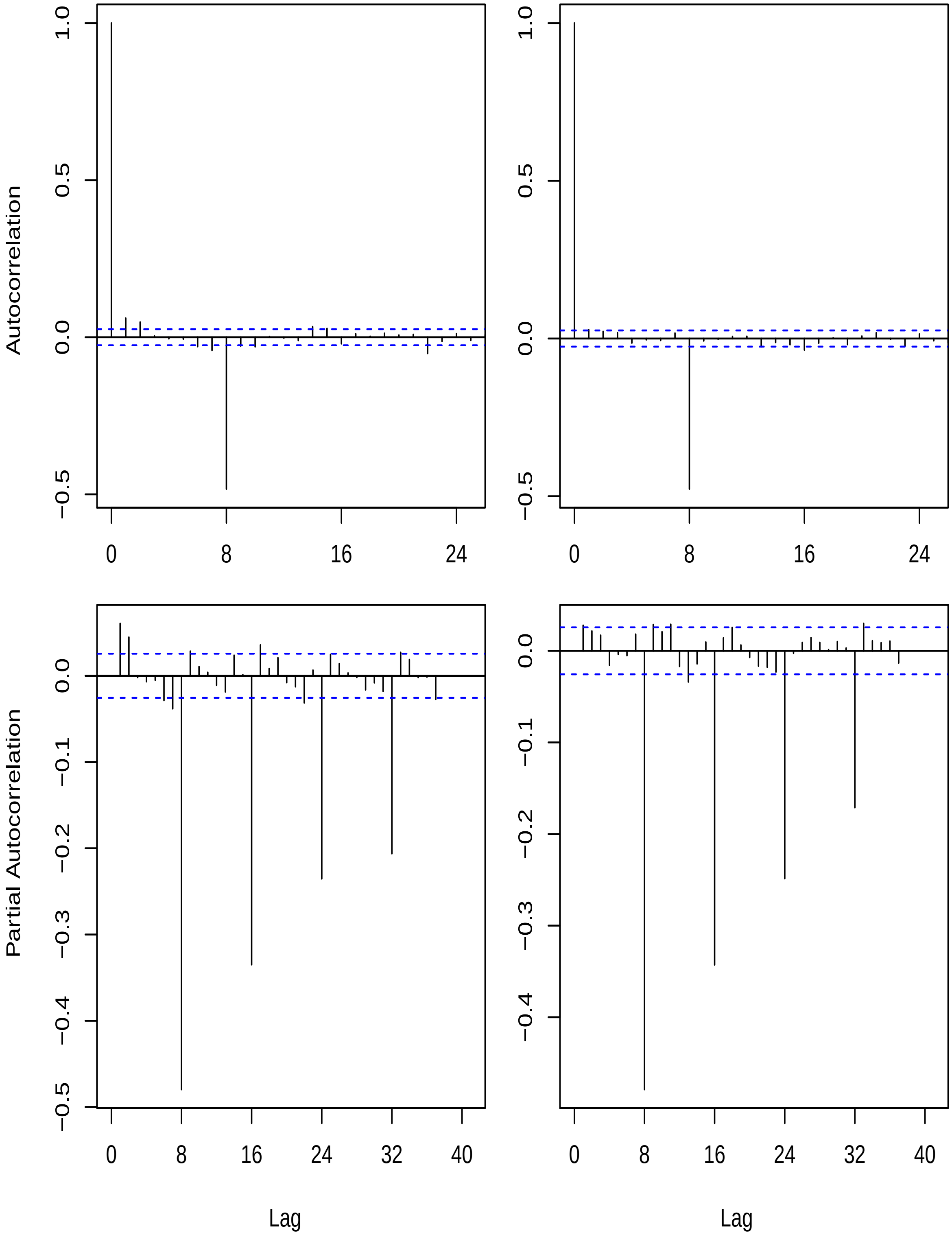


Figure 4.10: ACF of the seasonally difference (order 8) high and low acuity block arrivals (top row) and PACF of the seasonally difference (order 8) high and low acuity block arrival (bottom row), respectively.

Table 4.8: Performance of different seasonal ARIMA models for high and low acuity arrivals.

	high				low			
Models	AIC_c	RMSE	MAE	MAPE	AIC_c	RMSE	MAE	MAPE
$ARIMA_{(0,0,0)(4,1,1)_8}$	22535.37	2.013	1.577	91.803	28809.23	3.845	2.928	99.795
$ARIMA_{(1,0,0)(4,1,1)_8}$	22523.21	2.013	1.578	91.803	28810.91	3.852	2.937	99.795
$ARIMA_{(0,0,1)(4,1,1)_8}$	22509.62	2.013	1.578	91.803	28793.15	3.843	2.927	99.795
$ARIMA_{(1,0,1)(4,1,1)_8}$	22501.77	2.013	1.578	91.803	28748.76	3.836	2.927	99.795
$ARIMA_{(1,0,2)(4,1,1)_8}$	22501.37	2.013	1.578	91.803	28743.26	3.835	2.928	99.795
$ARIMA_{(2,0,1)(4,1,1)_8}$	**22500.80**	**2.012**	**1.577**	**91.803**	28744.24	3.853	2.935	99.795
$ARIMA_{(2,0,2)(4,1,1)_8}$	22502.97	2.013	1.578	91.803	28745.04	3.835	2.925	99.795
$ARIMA_{(1,0,2)(4,1,2)_8}$	22504.26	2.013	1.578	91.803	28743.94	3.835	2.928	99.795
$ARIMA_{(1,0,2)(4,1,3)_8}$	22505.20	2.012	1.578	91.803	**28690.81**	**3.802**	**2.887**	**99.795**
$ARIMA_{(1,0,2)(4,1,4)_8}$	22507.59	2.012	1.579	91.803	48745.21	3.848	2.931	99.795
$ARIMA_{(1,0,2)(5,1,3)_8}$	22507.32	2.013	1.578	91.803	28747.57	3.835	2.929	99.795
$ARIMA_{(1,0,2)(5,1,2)_8}$	22505.22	2.013	1.578	91.803	28745.66	3.836	2.929	99.795
$ARIMA_{(1,0,2)(5,1,1)_8}$	22507.51	2.013	1.578	91.803	28744.83	3.837	2.929	99.795

error term. Similarly for the best fitted SARIMA model for the low acuity patient arrivals, the $ARIMA_{(2,0,1),(4,1,1)_8}$ model means that the data needs to be differenced seasonally (at lag 8 for block arrivals) once and the response series (high acuity arrivals) depends on four seasonally lagged previous observations, three seasonally lagged error terms, one previous observations, and two previous error terms.

Table 4.9: Parameter estimates of seasonal $ARIMA_{(2,0,1)(4,1,1)_8}$ model for high acuity arrivals.

	ϕ_1	ϕ_2	θ_1	Φ_1	Φ_2	Φ_3	Φ_4	Θ_1
Coefficient	0.1552	0.0398	-0.0820	0.0061	-0.0291	-0.0136	-0.0297	-0.9895
St. Error	0.0873	0.0135	0.0131	0.0142	0.0132	0.0139	0.0132	0.0025

Table 4.10: Parameter estimates of seasonal $ARIMA_{(1,0,2)(4,1,3)_8}$ model for low arrivals.

	ϕ_1	θ_1	θ_2	Φ_1	Φ_2	Φ_3	Φ_4	Θ_1	Θ_2	Θ_3
Coefficient	0.9942	-0.9504	-0.0285	-0.4172	-0.9802	0.0249	-0.0003	-0.5539	0.5416	-0.9755
St. Error	0.0038	0.0141	0.0138	0.0174	0.0168	0.0169	0.0146	0.0110	0.0090	0.0078

Forecasts for future high and low acuity block arrivals are made using the above parameter estimates and corresponding values for lagged arrivals and error terms.

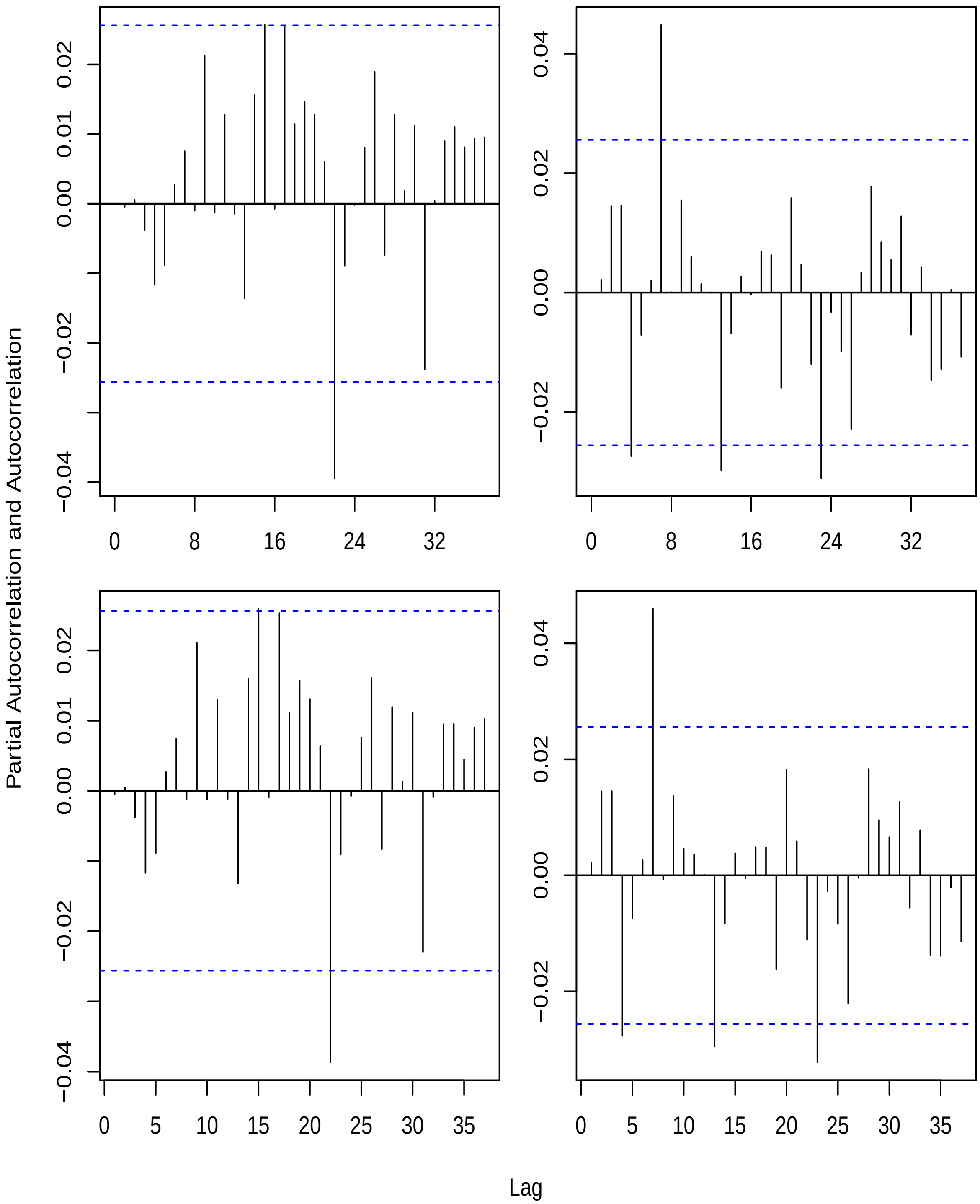


Figure 4.11: ACF of the residuals of the final model for high and low acuity block arrivals (top row) and PACF of the residuals of the final model for high and low acuity block arrivals (bottom row).

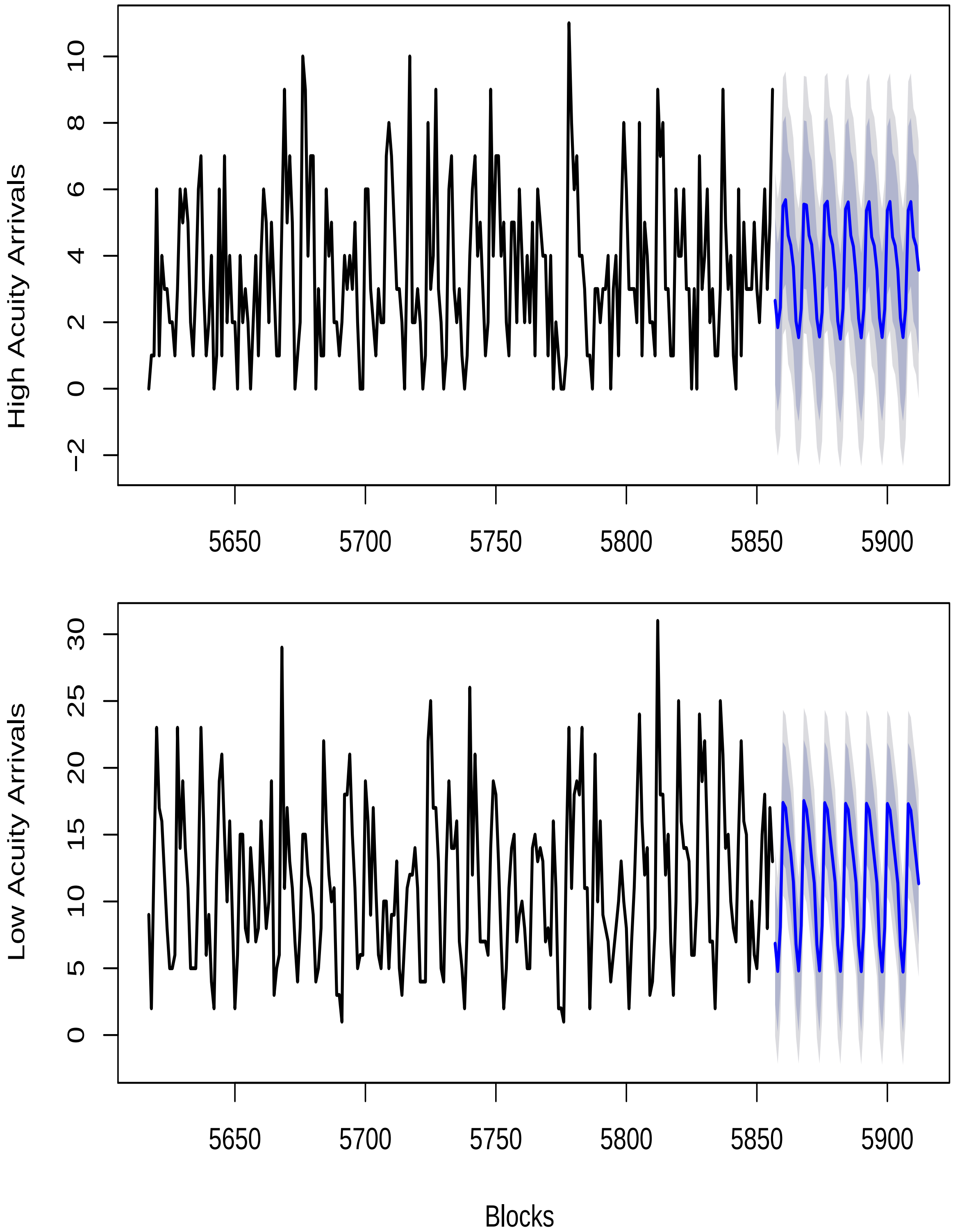


Figure 4.12: One month of preceding block arrival data with 56 blocks (7 days) ahead forecast with 95% confidence interval for high (top) and low (bottom) acuity arrivals.

4.4.4 Forecasting Outputs for GLM with Calendar Variables

We modeled here the log of the expected mean arrival per block as a function of calendar variables, such as, block of the day, day of the week, and months of the year. The reason for using the log link function is to safeguard the expected arrival value from producing a negative value. Since the dependent variable is a count, we used both Poisson and Negative Binomial regression. The accuracy of the Poisson and negative binomial regression models for the high and low acuity arrivals were compared in terms of AIC, RMSE, MAE, and MAPE in table 4.11. Both the models produced similar results, with the Poisson model marginally superior than the Negative Binomial model. Moreover, Poisson regression is a one parameter model and is less complicated than the negative binomial model. Therefore, I advocate the use of the Poisson regression model as long as both models produce similar results. Comparison of the GLM models with SARIMA and GLARMA models will be discussed later in this chapter.

Table 4.11: Performance comparison of GLM models.

	high				low			
Models	AIC	RMSE	MAE	MAPE	AIC	RMSE	MAE	MAPE
Poisson	21525	2.040	1.624	91.820	28049	3.763	2.838	99.795
Negative Binomial	21528	2.040	1.624	91.820	28046	3.763	2.839	99.795

Results of the parameter estimates along with their respective standard errors and p-values for both the Poisson and negative binomial regression models for the high and low acuity arrivals are presented in table 4.12 and 4.13, respectively. For the block variables I have considered the midnight block (12 A.M. - 3 A.M.) as the reference category so that every other block can be compared with the midnight block. Similarly for the variables day of the week, month of the year, and year, Sunday, January, and the year 2012 were used as reference categories, respectively.

When comparing with Block 1 (12 A.M. - 3 A.M.), for both the high and low acuity arrivals, all other blocks but block 2 (12 A.M. - 3 A.M.) were significant with p-values (< 0.001). Except for block 2 (12 A.M. - 3 A.M.), the remaining blocks have positive coefficients meaning

Table 4.12: Parameter estimates of Poisson and Negative Binomial model for high acuity arrivals.

Regressors	Poisson			Negative Binomial		
	Coeff	SE	p-value	Coeff	SE	p-value
Intercept	0.818	0.039	< .001	0.818	0.039	< .001
3 A.M.-6 A.M.	-0.279	0.040	< .001	-0.279	0.040	< .001
6 A.M.-9 A.M.	0.063	0.036	0.080	0.063	0.036	0.080
9 A.M.-12 P.M.	0.697	0.031	< .001	0.697	0.031	< .001
12 P.M.-3 P.M.	0.759	0.031	< .001	0.759	0.031	< .001
3 P.M.-6 P.M.	0.608	0.031	< .001	0.608	0.031	< .001
6 P.M.-9 P.M.	0.551	0.032	< .001	0.551	0.032	< .001
9 P.M.-12 A.M.	0.375	0.033	< .001	0.375	0.033	< .001
Monday	0.142	0.027	< .001	0.142	0.027	< .001
Tuesday	0.110	0.027	< .001	0.110	0.027	< .001
Wednesday	0.119	0.027	< .001	0.119	0.027	< .001
Thursday	0.164	0.026	< .001	0.164	0.026	< .001
Friday	0.073	0.027	0.007	0.073	0.027	0.007
Saturday	0.023	0.027	0.391	0.023	0.027	0.392
February	0.008	0.034	0.806	0.008	0.034	0.806
March	0.012	0.034	0.718	0.012	0.034	0.718
April	0.058	0.033	0.084	0.058	0.033	0.084
May	0.053	0.033	0.111	0.053	0.033	0.111
June	0.117	0.033	< .001	0.117	0.033	< .001
July	0.079	0.033	0.017	0.079	0.033	0.017
August	0.060	0.033	0.073	0.060	0.033	0.073
September	0.052	0.034	0.120	0.052	0.034	0.120
October	0.051	0.033	0.124	0.051	0.033	0.124
November	0.046	0.041	0.260	0.046	0.041	0.260
December	0.049	0.041	0.229	0.049	0.041	0.229
Year 2013	-1.076	0.039	< .001	-1.076	0.039	< .001

Table 4.13: Parameter estimates of Poisson and Negative Binomial model for low acuity arrivals.

Regressors	Poisson			Negative Binomial		
	Coeff	SE	p-value	Coeff	SE	p-value
Intercept	1.834	0.022	< .001	1.835	0.023	< .001
3 A.M.-6 A.M.	-0.399	0.023	< .001	-0.399	0.023	< .001
6 A.M.-9 A.M.	0.155	0.020	< .001	0.155	0.020	< .001
9 A.M.-12 P.M.	0.932	0.017	< .001	0.931	0.018	< .001
12 P.M.-3 P.M.	0.925	0.018	< .001	0.925	0.018	< .001
3 P.M.-6 P.M.	0.790	0.018	< .001	0.790	0.018	< .001
6 P.M.-9 P.M.	0.677	0.018	< .001	0.677	0.018	< .001
9 P.M.-12 A.M.	0.544	0.018	< .001	0.544	0.019	< .001
Monday	0.111	0.015	< .001	0.110	0.015	< .001
Tuesday	0.076	0.015	< .001	0.074	0.015	< .001
Wednesday	0.035	0.015	0.021	0.033	0.015	0.031
Thursday	0.054	0.015	< .001	0.053	0.015	< .001
Friday	0.035	0.015	0.022	0.033	0.015	0.032
Saturday	-0.018	0.015	0.233	-0.018	0.015	0.241
February	-0.061	0.019	0.001	-0.062	0.020	0.002
March	0.042	0.018	0.021	0.042	0.019	0.025
April	-0.018	0.019	0.332	-0.018	0.019	0.336
May	-0.000	0.018	0.984	-0.000	0.019	0.982
June	-0.006	0.019	0.715	-0.007	0.019	0.718
July	0.085	0.018	< .001	0.086	0.019	< .001
August	0.059	0.018	0.001	0.058	0.019	0.002
September	0.014	0.019	0.447	0.014	0.019	0.460
October	0.007	0.018	0.700	0.006	0.019	0.720
November	-0.009	0.023	0.691	-0.009	0.023	0.695
December	0.036	0.022	0.114	0.036	0.023	0.118
Year 2013	-0.001	0.014	0.972	-0.001	0.014	0.965

that arrivals during those blocks are higher on average relative to block 1. For days of the week, all the weekdays appeared to be significant with p-values (< 0.001) when compared with Sunday. Note that Saturday was not significantly different than Sunday. Positive coefficient values for weekdays (Monday to Friday) suggest that average arrivals on those days are higher compared to that of Sundays. The months of June and July were significant for both the high and low acuity arrivals, while the month of February was only significantly different for the low. Average arrivals during July and August are higher relative to January, whereas the average arrival for February is lower than that of January. This is partly due to the month of February having comparably fewer days than other months of a year. Arrivals for the year 2013 were significantly different than for year 2012 for high acuity arrivals but were not significantly different for low acuity arrivals. This implies that even though population growth and aging population contributes to an increase in average arrivals year after year, this may not be to a statically significant degree between any two consecutive years.

Tables 4.12 and 4.13 shows the effect of block of day, day of week, and month of year on the mean high and low acuity patient arrivals, respectively. A significant positive (negative) coefficient value indicates an increase (decrease) in the mean arrivals of that particular category as compared to the reference category. For instance, from Table 4.12, coefficient for Monday (0.142) tells us that compared to Sunday, on average Monday will have approximately 15.26% ($e^{0.142} - 1 = 0.1526$) more high acuity patient arrivals, provided other variables remain constant. Similarly from table 4.13, coefficient for February (-0.061) tells us that compared to January, on average February will have 5.92% ($1 - e^{-0.061} = 0.0592$) less low acuity patient arrivals, provided other variables remain constant.

We presented diagnostic plots for Poisson regression only. Figure 4.14 presents four subplots for model diagnostics for each of high and low acuity arrivals, respectively. I observed that the Poisson model is a good predictive model for our data. All the model assumptions are also satisfied, as is evident from Figure 4.14. No significant pattern in the residual versus predicted plot suggests that homogeneity of variance assumption is satisfied. The Normal Q-Q plot demonstrated that the normality assumption is satisfied, because all the points are on or

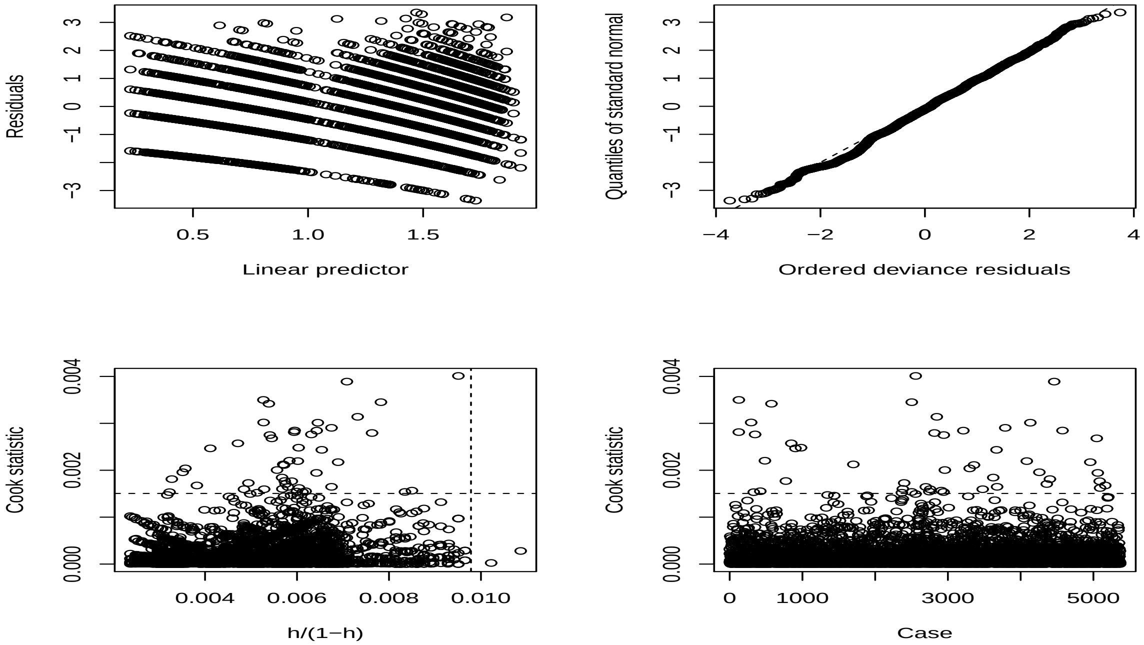


Figure 4.13: Residual analysis of the fitted Poisson model for the high acuity arrivals.

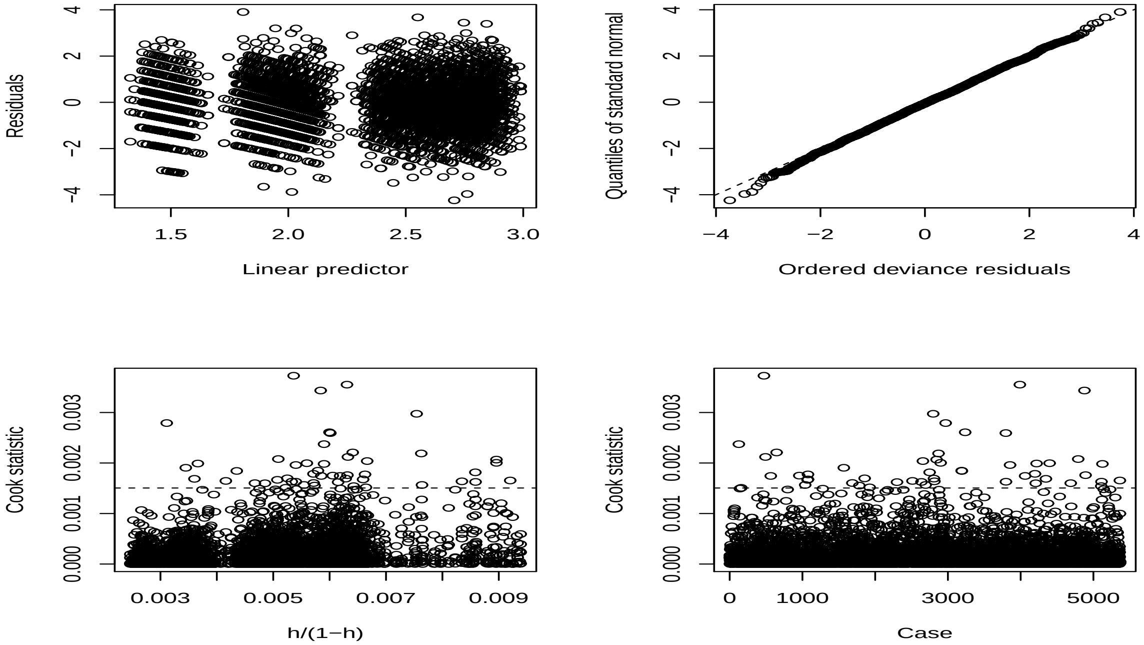


Figure 4.14: Residual analysis of the fitted Poisson model for the low acuity arrivals.

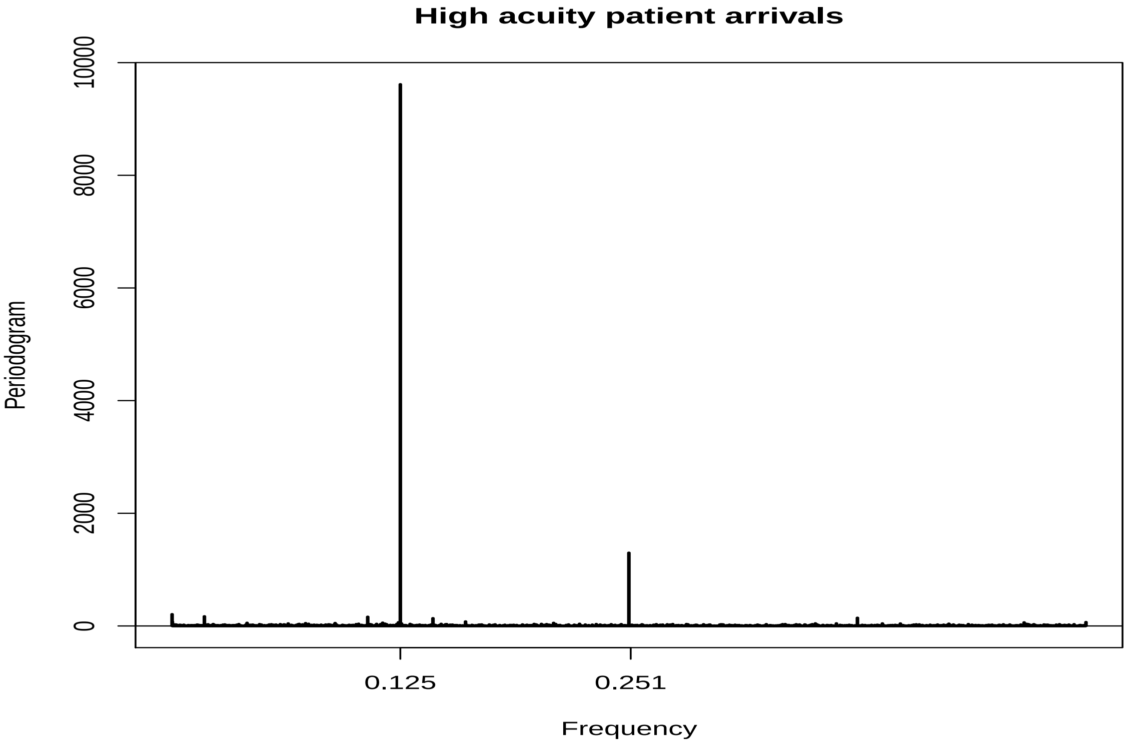
near the straight line. The Cook statistic shows what are the influential observations in the data set.

4.4.5 Forecasting Outputs for Harmonic Regression

In this section, we present the results from fitting the Harmonic regression model to our time series data. Both the GLM (Poisson) and the ARIMA regression models were fitted considering the sine and cosine functions as covariates to predict the high and low acuity block arrivals. It turns out that the use of Fourier terms and the calendar variables as predictors serves the same purpose, i.e., both sets of predictor capture the seasonal variability that present in our data. Therefore they are multi-collinear (correlated predictors), and both sets of predictors in the model add no extra information as compared to the one set in the model. The output for Harmonic Regression thus omits calendar variables as predictors.

The first step of the Harmonic regression is to determine an appropriate set of frequencies (wavelengths) that are required to complete a full seasonal cycle. The Pediogram is a very useful tool in describing the frequencies of a time series data. It is a Fast Fourier Transform (FFT) based non parametric method which is used to identify the dominant periods (or frequencies) of a time series. The Periodogram plot shows a high peak at a frequency if the time series has a strong sinusoidal signal at that frequency.

Figure 4.15 represents the periodogram plots for high and low acuity block arrivals, respectively. A dominant long spike at frequency 0.125 and a short spike at frequency 0.251 were manifested for both the high and low acuity arrivals. This means that strong sinusoidal signal appeared to be present at those frequencies. The reciprocal of the frequencies determines the period required to complete a full sinusoidal cycle. Therefore, I used 8 and 4 as our periods in the sine and cosine terms of our Harmonic regression model. The 8-period (24-hour) cycle is a-priori expected because of the day of the week effect, however, the more interesting is the 4-period (12-hour) cycle. I suspect this is due to the fact that most blocks 12 hours apart are subject to very different levels of congestion.



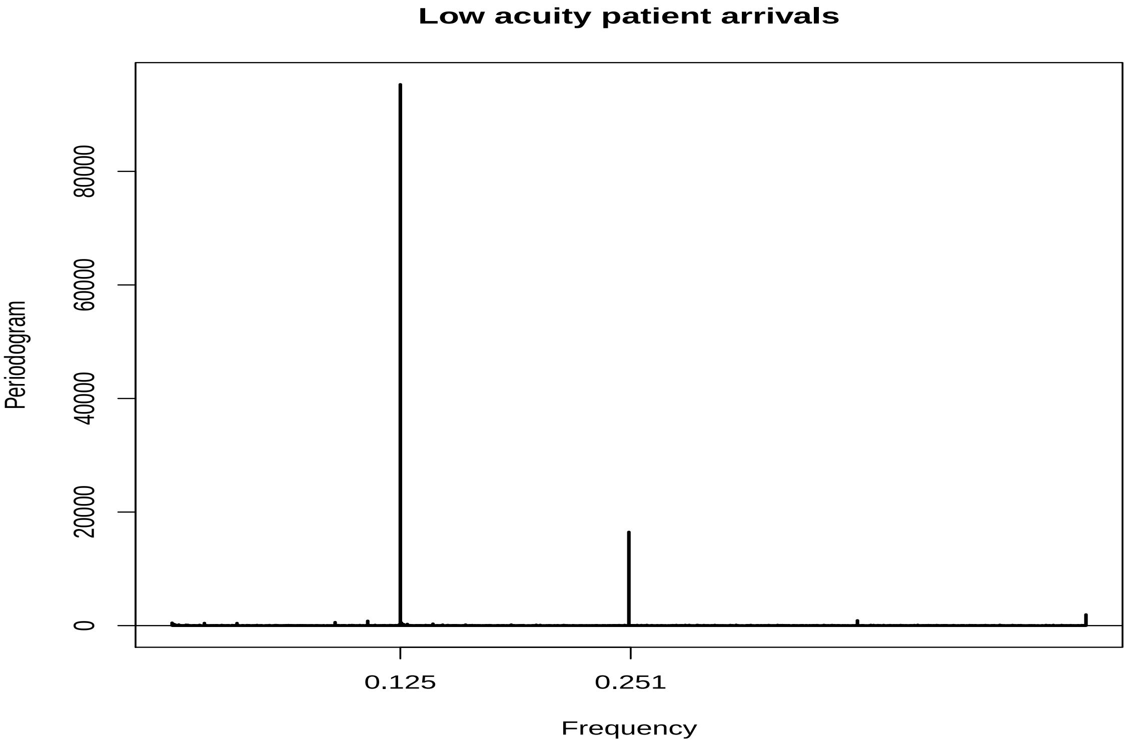


Figure 4.15: Periodogram plots for the high and low acuity arrivals.

The accuracy measures to compare relative performances of the Harmonic regression based on GLM and time series models for the high and low acuity arrivals were presented in table 4.14. Although accuracy measures are very similar for both modeling strategies, Harmonic regression based on GLM is slightly superior to Harmonic regression based on the time series model for our block arrival data.

Table 4.14: Performance comparison of the GLM and ARIMA (including harmonic terms).

		high			low	
Models	RMSE	MAE	MAPE (%)	RMSE	MAE	MAPE (%)
GLM (Poisson)	1.985	1.551	91.803	3.891	2.954	99.795
ARIMA	1.992	1.558	91.803	3.914	2.997	99.795

The parameter estimates, standard errors, and the significance levels of the Fourier terms of our Harmonic regression model based on GLM appear in table 4.15. Although all the coefficients were statistically significant, the coefficient for sin($2\pi t/4$) is marginally significant while other coefficients were highly significant. Similar output for the Harmonic regression based on ARIMA model for the high and low acuity arrivals were presented in tables 4.16 and 4.17.

Table 4.15: Parameter estimates of Poisson model with harmonic terms as predictors for high and low arrivals.

Regressors		high			low	
	Coeff	SE	p-value	Coeff	SE	p-value
Intercept	1.214	0.007	$< .001$	2.345	0.004	$< .001$
sin($2\pi t/8$)	-0.541	0.011	$< .001$	-0.547	0.006	$< .001$
cos($2\pi t/8$)	-0.150	0.010	$< .001$	-0.173	0.005	$< .001$
sin($2\pi t/4$)	0.030	0.010	0.003	0.011	0.005	0.058
cos($2\pi t/4$)	0.229	0.010	$< .001$	0.266	0.006	$< .001$

Table 4.16: Parameter estimates of $ARIMA_{(1,0,2)}$ model with Fourier terms as predictors for high arrivals.

	ϕ_1	θ_1	θ_2	Intercept	$\sin(2\pi t/8)$	$\cos(2\pi t/8)$	$\sin(2\pi t/4)$	$\cos(2\pi t/4)$
Coefficient	0.199	-0.124	0.042	3.664	-1.711	-0.621	0.251	0.604
St. Error	0.218	0.217	0.023	0.031	0.040	0.040	0.036	0.036

Table 4.17: Parameter estimates of $ARIMA_{(1,0,2)}$ model with fourier terms as predictors for low arrivals.

	ϕ_1	θ_1	θ_2	Intercept	$\sin(2\pi t/8)$	$\cos(2\pi t/8)$	$\sin(2\pi t/4)$	$\cos(2\pi t/4)$
Coefficient	0.993	-0.959	-0.018	11.436	-5.262	-2.197	0.637	2.300
St. Error	0.003	0.014	0.013	0.158	0.069	0.069	0.068	0.068

4.4.6 Forecasting Outputs for the Generalized Linear Autoregressive Moving Average model

Under the GLARMA model framework, I fit a Poisson regression model with calendar variables as predictors with different combinations of autoregressive and moving average terms. There are two significant autoregressive terms and one significant moving average term included in the model for high acuity arrivals (GLARMA(2,1)). However, for the low acuity arrivals only one significant autoregressive term and two significant moving average terms were included (GLARMA(1,2)). The full model and reduced model outputs have been documented in tables 4.19 and 4.20, respectively. I observed that the same predictors were significant as seen in ARIMA or GLM models, with the exception of the autoregressive and moving average term included in the model.

For high acuity arrivals, a significant Likelihood Ratio test (p-value$<$ 0.001) and a Wald test (p-value$<$ 0.001) ensures the appropriateness of GLARMA model over GLM having the same structure of the covariates. GLARMA(2,1) was the best fitted model for high acuity arrivals with lowest AIC value (20495). Similarly, for low acuity arrivals, significant Likelihood Ratio test (p-value=0.007) and Wald test (p-value=0.008) ensures the appropriateness of the

GLARMA model over the GLM having the same structure of the covariates. GLARMA(1,2) was the best fitted model for low acuity arrivals with lowest AIC value (20675).

Forecast accuracy measures for the high and low acuity GLARMA models were documented in table 4.18. Comparing accuracy measures of GLARMA model with that of SARIMA, GLM, and Harmonic regression, we observed that the GLARMA model produced the lowest values for AIC, MAE, and RMSE. Therefore, GLARMA model should be considered the best model to produce forecast for our ED block arrival data.

All the mean block arrivals were significant when compared with the reference midnight block 1 (12 A.M.-3 A.M.) for both the high and low acuity arrivals. Compared to Sunday, all the weekdays were significantly different in mean high and low acuity arrivals but not the Saturday. Therefore, as suggested in our initial analysis, the day of the week variable was classified as Monday, weekend, with the other weekdays aggregated in the reduced model. Four of the months appeared to be significant relative to January, and among them, two were winter months (February and March) and the other two were summer months (July and August). The month was classified as February, July, and an aggregate for the rest of the year on the final model. Thus following our initial analysis, the year variable will be removed as a predictor in our reduced model. Therefore, the predictors for our final model consist of all categories of block variable, day of the week variable categorized as Monday, weekend, and other weekdays, month of the year variable categorized as February, July, and rest of the year, and one moving average term.

Table 4.18: Forecast accuracy measures for best fitted high and low acuity GLARMA models.

	high				low			
Models	AIC	RMSE	MAE	MAPE	AIC	RMSE	MAE	MAPE
GLARMA(2,1)	20495	1.070	0.827	91.801	–	–	–	–
GLARMA(1,2)	–	–	–	–	20675	1.139	0.861	99.792

Table 4.19: Parameter estimates of the full GLARMA model for high and low acuity arrivals.

		high			low	
Regressors	Coeff	SE	p-value	Coeff	SE	p-value
Intercept	0.848	0.029	< 0.001	1.853	0.022	< 0.001
3 A.M.-6 A.M.	-0.326	0.038	< 0.001	-0.386	0.022	< 0.001
6 A.M.-9 A.M.	0.069	0.034	0.042	0.164	0.019	< 0.001
9 A.M.-12 P.M.	0.854	0.029	< 0.001	0.943	0.016	< 0.001
12 P.M.-3 P.M.	0.920	0.029	< 0.001	0.930	0.017	< 0.001
3 P.M.-6 P.M.	0.757	0.030	< 0.001	0.796	0.017	< 0.001
6 P.M.-9 P.M.	0.678	0.030	< 0.001	0.679	0.017	< 0.001
9 P.M.-12 A.M.	0.490	0.031	< 0.001	0.542	0.017	< 0.001
Monday	0.141	0.027	< 0.001	0.112	0.015	< 0.001
Tuesday	0.105	0.027	< 0.001	0.074	0.015	< 0.001
Wednesday	0.130	0.027	< 0.001	0.033	0.015	0.028
Thursday	0.176	0.027	< 0.001	0.054	0.015	< 0.001
Friday	0.079	0.028	0.004	0.032	0.015	0.034
Saturday	0.038	0.028	0.169	-0.020	0.015	0.181
February	0.028	0.037	0.437	-0.063	0.020	0.002
March	0.016	0.036	0.644	0.042	0.019	0.030
April	0.077	0.036	0.033	-0.019	0.020	0.340
May	0.071	0.036	0.047	-0.001	0.019	0.953
June	0.154	0.035	< 0.001	-0.007	0.020	0.719
July	0.093	0.035	0.008	0.085	0.019	< 0.001
August	0.085	0.035	0.017	0.058	0.019	0.002
September	0.069	0.036	0.057	0.013	0.019	0.486
October	0.057	0.036	0.112	0.006	0.019	0.758
November	0.053	0.036	0.142	-0.014	0.020	0.460
December	0.023	0.036	0.527	0.020	0.019	0.286
Year 2013	0.082	0.014	< 0.001	-0.046	0.008	< 0.001
AR(1)	–	–	–	0.011	0.003	0.003
AR(2)	0.015	0.006	0.021	–	–	–
MA(1)	0.024	0.006	< 0.001	–	–	–
MA(2)	–	–	–	0.004	0.003	0.278

Table 4.20: Parameter estimates of the reduced GLARMA model for high and low acuity arrivals.

		high			low	
Regressors	Coeff	SE	p-value	Coeff	SE	p-value
Intercept	0.702	0.028	< 0.001	1.830	0.016	< 0.001
3 A.M.-6 A.M.	-0.325	0.037	< 0.001	-0.386	0.022	< 0.001
6 A.M.-9 A.M.	0.069	0.034	0.042	0.165	0.019	< 0.001
9 A.M.-12 P.M.	0.853	0.029	< 0.001	0.943	0.017	< 0.001
12 P.M.-3 P.M.	0.920	0.029	< 0.001	0.930	0.017	< 0.001
3 P.M.-6 P.M.	0.757	0.030	< 0.001	0.796	0.017	< 0.001
6 P.M.-9 P.M.	0.678	0.030	< 0.001	0.679	0.017	< 0.001
9 P.M.-12 A.M.	0.490	0.031	< 0.001	0.542	0.017	< 0.001
Monday	0.102	0.017	< 0.001	0.058	0.009	< 0.001
Weekend	0.119	0.024	< 0.001	0.121	0.013	< 0.001
February	-0.032	0.028	0.249	-0.073	0.016	< 0.001
July	0.031	0.026	0.232	0.075	0.014	< 0.001
AR(1)	–	–	–	0.013	0.003	< 0.001
AR(2)	0.020	0.006	0.001	–	–	–
MA(1)	0.030	0.006	< 0.001	–	–	–
MA(2)	–	–	–	0.007	0.003	0.066

A pictorial representation of the residual analysis of the fitted GLARMA model for high and low acuity arrivals are provided in Figures 4.16 and 4.17. The first subplot (top-left) is the ACF plot of the residuals, which indicates that the serial correlation present in the data has adequately been dealt with (all spikes within control limit). The residual versus time plot (top-right) satisfies the stationarity assumption of the time series. The Probability Integral Transform (PIT) histogram (bottom-left) suggests that the Poisson model for the counts is appropriate for this data, as all the bars are clustered around 1. The Q-Q plot (bottom-right) shows reasonable conformity with normality.

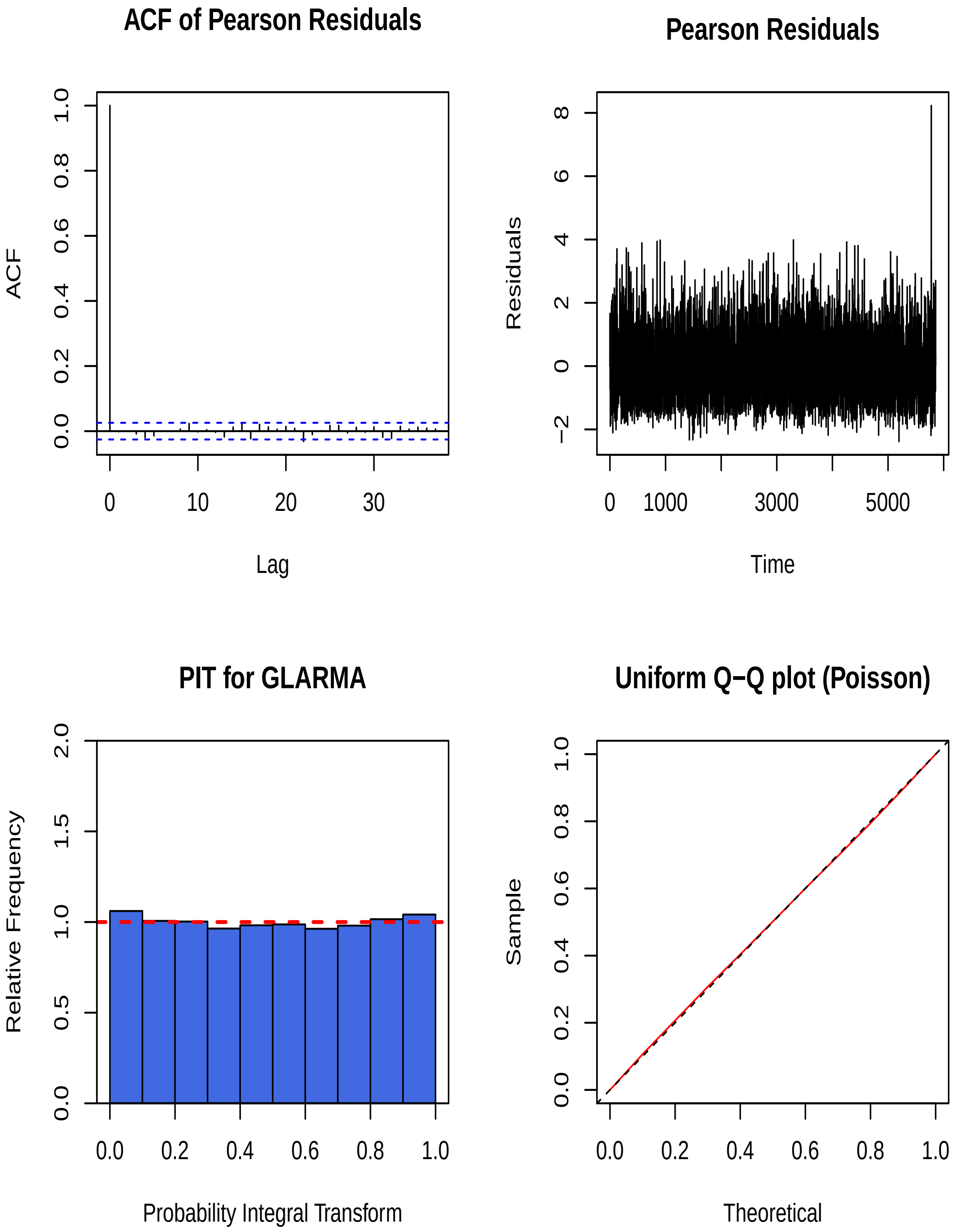


Figure 4.16: Residual analysis of the fitted GLARMA model for high acuity arrivals.

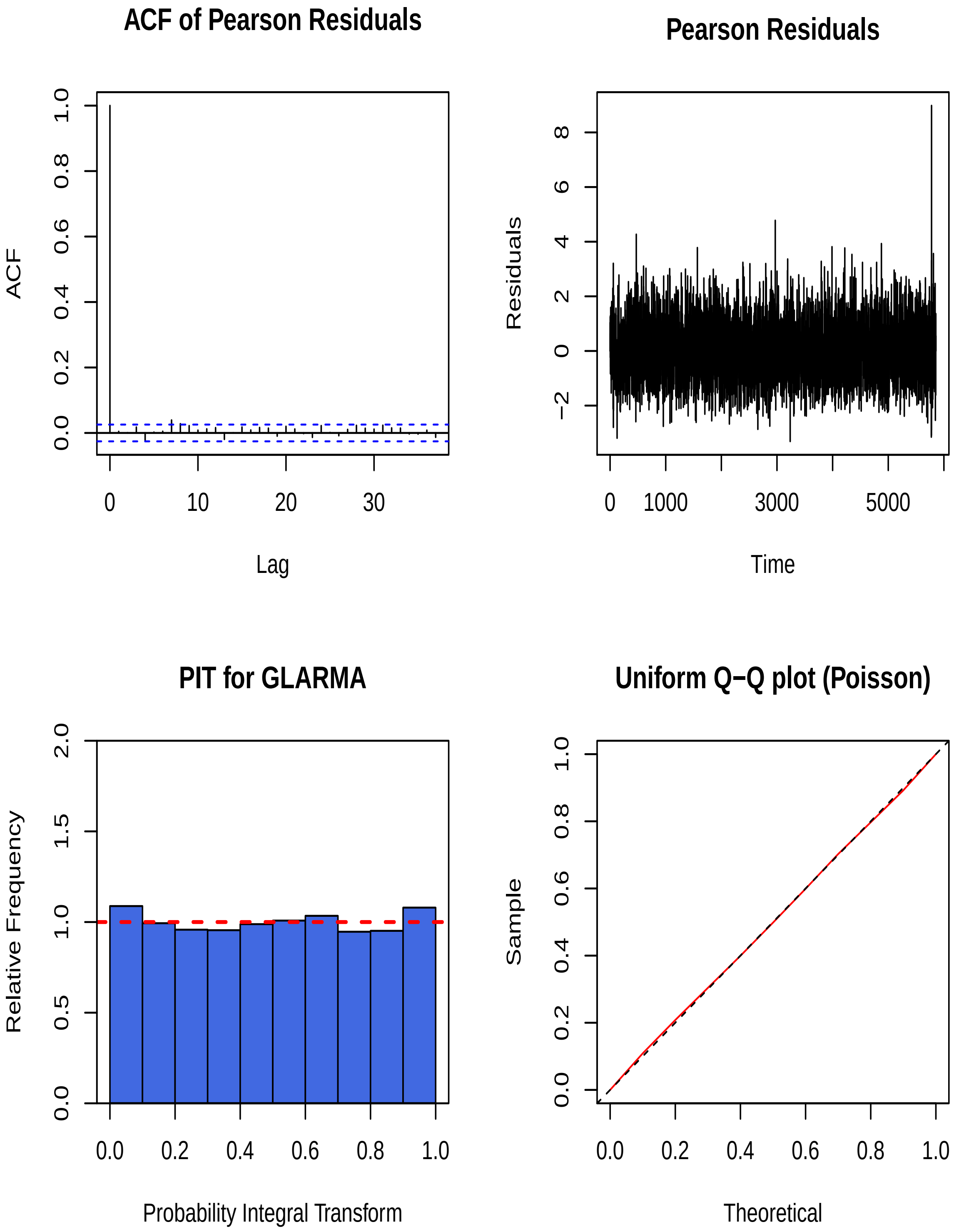


Figure 4.17: Residual analysis of the fitted GLARMA model for low acuity arrivals.

4.4.7 Rolling Horizon Approach to Produce Forecast

The validity of a forecasting model is assessed by comparing the forecast values with the actual values. Therefore, the sample data is first divided into two datasets, namely, training and test data. Forecasting models are fitted using training data and forecasts are then made for the length of the test data. Using the difference between the actual test data values and the forecast, certain methods were developed to measure the accuracy of the forecasting model.

There are no defined rules on how to divide the sample data into training and test data. However, there is a consensus among researchers about the use of as many data points for the model fitting purposes as possible, which leads to a notably larger amount of data points in the training dataset than in the test dataset. Initially I divided two years worth of data into training data, which consists of 22 months of data points and the test data which contains 2 months of data points. But our ED data shows seasonal variability by hours of the day, days of the week, and months of the year. For instance, a similar number of arrivals are expected at midnights everyday, every Mondays of a week, and the month of January every year. Therefore, to include seemingly similar values, a better way would be to separate the sample data such that the training data completes a cycle. I adopted the "Rolling Horizon" approach from operations management literature which serves our purpose. A description of the Rolling Horizon approach is documented in the method section (4.3.5).

Several accuracy measures were calculated using monthly forecast values for all 12 months of the year 2013, one month at a time based on the immediately preceding 12 months of data. A step-by-step approach to calculate rolling horizon forecast along with different accuracy measures are described below,

(i) A GLARMA model was fitted to the block arrival data from January, 2012 to December, 2012.

(ii) Forecast values for block arrivals for the month of January, 2013 were calculated based on the best fitted GLARMA model.

(iii) Forecast values for the month of January, 2013 were then compared with the observed values by calculating RMSE, MAE, maximum difference, correlation, and what percentage of observed values falls within the 80% and 95% confidence bound of the forecast values.

(iv) The training data was then move forward one month and a new GLARMA model was fitted to the data from February, 2012 to January, 2013 followed by calculation of the accuracy measures based on observed versus forecast values.

(v) Steps (ii) to (iv) were repeated until calculation for all the months of year 2013 is completed.

Tables 4.21 and 4.22 shows the results of all such calculations for the high and low acuity arrivals, respectively.

Table 4.21: Accuracy measures of 12 months rolling horizon forecasts for high acuity arrivals.

Months	Measures		% Within		Max Diff [†]	Corr[‡]
	RMSE	MAE	95%	80%		
Jan	1.918	1.496	96.774	81.854	7.759	0.545
Feb	2.151	1.655	93.300	82.14	8.405	0.562
Mar	1.991	1.550	93.951	82.661	6.860	0.598
Apr	2.078	1.594	95.833	82.083	9.00	0.593
May	2.029	1.563	94.758	84.677	8.475	0.602
Jun	2.095	1.642	95.000	77.083	6.589	0.579
Jul	2.175	1.667	93.950	81.040	10.531	0.595
Aug	1.758	1.408	95.967	76.612	5.562	0.635
Sep	2.059	1.609	95.397	85.774	7.911	0.609
Oct	1.944	1.580	96.370	79.435	6.010	0.645
Nov	1.831	1.454	93.333	76.666	4.927	0.639
Dec	2.092	1.629	94.3548	82.258	9.507	0.525

[†] maximum of absolute differences between actual and forecast arrivals.

[‡] correlation between actual and forecast arrivals.

One observes that the correlation between rolling horizon forecasts and the actual values are around 60% for the high acuity arrivals whereas for low acuity arrivals it is around 80%.

Table 4.22: Accuracy measures of 12 months rolling horizon forecasts for low acuity arrivals.

Months	Measures		% Within		Max Diff †	Corr‡
	RMSE	MAE	95%	80%		
Jan	3.288	2.677	96.370	79.838	11.533	0.799
Feb	3.750	2.887	92.857	76.785	9.282	0.746
Mar	3.204	2.464	94.354	80.241	7.692	0.795
Apr	3.572	2.854	94.166	79.166	12.236	0.770
May	3.401	2.638	95.967	79.838	14.388	0.797
Jun	3.608	2.819	95.000	77.916	10.431	0.764
Jul	3.233	2.475	95.564	79.032	11.302	0.805
Aug	3.474	2.828	95.967	78.629	12.031	0.809
Sep	3.459	2.623	95.397	81.171	14.321	0.795
Oct	3.509	2.702	94.758	82.661	11.409	0.810
Nov	3.354	2.647	94.583	80.833	9.564	0.803
Dec	4.116	3.055	94.354	81.451	12.838	0.743

† maximum of the differences between actual and forecast arrivals.

‡ correlation between actual and forecast arrivals.

The maximum difference between actual and forecast values for any block during the months of year 2013 ranges from 5 to 9 for high acuity arrivals and 9 to 14 for the low acuity arrivals. The lowest RMSE and MAE value for high acuity arrivals were observed for the rolling horizon month of August, 2013 whereas that of low acuity arrivals were observed for the rolling horizon month of March, 2013. The percentage of actual values that fall inside the 80% and 95% forecast bands conforms with the stated percentage.

The plot for the forecast values and the observed values for high acuity arrivals for the first six rolling horizon months are presented in Figure 4.18 with the same for the remaining six months shown in Figure 4.19. Similar plots for low acuity arrivals are displayed in 4.20 and 4.21. The blue dots represent the actual values while the red line represent the forecast values. I see close resemblance in each of these 12 graphs.

Finally, the density plots for the high and low acuity arrivals for the residuals are displayed in Figures 4.22 and 4.23, respectively. The residuals appear to be normally distributed for all the months which satisfies the assumption of the GLARMA model fit.

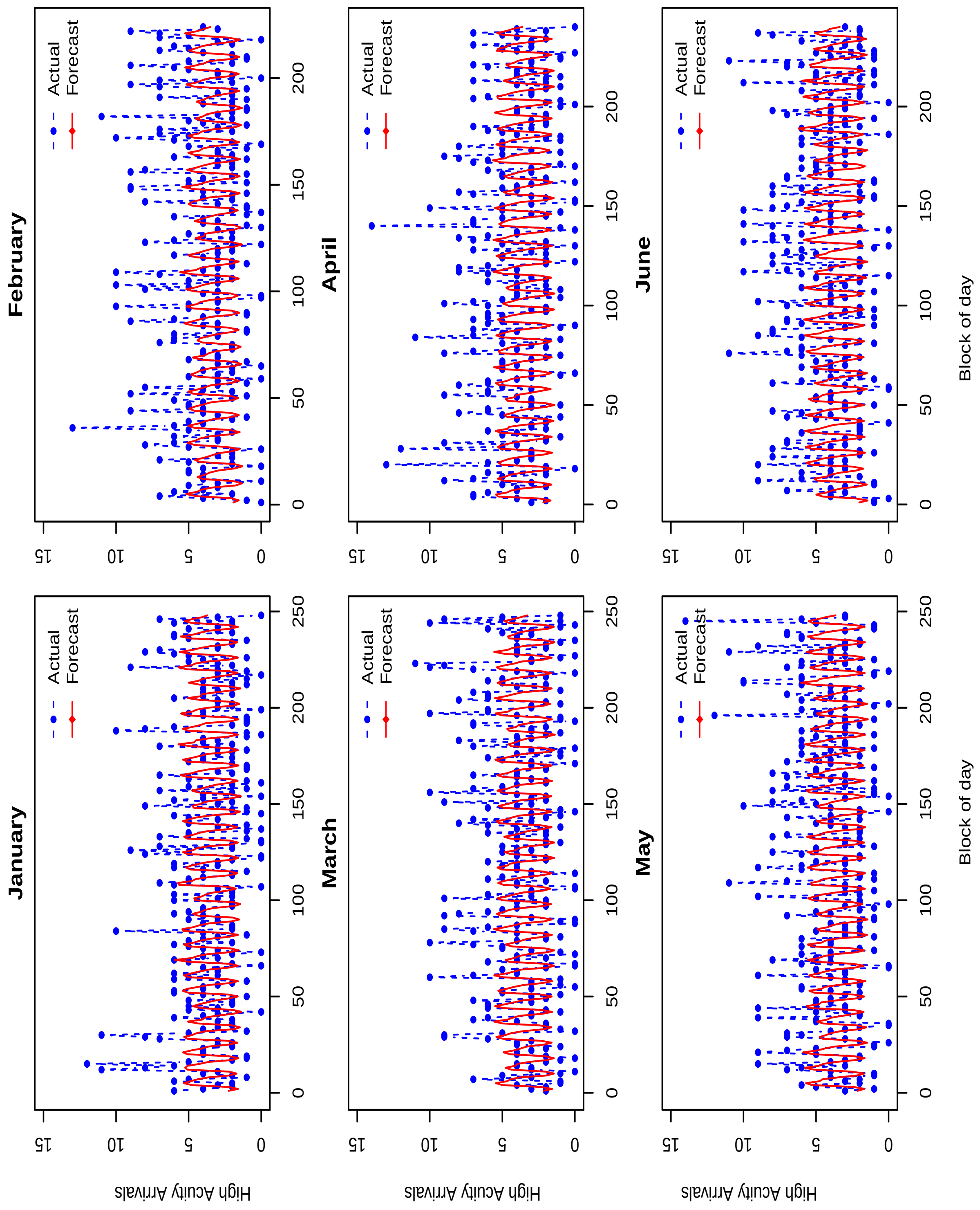


Figure 4.18: Rolling horizon forecasts for high acuity arrivals from months January to June, 2013.

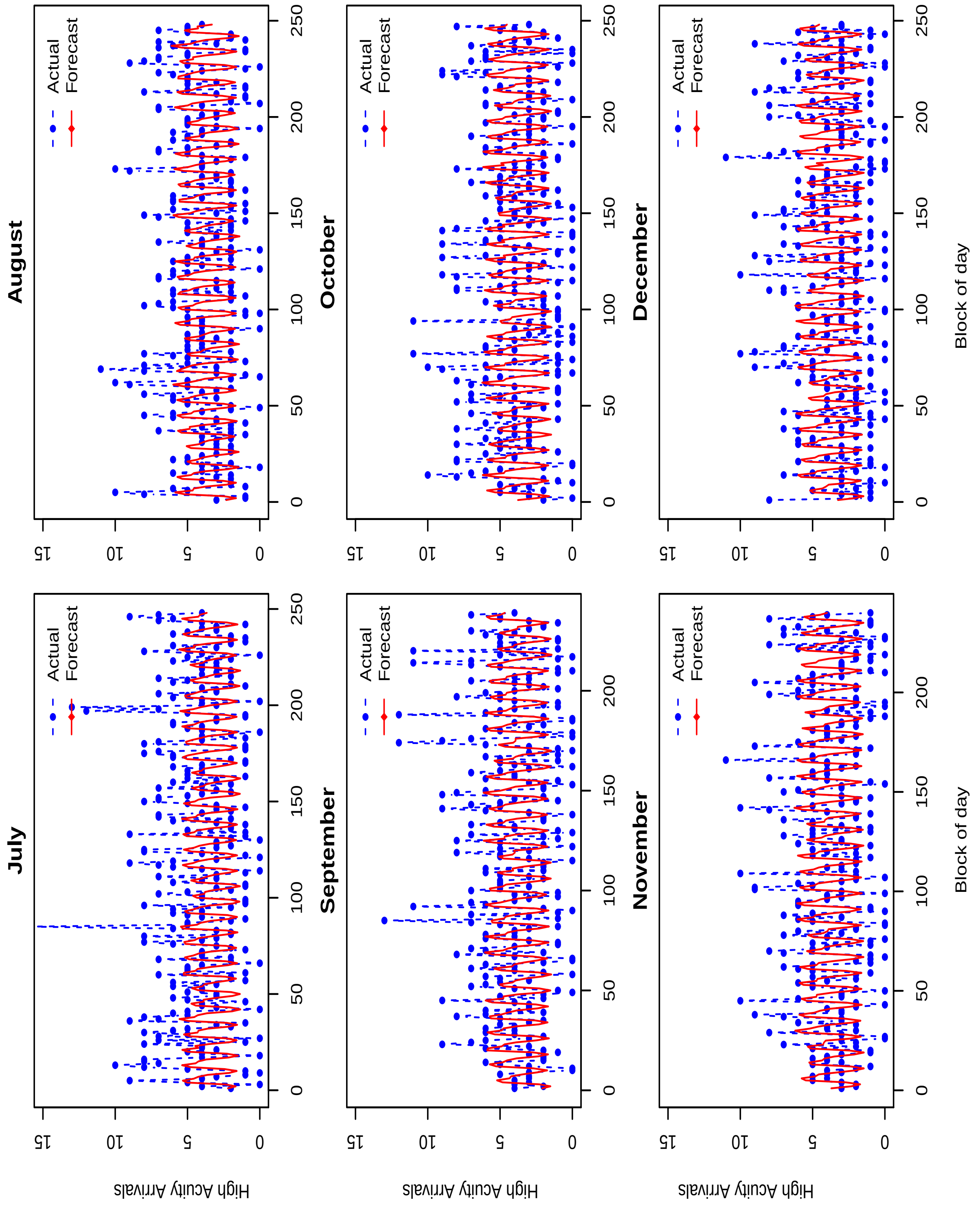


Figure 4.19: Rolling horizon forecasts for high acuity arrivals from months July to December, 2013.

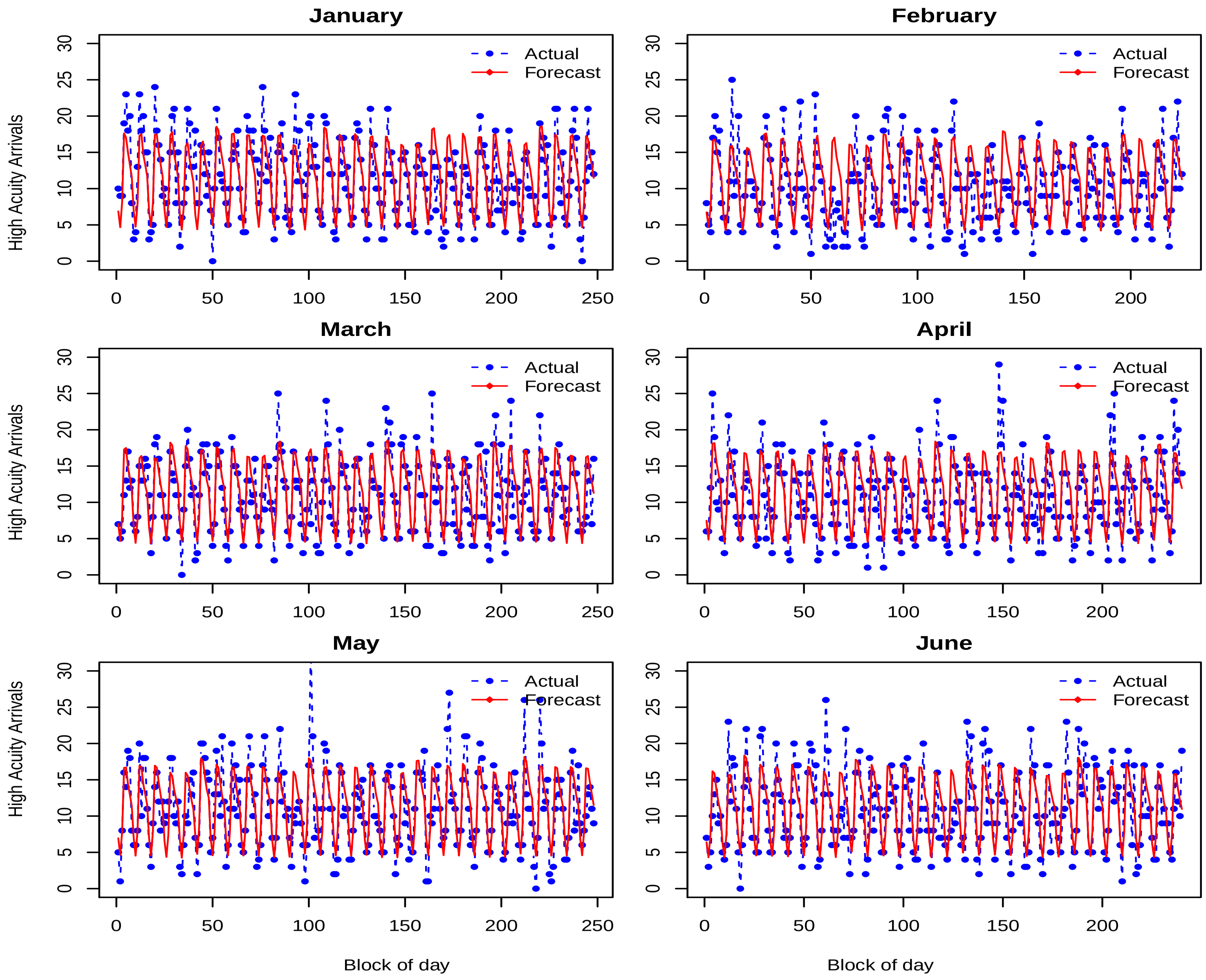


Figure 4.20: Rolling horizon forecasts for low acuity arrivals from months January to June, 2013.

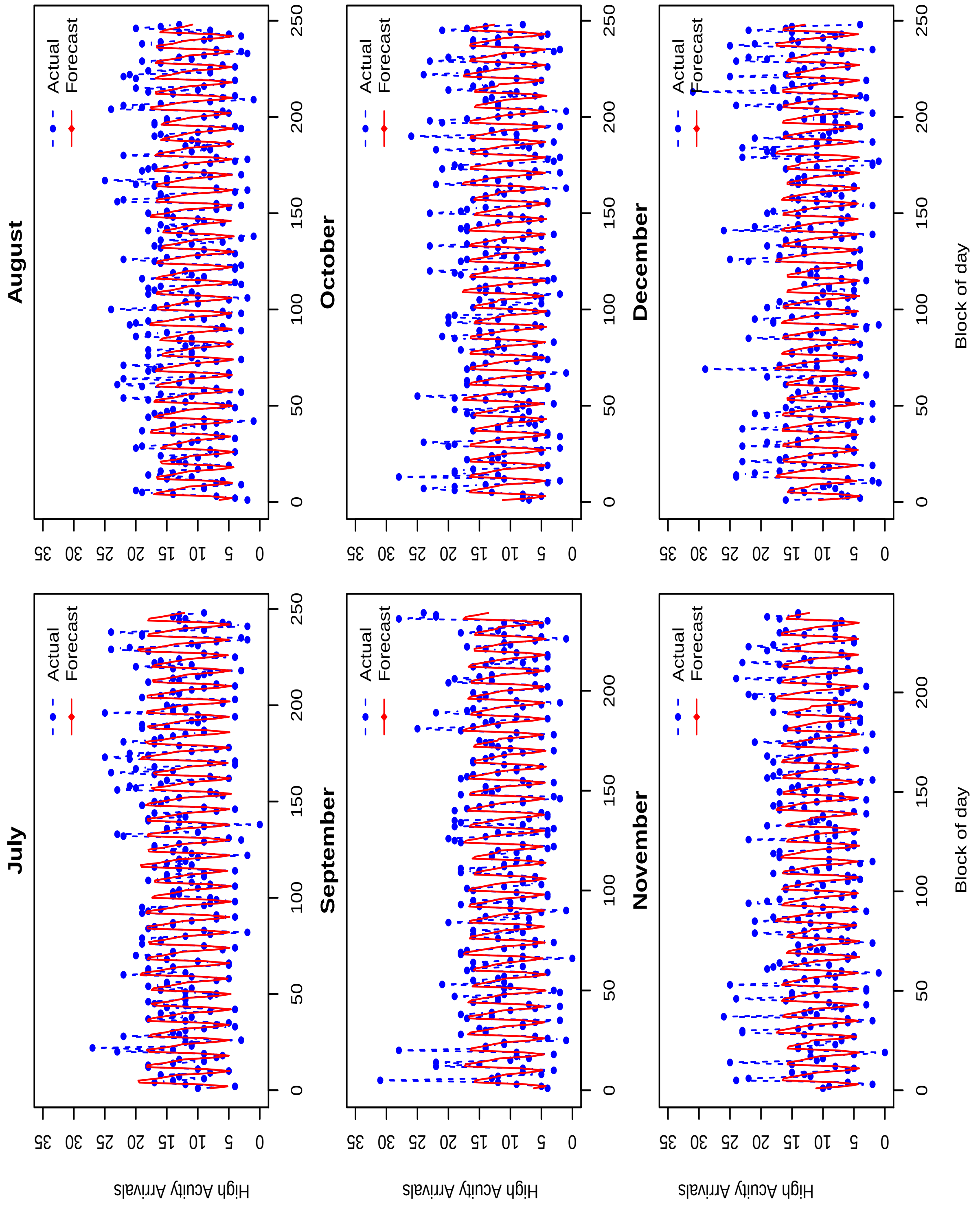


Figure 4.21: Rolling horizon forecasts for low acuity arrivals from months July to December, 2013.

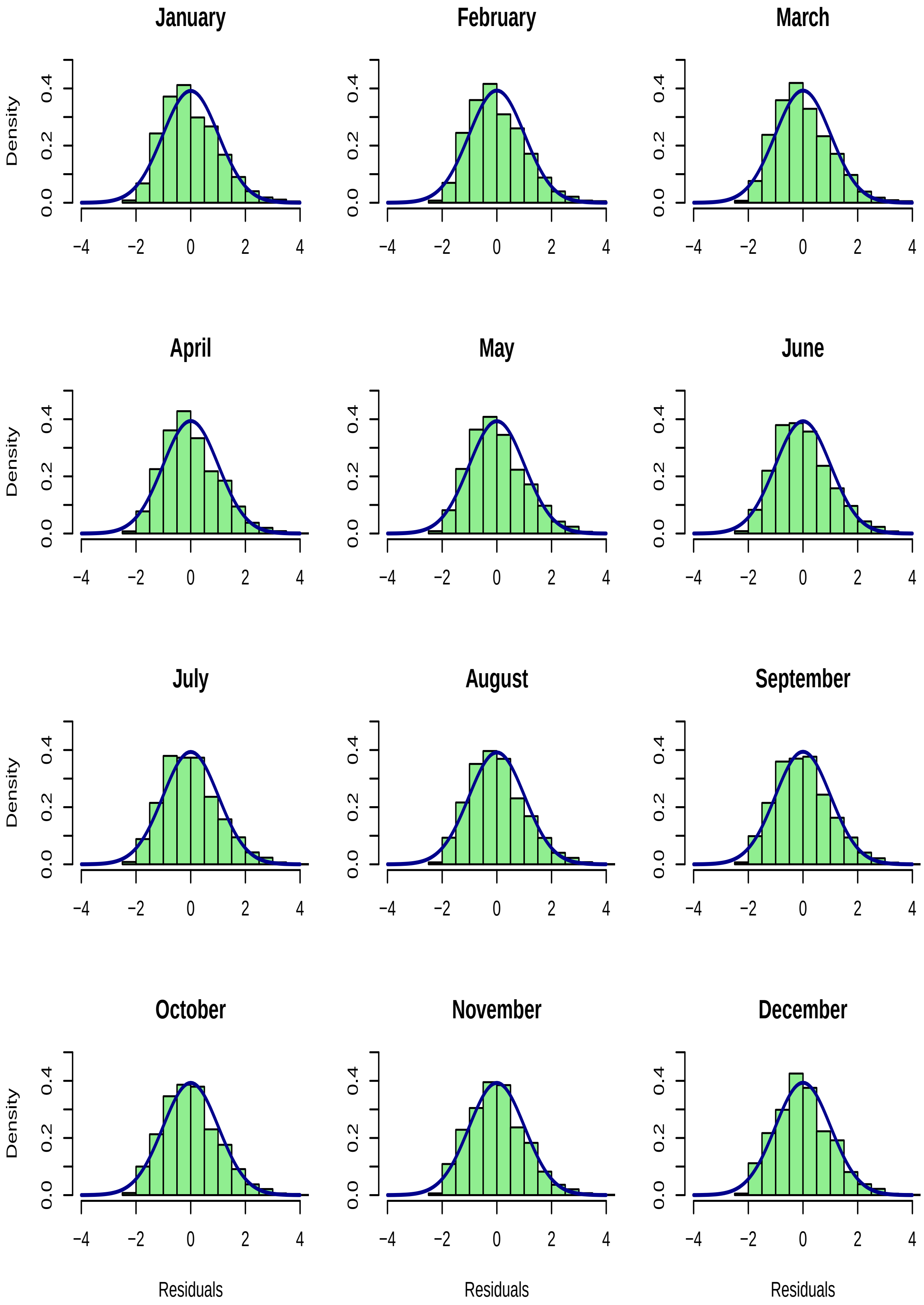


Figure 4.22: Density plot of residuals for the high acuity rolling horizon forecast .

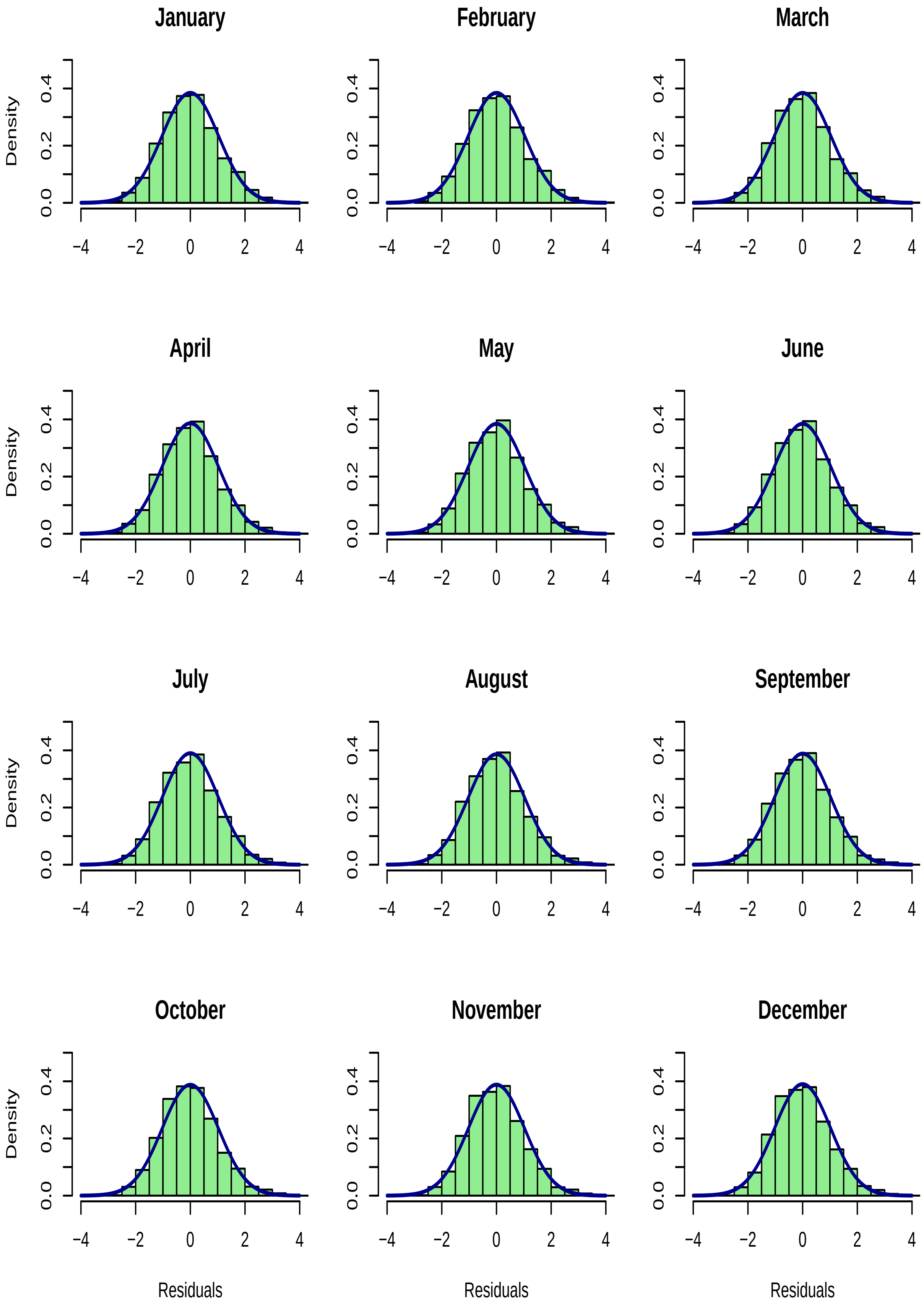


Figure 4.23: Density plot of residuals for the low acuity rolling horizon forecast.

A sample of the first 24 blocks from the month of January, 2013 have been selected to show how the actual values differs from the forecast values, and how many of the actual values are contained in the 80% and 95% forecast confidence intervals. It has been observed from table 4.23 that the actual values for blocks 12 and 15 are not contained in the 95% forecast confidence interval, whereas those observed for all other blocks are. Two out of 24 actual values outside 95% confidence limit is equivalent to approximately 8% out. Therefore, 95% of the actual values are not contained in the confidence interval as proposed. However, this is an expected randomness phenomenon, and the confidence intervals are defined under a probability construct which remains true only if the experiment is continued for a long period of time.

A counter intuitive example is observed from table 4.24 where all of the 24 actual values for low acuity arrivals are contained in the 95% confidence interval. Although only 95% of the actual values are expected to fall within these confidence intervals, from the selected 24 blocks, 100% of the actual low acuity arrivals are within 95% confidence interval by random chance. However, if the experiment were to be run for a sufficiently long time period, approximately 95% of the actual values will be contained in the proposed 95% confidence interval.

Table 4.23: Point and interval forecast for the first 24 blocks of January based on past 12 months of fitted data.

Blocks	Arrivals	Forecast	95% CI		80% CI	
			Lower	Upper	Lower	Upper
1	6	2.25	0.00	5.99	0.00	4.69
2	4	1.65	0.00	5.39	0.00	4.09
3	2	2.13	0.00	5.87	0.00	4.57
4	3	4.88	1.14	8.62	2.44	7.32
5	2	5.25	1.52	8.99	2.81	7.70
6	6	4.69	0.95	8.43	2.25	7.13
7	5	4.14	0.40	7.88	1.70	6.58
8	1	3.54	0.00	7.28	1.10	5.98
9	2	2.08	0.00	5.82	0.00	4.52
10	4	1.56	0.00	5.30	0.00	4.00
11	3	2.19	0.00	5.92	0.00	4.63
12	11	4.87	0.00	8.61	2.43	7.32
13	8	5.40	1.66	9.14	2.96	7.84
14	6	4.50	0.76	8.24	2.06	6.94
15	12	4.39	0.65	8.13	1.95	6.83
16	5	3.63	0.00	7.37	1.19	6.07
17	2	2.08	0.00	5.82	0.00	4.52
18	1	1.59	0.00	5.33	0.00	4.03
19	1	2.18	0.00	5.92	0.00	4.63
20	4	5.08	1.34	8.82	2.64	7.52
21	4	5.46	1.72	9.20	3.02	7.90
22	4	4.93	1.19	8.67	2.49	7.37
23	4	4.60	0.87	8.34	2.16	7.05
24	2	3.47	0.00	7.21	1.03	5.91

Table 4.24: Point and interval forecast for low acuity arrivals for the first 24 blocks of January based on past 12 months of fitted data.

Blocks	Arrivals	Forecast	95% CI		80% CI	
			Lower	Upper	Lower	Upper
1	10	6.88	0.00	13.79	2.36	11.39
2	9	4.66	0.00	11.57	0.14	9.17
3	9	8.19	1.27	15.10	3.67	12.70
4	19	17.63	10.72	24.55	13.12	22.15
5	23	17.04	10.13	23.95	12.52	21.55
6	18	15.03	8.12	21.94	10.51	19.54
7	20	13.33	6.42	20.24	8.82	17.85
8	8	11.63	4.71	18.54	7.11	16.14
9	3	6.93	0.02	13.85	2.42	11.45
10	4	4.73	0.00	11.65	0.22	9.25
11	13	7.87	0.96	14.78	3.36	12.39
12	23	17.27	10.36	24.18	12.75	21.78
13	18	17.53	10.62	24.45	13.02	22.05
14	20	14.67	7.76	21.59	10.16	19.19
15	15	13.27	6.36	20.18	8.76	17.79
16	15	11.73	4.82	18.64	7.21	16.25
17	3	6.85	0.00	13.76	2.33	11.36
18	4	4.71	0.00	11.63	0.20	9.23
19	5	8.10	1.19	15.02	3.59	12.62
20	24	17.45	10.53	24.36	12.93	21.96
21	18	17.48	10.57	24.39	12.97	22.00
22	16	15.09	8.17	22.00	10.57	19.60
23	14	13.33	6.41	20.24	8.81	17.84
24	9	11.76	4.85	18.68	7.25	16.28

4.5 Discussion

In this chapter, several forecasting models were investigated to identify the appropriate model for accurate short term ED arrival prediction. Separate forecasting models for high and low acuity patients were developed due to the fact that the arrival pattern as well as the resources to serve them vary between these two categories. Four different forecasting models, namely, SARIMA, GLM, Harmonic regression, and GLARMA models were considered.

Our results suggest that GLARMA is the most appropriate forecasting model to accurately forecast short-term ED arrival counts. Our finding also indicates that when an appropriate model is used, calendar variables alone are capable of producing accurate short-term forecast for ED arrivals. Accurate prediction of ED arrivals can have a direct implication on ED administrative decisions. At any point in time, having a particular workload, if an ED administrator can predict how much arrivals are expected in 3, 6, 9, and 12 hours in future, then s/he can take steps to have more resources in order to meet future needs.

Several other authors (Rotstein *et al.* (1997), Marcilio *et al.* (2013)) advocated the sole use of calendar variables in predicting ED patient visits. However, the method proposed here differs in that it can model count responses after controlling for serial correlation present in time series data. The GLARMA model, which is the mixture of a generalized linear model and a time series model, demonstrated the capability of making inferences based only on calendar variables (predictors) while taking into consideration the serial dependence among subsequent ED arrivals through incorporating autoregressive and/or moving average components into the model.

Few other authors (Reis & Mandl (2003), Chan *et al.* (2011), Au-Yeung *et al.* (2009)) considered fitting separate models dividing ED arrivals by some categories other than by their respective acuity. Based on the arrival pattern, I am aware of only one study (Sun *et al.* (2009)) that uses acuity categories as the basis for model fitting categorization. However, they did not aggregate the acuity categories based on how patients are being served in an ED.

It is to be noted that ED workload is very difficult to estimate and depends on a multitude of factors. Forecasting arrivals is a necessary component to predict workload, but may not be sufficient. There are other factors that affect ED workload. These include activities elsewhere. For instance, long-term care beds in the community for patients being discharged from in-patient units. Nurse staffing level have an impact as well. We have also seen that it is impractical to have physicians, nurses, diagnostic services, and so forth on stand-by (salaried or non-salaried). One solution would be to allocate more employees when more arrivals are expected or ask for overtime either before or after a shift to alleviate a surge.

While physician staffing is more directly related to new arrivals, nursing staff have to provide inpatient nursing care to the patients who are scheduled to be admitted to the hospital but for whom no bed is yet available. Therefore, in order to accurately predict workload for ED staff, one has to include the workload of those patients who are remaining in the ED while new arrivals continue to register. It is very difficult to predict how many of these patients will be there at any given point in time, there are multitude of factors impacting ED workload, and I intend to work on many of them in future. However, I have addressed some of these aspects in chapter 5, where I empirically investigated the effect of workload on ED performance measures, such as, number discharged, service time, waiting time, and length of stay.

4.6 Limitations

An obvious limitation of our study is the degree to which it can be generalized. I have collected data from a single hospital whose characteristics (patient population, severity of illness, etc) may vary in many ways from other hospitals. Therefore the exact same model may not prove to be applicable to other hospital settings. However, if the variable of interest is measured as counts and successive responses are serially correlated then GLARMA is the more appropriate model to use.

There were 42 missing arrival entries out of 88189 ED arrivals which contributes to only

0.05% of the collected data. This would have been a limitation if we would have observed higher percentages of missing values. Moreover, mean imputation was done to fill-up those missing values after controlling for the block of day, day of week, and month of year.

I did not have access to other socio-demographic variables which may affect ED arrivals. If I could have obtained such data, their effects on ED arrivals may have been tested to supplement or add value to previous researches.

Even though inclusion of climate variables will make the forecasting model more complete, it will simultaneously make the model complicated as well. Historically climate variables were shown to have statistically insignificant predictability on ED arrival prediction. However, in our particular case of ED arrival forecasting, the effect of climate variables on future ED arrivals could be explored.

4.7 Conclusion and future research

Our analysis suggests the use of the GLARMA model to accurately predict short term ED arrivals which in turn help ED managers to efficiently allocate ED resources to meet future demands. Although a relatively high forecast error was detected occasionally for particular blocks, calendar variables alone can reasonably predict future ED arrivals when serial correlation among successive block arrivals is accounted for. Thus, a reliable forecasting model for short term ED arrival forecasts can be formulated and implemented with an automated system for better management of ED resources.

Forecasting models for ED patient arrivals based on calendar variables are easy to formulate and can be automated for future use. A more complicated time series model that takes fractional differences to make the time-series stationary could be constructed (Granger & Joyeux (1980)). One may also consider using estimating equation techniques (Liang & Zeger (1986)) to forecast ED arrivals. Generalized estimating equation, which does not require the response distribution to be known, is one such technique where first two moments are enough to fully specify the likelihood (quasi) function to estimate the parameters of the model. A robust sandwich estimator for the variance is also available to construct confidence intervals for the parameters of interest.

Chapter 5

Empirical Analysis of an Emergency Department Service Process

5.1 Introduction

The efficient operation of an emergency department (ED)'s service process is necessary to avoid overcrowding and other undesirable consequences. Patient care delivery in an emergency department (ED) depends on the prompt yet accurate operation of many interconnected activities, such as registration, triage, nurse or physician assessment, diagnosis, consultation and treatment from specialists if necessary, and ultimately, discharge home or admission to a hospital bed. EDs possess a service system where multiple streams of diverse patient types appear for service, and for which the servers have multitasking capabilities. The servers also have discretion over to whom, when, and how to provide such services.

Emergency department resources and capacity are usually limited. However, the demand for ED services is gradually increasing over time due to an aging population and the availability of new technologies. The Ontario Ministry of Health and Long-Term Care (2010) observed an approximate 6% increase in the number of emergency department visits from 2004/05 through 2008/09, while operating costs increased by about 28%. A report by the Canadian Institute for Health Information (2014) shows that patients who are admitted to hospital spend almost five times longer in ED than their non-admitted counterparts. In fact, 10% spend more than 28 hours in the ED after admission while waiting for beds. Such discharge delays naturally contribute to ED overcrowding.

This chapter presents the results of an empirical investigation applied to an ED of an Ontario hospital. An empirical investigation of such ED performance measures will help ED administrators to better understand how ED behaves under different conditions. Successful application of operations management principles can help service managers to manage and allocate resources efficiently. Operations management researchers have advocated the need for more empirical studies to investigate the extent to which human behavior may alter the relationship between operational variables and performance (eg., Boudreau *et al.* (2003), Jouini *et al.* (2008)). Particularly in healthcare, to clearly understand the relationship between server response to workload and productivity, detailed empirical investigations are necessary (KC & Terwiesch (2009), KC & Terwiesch (2012)).

Failure to account for the effect of workload on productivity can result in poor performance (Kuntz *et al.* (2014), Tan & Netessine (2014)), imbalance in physical resources (Green & Nguyen (2001)) and in labour (Green *et al.* (2013)). Therefore, proper understanding of the relationship between workload and productivity will aid in better management of the ED resources and thus will maximize utilization and minimize waste of expensive resources. The goal of this chapter is thus to contribute to the emergency medicine and operations management literature by empirically investigating the impact of patient workload on four measures of services: number discharged, ED LOS, service time, and waiting time. Our study will have an emphasis on the effect of high and low acuity workload on service measures for both high and low acuity patients.

This chapter is organized as follows: in Section 5.2, the related literature is reviewed. A description of the hypotheses under study and their development are discussed in Section 5.3. Research settings and the statistical model specifications to address the hypotheses are described in Section 5.4. In Section 5.5, we report the results from the analysis of an emergency department patient flow. Discussion, concluding remarks and future research directions appear in Sections 5.6 and 5.7.

5.2 Related Literature

We divided the existing body of literature into two parts: literature related to ED and literatures not related to the ED.

5.2.1 Empirical Literature Related to ED

5.2.1.1 Emergency Medicine Literature

Schull *et al.* (2007) studied the effect of low acuity ED patients on the waiting time of medium and high acuity patients. Applying an autoregression model to one year of ED visit data from Ontario hospitals, the authors observed that low acuity ED patients are associated with a negligible increase in ED LOS and time to first physician contact for medium and high acuity waiting time. On the other hand, McCarthy *et al.* (2009) used discrete time survival analysis on one year of data from 5 EDs to conclude that ED crowding statistically delayed patients' waiting and boarding times, but not the treatment time. To determine the association between the hospital census variables and ED LOS, Lucas *et al.* (2009) applied a stepwise multiple regression model to five weeks of data from five hospitals and observed a significant positive relationship between ED LOS and hospital level census. However, the relationship of the ED LOS and ED census was marginally significant.

5.2.1.2 Operations Management Literature

Empirically investigating the effect of patient census on total service rate and service rate per patient, Armony *et al.* (2011) observed evidence of a state dependent service rate (which involved both speed-up and slow-down mechanisms). Using 5 years of data from an Israeli hospital, the authors found that, as a function of patient census, the service rate first decreases then increases then again decreases, with some noise in the tails due to small sample sizes. Batt & Terwiesch (2012) identified the mechanisms that lead to state dependent services. Identifying task reduction and early task initiation as speed-up mechanism and multitasking and interference as slow-down mechanisms, the authors obtained evidence of the existence of such mechanisms while analyzing 3 years of ED visit data from an urban teaching hospital using survival analysis and a count regression model. The authors also developed a discrete event simulation model, and used it to establish that ignoring state-dependent service times leads to modeling errors and could cause hospitals to over-invest in human and physical resources. In contrast, KC (2013) looked at the multitasking behavior of physicians in a busy ED using overall performance measures such as processing time, throughput rate, and output quality. Defining the busy period as the amount of time needed to serve a given number of patients without interruption, KC (2013) showed that busy period has a U-shaped response to the level of physician multitasking. The term "U-shaped relationship" means that an increase in the explanatory variable is associated with an initial reduction in the response, followed by an increase after a particular value of that explanatory variable.

In a discretionary work setting, Berry Jaeker *et al.* (2013) tested the hypothesis that resource availability induces demand based on 2 years and 7 months of patient-level data from two academic hospitals. The authors had shown that increased capacity is associated with an increased wait time, and an increase in physician's capacity for ordering diagnostic tests increased the probability of more diagnostic tests being ordered. In contrast, to investigate the effect of workload on system productivity, Green *et al.* (2013) investigated how fluctuations in nurse workload due to irregularities in scheduling and/or unpredictable demand had a direct impact on absenteeism. The authors observed that there is a positive association between high demand and nurse absenteeism. Using 10 months of ED data from a large New York City hospital, Green *et al.* (2013) found that the uncertain shortage of service capacity created by nurse absenteeism should be taken into account when matching supply with anticipated demand.

Song *et al.* (2015) carried out an empirical investigation of the impact of queue pooling on throughput times in a discretionary task setting (servers have freedom to choose whom, when, and how to serve). Four years (2007-2010) of patient level emergency department (ED) data were used to show that patients' lengths of stay are longer when physicians are assigned patients from a pooled queue, compared to when each physician has a dedicated queue. In order to support their findings, Song *et al.* (2015) have used "social loafing theory", which suggests that when work is shared individuals subconsciously exert less effort, to propose that dedicated queues provide physicians with greater ownership concerning their workload than pooled queues, enabling them to more effectively manage their work.

5.2.2 Other Related Empirical Literature

Using data on patient transport services and cardiothoracic surgery, KC & Terwiesch (2009) conducted an econometric analysis to observe the impact of workload on service time and patient safety. The authors argued that the workers accelerate the service rate as the workload increases. However, the authors also noticed that long periods of increased workload leads to decrease in service rate and quality of care. In a separate study KC & Terwiesch (2012) found that when the ICU occupancy is high, patients are likely to be discharged early, which in turn increases the likelihood of patient readmission. Bartel *et al.* (2014) investigated the role of inpatient and outpatient care in reducing readmission and mortality.

Berry Jaeker *et al.* (2012), using data from 283 hospitals, observed that high congestion increases patients' hospital length of stay. Berry Jaeker & Tucker (2013), on the other hand, examined the impact of patient load on Length of Stay (LOS), with a special focus on the "spillover" effect of patient load across the two main types of patients within a hospital: medical and surgical. Separating the congestion effect (increased LOS) from the workload smoothing effect (decreased LOS), Berry Jaeker *et al.* (2013) showed that when hospitals become busy, workers prioritize some types of patients by systematically shortening the LOS of other types of patients. In the context of an Intensive Care Unit (ICU), Chan *et al.* (2013) measured how congestion in an ICU can lead to delays in boarding (transferring patient) from the ED to the ICU and consequently impact patients' ICU LOS.

Beyond healthcare related studies, using a laboratory experiment on inventory lines, Schultz *et al.* (1998) showed that the workers' processing times are not independent of the state of the system. Based on the data from a restaurant chain Tan & Netessine (2014) observed that when overall workload is small, servers expend more and more sales effort with the increase in workload at a cost of slower service speed. However, after a certain workload threshold, servers start to reduce their sales efforts and work more promptly with the further rise in workload.

Important differences between our study and most related studies are described below:

(i) Batt & Terwiesch (2012) address speed-up and slow-down mechanisms with a definition of workload based on diagnostic procedures, whereas we define workload as the census at time t divided by available resources at time t. Moreover, our objective is to show how the speed-up and slow-down mechanisms work separately for high and low acuity patients.

(ii) Song *et al.* (2015) compared ED performance measures based on whether a patient was served under a dedicated queue versus a pooled queue with a fair patient routing constraint, whereas we are interested in studying the effect of workload on productivity. Although our ED has dedicated queues for high and low acuity patients, resources are often shared across these queues, and

(iii) Berry Jaeker & Tucker (2013) studied the effect of workload of one group to the productivity of the other group (spillover effect) by separating hospital patients into two groups; medical and surgical. We focus on the spillover effect of workload on productivity by separating ED patients into two groups; high and low acuity.

5.3 Hypotheses Development

We grouped resuscitation and emergency patients (i.e., CTAS categories 1 and 2) together to form the high acuity patient category and urgent, less-urgent, and non-urgent patients to form the low acuity patient category. When comparing high exposure to high response or low exposure to low response, we use the word “same type” of patients. Similarly, when comparing low exposure to high response or high exposure to low response, we use the word “other type” of patients. Since we are interested in investigating the impact of increased workload on four different ED service measures (namely, number discharged, length of stay, service time, and waiting time), we have grouped our hypotheses accordingly.

5.3.1 Expected Patient Discharge

Previous operations management literature suggests that physicians’ medical decisions and speed can be influenced by how many patients are currently in the hospital (Batt & Terwiesch (2012), KC & Terwiesch (2012), Kuntz *et al.* (2014)). Patients in our study ED are treated in two separate areas depending on their severity (i.e., in the high or low acuity group). Therefore, we investigate the speed-up and slow-down effects of high and low acuity census on the average number of high and low acuity patient discharges, respectively.

HYPOTHESIS 1: *The mean number of patients discharged increases with an increase in the same type of patient census.*

Berry Jaeker & Tucker (2013) tested the “spillover” effect of medical and surgical workload on the other type of patients LOS. Although physicians in our study ED are assigned exclusively to serve a stream of patients in a dedicated area, there may be crossover due to medical emergencies or to minimize physicians’ idle time. Thus we expect the following:

HYPOTHESIS 2: *The mean number of patients discharged increases with the increase in the other type of patient census.*

5.3.2 Patient Length of Stay

Next, we discuss the hypotheses concerning the effect of the state of the system on patient length of stay (LOS). Lucas *et al.* (2009) studies the effect of hospital census variables on ED LOS and observed that ED census is significantly positively correlated with ED LOS. However, significant association between ED LOS and ED census was not observed from multiple regression output when controlled for other census variables (i.e., hospital, ICU, Telemetry). In a hospital setting, Berry Jaeker *et al.* (2012) demonstrated that controlling for patient conditions, high current patient inventory (congestion) is associated an increased LOS. Additionally, Berry Jaeker & Tucker (2013) showed that when services are shared, additional demand of one type of patient can create congestion for all types of patients. Thus we expect,

HYPOTHESIS 3: *Increased census of the same or other type of patient is associated with an increased LOS.*

Using data from a large Israeli hospital Armony *et al.* (2011) studied the cause of long delays in transfers from the ED to an inpatient ward (IWs) and concluded that those delays are not only caused by a lack of beds, but also by 13 other plausible causes. The authors observed that on average it takes 3.2 hours to transfer a patient from an ED to an IW, and pointed out that the admission process as well as the wait for the respective physician and nurse are the significant contributors to these delays. Thus we expect,

HYPOTHESIS 4: *Patients who are admitted to hospital beds have longer LOS than patients who are discharged home.*

(Support for this hypothesis is also found in the report by the Ontario Ministry of Health and Long-Term Care (2010)).

5.3.3 Patient Waiting Time

In an ER, almost everyone is delayed. Only CTAS 1 patients tend to be seen immediately. Thus, patients' waiting time are longer if the ED is congested. The congestion occurs despite reduced service time for the service of interest because customers' total LOS become longer Berry Jaeker & Tucker (2013). Therefore, if the workload increases, this will result in a higher probability of waiting, and an increase in average waiting time for the patients. Thus we expect,

HYPOTHESIS 5: *Increased workload of the same type of patient is associated with an increased waiting time.*

HYPOTHESIS 6: *Increased workload of the other type of patient is associated with an increased waiting time.*

5.3.4 Patient Service Time

We constructed several hypotheses concerning the effect of workload on the patients' service time. Classical queueing theory generally assumes that the service time is independent of the state of the system (Wolff (1989)). This assumption does not hold for many service systems (i.e., hospitals, outbound call centers) where servers have discretion over whom and how to provide services. The existence of the speedup and slowdown mechanisms in an ED was considered by Armony *et al.* (2011). Analyzing the effect of workload on service time, Batt & Terwiesch (2012) tested these hypotheses and obtained evidence in favor of these hypotheses. Since resources are shared, our interest is to look at the effect of high and low acuity workload on the high and low acuity service times. For a shared service, Berry Jaeker & Tucker (2013) showed that when services are shared, additional demand of one type can create delay for all types of customers. Thus we expect,

HYPOTHESIS 7: *As workload of the same type of patient increases, patient service times initially decrease then increase.*

HYPOTHESIS 8: *As workload of the other type of patient increases, patient service times* *initially decrease then increase.*

A summary of our hypotheses and proposed models are summarized in the following table.

Table 5.1: Summary of the hypotheses and proposed models.

Metric	Hypotheses	Model
Number Discharged	1. The mean number of patients discharged increases with an increase in the same type of patient census.	Poisson Regression
	2. The mean number of patients discharged increases with the increase in the other type of patient census.	Poisson Regression
Length of Stay	3. Increased census of the same or other type of patient is associated with an increased LOS.	Cox-PH Model
	4. Patients who are admitted to hospital beds have longer LOS than patients who are discharged home.	Cox-PH Model
Waiting Time	5. Increased workload of the same type of patient is associated with an increased waiting time.	Cox-PH Model
	6. Increased workload of the other type of patient is associated with an increased waiting time.	Cox-PH Model
Service Time	7. As workload of the same type of patient increases, patient service time initially decreases then increases.	Cox-PH Model
	8. As workload of the other type of patient increases, patient service time initially decreases then increases.	Cox-PH Model

5.4 Model Specifications

We first introduce and define necessary variables required for our model formulation. Different response, explanatory, and control variables used in our models are presented in the Tables 5.2, 5.3, and 5.4, respectively.

Table 5.2: Definitions of response variables for operational performance measure.

Measure	Description and coding
DischHigh	Number of high acuity patients discharged from ED per block.
DischLow	Number of low acuity patients discharged from ED per block.
LOSHigh	High acuity patient length of stay (registration to discharge).
LOSLow	Low acuity patient length of stay (registration to discharge).
STHigh	High acuity patient service time (initial assessment to discharge).
STLow	Low acuity patient service time (initial assessment to discharge).
WTHigh	High acuity patient waiting time (registration to initial assessment).
WTLow	Low acuity patient waiting time (registration to initial assessment).

Table 5.3: Definitions of explanatory variables for operational performance measure.

Measure	Description and coding
CenHigh	Number of high acuity patients in the system (counts).
CenLow	Number of low acuity patients in the system (counts).
CenHigh^2	Square of number of high acuity patients in the system (counts)
CenLow^2	Square of number of high acuity patients in the system (counts)
CenHighLag1	Number of high acuity patients in the system at previous block (counts)
CenLowLag1	Number of low acuity patients in the system at previous block (counts)
HighWorkload	Number of high acuity patients in the system divided by total number of beds for high acuity patients (ratio)
LowWorkload	Number of low acuity patients in the system divided by total number of beds for low acuity patients (ratio)
BusyHigh	Indicator variable takes 1 if high census is greater than the number of high beds
BusyLow	Indicator variable takes 1 if low census is greater than the number of low beds
Admitted	Binary variable indicating admitted to hospital beds or discharged home.
Age	Age of the patient (continuous variable).

Different indicies and parameters used in our models are listed in the following Table 5.5.

Table 5.4: Definitions of control variables for operational performance measure.

Measures	Description and coding
BoD (1-24)	Block of day variables are defined using 23 dummy variables
DoW (Mon-Sun)	Day of week variables are defined using 6 dummy variables
MoY (Jan-Dec)	Month of year variables are defined using 11 dummy variables

Table 5.5: Definitions of the indicies and parameters used in our models.

Notations	Description and coding
i	Index for patient observations (total 88189).
b	Index for hourly blocks (total 24 × 731, for 2 years of data).
α	Coefficient for covarites related to high acuity response variable.
β	Coefficient for covarites related to low acuity response variable.
Λ	Vector of coefficients for calendar variables related to high acuity response variable.
Γ	Vector of coefficients for calendar variables related to low acuity response variable.
ϵ	Errors for high acuity.
ε	Errors for low acuity.

5.4.1 Model for Hypotheses 1 and 2 : Number Discharged & Census

We are interested to see how workload impacts the amount of discharge per 1-hour block. The number discharged per block is a count random variable, and generalized linear models are used to predict the expected number of discharges as a function of various predictor variables. In the formulation below, the log link function has been used, i.e., log of the number of expected discharge per block is linearly related to the predictor variables. Positive β values indicates an increase in the number discharged as a result of an increase in the predictor variables.

$$\begin{aligned}
\log(\mathrm{E}(DischHigh_b)) &= \alpha_0 + \alpha_1 CenHighlag1_b + \alpha_2 CenLowlag1_b + \alpha_3 CenHighLag2_b + \\
&\quad \alpha_4 CenLowLag2_b + \Lambda\ controls_b + \epsilon_b \\
\log(\mathrm{E}(DischLow_b)) &= \beta_0 + \beta_1 CenHighlag1_b + \beta_2 CenLowlag1_b + \beta_3 CenHighLag2_b + \\
&\quad \beta_4 CenLowLag2_b + \Gamma\ controls_b + \varepsilon_b
\end{aligned}$$

where the subscript b denotes the hourly block of the day.

5.4.2 Model for Hypotheses 3 and 4 : LOS & Census

We used survival models to test our hypotheses regarding LOS. Patient LOS can be thought of as a time-to-event data where event is the discharge from an ED. Survival analysis allows us to predict a patient's likelihood of LOS in any given block (hour) based on an underlying hazard function scaled by the patient's and ED's characteristics (similar to Berry Jaeker & Tucker (2013)). We used time-varying baseline hazard function (as used in KC & Terwiesch (2012)) and included calender variables to account for seasonal variability.

For the ease of interpretation, we report the hazard ratio which is the ratio of two hazard rates. The value of a hazard ratio more than 1 indicates that as the workload increases, patients are more likely to be discharged compared to the baseline condition, while a hazard ratio less than 1 means patients are less likely to be discharged as the workload increases. For instance, a hazard ratio of 1.05 states that a 1% increase in the workload is associated with a 5% increase in the expected number of discharges after controlling for all the covariates.

$$\begin{aligned}
\log(h(LOSHigh_i)) &= \alpha_0 + \alpha_1 CenHigh_i + \alpha_2 CenLow_i + \alpha_3 CenHigh_i^2 + \alpha_4 CenLow_i^2 + \\
&\quad \alpha_5 CenHighLag1_i + \alpha_6 CenLowLag1_i + \alpha_7 BusyHigh_i + \\
&\quad \alpha_8 BusyLow_i + \alpha_9 AdmitHosp_i + \Lambda Controls_i + \epsilon_i \\
\log(h(LOSLow_i)) &= \beta_0 + \beta_1 CenHigh_i + \beta_2 CenLow_i + \beta_3 CenHigh_i^2 + \beta_4 CenLow_i^2 + \\
&\quad \beta_5 CenHighLag1_i + \beta_6 CenLowLag1_i + \beta_7 BusyHigh_i + \\
&\quad \beta_8 BusyLow_i + \beta_9 AdmitHosp_i + \Gamma Controls_i + \varepsilon_i
\end{aligned}$$

5.4.3 Model for Hypothesis 5 and 6 : Waiting Time & Workload

Similar to the previous subsection, we used the Cox-PH model to test the hypotheses regarding the impact of workload on ED waiting time as follows:

$$\begin{aligned}
\log(h(WTHigh_i)) &= \eta_0 + \eta_1 CenHigh_i + \eta_2 CenLow_i + \eta_3 CenHigh_i^2 + \eta_4 CenLow_i^2 + \\
&\quad \eta_5 CenHighLag1_i + \eta_6 CenLowLag1_i + \eta_7 BusyHigh_i + \\
&\quad \eta_8 BusyLow_i + \eta_9 Admitted_i + \eta_{10} Age_i + \Lambda Controls_i + \epsilon_i \\
\log(h(WTLow_i)) &= \eta + \eta_1 CenHigh_i + \eta_2 CenLow_i + \eta_3 CenHigh_i^2 + \eta_4 CenLow_i^2 + \\
&\quad \eta_5 CenHighLag1_i + \eta_6 CenLowLag1_i + \eta_7 BusyHigh_i + \\
&\quad \eta_8 BusyLow_i + \eta_9 Admitted_i + \eta_{10} Age_i + \Gamma Controls_i + \varepsilon_i
\end{aligned}$$

where the subscript i denotes individual patients. A positive η coefficient relates an increase in mean service time with an increase in the covariate values.

Maximum likelihood methods have been used to estimate the model parameters.

5.4.4 Model for Hypotheses 7 and 8 : Service Time & Workload

To test the hypotheses about the impact of workload on ED service time, we again used the Cox-PH model.

$$\begin{aligned}
\log(h(STHigh_i)) &= \theta_0 + \theta_1 CenHigh_i + \theta_2 CenLow_i + \theta_3 CenHigh_i^2 + \theta_4 CenLow_i^2 + \\
&\quad \theta_5 CenHighLag1_i + \theta_6 CenLowLag1_i + \theta_7 BusyHigh_i + \\
&\quad \theta_8 BusyLow_i + \theta_9 Admitted_i + \theta_{10} Age_i + \theta_{11} Controls_i + \epsilon_i \\
\log(h(STLow_i)) &= \theta + \theta_1 CenHigh_i + \theta_2 CenLow_i + \theta_3 CenHigh_i^2 + \theta_4 CenLow_i^2 + \\
&\quad \theta_5 CenHighLag1_i + \theta_6 CenLowLag1_i + \theta_7 BusyHigh_i + \\
&\quad \theta_8 BusyLow_i + \theta_9 Admitted_i + \theta_{10} Age_i + \theta_{11} Controls_i + \varepsilon_i
\end{aligned}$$

where the subscript i denotes individual patients. A positive θ coefficient indicates an increase in mean service time with an increase in the covariate values. We have conducted all our statistical analyses using the free statistical software package R (R Core Team (2013)).

5.5 Results and Interpretation

5.5.1 Descriptive analysis

We display descriptive statistics for our continuous response variables LOS, waiting time, and service time through box plots in Figure 5.1. It is apparent from the box plots that all three duration measures are right skewed. There are some excessively high values that can be considered as outliers. Therefore, we considered the values within three times the interquartile range as the legitimate observations. We then produced box plots for high and low acuity patients separately for patient LOS, waiting time, and service time, respectively (Figure 5.2). Right skewness is still present in all cases for both the high and low acuity patients. This is due to the fact that a small number of individuals are experiencing excessive delays.

There are differences in median LOS, waiting time, and service time between high and low acuity patients (Figure 5.2). In our ED, the ratio of high to low acuity patient presentations is about 1:3. The median LOS for high acuity patients is significantly higher than that of the low acuity patients. This is due to the fact that high acuity patients require significant amount of time and resources to render the appropriate treatment. Additionally, time in the ED will also be higher for the high acuity patients, because they are much more likely to need admission and are therefore likely to suffer significant delays in the ED. Median service times, on the other hand, are higher for high acuity patients solely due to higher requirement for resources and time. A lower median waiting time for high acuity patients is to be expected, as they have higher priority to get served, and their patient volume is substantially smaller as well.

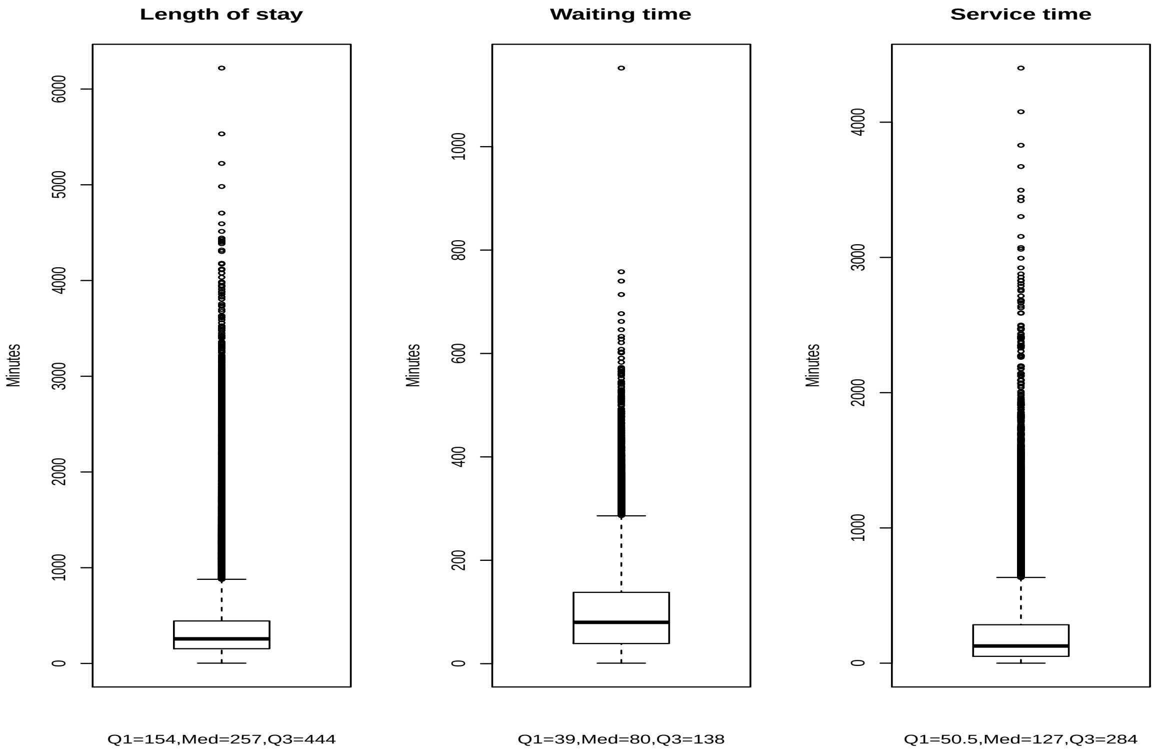


Figure 5.1: Boxplots of length of stay, waiting time and service time, respectively, from left to right, with high and low acuity patients merged into a single class.

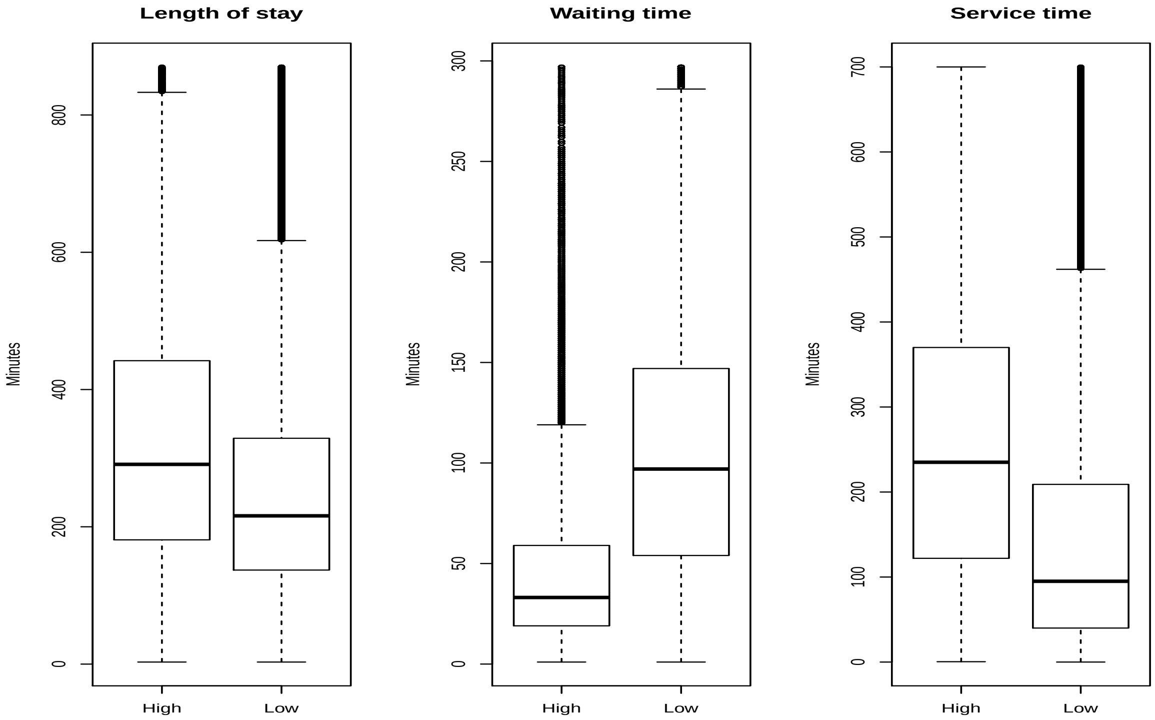


Figure 5.2: Boxplot of high and low acuity patients separately for length of stay, waiting time and service time, respectively, from left to right.

5.5.2 Statistical Assessment of the Hypotheses

5.5.2.1 Hypotheses 1 and 2: Number Discharged and Census

The number of discharged patients per block is a count random variable which is skewed to the right. We ran a goodness-of-fit test to see which of the theoretical distributions is best fit for our data. We used generalized additive models for location, scale, and shape (GAMLSS) to fit different candidate distributions to the number discharged per block. GAMLSS is a broader regression type model framework that can handle highly skewed and kurtotic discrete and continuous distributions, as well as the exponential family of distributions to describe the response variable. A detailed description of the GEMLSS models can be obtained in Stasinopoulos & Rigby (2007). We made use of the goodness-of-fit capability of the 'gamlss' package in R to obtain the best fitted theoretical distribution to our data.

Among all the candidate models, we tabulated and plotted the four best models that fit our block number of discharge data for high and low acuity patients, as assessed by the Akaike Information criterion (AIC). Table 5.6 presents the AIC values for the possible models that performed best. Graphical representation of the empirical and theoretical probability distributions for the high and low acuity patients are displayed in Figures 5.3 and 5.4, respectively. The smallest AIC value in Table 5.6 corresponds to the negative binomial distribution for both the high (51147.59) and low (78060.44) acuity patients. Additionally, the Figures 5.3 and 5.4 indicate that negative binomial is the best fitted distribution for both of our high and low acuity discharge data. Therefore, we fit the GLM where the response variable (number discharged) is distributed as a negative binomial distribution with the link function being the log. The degrees of freedom in Table 5.6 equals the number of parameters estimated for each of these distributions.

The effect of high and low acuity census on the high and low acuity discharges (hypotheses 1 and 2) is presented in Table 5.7. It is to be noted that the current block's census has not been considered as a predictor for the current block's discharge. This is due to the fact that the surge

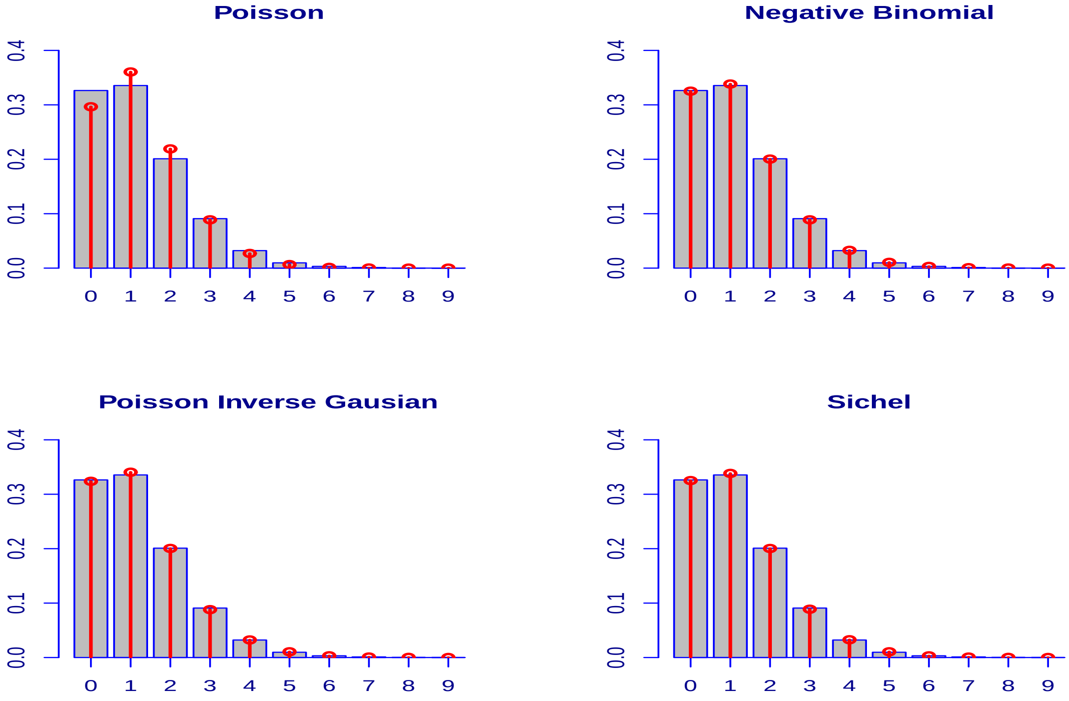


Figure 5.3: Fitted distributions for high acuity patient discharge where solid bars represents empirical distribution and the line bars with dot represents theoretical distribution.

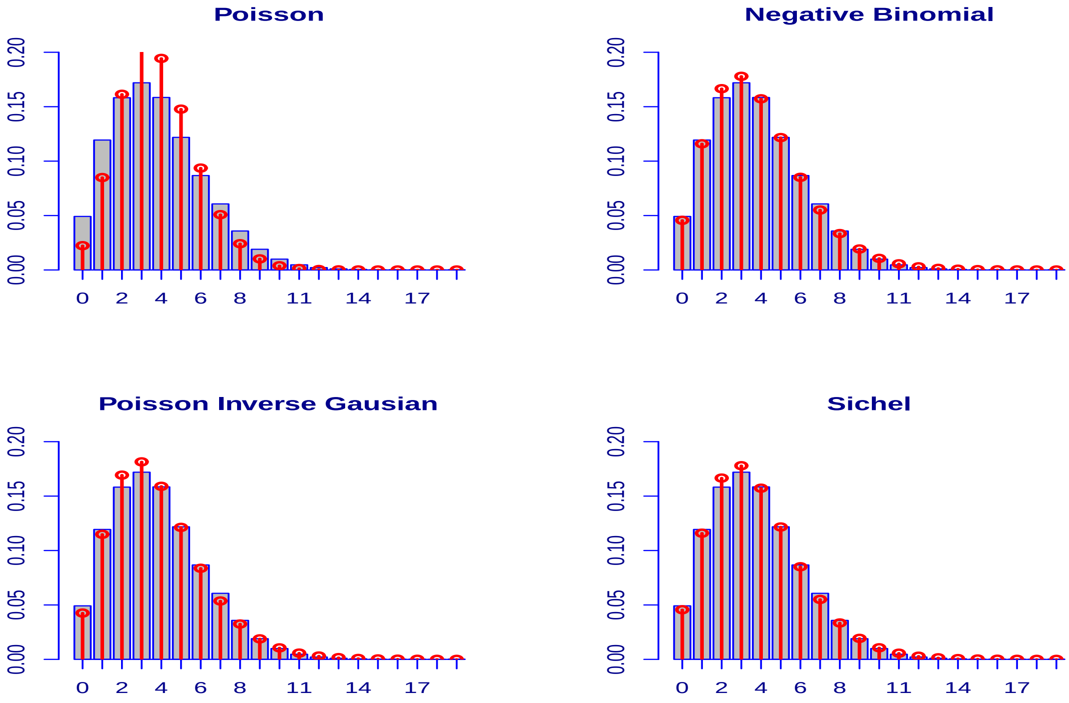


Figure 5.4: Fitted distributions for low acuity patient discharge where solid bars represents empirical distribution and the line bars with dot represents theoretical distribution.

Table 5.6: Model selection using Akaike Information Criterion for the high and low acuity patients, respectively.

Distributions	High Acuity Discharge		Low Acuity Discharge	
	degrees of freedom	AIC	degrees of freedom	AIC
Poisson	1	51359.45	1	79579.50
Negative Binomial	2	**51147.59**	2	**78060.44**
Poisson Inverse Gaussian	2	51149.35	2	78103.69
Sichel	3	51149.59	3	78062.44

in current block census is not manifested until the service process proceeds to the next time block.

We find support for hypothesis 1 (Table 5.7). An increase in the high (low) acuity census in earlier blocks does increase the expected number of high (low) discharges at the current block (p-value < 0.001). Holding everything else constant, an increase of one patient in the high acuity census during the previous block is associated with an 5.6% increase in the expected number of high acuity discharge ($e^{0.055} = 1.056$).

Table 5.7: Negative binomial regression output for high and low acuity patient discharges based on high and low acuity censuses, after controlling for calendar variables.

Predictors	High Acuity			Low Acuity		
	Coef	St. Err	p-value	Coef	St. Err	p-value
Intercept	-0.973	0.061	< .001	0.796	0.034	< .001
High Census (lag1)	**0.055**	0.004	**< .001**	**-0.013**	0.002	**< .001**
Low Census (lag1)	0.003	0.002	0.168	**0.022**	0.001	**< .001**
High Census (lag2)	**0.024**	0.004	**< .001**	0.003	0.002	0.165
Low Census (lag2)	-0.002	0.002	0.356	**0.008**	0.001	**< .001**
High Census (lag1)^2	-0.003	0.000	< 001	0.001	0.000	0.991
Low Census (lag1)^2	0.002	0.000	0.630	-0.003	0.000	< 001
BoD(Dummy)	Control	Control	Control	Control	Control	Control
DoW(Dummy)	Control	Control	Control	Control	Control	Control
MoY(Dummy)	Control	Control	Control	Control	Control	Control

The support for hypothesis 2 is mixed. Although the high acuity census significantly decreases the expected number of low acuity discharges (-0.013, p-value < 0.001), the low acuity census does not significantly decrease the high acuity expected number of discharge (0.003, p-value = 0.168). Holding everything else constant, a unit increase in high census at previous block is associated with a 1.3% decrease in the expected number of low acuity discharge ($e^{-0.013} = 0.987$).

5.5.2.2 Hypotheses 3 and 4: Length of Stay and Workload

We used the Cox-Proportional Hazard (CoxPH) model to investigate the effect of census and admission status indicator on the patients LOS (hypotheses 3, and 4). A positive coefficient indicates that the hazard rate is increasing, and consequently the survival time is shortening. An opposite interpretation for the hazard and survival functions hold for a negative coefficient value. Although the sign of the coefficients are easily interpretable, the same is not true for the magnitudes. Rather, we interpret exponentiated coefficients (hazard ratios) which represent a multiplicative effect in the model.

Based on Table 5.8, we do not find support for hypothesis 3. An increase in same type of patient census does not significantly increase same type of patient LOS. For instance, holding everything else constant, a unit increase in high acuity census at departure increases the hazard of high acuity LOS by 1.7% $\left(1 - e^{0.017}\right)$, i.e., LOS decreases. A possible explanation for such a phenomenon may be the following: as the congestion at departure increases more patients are discharged from the ED to accommodate incoming patients which reduces the overall LOS (or, servers are working at a greater speed to compensate for the factors that push the LOS upward). While looking at the effect of other type of census on other type of patient LOS, only high acuity census increases the low acuity LOS, but not vice versa. We observed that a unit increase in high acuity census is associated with a 1.5% increase in the low acuity LOS. We also noticed that the square terms are either not significant or produced very small coefficient values, which refutes the possibility of the increase (or decrease) of LOS followed by a decrease (or increase).

Finally, We do find support for hypothesis 4 (Table 5.8). Patients admitted to hospital beds have higher LOS than those who are discharged to home. This is true for both the high and low acuity patients (negative coefficients -0.992 and -1.701 for high and low acuity patients, respectively). Compared to discharged patients, there is a 60.3% reduction in hazard of LOS for admitted patients, holding everything else constant.

5.5.2.3 Hypotheses 5 and 6: Waiting Time and Workload

We tested here the effect of high and low acuity patient workload on the high and low acuity patient waiting times. We find support for hypothesis 5 (5.9). An increase in the same type of patient workload is associated with an increase in the waiting time. We observed from Table 5.9 that the current and lag workload values for high acuity patients are significantly affecting the high acuity patient waiting times (p-value < 0.01). Similarly, the current and lag low acuity patient workload are significantly affecting the low acuity patient waiting times (p-value < 0.01). The magnitude of the effect is comparatively high, i.e., a unit increase in workload values for high and low acuity patients is likely to increase high and low acuity patient waiting times by approximately 60 and 20 percent, respectively.

However, we do not find complete support for hypothesis 6. High acuity workload significantly affects low acuity waiting time, but the opposite is not true. This is the reverse of what we found for the service time. Quadratic workload terms are found to be not significantly affecting the waiting times, and the magnitude of effect is minimal. Therefore, according to the data, the patient waiting times continues to increase with an increase in census .

Table 5.8: Predicting patients LOS based on census at departure after controlling for calendar variables using Cox-Proportional Hazard models.

Predictors	High Length of Stay				Low Length of Stay			
	Coef	St. Err	*p*-value	HR	Coef	St. Err	*p*-value	HR
Census High	0.017	0.007	0.023	1.017	-0.014	0.004	0.002	0.985
Census Low	0.003	0.001	0.074	1.003	0.037	0.002	< .001	1.037
Census High (Lag 1)	0.013	0.004	0.001	1.013	-0.006	0.001	< .001	0.993
Census Low (Lag 1)	0.004	0.000	< .001	1.004	0.002	0.000	< .001	1.002
Census High (quad)	-0.000	0.000	0.021	0.999	-0.000	0.000	0.182	0.999
Census Low (quad)	-0.000	0.000	0.001	0.999	0.000	0.000	< .001	1.000
Admited	-0.922	0.015	< .001	0.397	-1.701	0.013	< .001	0.182
Age	-0.011	0.000	< .001	0.988	-0.012	0.000	< .001	0.987
Block of Day(Dummy)	Control	Control	Control	Control	Control	Control	Control	Control
Day of Week(Dummy)	Control	Control	Control	Control	Control	Control	Control	Control
Month of Year(Dummy)	Control	Control	Control	Control	Control	Control	Control	Control

Table 5.9: Cox Proportional Hazard model output for patient waiting time based on ED workload, after controlling for calendar variables.

Predictors	High Acuity				Low Acuity			
	Coef	St. Err	*p*-value	HR	Coef	St. Err	*p*-value	HR
High Workload	-0.914	0.250	**< .001**	0.400	-0.516	0.143	**< .001**	0.596
Low Workload	-0.020	0.056	0.714	0.979	-0.206	0.032	**< .001**	0.813
High Workload (Lag1)	-0.210	0.056	**< .001**	0.810	-0.000	0.032	0.988	0.999
Low Workload (Lag1)	-0.010	0.010	0.327	0.989	-0.021	0.006	**< .001**	0.979
High Workload (sq)	0.413	0.303	0.173	1.512	-0.024	0.173	0.886	0.975
Low Workload (sq)	0.000	0.012	0.980	1.000	-0.008	0.007	0.265	0.991
Busy High	-0.363	0.238	0.127	0.695	-0.423	0.301	0.160	0.654
Busy Low	-0.097	0.049	0.046	0.907	-0.130	0.035	< .001	0.877
Age	0.000	0.000	< .001	1.000	-0.001	0.000	< .001	0.998
BoD (Dummy)	Control	Control	Control	Control	Control	Control	Control	Control
DoW (Dummy)	Control	Control	Control	Control	Control	Control	Control	Control
MoY (Dummy)	Control	Control	Control	Control	Control	Control	Control	Control

5.5.2.4 Hypotheses 7 and 8: Service Time and Workload

In order to model the service times (time it takes to treat a patient), we used the Cox-PH model. We wish to reiterate that positive coefficients are indicative of increasing hazard and shortening survival times, whereas negative coefficients are suggestive of decreasing hazard and expanding survival times. We do find support for hypothesis 7. For the same type of patient, service times initially decrease as a function of patient workload, followed by an increase (p-value < 0.01). For example, one unit increase is high workload is associated with a 82.7% increase in hazard of service time, holding everything else constant. Consequently, a unit increase in the square of the high acuity census is associated with a 60.6% reduction in hazard of service time.

We do not find complete support for hypothesis 8. The low acuity workload significantly affect the high acuity service time (p-value < 0.01), but the opposite is not true (p-value$=$ 0.138). A possible explanation is that when low acuity workload is low, low acuity resources may be used in the high acuity area which is less likely to happen when the low acuity workload is high.

Turning our attention to low acuity patient service times, we noticed that only the low acuity census (linear and quadratic) terms are found to be significant. An additional increase in low acuity workload is associated with a 26.7 percent decrease in the low acuity patient service times, keeping all other factors constant. Conversely, after controlling other factors, low acuity service times are decreased by 5.3 percent with an additional increase in the square of the low acuity patient workload.

Table 5.10: Cox Proportional Hazard model output for patient service time based on ED workload, after controlling for calendar variables.

Predictors	High Acuity				Low Acuity			
	Coef	St. Err	*p*-value	HR	Coef	St. Err	*p*-value	HR
High Workload	0.925	0.251	**< .001**	1.827	0.213	0.144	0.138	1.237
Low Workload	0.311	0.056	**< .001**	1.365	0.236	0.032	**< .001**	1.267
High Workload (Lag1)	0.074	0.056	0.184	1.077	0.016	0.031	0.601	1.016
Low Workload (Lag1)	0.072	0.010	**< .001**	1.075	0.011	0.006	0.063	1.011
High Workload (sq)	-0.929	0.302	**0.002**	0.394	-0.272	0.175	0.120	0.761
Low Workload (sq)	-0.061	0.012	**< .001**	0.940	-0.053	0.007	**< .001**	0.947
Busy High	0.814	0.243	**< .001**	2.258	-0.210	0.302	0.486	0.810
Busy Low	-0.077	0.048	0.114	0.925	-0.012	0.034	0.719	0.987
Age	-0.010	0.000	< .001	0.989	-0.016	0.000	< .001	0.983
BoD (Dummy)	Control	Control	Control	Control	Control	Control	Control	Control
DoW (Dummy)	Control	Control	Control	Control	Control	Control	Control	Control
MoY (Dummy)	Control	Control	Control	Control	Control	Control	Control	Control

5.6 Discussion

In this chapter, we empirically investigated the effect of workload on server productivity based on 2 years of hospital ED data. Four performance measures(the number of patients discharged per hour, as well as the patient LOS, service times, and waiting times) are used to test the hypotheses regarding the effect of workload, patient census, system congestion, and patients who require admission. In addition, we divided our patient population between a high and a low acuity category, and looked at the effect of the same and opposite type of patient workload on patient service performance for both classes.

Our analyses revealed that server productivity increased with an increase in the system workload. We observed that an increase in the same type of patient census is associated with an increase in the average number discharged for both the high and low acuity patients. However, such an increase in the average discharge was not observed for an increase in the opposite type of patient census. Furthermore, an increase in the high acuity census was associated with a decrease in the average low acuity number discharged. This may well be due to the fact that resources may sometimes be pulled out from the low acuity area in order to relieve high acuity patient demand. We also observed that an increase in workload of the same or opposite type of patient is associated with an initial decrease in the LOS and service time; however, in the case of a sustained increase in workload, there is a subsequent decrease in these measures. Patients who required admission were found to have longer LOS than those who were discharged home.

We observed that patient treatment times were affected by the census level in our ED. Since an increase in the same type of patient census increases the average number of patient discharged, a decision regarding the addition of resources should not only depend upon the number of patient waiting for service, but also upon how fast the servers are serving them. At the same time, ED managers also need to ensure that speedy service does not compromise the quality of care. Therefore, our study suggests that ED resource allocation should take into account the workload dependent service processes we have observed.

Our analyses suggest that increased workload is associated with an increased productivity (reduction in patient LOS and service time). However, the increased productivity is shown to be constrained by a threshold point after which the productivity diminishes. Server fatigue due to overwork was considered as a potential factor responsible for decreasing productivity in previous literature. We contrasted the number of patients discharged at the end of the shift with the same quantity at the beginning of the shift, and observe a significantly lower number of patients discharged at the end of the shift. Therefore, if an increase in demand for services is persistent, ED management should call for additional resources during a physician's work shift rather than at the end of his/her shift so long as this is an actionable decision.

Our work contributed to the emergency medicine and operations management literature in several ways. Firstly, we showed that, when workload increases, in addition to the speed-up and slow-down of the ED services, there exists cross-over effects from the high and low acuity patient workload on the low and high patient queue performance measures. Secondly, we demonstrated that under persistent high workload, server performance at the beginning of the shift is significantly higher than that at the end of the shift. In order to maintain the service level, additional resources should be called upon to supplement the current server. Finally, we revealed that ED patients who require admission spend more time in the ED as compare to those who are discharged.

Speed-up and slow-down mechanisms are manifested in all our data analyses, which is consistent with the previous research (Batt & Terwiesch (2012)). We used system load and census variables to predict four different ED queue measures, whereas Batt & Terwiesch (2012) used system load and number of diagnostic tests ordered as covariates to predict task time and service time. The ”quicker at first, slower later” service phenomenon was also observed by KC (2013) while analyzing the effect of physician multitasking on ED performance improvement.

Similar to Berry Jaeker & Tucker (2013), but in a different context, we observed the existence of spillover effect of high and low workload and census on ED queue performance measures for the opposite class of patients. However, the effect was sometimes observed in one direction only. For example, both the high and low acuity census significantly affects the high acuity service times, but only low acuity patient census affect low acuity patient service times. Our observation of longer LOS for admitted patients than for those discharged in consistent with the previous work of Armony *et al.* (2011).

5.7 Limitations and Future Work

Our dataset is comprised primarily of time stamps during the ED service process; it is lacking patient demographic information. More precise results could have been obtained if patient demographics were controlled for in our statistical analyses. Furthermore, our results are based on a single hospital ED. Studies based on multiple EDs would have provided a more generalizable result. Finally, one could contemplate a more extensive statistical model which includes the impact of interaction between covariates on the service quality measures.

Data from other hospitals might suggest that the number discharged follows a different distribution than either the Poisson or Negative Binomial, and hence a different GLM model might need to be fitted under such circumstances. Parametric survival models, such as, accelerated failure time (AFT) models are natural alternatives to the Cox-PH models (Kalbfleisch & Prentice (2011)). AFT models can be used to model time to event data, such as, patient LOS, service time, and waiting time.

Chapter 6

A multi-class multi-server accumulating priority queue with application to health care

6.1 Introduction

One way of dealing with waiting line problems in the presence of diverse client needs is a priority queueing mechanism. A practical example from the field of health care would be the acuity rating systems which have been employed in many countries to classify emergency patients according to their level of severity. In the context of emergency medicine, the Canadian Triage and Acuity Scale [CTAS (2005)] and the Australian Triage Scale [ATS (2000)] (on which CTAS is based) are two examples, where patients are classified into five priority classes (see Table 6.1, below). Each class is associated with a specified performance target assessed in terms of a set of Key Performance Indicators (KPIs). Each KPI comprises a threshold time standard, along with the proportion of patients who should not exceed that time standard. These standards are ostensibly based upon clinical need, although the case can be made that for the lower acuity classes, the KPIs reflect performance benchmarks more than clinical need.

Table 6.1: CTAS Key Performance Indicators

Category	**Classification**	**Access**	**Performance Level**
1	Resuscitation	Immediate	98%
2	Emergency	15 minute	95%
3	Urgent	30 minute	90%
4	Less urgent	60 minute	85%
5	Not urgent	120 minute	80%

A different situation where a prioritized system arises in health care is in hip and knee replacement surgery (Arnett *et al.* (2003)), where distinctions are made among various elective classes (see Table 6.2 below).

Table 6.2: Key Performance Indicators for Hip and Knee Replacement Surgery, Canada

Category	**Wait Time Target**
Emergency	Immediate to 24 hours
Urgent, priority 1	Within 30 days
Urgent, priority 2	Within 90 days
Scheduled	Consultation within 3 mos, Treatment in next 6 mos

There is no reason to presume, a priori, that the stipulated KPIs for each customer class will be met under a classical priority service discipline, for any given set of patient presentation rates. It might well be the case in a two-class system, for instance, that high-priority patients may receive better service than their specified target, while the service level of the low priority patients misses its target. This indicates the need for a priority mechanism that can provide the extra degree of flexibility required to align the observed performance levels with the specified KPIs. The first model to do this was due to Kleinrock (1964), who let customers from a given class (say, k; $k \in \{1, 2, \ldots, K\}$ where K denotes the number of classes) accumulate 'priority' at a rate $b_k > 0$, where $b_k > b_j$ for $k < j$. In this way, a customer from a non-urgent class who experiences a very long wait will eventually accumulate sufficient priority to access the server even when some customers from a more urgent class may be present, and at an earlier time point than if a static priority mechanism were in place.

Kleinrock (1964)'s analysis gave a set of recursive formulae for the mean waiting time before service for each class. However, as illustrated in the two previous examples, the performance of a queueing system in a health care setting is usually specified by the tail of the waiting time distribution for each class, and not by the average waiting times. With this in mind, Stanford *et al.* (2013) recently reconsidered Kleinrock's model, which they renamed the "accumulating priority queue" (APQ), and obtained the waiting time distribution for each priority class in the single server setting.

It can be argued that a variant of the accumulated priority approach is being used already in certain priority health care settings, on an implicit level at least, whenever the deciding health care professional factors the time spent waiting as well as the patient's acuity level in the decision to select the next patient for treatment. In fact, Hay *et al.* (2006) in their simulation model present an approach employing what they call "operational priority", in which each patient is assessed, and assigned an initial priority score which then increases over time.

This note is the first to present distributional results of a queueing-theoretic form for an APQ in a multi-server setting. The results that we present have restricted applicability, in that they require us to assume that all treatment times are exponentially distributed with the same mean. As such, they could be applied in settings such as hip and knee surgery, where treatment durations are comparable for all patient groups (except for Emergency cases such as hip fractures, which are handled separately). The present model cannot be applied in an Emergency Department setting, where treatment times are clearly different for patients of the various acuity levels. (This case is the subject of ongoing work, for which substantial further analytical effort is required.)

The purposes of this note are two-fold. In the first instance, we wish to present the exact transform of the waiting time distribution for each class in the case where treatment times are identical. The second purpose is to carry out numerical investigations to assess the performance of the multi-server APQ model. Typically, for a multi-server system with two or three classes and KPIs with a doubling time benchmark (such as was seen for the lower classes under CTAS and ATS), we are interested in addressing questions such as which are the limiting KPIs, what are the optimal accumulation rates to assign, and what is the maximal traffic load that can be accommodated by a given number of servers.

The remainder of the paper is arranged as follows. In the next section we describe the model. Section 3 contains our derivation of the waiting time distribution for each class. A series of numerical investigations are reported in section 4, where we also present a method for choosing the optimal value of the accumulation rates to satisfy given performance objectives in the two-class case. The final section of the paper gives conclusions and future research directions.

6.2 Description of the Model

The model considered in this note is essentially that in Kleinrock (1964) and Stanford *et al.* (2014), but with $c > 1$ servers, and it is restricted to the case of a common exponential service time distribution for all classes. Customers of class-k, $k = 1, 2, \ldots, K$ arrive to the queue according to a Poisson process with rate λ_k. If a server is free when the customer of class-k arrives, then that customer enters service immediately. Otherwise, they wait in the queue for service, accumulating priority at rate b_k where $b_1 > b_2 > \ldots > b_K$, so class-1 here is the highest priority class, and class-K the lowest. Thus a customer of type k arriving at time t will have accumulated priority $b_k(t' - t)$ by time t'. If all servers are busy, then at the time of the next service completion, the customer that enters service will be the one with the highest accumulated priority at that instant. The common exponential service time distribution has mean $1/\mu$ and Laplace Stieltjes Transform (LST) $\widetilde{B}(s) = \mu/(\mu + s)$. All inter-arrival times and service times are independent of one another. As in Stanford *et al.* (2014), throughout this note, the LST of a random variable with distribution function F will be denoted by $\widetilde{F}$.

In the interests of tractability, we restrict ourselves to the case where the service times are exponentially-distributed with a common mean. Whereas, in a single server queue, the commencement of service for a waiting customer occurs when the service of the preceding customer is completed, in multi-server queue, it occurs when one of the servers becomes free. In the single-server case there are no ongoing services to worry about but, in the multi-server case, the future evolution of the queue will depend on the stage of service of those customers whose service is continuing.

Specifically we need to know the minimal residual service time among the continuing customers, in order to specify when the next customer can move into service. For non-exponential distributions, this task is tractable only when the number of servers is small, and the service time distributions are simple extensions of the exponential, such as Erlang distributions of low order, or mixtures of two exponentials. Even in the case where the other service times are exponential, but with class-dependent means, to characterize the minimum residual time we need to know the mix of continuing customers, and the different possibilities for such a mix make the analysis, at least, very complicated. Furthermore, it is at present unclear how the reordering of service times in an APQ setting affects the duration of the busy periods.

For these reasons, we have opted to solve the common exponential case first and, as we have already noted, it can be a good model for situations such as hip and knee surgery. We are pursuing the non-identical service time case in ongoing work, both analytically and, as in Xiong *et al.* (2013) via a near-perfect simulation approach which can be applied to this situation.

6.3 Waiting Time Distributions

We turn now to finding the distribution of the waiting time before service commences for the various classes. Let $\widetilde{W}^{(k)}(s)$ denote the Laplace transform of the stationary waiting time distribution for customers of class-k; $k = 1, 2, \ldots, K$. We begin by observing that the waiting time prior to service is strictly positive only if an arrival finds all c servers busy, and otherwise it is 0.

Any priority mechanism that selects among waiting customers with service time requirements drawn from the same distribution will have no impact upon the chance that an arrival finds all of the servers busy, which can be identified from the corresponding $M/M/c$ queue.

With $C(A, c)$ being the probability that all servers are simultaneously busy in a stationary $M/M/c$ queue with $A = \lambda/\mu$ and $\lambda = \sum_{i=1}^{K} \lambda_i$, it immediately follows that

$$\widetilde{W}^{(k)}(s) \quad = \quad (1 - C(A, c)) + C(A, c)\widetilde{W}_{+}^{(k)}(s);\ k = 1, 2, \ldots, K. \tag{6.1}$$

where $\widetilde{W}_{+}^{(k)}(s)$ is the LST of the class-k waiting time distribution, conditional on it being positive, that is, conditional on a class-k customer arriving to find all servers busy.

Thus we need to find $\widetilde{W}_{+}^{(k)}(s)$, the LST of the class-$k$ waiting time distribution, conditional on an arrival of class-k finding all servers busy. In the following lemma we will denote this by $\widetilde{W}_{+}^{(k)}(s; \mu, c)$, to explicitly state the dependence of the results on the number of servers c and the common service rate μ for all classes.

Lemma 6.3.1 *Consider the accumulating priority queue where all classes have exponentially distributed service times, with common mean* $1/\mu$. *Then*

$$\widetilde{W}_{+}^{(k)}(s; \mu, c) \quad = \quad \widetilde{W}_{+}^{(k)}(s; c\mu, 1);\ k = 1, 2, \ldots, K. \tag{6.2}$$

Proof: When all servers are busy, the times between service completions are exponentially distributed with parameter $c\mu$ since there are c exponential servers, each serving at rate μ. Thus, the times between service completions have the same distribution as if there were a single exponential server working at rate $c\mu$. Service completion times correspond to times at which the next customer is selected for service, which we can think of as the service selection process. Since the service selection processes have the same distribution in both cases, the distributions of the waiting times until service commences for all classes will also be identical in both cases. □

It follows directly from Lemma 6.3.1 that the waiting time LST for delayed customers in the multi-server case can be obtained directly from the single-server results with service rate $c\mu$. We give below the critical formulas from Stanford *et al.* (2014) for the single-server model, to provide an intuitive appreciation for the nature of the algorithm. The interested reader is directed to Stanford *et al.* (2014) for the detailed analysis.

The results in Stanford *et al.* (2014) rely on the concept of a customer 'becoming accredited at level k', and this, in turn, is expressed in terms of 'the maximum priority process' $(M_1(t), \ldots, M_K(t))$ which is defined precisely there. Intuitively speaking, $M_k(t)$ is the maximum possible priority that a customer of class k can have at time t, given the history of the process up to the time that the current customer entered service. If there is no customer in service at time t, then $M_k(t) = 0$ for all $k = 1, \ldots, K$. We say that a customer of class $i \leq k$ becomes 'accredited at level k' when its priority exceeds $M_{k+1}(t)$. Once this happens, it is not possible for any customer of class $j > k$ to enter service before the customer that has become accredited at level k. In Corollary 7.3 of Stanford *et al.* (2014), the authors establish that customers of class $i \leq k$ become accredited at level k according to a Poisson process at rate $\lambda_i(b_i - b_{k+1})/b_i$.

Thus in particular, class-k become accredited at level k according to a Poisson process at rate $\lambda_k(1 - b_{k+1}/b_k)$. Since any such customer cannot become accredited at a higher level, this fraction $(1 - b_{k+1}/b_k)$ of class-k customers are served during a cycle related to customers who have managed to gain accreditation at level k. The remaining fraction (b_{k+1}/b_k) of class-k customers enter service at one of accreditation levels $k+1$ through n, all of which are achievable by class-$(k + 1)$ customers. It is established in Stanford *et al.* (2014) that the class-k customers who fail to become accredited at level k perceive a scaled version of the class-$(k + 1)$ waiting time distribution.

We can now summarize the results for the waiting time distributions. Let $\rho_i = \lambda_i/c\mu$ for $1 \leq i \leq K$, and $\rho = \sum_{i=1}^{K} \rho_i = \lambda/c\mu$, and $\sigma_k = \sum_{j=1}^{k} \rho_j(b_j - b_{k+1})/b_j$. Finally, we define $\widetilde{W}_{acc}^{(k)}(s)$ to be the LST of the waiting time of a class-k customer who is served at accreditation level k.

The LSTs $\widetilde{W}_{+}^{(k)}(s)$ are calculated for $k = K - 1, K - 2, ..., 1$ starting from $\widetilde{W}_{+}^{(K)}(s) = \widetilde{W}_{acc}^{(K)}(s)$

using the recursion

$$\widetilde{W}_{+}^{(k)}(s) = \left(\frac{b_{k+1}}{b_k}\right)\widetilde{W}_{+}^{(k+1)}\left(\frac{b_{k+1}}{b_k}s\right)+\left(1-\frac{b_{k+1}}{b_k}\right)\widetilde{W}_{acc}^{(k)}(s) \tag{6.3}$$

where $\widetilde{W}_{acc}^{(k)}(s)$ is given by

$$\begin{aligned}\widetilde{W}_{acc}^{(k)}(s) &= \left[\frac{(1-\rho)}{(1-\sigma_k)}+\widetilde{W}_{+}^{(k+1)}\left(\frac{b_{k+1}}{b_k}s\right)\sum_{j=1}^{k}\frac{\rho_j(b_{k+1}/b_j)}{(1-\sigma_k)}\right.\\ &\quad\left.+\sum_{j=k+1}^{K}\frac{\rho_j}{(1-\sigma_k)}\widetilde{W}_{+}^{(j)}\left(\frac{b_j}{b_k}s\right)\right]\widetilde{W}_{acc}^{(k,0)}(s).\end{aligned} \tag{6.4}$$

The term $\widetilde{W}_{acc}^{(k,0)}(s)$ above is defined in Stanford *et al.* (2014), and requires further substitutions in terms of other random variables which the interested reader will find defined there. We list the necessary formulas for the determination of $\widetilde{W}_{acc}^{(k,0)}(s)$ in the Appendix.

6.4 Numerical Investigations

Our purpose in working with the Accumulating Priority Queue (APQ) is to identify a selection scheme that can be applied to different acuity-based triage categories, which factors patients' waiting times into the decision of who to treat next, and which is flexible enough to enable a collection of Key Performance Indicators (KPIs) written in terms of distributional tails to be met. A classical priority mechanism does not achieve this: either the collection of KPIs is met for a given set of patient presentation rates, or it is not. What makes the APQ more flexible in this regard is that we are free to select the accumulation rates $b_k; k = 1, 2, \ldots, K$. That is, we see the priority accumulation rates in a performance setting, as values to be determined that could be adjusted in the event of non-compliance, precisely so as to achieve compliance of all KPIs at a given staffing level, if it is at all possible to achieve compliance. Such a view will be medically appropriate as long as the triage categories and the KPIs themselves have been chosen well to ensure proper care for all patient acuity levels.

We caution that the mechanism is not intended to be used in a situation where the b_k's are selected for individual patients by medical professionals in an attempt to reflect clinical perceptions of the severity of that patient's condition. It is the role of the allocation to triage categories to ensure that patients receive appropriate treatment.

Our first numerical example is loosely drawn from the CTAS model, except for the fact that we assume that all patient classes have the same treatment time distribution, which is unlikely to occur in reality in an Emergency department. We focus on three priority classes based on the KPIs for the urgent (CTAS-3), less urgent (CTAS-4) and non-urgent (CTAS-5) patients with the APQ mechanism. The key performance indicator (KPI) for CTAS-3 (urgent) patients is that they will have to be seen within 30 minutes at least 90% of the time. The KPI requirements for CTAS-4 (less urgent) and CTAS-5 (non urgent) patients require the commencement of treatment within one hour at least 85% of the time and within two hours for at least 80% of the time, respectively.

The Gaver-Stehfest (GS) algorithm (Gaver (1966) and Stehfest (1970)) with 8 points has been used to invert the LSTs of the waiting time distributions. As a check, since the Gaver-Stehfest numerical inversion is in fact an approximation, the results were verified with lengthy simulation experiments, with each simulation run being carried out for one million customers.

Our initial results are for a two-class, two-server APQ model comparing CTAS-3 KPI for class-1 with the CTAS-4 KPI for class-2, based on the following parameters: the arrival rates are $\lambda_1 = 0.9$ and $\lambda_2 = 0.8$ for class-1 and class-2 patients, respectively, while service times for both classes are exponentially distributed with mean $1/\mu = 1$. The accumulation rate, b_1, for class-1 (CTAS-3) is set to one, whereas the accumulation rate for class-2 (CTAS-4) assumes one of three values: $b_2 = 0, 0.5$ and 1.

Figures 6.1 and 6.2 present the resulting class-1 and class-2 waiting time distributions respectively, where Gaver-Stehfest evaluations overlay the simulation results. The fact that no discrepancy can be discerned between the graphs underscores the accuracy of the GS inversion with only 8 points. We see from Figure 6.1 that the class-1 requirements were met for values of $b_2 \leq 0.17$. If class-2 patients were to accumulate priority at a higher rate, the class-1 KPI would fail to be met. Similarly, from Figure 6.2, we see that as long as class-2 patients are accumulating priority at rates b_2 sufficiently close to 1, they are complying with the class-2 KPI. The combination of these statements means that for the given scenario, there is no class-2 accumulation rate b_2 that simultaneously satisfies both KPIs. In order to do so, an additional server would be required.

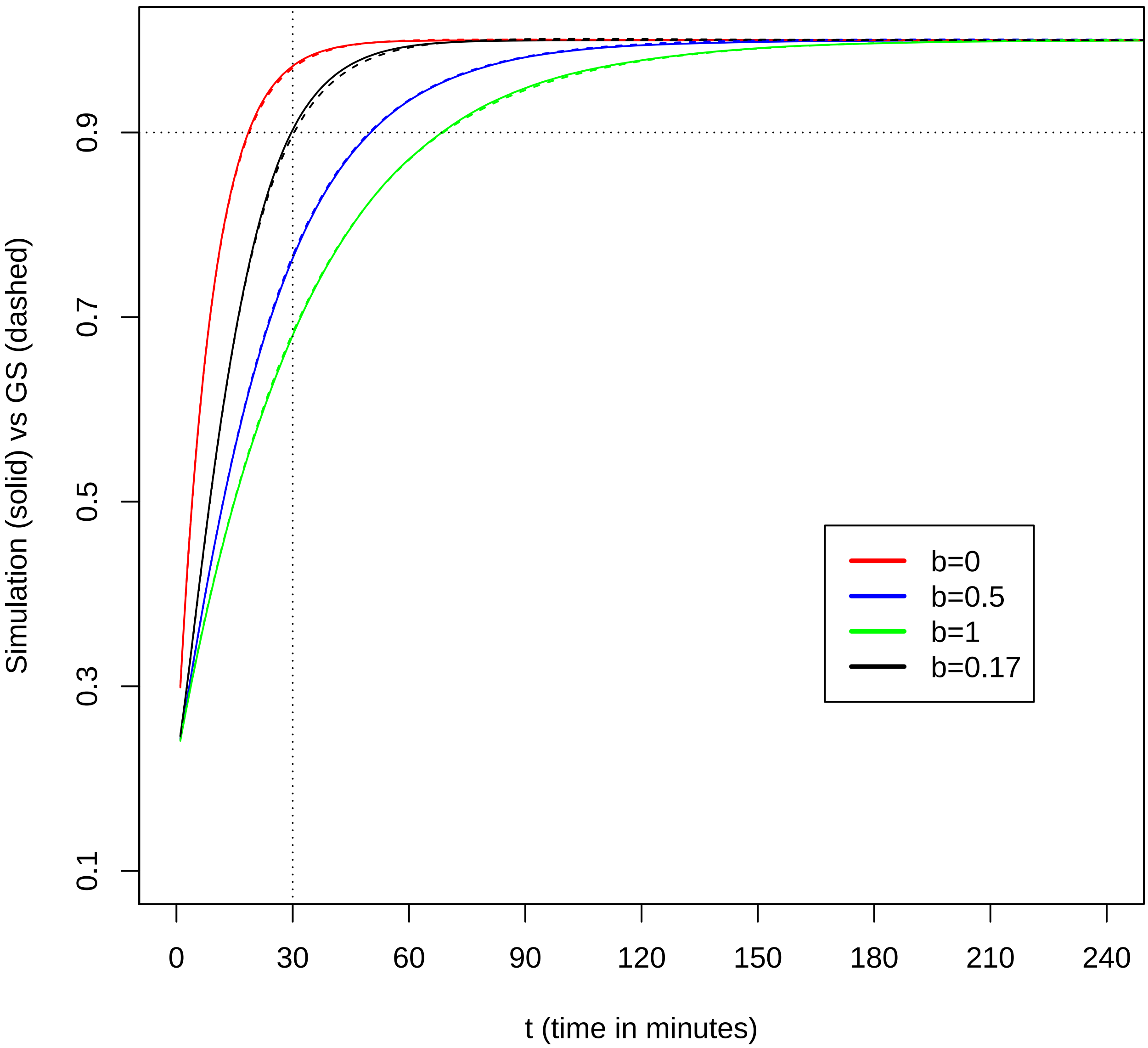


Figure 6.1: The waiting time distribution function for class-1, with arrival rates $\lambda_1 = 0.9$ and $\lambda_2 = 0.8$ and accumulation rates $b_1 = 1$ and $b_2 = 0, 0.5, 1$; simulation (solid line) and Gaver-Stehfest 8 point evaluation (dashed line).

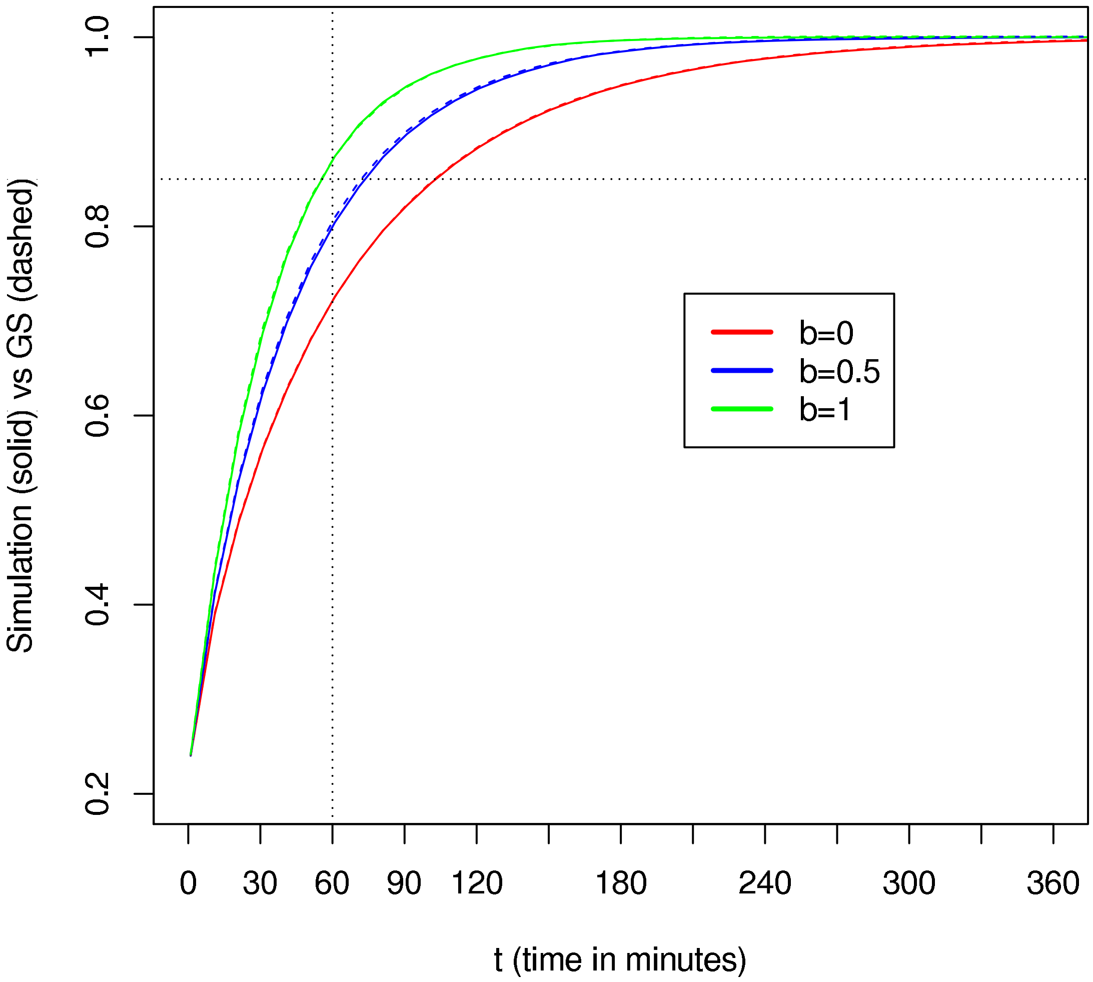


Figure 6.2: The waiting time distribution function for class-2, with arrival rates $\lambda_1 = 0.9$ and $\lambda_2 = 0.8$ and accumulation rates $b_1 = 1$ and $b_2 = 0, 0.5, 1$; simulation (solid line) and Gaver-Stehfest 8 point evaluation (dashed line).

Determining the Limiting Key Performance Indicator

For any given set of arrival rates λ_i; $i = 1, 2$, service rate μ, accumulation rates b_2 (with b_1 arbitrarily set to 1) and number of servers c that correspond to a stable ($\rho < 1$) queueing system, one can determine the waiting time distribution functions for the various classes as in the two foregoing Figures (6.1, 6.2). We seek now to address a more comprehensive question: at what utilisation levels (ρ), and for what values of $b_2 = b$, do the different classes of customers meet their KPI targets? We have prepared a set of figures to respond to this question, with the total utilisation level on the x-axis and permissible $b_2 = b$ values for a specified class along the y-axis. The goal is to determine a "feasible region" comprising a combination of overall utilisation and specified accumulation rate, for which all classes of customer meet their KPI targets. Initially considering the two-class case where both the class-1 and class-2 customers meet their targets, we varied our utilisation rates to see for what values of b_2 both the classes meet their targets, with b_1 fixed at unity, for differing numbers of servers. The feasible region comprises those points below the red curve (indicating the maximal rate b_2 for which the class-1 KPI is still met), and above the blue curve (indicating the minimal rate b_2 for which the class-2 KPI is still met). On the basis of the traffic patterns, one may use these or similar graphs to identify an appropriate b_2 and/or the number of servers to ensure compliance with both targets.

In Figure 6.3, we have plotted four graphs for the single server, two server, three server and five server cases, respectively. The utilisation levels vary along the horizontal axis while the b_2 values vary along the vertical axis. If we look at each graph closely, it is evident that above a certain utilisation level, there is no value of b_2 such that both classes simultaneously meet their KPI targets. For example, in the single server case, any utilisation level above 82% is such that either class-1 or class-2 patients will fail to meet their targets for any value of b_2 , and there are values of b_2 for which both classes fail to meet their targets. One observation we have from the sequence of graphs is that as the number of servers increases, the highest utilisation level for which both classes meet their targets for some value of b_2 also increases.

This result is anticipated, as it reflects a well-known property of multi-server queues: as the number of servers increases, the utilisation level required to produce the same level of delay also increases.

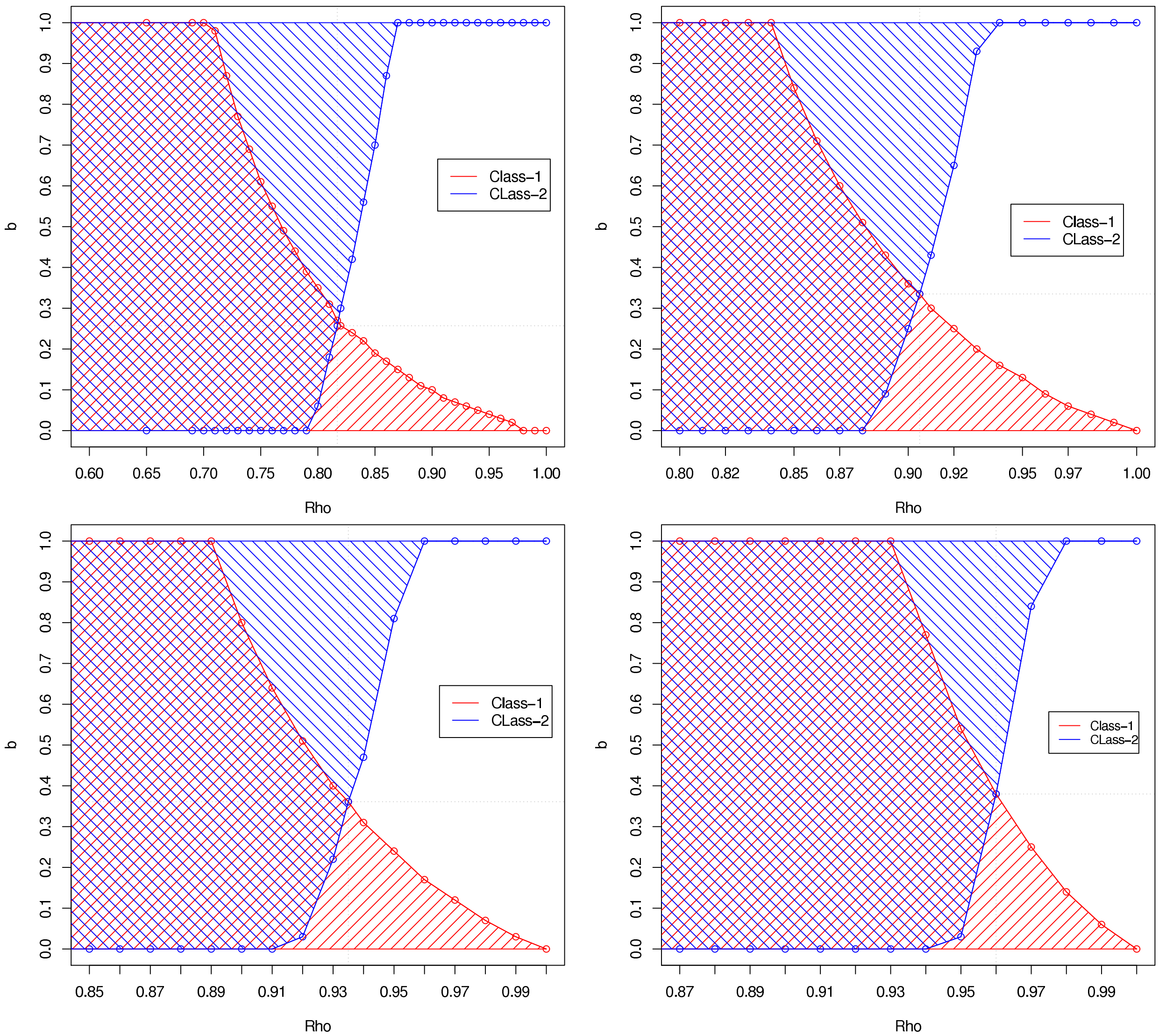


Figure 6.3: Permissible range of values of $0 < b_2 = b < 1$ to meet the class-1 and class-2 KPI where b_1 is set to 1 and arrival rates for both classes are considered equal, for one, two, three and five server cases, respectively (left to right, top to bottom).

Algorithm for finding maximum ρ and optimum b

We observe that in the foregoing examples, depending upon our choice of b_2, the range of utilizations satisfying both KPIs changed. We consider the optimal value of b_2 to be the one for which both KPIs are satisfied for the largest range of the utilization ρ, with the ratio between the arrival rates held constant. The shape of the graphs in Figure 6.3 suggests a procedure for finding the optimal value of b_2 for a given pair of class-1 and class-2 KPIs and a given number of servers. For example, if we choose b_2 equal to about 0.26 in the single-server case, both KPIs will be satisfied for $\rho \leq 82\%$ roughly, while in the five server case, we should choose b_2 to be about 0.39. A modified bisection algorithm that can be used to find the optimal value of b_2 for any two-class system is provided in the appendix. The algorithm, when run for the same parameters as were used in Figure 6.3, leads to the pairs of maximum utilisation ρ_{max} and optimal accumulation rate b_2^* given in Table 6.3.

Table 6.3: Maximum ρ and optimal b_2

Number of Servers	$\rho_{\mathbf{max}}$	b_2^*
1	81.7%	0.257
2	90.5%	0.335
3	93.5%	0.361
5	96.0%	0.380

For a given number c of servers, and specified KPIs, the value of b_2^* that is reported by the above algorithm will ensure that both KPIs are met for the widest range of traffic load possible. In a situation where, at any time, we won't know the traffic parameters precisely, we would argue that this is the best choice of b_2. Furthermore, although it is highly unlikely that the arrival and service rates and number of servers would all conspire to yield a precise utilisation of ρ_{max} in any given application, a choice of $b_2 = b_2^*$ will allow for the maximal possible increase in future demand before another server would be required.

Table 6.3 reveals that as the number of servers increases, both the maximum utilization rate ρ_{max} and the optimal accumulation rate b_2^* increase as well. From Figure 6.3 it appears to be the case that as the number of servers increases, the accumulation rate of the bounding curve for KPI 2 becomes steeper, leading to the higher point of intersection.

We now turn our attention to the development of similar graphs in some three class, two server scenarios. We have considered equal arrival rates (33% for each) for Figures 6.4 and 6.5

and unequal arrival rates (37.5%, 50%, and 12.5%, for class-1, class-2, and class-3 patients, respectively) for Figures 6.6 and 6.7. The service rate was set to unity as before. Utilisation levels were varied on the horizontal axis while b_3 was varied over the vertical axis, subject to $0 < b_3 < b_2$. For Figures 6.4 and 6.6, we have fixed $b_1 = 1$, $b_2 = 0.9$ and let b_3 vary accordingly. However, for Figures 6.5 and 6.7, we have fixed $b_1 = 1$, $b_2 = 0.4$ so that b_3 can only vary between 0 and 0.4. In graphs 6.4 and 6.6, a utilisation is reached such that no value of b_3 can lead to compliance with the class-1 KPI, prior to any limitation being observed for the class-3 KPI. Even in a queue with absolute priority for class-3 (i.e. $b_3 = 0$), its KPI would be met, while having a positive value of b_3 adversely affects the quality of service seen by the higher priority customers.

In Figures 6.5 and 6.7, we observe that the situation is less severe overall, since $b_2 = 0.4$. It is still the case, however, that an impact is seen first for the class-1 KPI as ρ increases.

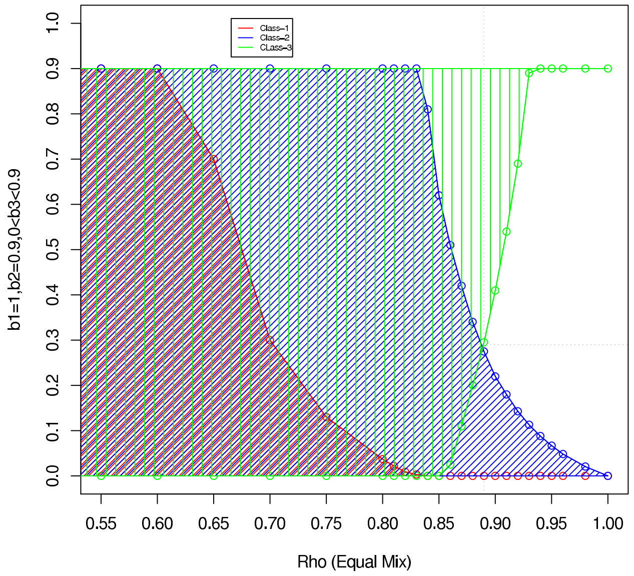


Figure 6.4: Permissible range of values of $0 < b_3 < 0.9$ to meet the class-1, class-2 and class-3 KPI where $b_1 = 1$ and $b_2 = 0.9$ and arrival rates for all classes are considered equal.

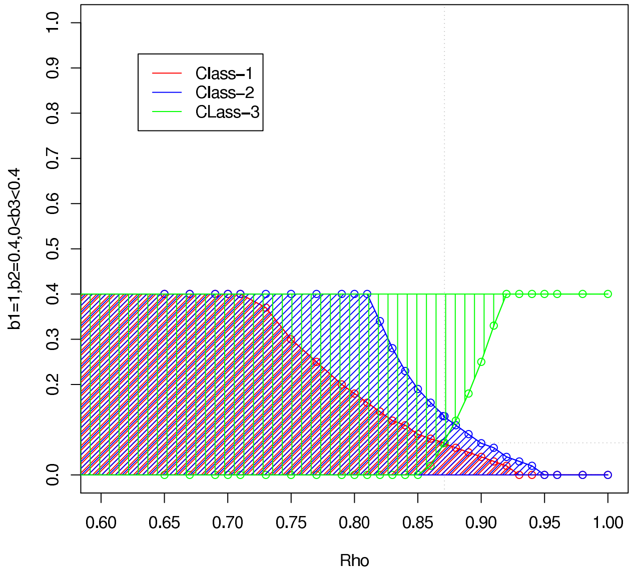


Figure 6.5: Permissible range of values of $0 < b_3 < 0.4$ to meet the class-1, class-2 and class-3 KPI where $b_1 = 1$ and $b_2 = 0.4$ and arrival rates for all classes are considered equal.

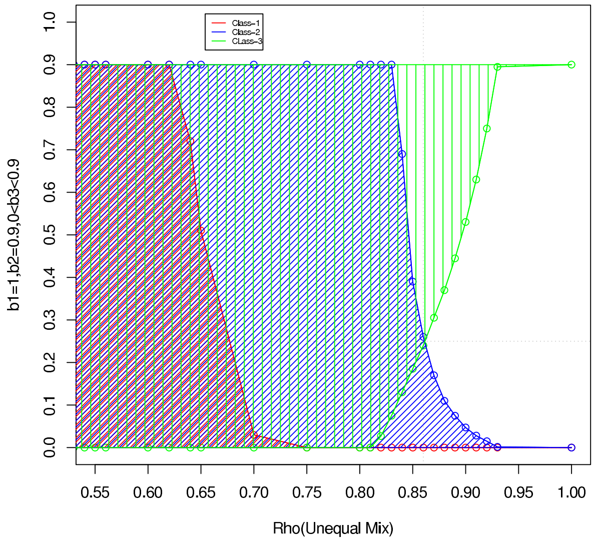


Figure 6.6: Permissible range of values of $0 < b_3 < 0.9$ to meet the class-1, class-2 and class-3 KPI where $b_1 = 1$ and $b_2 = 0.9$ and arrival rates are considered unequal (37.5%,50%,12.5%, respectively).

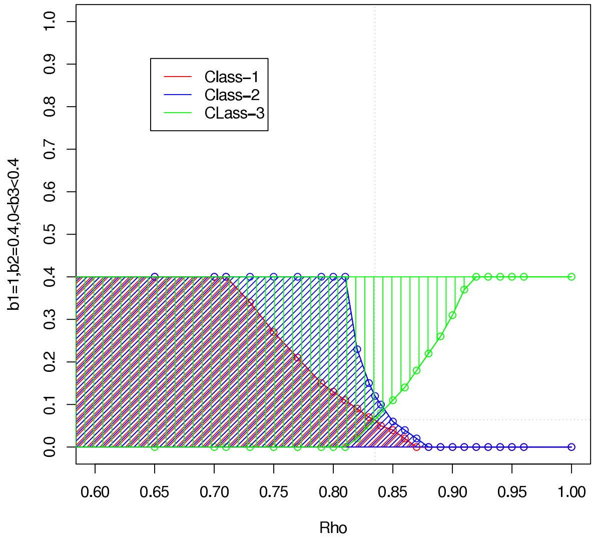


Figure 6.7: Permissible range of values of $0 < b_3 < 0.4$ to meet the class-1, class-2 and class-3 KPI where $b_1 = 1$ and $b_2 = 0.4$ and arrival rates are considered unequal (37.5%,50%,12.5%, respectively).

Applying the bisection algorithm to the 3-class, two server cases considered here, with $b_2 = 0.4$ and b_3 allowed to vary, one obtains the results listed in Table 6.4.

The use of these and similar graphs provides insight into the behavior of the multi-server, multi-class APQ. Depending on the service requirements of different classes of customers, one may fine-tune their respective accumulation rates to try to achieve their performance standard.

Table 6.4: Maximum ρ and optimal b_3

Three Classes (Two Servers)	$\rho_{\max}$	b_3^*
Equal arrival rate	87.1%	0.071
Unequal arrival rate	83.5%	0.064

6.5 Conclusion and Future Research

We have presented a multi-server APQ model for an arbitrary number of customer classes, for which all service times are selected from a common exponential distribution. An obvious extension of our current model is to the situation where the rate of service differs for different classes of patients. However, such a problem is far from trivial analytically, since the duration of a queueing period in a multi-server context will depend on the mix of customers that is in service when the queueing period starts, via the fact that this mix affects the distribution of the time the initial service completions occur.

The extension to non-exponential service time distributions can be considered as effectively intractable in the multi-server context, although some results might be possible in the case of a small number of servers for small-order Erlang distributions, which are built upon the exponential.

The present APQ models, both single-server and multi-server, enable one to ascertain whether a given set of accumulation rates will enable compliance with a given set of KPIs for a given traffic pattern of patient arrival and service rates. This may not be an absolute criteria in a clinical settings (e.g. ED) where either the KPI's are not scientifically defined or the waiting time credits are not allowed to the lower acuity patients (CTAS 4 and 5) as compared to higher acuity patients (CTAS 4). However, the APQ models can help identify other, better time limits for the KPI based performance systems.

We have also considered a particular optimization problem related to the model, namely, the selection of the "best" accumulation rate which allows for the maximum possible utilisation that complies with the KPIs. Another pertinent optimization problem relates to an inherent flaw in KPIs based solely on a target time and compliance probability: patients who exceed the standard are of no consequence, so long as they are few enough in number that the relevant KPI is met. In fact, when such rare events do occur, one would wish to minimize any additional amount of time such patients wait. One such criterion would be the minimization of the total or average amount of excess waiting beyond the specified targets. This topic is among those we hope to address in future work.

Implementation of our model in an emergency department suffers from two clinical difficulties. First, patients in each CTAS category may take variable amounts of time to be assessed and treated. Second, different physicians (servers) may take different amounts of time to assess and treat the same individual. In order to overcome those difficulties one should employ a simulation model where different service time distributions for each CTAS class patients can be assigned and an estimate (along with variability) of the time a server takes to treat a particular CTAS class patient may be obtained from the POWER study (Dreyer *et al.* (2009)). This is another avenue of future reasearch that we hope to pursue.

6.6 Acknowledgments

The work of Azaz Bin Sharif and David Stanford has been supported by the latter's NSERC Discovery grant. Peter Taylor would like to thank the Australian Research Council for supporting his work through Linkage Grant LP0349153 and Discovery Grant DP110101663. Ilze

Ziedins' work was funded by a grant from the Science FRDF at the University of Auckland.

6.7 Appendices

Appendix A: Relevant waiting time formulas

We present below the remaining formulas needed for the determination of $\widetilde{W}^{(k,0)}_{acc}(s)$ which appears in (6.4), for reasons of completeness. The derivation of these expressions can be found in Stanford *et al.* (2013).

The terms $\widetilde{W}^{(k,0)}_{acc}(s)$ are given by

$$\widetilde{W}^{(k,0)}_{acc}(s) = \frac{\{\mu_{\theta_{k-1}} - \sum_{i=1}^{k} \lambda_i (b_k - b_{k+1})/b_i\}\{\widetilde{\gamma}_k(sb_{k+1}/b_k) - \widetilde{\theta}_{k-1}(s)\}}{\{1 - b_{k+1}/b_k\}\{s - (\sum_{i=1}^{k} \lambda_i b_k/b_i)(1 - \widetilde{\theta}_{k-1}(s))\}}. \tag{6.5}$$

where $\widetilde{\theta}_k(s)$, $k = 1, 2, \ldots, K$ is the LST of the duration of a busy period during which customers "gain accreditation relative to class-$k+1$" (see Stanford *et al.* (2013)) and

$$\frac{1}{\mu_{\theta_k}} = \left.\frac{d\widetilde{\theta}_k(s)}{ds}\right|_{s=0} \tag{6.6}$$

is the mean of a random variable with LST $\widetilde{\theta}_k(s)$.

Due to the assumption that customers have a common exponential service time distribution,

$$\widetilde{\theta}_k(s) = \frac{(c\mu + s + \Lambda_k) - \sqrt{(c\mu + s + \Lambda_k)^2 - 4\Lambda_k c\mu}}{2\Lambda_k} \tag{6.7}$$

where

$$\Lambda_k = \sum_{i=1}^{k} \lambda_i (1 - b_{k+1}/b_i). \tag{6.8}$$

Finally, $\widetilde{\gamma}_k(s)$ is the solution to the functional equation $\widetilde{\gamma}_k(s) = \widetilde{\theta}_{k-1}(s + (\sum_{i=1}^{k} \lambda_i (b_k - b_{k+1})/b_i)(1 - \widetilde{\gamma}_k(s))$.

Appendix B: Algorithm for finding maximum ρ and optimum b

1. Set $(\rho_L, b_L) = (0, 0.5)$ and $(\rho_U, b_U) = (1, 0.5)$. It is necessarily the case that both the class-1 and class-2 KPIs must be satisfied when the load is given by ρ_L and the accumulation rate of class-2 traffic is given by b_L, so we know that $\rho_L < \rho_{max}$. It is possible that the class-1 KPI is satisfied when the load is given by ρ_U for some low values of b_U. However the class-2 KPI will certainly not be satisfied for any value of b_U when $\rho_U = 1$, because a non-priority queue with this load is unstable. So we know that $\rho_U > \rho_{max}$.

2. Set $\rho_M = (\rho_L + \rho_U)/2$ and $b_M = (b_L + b_U)/2$.

3. Evaluate whether one, both, or neither of the KPIs is met with utilisation level ρ_M and accumulation rate b_M. If neither is satisfied, then we know that $\rho_M > \rho_{max}$ and so we set $(\rho_U, b_U) = (\rho_M, b_M)$; if both are satisfied, then we know that $\rho_M < \rho_{max}$ and so we set $(\rho_L, b_L) = (\rho_M, b_M)$. In both cases, return to Step 2.

4. In the case where one KPI is satisfied at the point (ρ_M, b_M), while the other is not, we need to search for an appropriate new value of b_M. This is done as follows:

   (a) If KPI 1, but not KPI 2, is met, then set $b_M^L = b_M$ and $b_M^U = 1$. If KPI 2, but not KPI 1, is met, then set $b_M^L = 0$ and $b_M^U = b_M$.

   (b) Set $b_M^M = (b_M^L + b_M^U)/2$. If neither KPI is satisfied at the point (ρ_M, b_M^M), then we know that $\rho_M > \rho_{max}$ and so we set $(\rho_U, b_U) = (\rho_M, b_M^M)$; if both KPIs are met at the point (ρ_M, b_M^M), then we know that $\rho_M < \rho_{max}$ and so we set $(\rho_L, b_L) = (\rho_M, b_M^M)$. In both cases, return to Step 2.

   If KPI 1, but not KPI 2, is met at the point (ρ_M, b_M^M), then set $b_M^L = b_M^M$. Return to Step 4(b).

   If KPI 2, but not KPI 1, is met at the point (ρ_M, b_M^M), then set $b_M^U = b_M^M$. Return to step 4(b).

5. Once $\rho_U - \rho_L < \epsilon$, for some precision level ϵ, the algorithm reports $(\rho_{max}, b_2^*) = (\rho_M, b_M)$.

Chapter 7

An Optimization Problem for Queues Operating under Waiting Time Targets

7.1 Introduction

In many situations, a health care or other service system attends to a number of distinct customer populations with differing urgencies for commencement of service. One ready example arises in the field of emergency medicine, which serves the needs of patients whose lives are in imminent danger, those of moderate urgency, and others with comparatively minor complaints. The Canadian Triage and Acuity Scale CTAS (2005), (see Table 7.1, below) as well as the Australasian Triage Scale ATS (2000) on which it was based, identify five distinct patient populations, and sets a service standard for commencement of service for each group. These standards specify a delay target and a corresponding compliance probability p for each class, such that the chance a patient from the given class will be seen by the delay target (i.e., commence treatment) is at least p.

Table 7.1: CTAS Key Performance Indicators

Category	Classification	Access	Performance Level
1	Resuscitation	Immediate	98%
2	Emergency	15 minute	95%
3	Urgent	30 minute	90%
4	Less urgent	60 minute	85%
5	Not urgent	120 minute	80%

CTAS is an example of a set of Key Performance Indicators (KPIs) comprising delay targets d_i and corresponding compliance probabilities p_i; $i = 1, 2, \ldots, 5$. KPIs are widely used in health care, both for "visible" queues such as those in Emergency departments, as well as "invisible" queues or waiting lists, such as in (Arnett *et al.* (2003)), which pertains to hip and knee replacement surgery. Britain's National Health Service (NHS) uses KPIs for a diverse range of health services, including, for example, mental health (Dodd (2011)), Accident & Emergency department (NHS-Stockport (2014)), Cancer time to treatment (NHS-Leeds (2013)), and Diagnostic tests (TheGuardian (2011)), to name a few. Similar trends can be found in many Western countries.

At face value, there appears to be a close link between systems operating to KPI delay targets and service systems offering specified lead times for the delivery of a particular service (see, for instance, Keskinocak *et al.* (2001), Çelik & Maglaras (2008), Akan *et al.* (2012)). However, lead time problems are typically characterized by a revenue stream, and are typically concerned with the "right" lead time to offer for a specified request as a function of the orders presently in the service system in order to maximize profit or minimize a penalty function. In contrast to this, the delay targets in KPI problems are fixed, and in the health care field where they predominate, they usually having been set by medical professionals in response to the perceived clinical need of the various patient classes. Furthermore, they are typically set prior to any consideration of the traffic characteristics of the patient classes (frequency of demand, treatment time distributions, etc.). It then falls to the health care professionals responsible for the operation of the particular facility to determine a patient selection rule (in queueing terms, a service discipline) so that the KPIs are all met.

In essence, the selection of a suitable service discipline to accommodate the differing delay target needs of the various patient classes in a KPI-driven system is inherently challenging. Not only is a first-come, first served (FCFS) discipline ill-suited to the task, but in all likelihood a standard priority mechanism will fail as well: patients of lower priority are subject to repeated overtaking by recent higher priority arrivals, with no consequence being given for the time such lower priority patients have already spent in the system. What is needed is a system which factors both the time that customers have spent in the system as well as their acuity level so as to better adhere to the stated delay targets. In other words, patients from the various classes should be allowed to accrue priority credit while they wait, at a rate that reflects their relative urgency for treatment. The determination of the average waiting times for a queue operating under such a policy was first considered by Kleinrock (1964). Recently, Stanford *et al.* (2014) determined the waiting time distributions for the various patient classes in what they have called the Accumulating Priority Queue (APQ).

In a nutshell, the advantage of an APQ approach for systems operating under KPIs is that it enables each class of customers to progress fairly towards eventual access to a server by its own waiting time target. Customers can still be overtaken by others of greater urgency or acuity, but they will not be overtaken indefinitely, due to the growing accumulated priority the longer a customer waits.

At the same time, when seen from another perspective, systems designed to respond solely to stated KPIs suffer a fundamental flaw: no consideration is given for the consequences of those patients whose waiting time exceeds the standards. This inability of KPIs to reflect the increased (rather than diminished) urgency of patients whose wait exceeds the specified target was one of the points commented upon by Dr. Chris Baggoley of Australia's Expert Panel Review of Elective Surgery and Emergency Access Targets in a 2012 keynote address (Davies & Little (2012)).

This paper responds to this oversight by presenting an optimization model to minimize the total expected excess delay beyond the target delays for health systems operating under KPIs. We then generalize the model by seeking to minimize a weighted average of the total expected excess over all classes of customers. We seek solutions to the model using the Accumulating Priority Queue (APQ) service discipline. Loosely speaking, the ability to choose the priority accumulation rates for the various classes provide an extra margin of flexibility over the standard non-preemptive priority discipline to ensure that the traffic patterns observed in the system for the classes of customers adhere to the stated KPIs.

When minimizing the total expected excess waiting time, it will be seen that our numerical examples reveal that this leads to an easily implemented rule of thumb for the optimal priority accumulation rates which can have an immediate impact on health care delivery. The rule of thumb says that the patients from each class should accumulate priority at a rate in inverse proportion to their delay targets. When this is done, patients from any class whose waiting time is approaching their delay target will tend to have similar accumulated priorities to that point in time, thereby ensuring comparable likelihoods of waiting time excess.

It should be stressed that there is no obligation upon a health facility governed by a set of KPIs to implement the same strategy for all its patient classes. In Emergency departments, one would never make a Resuscitation or Emergency patient wait for someone of lower acuity. However, one could still envisage this as part of an APQ setting merely by allowing for infinitely-large accumulation rates for these patient classes.

The rest of the paper is organized as follows. In section 2, we present the optimization problems for both the general case, and its restriction to Accumulating Priority service discipline. We end the section by addressing the matter of convexity of the corresponding APQ optimization problem.

Section 3 relates the Laplace transforms of the waiting time distributions and expected excesses. We show that when one resorts to a numerical inversion of the pertinent waiting time transform to compute the probabilities, the corresponding numerical inversion of the expected excess waiting time is obtained with minimal additional effort. Due to its simplicity and ease of implementation, Section 4 presents our preferred numerical inversion method, Gaver-Stehfest algorithm (Gaver (1966), Stehfest (1970)), and present the functional relationships needed to obtain the numerical results that follow.

In section 5, we present a series of numerical examples which explore the optimal behaviour of the APQ. The nature and impact of our various results and insights are summarized in the Conclusions section.

7.2 Formulation of the Optimization Problem

Consider a queue featuring either a single server or many servers, which attend(s) to L independent classes of customers with distinct KPIs of the form discussed in the introduction. Arrivals for the nth class of customers are from a Poisson process at rate $\lambda_n; n = 1, 2, \ldots, L$. All service time distributions are known, and they may differ from class to class.

We presume that the queue is operating under a particular service discipline $\mathbf{a} \in \mathcal{A}$ (that is, a rule for selecting the next customer to enter service), where $\mathcal{A}$ denotes the set of permissible work-conserving (server is not idle unless queue is empty) disciplines. The set $\mathcal{A}$ includes FCFS, last-come, first served (LCFS), Random Order of Service (ROS) and both Non-preemptive (NP) and Preemptive Resume (PR) disciplines, among others. We presume that the queue has been operating sufficiently long to have reached stationarity.

Letting $\mathcal{W}_n$ denote the stationary class-n waiting time random variable, define the following for $n = 1, 2, \ldots, L$:

- $W_n(x) = P(\mathcal{W}_n \leq x)$ is the cumulative distribution function,
- $S_n(x) = P(\mathcal{W}_n > x) = 1 - W_n(x)$ is the survival function, and
- $w_n(x) = dW_n(x)/dx$ is the probability density function of the stationary class-n waiting time distribution.

We will denote the respective Laplace transforms of these quantities, and all others to be defined below, by a $\tilde{}$ throughout the paper.

The expected amount of excess waiting time $H_n(d_n)$ for a typical class-n customer beyond a specified delay d_n, henceforth abbreviated the "expected excess", can be determined from

$$H_n(d_n) = \int_{d_n}^{\infty} (x - d_n) w_n(x) dx = \int_{d_n}^{\infty} S_n(x) dx. \tag{7.1}$$

The latter integral above is obtained by integrating the former one by parts. The quantity $\lambda_n H_n(d_n)$ can be interpreted as the expected amount of excess waiting for class-n customers per unit of time.

For such a queue as described above, the general form of the optimization problem can be stated as follows. We seek to minimize a weighted average of the total expected excess waiting per unit time over all permissible customer selection strategies,

$$\begin{array}{lrcl} \min_{\mathbf{a}\in\mathcal{A}} & Z & = & \sum_{n=1}^{L} \alpha_n \lambda_n H_n(d_n) \\ \text{subject to} & W_n(d_n) & \geq & p_n \ ; \ n \ = \ 1,\ 2,\ \ldots,\ L; \end{array}$$

where, $\alpha_n, n = 1, 2, \ldots, L$ denote the respective weights for the excess waiting for class-n customers. A priori, we observe that the stated problem might be infeasible. If feasible, we observe that there might conceivably be more than one optimal solution.

7.2.1 Formulating the APQ Optimization Problem

In Stanford *et al.* (2014), the APQ model first introduced by Kleinrock (1964), was reconsidered, and the stationary waiting time distributions for each class of customer were determined. Kleinrock (1964)'s model presumed that waiting customers of the ith priority class accrue priority "credit" at a rate $b_i \geq 0$, where $b_i \geq b_j$ if class i is considered to be of higher priority to class j. In what follows, we presume that a lower class index means a higher priority, so that class 1 has top priority and class L has the lowest priority. We note that the APQ discipline is work-conserving.

The APQ is a simple yet flexible scheduling model which allows both the priority of the customer (or acuity of the patient in health care applications) as well as its time spent waiting to be factored into the decision as to which customer next enters service. Setting $b_i = B > 0\ \forall i$, a first-come first-served queue that aggregates all customer classes is obtained. Setting $b_L = 1$ while $b_i = b_{i+1} * M;\ i = 1, 2, \ldots, L-1$, a classical non-preemptive priority model is obtained in the limit as $M \to \infty$.

To date, analytical solutions for the waiting time distributions exist for two cases: Stanford *et al.* (2014) presents the solution for the single-server case where each class may have its own specified service time distribution, while Sharif *et al.* (2014) addresses the multi-server case when all service times are drawn from the same exponential service time distribution, with service rate μ.

Let $\mathcal{A}_{APQ}$ denote the set of APQ service disciplines. The resulting APQ optimization problem then reduces to the selection of the best accumulation rates $b_i;\ i = 1, 2, \ldots, L$ in order to minimize the weighted average excess:

$$\begin{array}{lrcl} \min_{\mathbf{a}\in\mathcal{A}_{APQ}} & Z & = & \sum_{n=1}^{L} \alpha_n \lambda_n H_n(d_n) \\ \text{subject to} & W_n(d_n) & \geq & p_n\ ;\ n\ =\ 1,\ 2,\ \ldots,\ L; \\ & b_n & \geq & b_{n+1}\ ;\ n\ =\ 1,\ 2,\ \ldots,\ L-1; \\ & b_L & \geq & 0; \end{array}$$

where, $\alpha_n, n = 1, 2, \ldots, L$ again denotes the respective weight for the excess waiting for class-n customers. It is to be noted that the role of α_n's is not to make higher priority customers gain priority faster (that's the role of b_n's), rather α_n's provide greater weight to waiting times of the given class of customers, and as such, assess a greater penalty for incurred waits by those customers.

7.3 Convexity in the Two-class APQ Optimization Problem

It is immediately a matter of interest as to whether the APQ optimization problem is convex in the decision variables b_i. Unfortunately, as even the simplest numerical examples in the 2-class case show later, the expected class-1 excess $H_1(d_1)$ can have a negative curvature over some or all of its range as a function of its parameters, so we have been unable to infer convexity of the resulting weighted average even in the simplest case based upon mixtures of convex functions. We can establish, however, that the expected class-2 excess $H_2(d_2)$ is convex in b_2 in a two-class problem, and we can infer that the same will hold true for the lowest class L as a function of its accumulation rate b_L in a multi-class APQ problem.

Since it is the ratio b_2/b_1 of the priority accumulation rates rather than their individual values that determine which customer will be selected in the two-class case, we can arbitrarily set $b_1 = 1$ and investigate the behaviour as a function of one variable $b_2 = b$ where $0 \leq b \leq 1$.

Henceforth, we state explicitly the dependence upon b of our preceding probability functions $W_n(b, x)$, $S_n(b, x)$, $w_n(b, x)$ and $H_n(b, d_n)$; $n = 1, 2, \ldots, L$, as our goal is to establish that $H_2(b, d_2)$ is a convex function of b. To do so, we first need to establish an important lemma, whose proof has been relegated to the appendix.

Lemma 1. The stationary waiting time random variable $\mathcal{W}_2$ for class-2 customers in a stable two-class APQ can be expressed as the sum of two dependent random variables

$$\mathcal{W}_2 = \mathcal{W} + Y \tag{7.2}$$

where $\mathcal{W}$ denotes the stationary waiting time random variable in the M/G/1 FCFS comparator queue, and Y refers to a compound Poisson random variable

$$Y = \eta_1 + \eta_2 + \ldots + \eta_N \tag{7.3}$$

where the random variable N denotes the number of class-1 customers to accredit during $\mathcal{W}$, and $\eta_i;\ i = 1, 2, \ldots$ denote a sequence of i.i.d. busy period random variables for an M/G/1 queue with arrivals at rate $\lambda_1(1-b)$ and whose service times are drawn from the class-1 service time distribution. (In (7.3), it is to be understood that if $N = 0$, then $Y = 0$.)

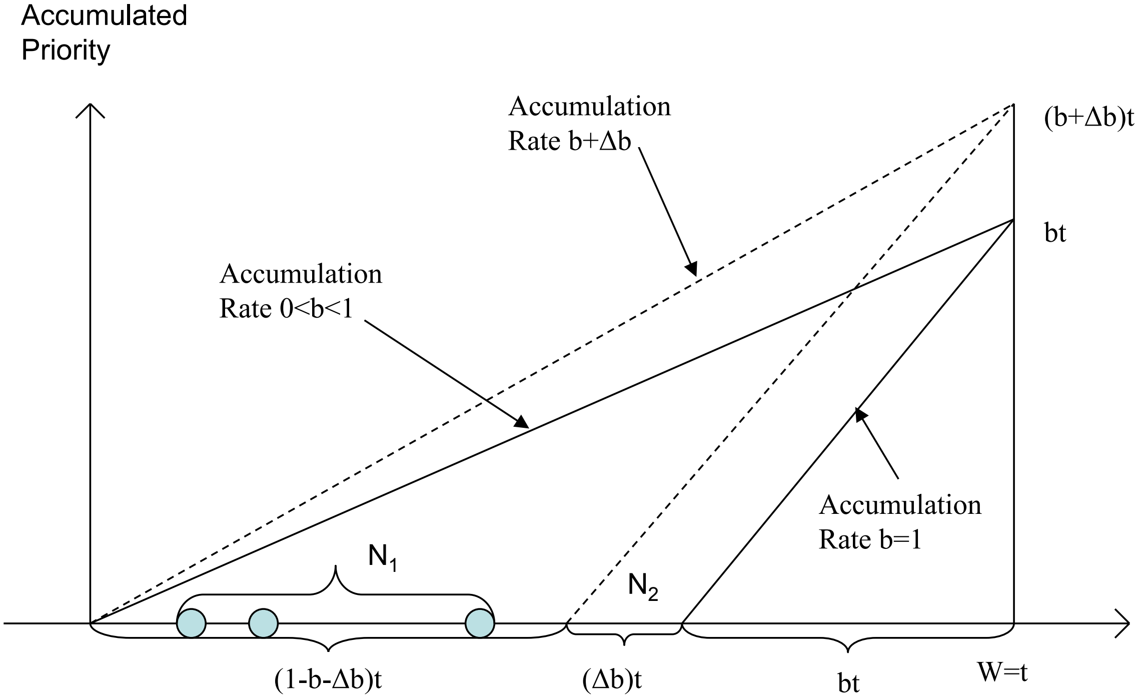


Figure 7.1: Impact of change in Accumulation Rate

Theorem 1. Given all the foregoing definitions, the function $S_2(b, x)$ is a monotonically decreasing convex function in b for $0 \leq b \leq 1, \forall x \geq 0$.

Proof. We proceed by calculating $(\partial/\partial b)S_2(b, x)$ from first principles:

$$(\partial/\partial b)[S_2(b, x)] = \lim_{\triangle b \to 0} \left\{ \frac{S_2(b + \triangle b, x) - S_2(b, x)}{\triangle b} \right\}$$

Reconsider the revised service discipline used in Lemma 1. From Figure 1, which presumes that we have conditioned upon $\mathcal{W} = t$, closer inspection of the accreditation process reveals that the N Poisson events underpinning Y are those that occur at rate λ_1 during the first $(1 - b)$ portion of $\mathcal{W}$. If the class-2 accreditation rate were to change to $b + \triangle b$, then only those events occurring during the first $(1 - b - \triangle b)$ portion of $\mathcal{W}$ would contribute to the compound Poisson process delaying our tagged customer.

Due to the independence of compound Poisson processes arising from Poisson events in non-overlapping intervals, we can write $Y = Y_1 + Y_2$ where Y_1 is the compound Poisson process corresponding to the N_1 accreditation events that occurred during the first $(1 - b - \triangle b)$ portion of $\mathcal{W}$, and Y_2 is the compound Poisson process corresponding to the N_2 accreditation events occurring during the subsequent $(\triangle b)$ portion of $\mathcal{W}$. Thus we obtain immediately

$$S_2(b, x) = P(\mathcal{W} + Y_1 + Y_2 > x); \;\; S_2(b + \triangle b, x) = P(\mathcal{W} + Y_1 > x).$$

Therefore

$$\begin{aligned} S_2(b + \triangle b, x) - S_2(b, x) &= -[P(\mathcal{W} + Y_1 + Y_2 > x) - P(\mathcal{W} + Y_1 > x)] \\ &= -[P(\mathcal{W} + Y_1 \leq x; \; \mathcal{W} + Y_1 + Y_2 > x)] \; < \; 0. \end{aligned}$$

Now for sufficiently small $\triangle b$, the probability of two or more events in a small period of duration $\triangle b \mathcal{W}$ is $o(\triangle b \mathcal{W})$. Similarly, the probability of one such event is $\lambda_1 \triangle b \mathcal{W} + o(\triangle b \mathcal{W})$. Substituting the terms as appropriate and removing the conditioning upon $\mathcal{W}$, one readily obtains

$$(\partial / \partial b)[S_2(b, x)] = -\lambda_1 E(\{\mathcal{W} \times I(\mathcal{W} + Y + \eta > x)\} \; < \; 0. \tag{7.4}$$

[In (7.4), $I(A)$ denotes the indicator function for the event A, equal to 1 if the event occurs, and 0 otherwise.]

Since the partial derivative is negative $\forall x$, this establishes that $S_2(b, x)$ is a monotonically decreasing function of b. Furthermore, as b increases, Y decreases in distribution, since the rate of the corresponding underlying Poisson process at rate $\lambda_1(1 - b)$ decreases. Consequently, the probability associated with the indicator function decreases, so that the derivative is a strictly increasing function of b, and the convexity immediately follows. □

Corollary 1. The excess function $H_2(b, d_2)$ above is monotonically decreasing and strictly convex in b for $0 \leq b \leq 1$ for every fixed $d_2 \geq 0$.

Proof. This follows immediately from Theorem 1 and $H_2(b, d_2) = \int_{u=d_2}^{\infty} S_2(b, u)du$. □

Corollary 2. The excess function $H_L(b_L, d_L)$ for the lowest priority class in a stable L-class APQ is strictly convex in b_L for $0 \leq b_L \leq b_{L-1}$, for every fixed $d_L \geq 0$.

Proof. The lowest priority class (class N) observes the unfinished workload in the corresponding M/G/1 FCFS queue at its arrival instants. By Corollary 7.3 in Stanford *et al.* (2014), the higher classes $1, 2, \ldots, L-1$ gain accreditation over class-L customers according to a Poisson process at rate $\sum_{i=1}^{L-1} \lambda_i(1 - b_L/b_i)$. The rest follows by direct analogy to the proof of the two-class case. □

While the class-1 expected excess function is not always convex, we can establish in the single-server case that it is monotonically non-decreasing in b for all values of d_1.

Theorem 2. For single-server systems, the function $S_1(b, x)$ is a monotonically non-decreasing function in b for $0 \leq b \leq 1, \forall x \geq 0$.

Proof. We compare sample paths in two stable APQ systems with fixed arrival rates, common service rate, and with the class-1 accumulation rate set to $b_1 = 1$. The sequences of arrival instants and service requirements for all customers in the two systems are the same. The only difference is that in the former case, the class-2 arrival rate is $b_2 = b$ where $0 \leq b \leq 1 - \Delta b$, whereas in the latter it is $b_2 = b + \Delta b \leq 1$, for small $\Delta b \geq 0$.

The busy periods of the queues in the two systems are the same, since the unfinished workload functions for both systems would be the same. The APQ service discipline merely rearranges the order in which some customers are served, relative to any other work-conserving discipline.

Consider a delayed class-1 customer in the former APQ system; we will refer to it as the "tagged" customer. In the latter system, all of the class-2 customers formerly served ahead of the tagged customer will still be served ahead of them. It may be the case that due to the increased rate of priority accumulation, some class-2 customer formerly present upon the arrival of the tagged class-1 customer might have completed service. However, all of the service times completed by the server in the former system prior to the tagged customer's entry would still be completed under the latter. The only difference is that there may be some more class-2 customer(s) who formerly were selected for service after the tagged customer, who now under the higher priority accumulation rate $b_2 = b + \Delta b$ might enter service ahead of it, leading to a longer waiting time. As the probability of such an event is positive, it follows that $S_1(b, x)$ is a monotonically non-decreasing function in b for $0 \leq b \leq 1, \forall x \geq 0.\square$

Corollary 3. For single-server systems, the excess function $H_1(b, d_1)$ above is monotonically non-decreasing in b for $0 \leq b \leq 1$ for every fixed $d_2 \geq 0$.

Proof. This follows immediately from Theorem 2 and $H_1(b, d_1) = \int_{u=d_1}^{\infty} S_1(b, u)du$. $\square$

Remark. The same line of proof cannot be used in the multi-server case, since the re-arranging of service times can lead to customers being served by different servers, and even to separate periods of time when all servers are busy. Nonetheless, we conjecture that the claim is still true in the multi-server case.

7.4 Relationship between the Transforms of the Waiting Time CDF and the Excess Wait

In this section, we once again consider an arbitrary service discipline for which the stationary waiting time distribution can be specified in terms of its Laplace-Stieltjes transform for each class of customers. While many such queueing systems have been analyzed successfully in this way, it is frequently the case that one cannot invert the LST analytically. In such cases, one commonly resorts to one of many effective procedures available (for example, Brigham & Morrow (1967), Gaver (1966), Stehfest (1970)) which numerically invert the LST to recover the desired probabilities. We establish in this section that the numerical evaluation of the expected excess per customer is a straightforward matter via numerical inversion for any service discipline for which the waiting time LSTs are readily available. We do so by establishing that the transform of the amount of excess waiting time can be expressed in terms of the waiting time LST which we use to evaluate KPI compliance. The implication is that the objective function which seeks to minimize the amount of excess can be evaluated with minimal extra effort beyond assessing whether KPI compliance has been achieved.

The Laplace transforms of $w_n(t)$, $W_n(t)$, $S_n(t)$, and $H_n(t)$ are given by

$$\begin{aligned}
\widetilde{w}_n(s) &= \int_{t=0}^{\infty} e^{-st} w_n(t)dt = \int_{t=0}^{\infty} e^{-st} dW_n(t) \\
\widetilde{W}_n(s) &= \int_{t=0}^{\infty} e^{-st} W_n(t)dt \\
\widetilde{S}_n(s) &= \int_{t=0}^{\infty} e^{-st} S_n(t)dt \\
\widetilde{H}_n(s) &= \int_{t=0}^{\infty} e^{-st} H_n(t)dt
\end{aligned}$$

Standard properties of Laplace transforms imply that

$$\widetilde{W}_n(s) = \frac{\widetilde{w}_n(s)}{s} \tag{7.5}$$

and

$$\widetilde{S}_n(s) = \frac{1}{s} - \widetilde{W}_n(s). \tag{7.6}$$

From (7.1), it follows that $H_n'(t) = -S_n(t)$, so that when integrating by parts,

$$\begin{aligned}\widetilde{H}_n(s) &= \int_{t=0}^{\infty} H_n(t)e^{-st}dt \\ &= \left[\frac{H_n(t)e^{-st}}{-s}\right]_0^{\infty} + \int_{t=0}^{\infty} \frac{e^{-st}}{s} H_n'(t)dt \\ &= \frac{H_n(0)}{s} - \frac{1}{s}\widetilde{S}_n(s).\end{aligned}$$

Let m_n = mean class-n waiting time. Since $H_n(0) = \int_{x=0}^{\infty} S_n(x)dx = m_n$,

$$\widetilde{H}_n(s) = \frac{m_n}{s} - \frac{1}{s^2} + \frac{\widetilde{w}_n(s)}{s^2}. \tag{7.7}$$

It is immediately apparent from (7.7) that the evaluation transform of the expected excess per customer $\widetilde{H}_n(s)$ is readily obtained for any value of s once the corresponding evaluation has been carried out for the transform of the waiting time distribution $\widetilde{w}_n(s)$. In the next section, we present the specific numerical inversion approach we use to achieve both tasks.

7.5 The Utility of Gaver-Stehfest Numerical Inversion for Evaluating Expected Excess

The numerical inversion of Laplace transforms has been an alternative to analytical inversion since the Fast Fourier Transform (FFT) technique gained popularity Brigham & Morrow (1967). Whereas the FFT can require hundreds of evaluations for a single time point of interest, there are alternatives that require only a handful of evaluations. The one we use has come to be known as Gaver-Stehfest numerical inversion; it is so named because the pioneering probabilistic work of Gaver (1966) was later refined algorithmically by Stehfest (1970).

Given a real-valued function $f(t); t \geq 0$ whose Laplace transform is $\widetilde{f}(s)$, then the Gaver-

Stehfest method for numerical Laplace transform inversion at the point t is given by the following (Gaver (1966), Stehfest (1970)):

$$f(t) = \frac{ln2}{t} \sum_{k=1}^{K} V_k \, \widetilde{f}\left(\frac{ln2}{t} \times k\right) \tag{7.8}$$

where the values V_k are the Gaver-Stehfest coefficients of order K (always even), half of which are positive and half negative numbers. These coefficients, as derived by Gaver, are combinatorial terms arising in order statistics, with the useful by-product that they always sum to 0. Typically $K = 8$ points provides two significant digits of accuracy, which is quite adequate for assessing waiting times. The table that provides the coefficients for $K = 2; 4; 6; 8$ is provided in the appendix.

In light of (7.5) and (7.8), the Gaver-Stehfest evaluation of the class-n waiting time distribution, $n = 1, 2, \ldots, N$ is achieved via

$$W_n(t) = \frac{ln2}{t} \sum_{k=1}^{K} V_k \, \frac{\widetilde{w}_n\left(\frac{ln2}{t} \times k\right)}{\left(\frac{ln2}{t} \times k\right)} = \sum_{k=1}^{K} \frac{V_k}{k} \, \widetilde{w}_n\left(\frac{ln2}{t} \times k\right). \tag{7.9}$$

Meanwhile, the Gaver-Stehfest numerical evaluation of the class-n expected excess per customer function, $n = 1, 2, \ldots, N$ is given by

$$\begin{aligned} H_n(t) &= \frac{ln2}{t} \sum_{k=1}^{K} V_k \, \widetilde{H}_n\left(\frac{ln2}{t} \times k\right) \\ &= \frac{ln2}{t} \sum_{k=1}^{K} V_k \left[\frac{m_n}{(\frac{ln2}{t} \times k)} - \frac{1}{(\frac{ln2}{t} \times k)^2} + \frac{\widetilde{w}(\frac{ln2}{t} \times k)}{(\frac{ln2}{t} \times k)^2}\right] \\ &= \sum_{k=1}^{K} \frac{V_k}{k} \left[m_n - \frac{1}{(\frac{ln2}{t} \times k)} + \frac{\widetilde{w}(\frac{ln2}{t} \times k)}{(\frac{ln2}{t} \times k)}\right]. \end{aligned} \tag{7.10}$$

It is readily apparent from a comparison of (7.9) and (7.10) that minimal extra effort is involved in determining the expected excess per customer beyond the specified thresholds $H_n(d_n); n = 1, 2, \ldots, N$ once the evaluations of the corresponding compliance probabilities $W_n(d_n)$ have been performed.

7.5.1 Necessary Formulae from the APQ Multi-server Model

To complete the optimization approach, we specify below the waiting time LSTs for each class of customers. We do this for the case of a multi-server APQ with c servers and $N = 2$ priority classes, where the service times for all classes are exponentially distributed at rate μ. We start with the specification of $\widetilde{w}_2(s)$.

Theorem 3. The Laplace-Stieltjes transform $\widetilde{w}_2(s)$ of the class-2 waiting time distribution can be written as

$$\widetilde{w}_2(s) = \widetilde{w}(s + \lambda_1(1 - b)(1 - \widetilde{\eta}_1(s))) \tag{7.11}$$

where $\widetilde{\eta}_1(s)$ is the Laplace-Stieltjes transform of the duration of a busy period of accrediting customers, and is obtained as the solution to the functional equation

$$\widetilde{\eta}_1(s) = c\mu/(c\mu + [s + \lambda_1(1 - b)(1 - \widetilde{\eta}_1(s))]). \tag{7.12}$$

Furthermore, the LST $\widetilde{w}(s)$ represents the LST of the waiting time in an M/M/c queue, and is given by

$$\widetilde{w}(s) = [1 - C(A, c)] + C(A, c) \left[(c\mu - \lambda)/(c\mu - \lambda + s)\right]. \tag{7.13}$$

where $C(A, c) = \frac{A^c}{c!(1-\rho)}/(\sum_{i=0}^{c-1} \frac{A^i}{i!} + \frac{A^c}{c!(1-\rho)})$ is the probability an arrival finds all c servers busy in an M/M/c queue with $A = \lambda/\mu$.

Proof. As a result of (7.2), the Laplace-Stieltjes transform of the class-2 waiting time distribution can be decomposed into the initial workload $\mathcal{W}$ followed by a delay represented by the compound Poisson process Y, where the number of terms N in the compound Poisson process represents the accrediting customers that arrive during $\mathcal{W}$. Now, the distribution of the initial workload $\mathcal{W}$ found by an arrival from a Poisson process is identical to that of the stationary waiting time in an M/M/c queue, whose LST is given by (7.13).

During periods when all c servers are busy, the duration of times between customers entering service are exponentially distributed at rate $c\mu$. As such, periods during which all servers are busy in an M/M/c queue with service at rate μ are indistinguishable, probabilistically, from busy periods in an M/M/1 queue with service at rate $c\mu$ (see, for instance, Sharif *et al.* (2014)). Thus, according to the revised service discipline introduced in the proof of Lemma 1, each of the accrediting customers to arrive during $\mathcal{W}$ adds an amount to the waiting time equal to a busy period of accrediting customers, and is thus equivalent to a busy period in an M/M/1 queue with service at rate $c\mu$ and arrivals at rate $\lambda_1(1-b)$. The functional equation for the LST $\widetilde{\eta}_1(s)$ of such a busy period is seen to be given by (7.12). Remembering that N is Poisson distributed during $\mathcal{W}$ at rate $\lambda_1(1-b)$, one readily obtains (7.11). □

The class-2 waiting time distribution will play a role in the determination of the class-1 waiting time distribution, so we choose to combine (7.11) and (7.13) to obtain the following expression for $\widetilde{w}_2(s)$:

$$\widetilde{w}_2(s) = [1 - C(A,c)] + C(A,c)\widetilde{w}_2(s|\text{busy}) \tag{7.14}$$

where $\widetilde{w}_2(s|\text{busy})$ denotes the LST of the class-2 waiting time distribution, conditional upon an arrival finding all servers busy, which is given by

$$\widetilde{w}_2(s|\text{busy}) = \left[\frac{c\mu - \lambda}{c\mu - \lambda + [s + \lambda_1(1-b)(1-\widetilde{\eta}_1(s))]}\right]. \tag{7.15}$$

The functional equation (7.12) gives rise to a quadratic equation in $\widetilde{\eta}_1(s)$ which can be factored to yield

$$\widetilde{\eta}_1(s) = \frac{(1 + \frac{s}{c\mu} + \rho_H) - [(1 + \frac{s}{c\mu} + \rho_H)^2 - 4\rho_H]^{1/2}}{2\rho_H} \tag{7.16}$$

where $\rho_H \equiv \lambda_1(1-b)/(c\mu)$. The equations (7.14) through (7.16) are the functional forms needed for the implementation of Gaver-Stehfest numerical evaluation for class-2 customers.

We conclude with the specification of $\widetilde{w}_1(s)$.

Theorem 4. The Laplace-Stieltjes transform $\widetilde{w}_1(s)$ of the class-1 waiting time distribution can be written as

$$\widetilde{w}_1(s) = [1 - C(A,c)] + C(A,c)\widetilde{w}_1(s|\text{busy}) \tag{7.17}$$

where $\widetilde{w}_1(s|\text{busy})$ denotes the LST of the class-1 waiting time distribution, conditional upon an arrival finding all servers busy. In turn, $\widetilde{w}_1(s|\text{busy})$ is given by

$$\widetilde{w}_1(s|\text{busy}) = \left[(1-b)\widetilde{w}_1^{(acc)}(s) + b\widetilde{w}_2(bs|\text{busy})\right] \tag{7.18}$$

where $\widetilde{w}_1^{(acc)}(s)$ denotes the LST of the waiting time of those class-1 customers who gain accreditation, and is given by

$$\widetilde{w}_1^{(acc)}(s) = \widetilde{w}_1^{(0)}(s)\left[\left(\frac{1-\rho}{1-\rho_H}\right) + \left(\frac{\rho-\rho_H}{1-\rho_H}\right)\widetilde{w}_2(bs|\text{busy})\right] \tag{7.19}$$

where

$$\widetilde{w}_1^{(0)}(s) = \frac{\mu(1-\rho_H)[\widetilde{\eta}_1(bs) - B(s)]}{(1-b)\times(s - \lambda_1 + \lambda_1 B(s))} \tag{7.20}$$

for $B(s) = c\mu/(c\mu + s)$.

Proof. Since service times for all customer classes follow the same exponential distribution, the re-ordering of customers does not change the long-run fraction of time that all c servers are busy, given by $C(A,c)$. Since class-1 arrivals constitute a Poisson process, $C(A,c)$ is also the chance such an arrival will find all servers busy (and hence be delayed). This justifies (7.17).

Turning to $\widetilde{w}_1(s|\text{busy})$, we employ the equivalent formulation during waiting times used in Theorem 3 above: namely, that of a single server with exponentially distributed service times at rate $c\mu$. We are then free to invoke Corollary 8.4 in Stanford *et al.* (2014), which leads directly to (7.18).

Equation (7.19) similarly comes from the last equation in Stanford *et al.* (2014) Corollary 8.4, and $\widetilde{w}_1^{(0)}(s)$ from earlier definitions therein. □

7.6 Numerical Experiments

A series of experiments were run for the APQ discipline, to determine the numerical behavior of the expected excess per customer functions for each class, as well as various weighted averages of the individual excess functions.

In the first set of experiments, the results of which are presented in Figures 2 and 3, we have considered a 2-class, 2-server APQ in which the customer arrival rates for both classes are kept the same, while the server occupancy changes from 80% in Figure 2 to 95% in Figure 3.

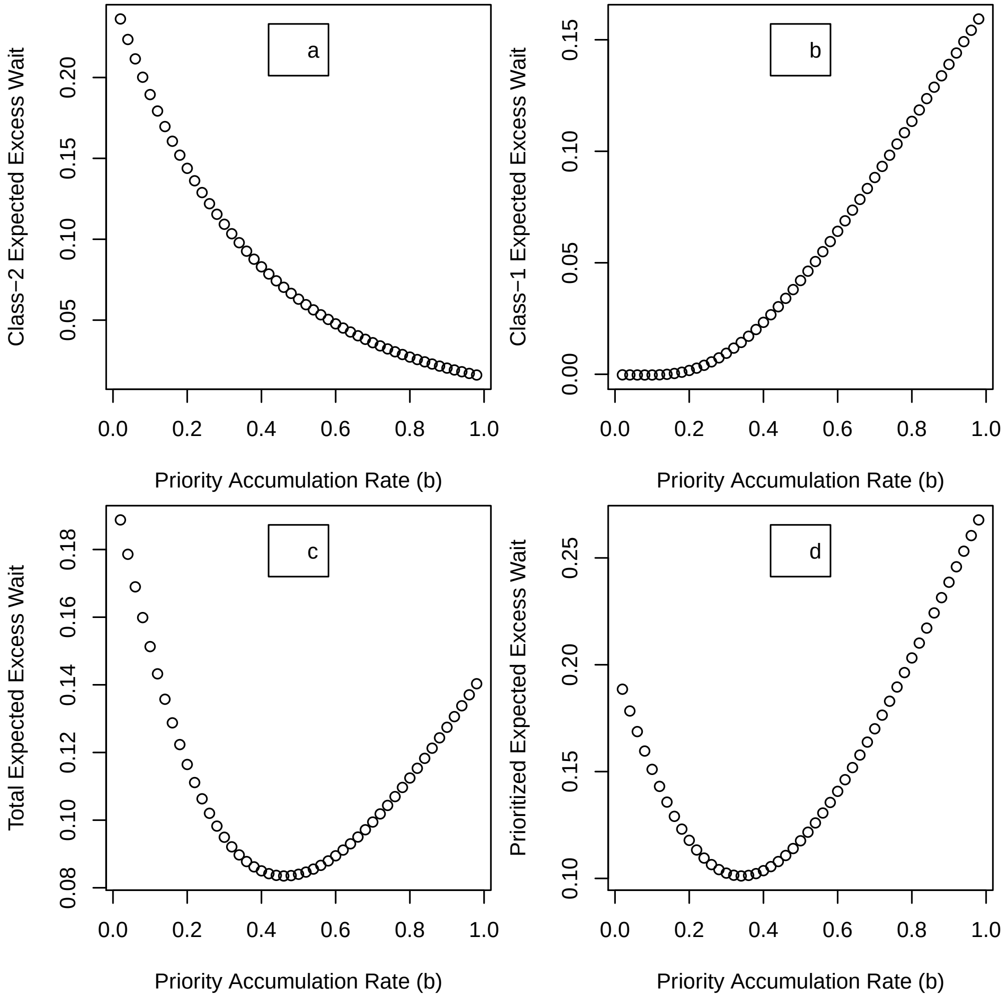


Figure 7.2: Two-class two-server APQ with equal arrival rates (80% occupancy)

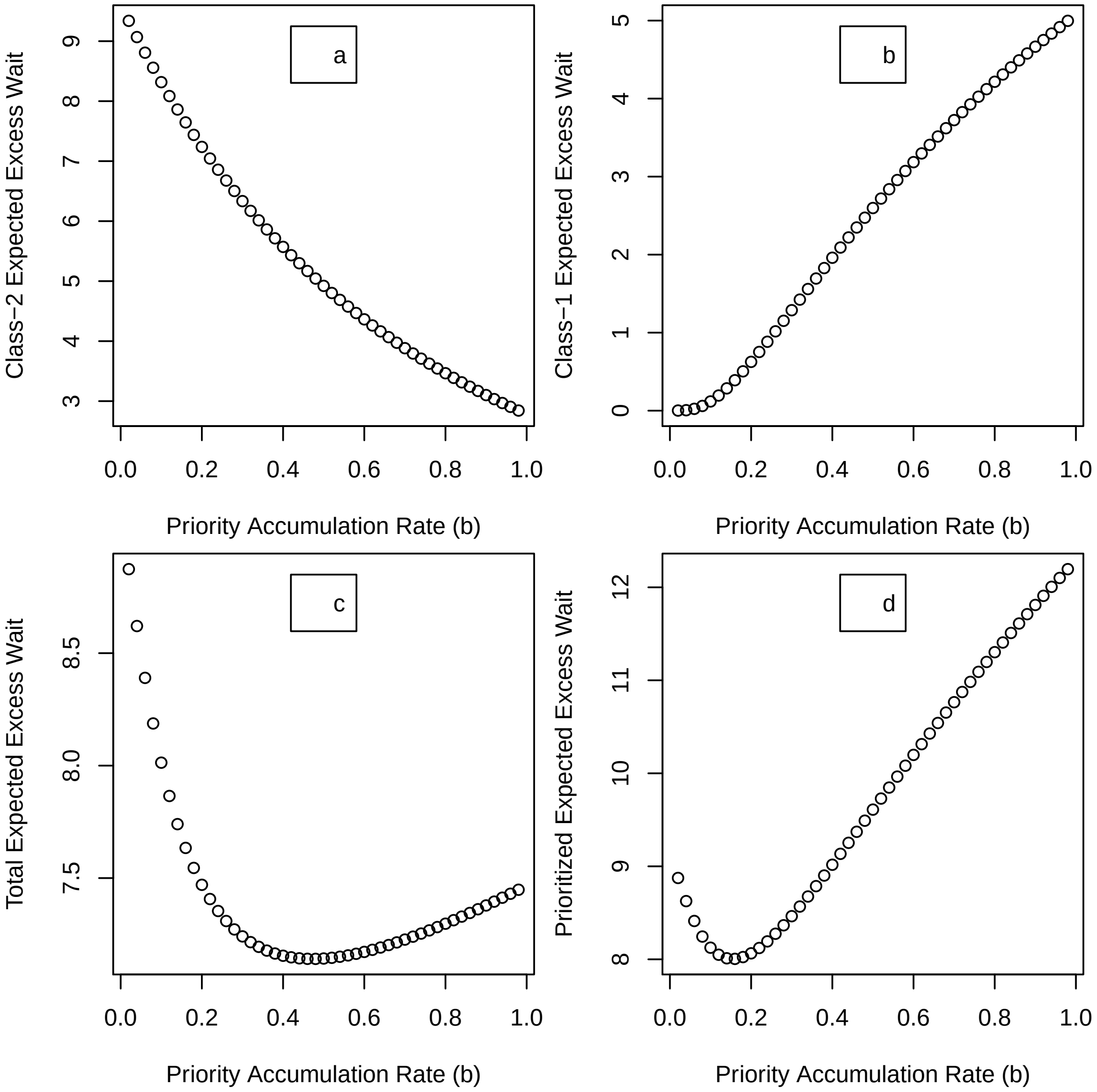


Figure 7.3: Two-class two-server APQ with equal arrival rates (95% occupancy)

Figures 2(a) and 3(a) present the expected excess waiting time per customer for a class-2 customer as a function of the priority accumulation rate for class-2 customers, *b*. The corresponding results for class-1 customers are presented in Figures 2(b) and 3(b). The expected excess increases with *b* for class-1 customers, as anticipated. The expected excess decreases for class-2 customers, and the class-2 figures are seen to be convex, both as established in Corollary 1.

Figures 2(c) and 3(c) display the total expected amount of excess waiting time (i.e. when the two classes are assigned equal weights $\alpha_1 = \alpha_2 = 1$). A comparison of the figures reveals that the optimal accumulation rate b is quite insensitive to the occupancy level, with the optimal value of b being very close to one half in all cases. Furthermore, generally speaking, the total expected excess function is reasonably flat in the vicinity of the optimal b value. The actual total expected excess wait itself varies considerably with a change in occupancy, from a minimum of the order of 0.1 at 80% occupancy, to a minimum on the order of 7 at 95% occupancy. The fact that the optimal b is slightly less than $1/2$ in all cases is not surprising, when one considers the fact that the class-2 target delay is twice as long for class-2 customers as it is for class-1 customers. What is perhaps not anticipated is just how insensitive the optimal b value is to a change in occupancy, and the flatness of the objective function (total expected excess) in the vicinity of the optimal value. Thus, when an equal weight is placed on excess waiting for both classes and the arrival rates are the same, a rule of thumb that works well is to set the priority accumulation rates in inverse proportion to the target delays. In this way, customers from the different classes approaching their respective delay targets will have accumulated comparable amounts of credit, as one would expect.

Figures 2(d)and 3(d) display the weighted total expected waiting time excess with $\alpha_1 = 2\alpha_2 = 2$. One observes that the occupancy definitely influences the optimal value of b. The optimal value of b is about $b = 0.34$ at 80% occupancy, and about $b = 0.16$ at 95% occupancy. Thus, no simple rule of thumb can be employed when differing weights are applied; a more detailed analysis using the methodology presented herein is needed to determine the optimal accumulation rates for a given configuration.

Figure 4 presents the comparable results when 80% of the arrivals belong to class 1, at an occupancy level of 95%. One notable change is that the average excess per customer for class-2 customers is typically an order of magnitude larger than that for class-1 customers. When the occupancy is large enough, as in Figure 4(d), then the weighted total expected amount of excess is minimized when b_2 is practically zero. In other words, in this case it is nearly optimal to resort to a standard non-preemptive priority queue.

Figures 5 presents the comparable results when 80% of the arrivals belong to class 2, at an occupancy level of 95%. Qualitatively, the results are consistent with those of Figures 2 through 4. The biggest difference seems to be the individual expected loss per customer curves for class-1 and class-2 customers. As the occupancy increases, these curves seem to be approaching a linear form in b.

At this point, we sought to verify that the rule of thumb did indeed manifest itself in other settings when the ratio of the delay targets was not a factor of 2. In Figure 6, we present the results for an equal mixture of class-1 and class-2 customers, at an occupancy level of 95%, when the class-2 delay target was three times as long as the class-1 delay target. As anticipated, we see that the minimum of the total expected excess waiting time occurred when b_2 is close to $1/3$, which is consistent with the rule of thumb. Furthermore, the optimal value of b_2 in the weighted total expected excess is smaller than in the pure expected excess scenario, just as had been the case with all the foregoing examples.

For a final set of examples, we present in Figures 7 and 8 the results we obtained for a three-class, two-server APQ with equal arrival rates at a 90% occupancy. The three delay targets satisfy $d_3 = 2d_2 = 4d_1$. We again have arbitrarily set $b_1 = 1$. We present the expected excess per customer for each of the three classes, as well as the total expected excess waiting time. In Figure 11, we have fixed $b_2 = 1/2$ to be consistent with the rule of thumb from the two-class case, and allowed b_3 to range over the interval $0 \leq b_3 \leq b_2 = 1/2$. We observed that the minimum total expected excess occurs when b_3 is close to $1/4$, as one would expect from the rule of thumb. In Figure 12, we reversed the situation, fixing $b_3 = 1/4$ and we allowed b_2 to vary over the interval $1/4 = b_3 \leq b_2 \leq 1$. Once again, as anticipated, the optimal accumulation rate for b_2 occurs close to $b_2 = 1/2$. We note in particular that the expected excess wait time per class-3 customer is not a convex function of b_2, which makes sense: as the priority accumulation rate for class-2 customers increases, it does so at the relative disadvantage to class-1 and class-3 customers, who in the former case are losing priority over

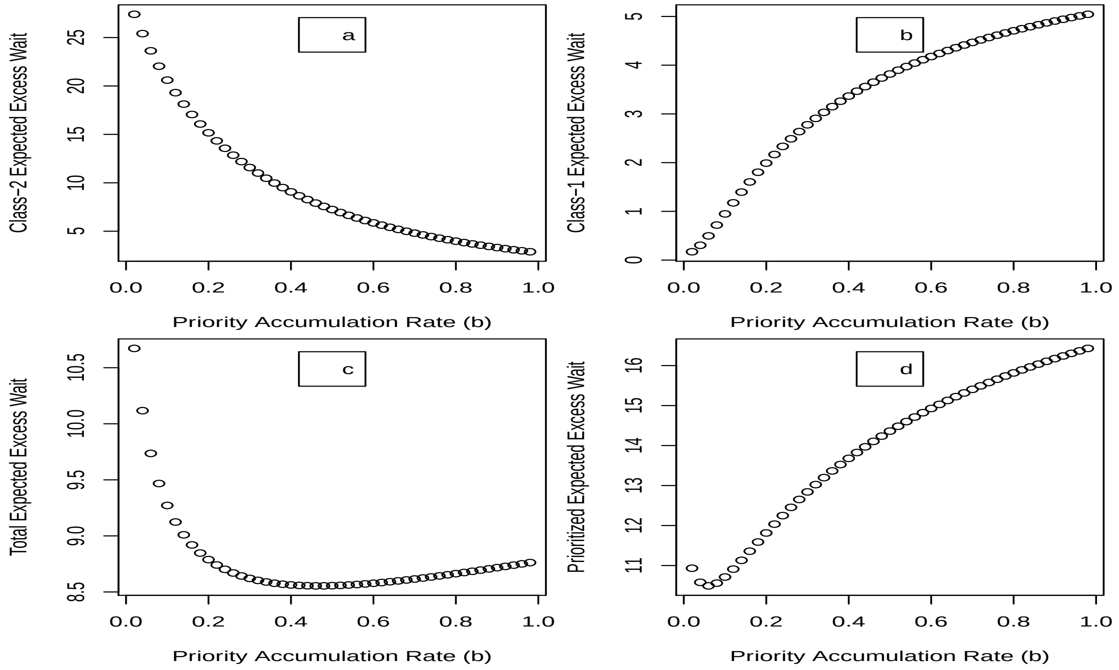


Figure 7.4: Two-class two-server APQ with 80% class-1 arrivals (95% occupancy)

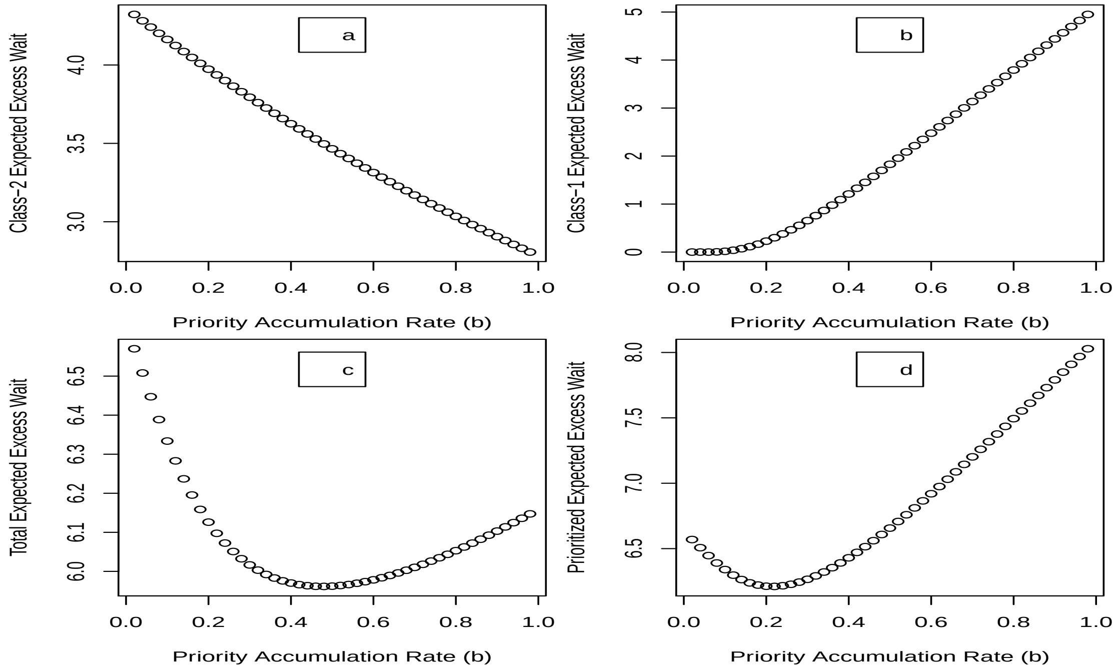


Figure 7.5: Two-class two-server APQ with 80% class-2 arrivals (95% occupancy)

class-2 customers, and in the latter case, are losing priority to class-2 customers.

We conclude this section by commenting on the utility of different KPI compliance stan-

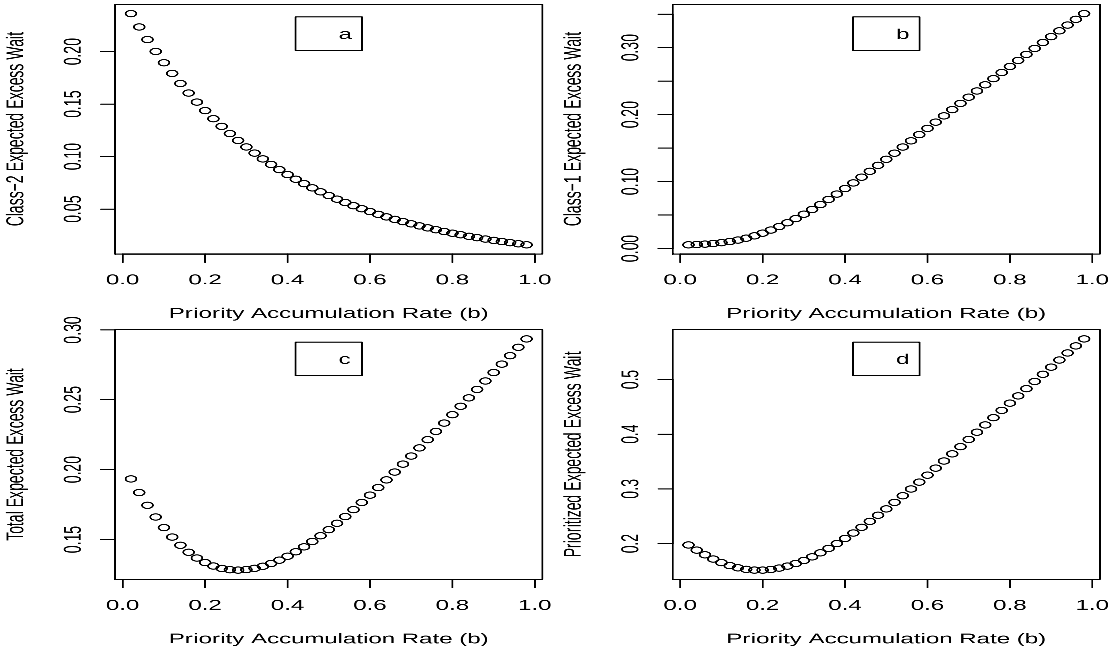


Figure 7.6: Two-class two-server APQ with equal arrivals (80% occupancy) where delay target for class-2 is 3 times that of class-1.

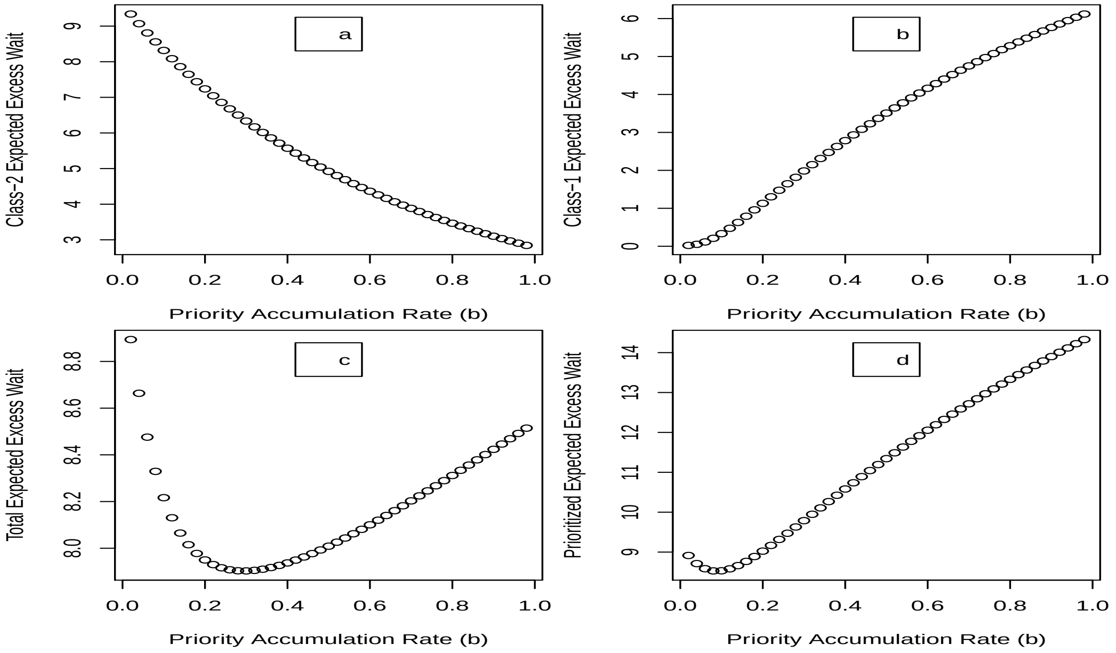


Figure 7.7: Two-class two-server APQ with equal arrivals (95% occupancy) where delay target for class-2 is 3 times that of class-1.

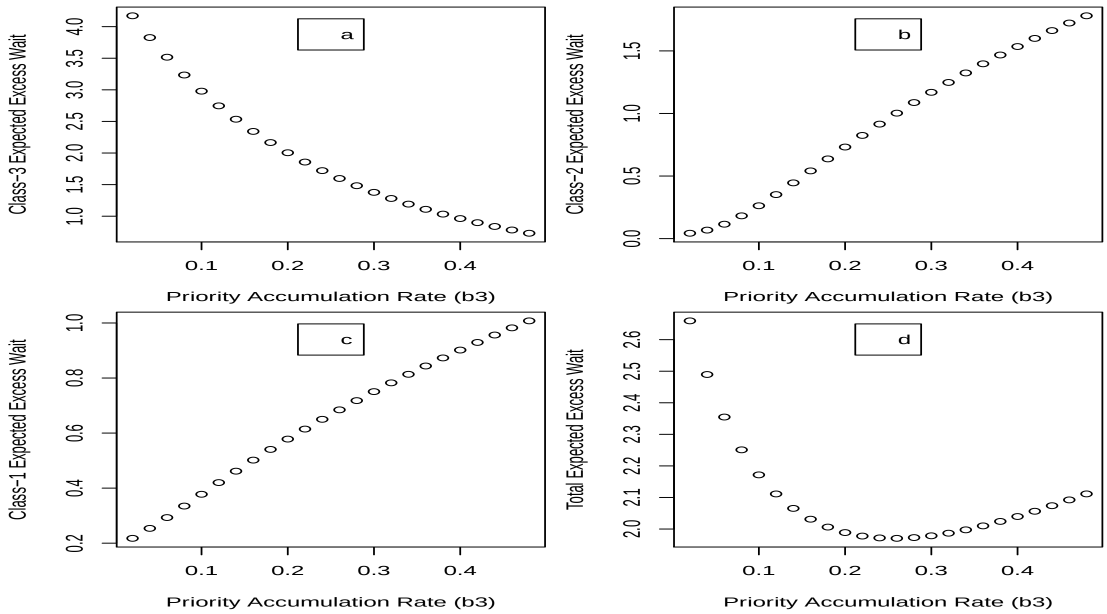


Figure 7.8: Three-class two-server APQ with equal arrivals (90% occupancy) where b_1 and b_2 are fixed at 1 and 1/2, respectively, while b_3 varies between 0 and 1/2, and the delay target for class-3 is 2 times that of class-2 and 4 times that of class-1.

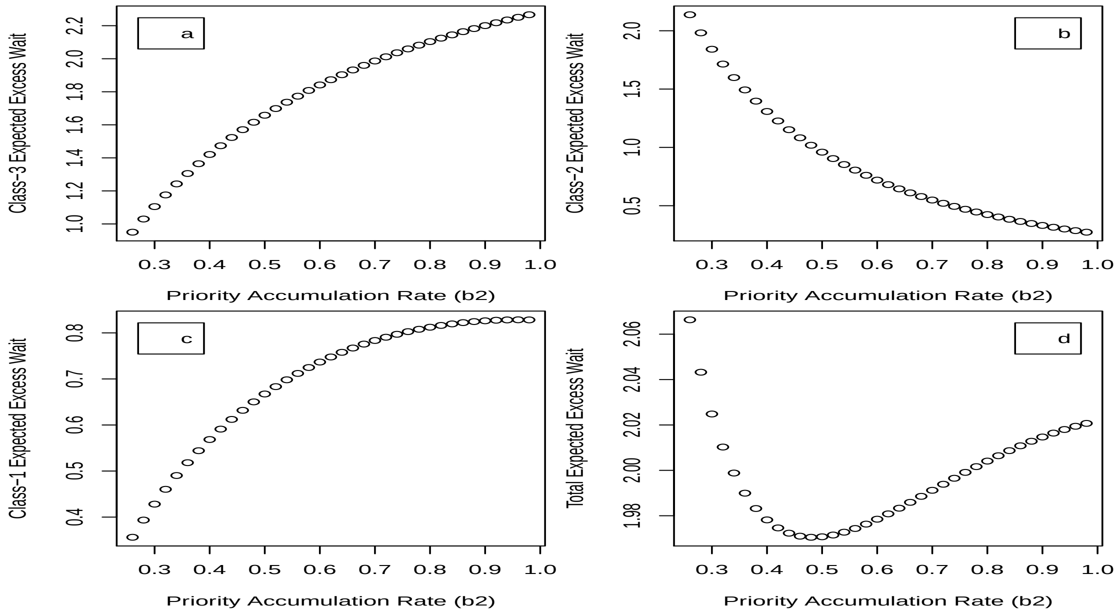


Figure 7.9: Three-class two-server APQ with equal arrivals (90% occupancy) where b_1 and b_3 are fixed at 1 and 1/4, respectively, while b_2 varies between 1/4 and 1, and the delay target for class-3 is 2 times that of class-2 and 4 times that of class-1.

dards in the total expected excess and the weighted total expected excess cases. As for the former, we have observed throughout our examples that the rule of thumb works to ensure that each class of customers has virtually the same chance of exceeding its delay target. In order for all of the KPIs to be satisfied, this means that the APQ must adhere to the most stringent compliance standard. In such a context, the other KPI compliance standards will be exceeded, and are in effect irrelevant, so that there is no need for lesser compliance standards than the most stringent one.

In the weighted total excess case, however, the optimal accumulation rates tend to be smaller than their counterparts in the total expected excess case. This means that the chance of non-compliance increases as the acuity decreases (i.e. the priority class index increases). In such a setting, more generous compliance standards for customers of lower acuity makes sense.

7.7 Conclusions and Future Research

In this paper, we have presented a general optimization problem for queueing systems operating under waiting time targets. Our optimization problem reflects the fact that customers who miss their waiting time targets are in fact of greater concern than those who meet them, and Key Performance Indicators (KPIs) on their own do not address this fact. Formally, our goal was to minimize a weighted average of the total expected excess waiting per unit time over all permissible customer selection strategies.

We have also established that the individual terms in our objective function, which represent the expected excess waiting time per customer for each class, are readily obtained as a by-product of the computation of the probabilities of KPI compliance whenever the latter are being evaluated via numerical inversion of their Laplace-Stieltjes transform.

While the aforementioned optimization problem can be applied to any work conserving queueing discipline, we have established that the Accumulating Priority Queue, where customers are selected for service based both on their priority class and the amount of time they have spent waiting for service, is well-suited to the goal of minimizing the excess waiting beyond the delay targets. In order to provide an intuition as to how the priority accumulation rates (b_n's) works under APQ, let us consider a two class case. In particular one should first establish that one can fix one rate and vary the other to explore all cases. Having fixed $b_1 = 1$, then $0 < b_2 < 1$ tells us where we are between the classical priority situation where waiting time has no impact ($b_2 = 0$) and the FCFS case where there are no distinction in the classes ($b_2 = 1$). In other words, it tells us where we are in the mix between classical priority and the FCFS, in the intermediate zone where class matters but waiting time, too.

By providing the decision maker with the added flexibility of determining the best priority accumulation rates for each class, one can achieve the desired balance between the amount of excess waiting that occurs in each class. When equal weight are used, the system seeks to minimize the total expected waiting time. Our numerical examples have established that the optimal APQ strategy in this case is well-approximated by a "rule of thumb" which accumulates priority for customers of a given class in inverse proportion to their delay target. This ensures that all customers approaching their waiting time targets will have accumulated comparable amounts of priority, so that each customer sees nearly the same probability of exceeding their respective targets. Consequently, all acuity classes will observe comparable levels of compliance. In the more general case where the weights placed on the excess waiting times are different, no rule of thumb exists; however, the optimal arrangement can nonetheless be determined according to the procedures herein.

One underlying constant in our studies has been the assumption of a common service time distribution for all customer classes. An obvious extension of our model would be to the situation where different classes have differing service time requirements. Until now, no suitable APQ model for the multi-class, multi-server APQ model with differing service time distributions has been been developed, but such work is under way at present.

There may well be appropriate settings for this model involving certain non-work-conserving queues, such as call centres with fully and partially trained agents. Customers with standard requirements could be seen by any agent, while customers with specialized needs (typically, these customers would be of higher priority) could only be served by a subset of the agents. The optimization problem in such a setting would rely upon the availability of the waiting time LSTs for the different classes of customer.

7.8 Acknowledgements

This research has been funded by the Discovery Grants program of Canada's Natural Sciences and Engineering Research Council (NSERC). We wish to thank Prof. Martin Utley of University College London for useful discussions on the secondary nature of the compliance probability in the KPI standard, and the use of the terminology to distinguish "visible" from "invisible" queues.

7.9 Appendices

Appendix A: Gaver-Stehfest Coefficients

Table 7.2: Coefficients for the Gaver-Stehfest Algorithm

V_2	V_4	V_6	V_8
2	-2	1	-1/3
-2	26	-49	145/3
	-48	366	-906
	24	-858	16394/3
		810	-43130/3
		-270	18730
			-35840/3
			8960/3

Appendix B: Proof of Lemma 1

We employ the same viewpoint to express the class-2 stationary waiting time $\mathcal{W}_2$ as was employed in Theorem 9.1 of Stanford *et al.* (2014). This is done by rearranging the service discipline in such a way that preserves the total waiting time experienced by a tagged arrival from that class.

Consider such a tagged class-2 customer. It will wait for all of the work present in the system upon its arrival, plus that which arrives later but will be served ahead of it. Since the tagged customer is from the lowest priority class, and is an arrival from a Poisson process, the distribution of the work it finds in system at such instants equals the distribution of the stationary unfinished workload, by the "Poisson-arrivals-see-time averages" (PASTA) property of the Poisson process (Wolff (1982)). However, the stationary unfinished workload is invariant for all work-conserving service disciplines. As shown in Stanford *et al.* (2014) Theorem 9.1, its distribution is the same as the stationary waiting time distribution for $\mathcal{W}$ in the M/G/1 FCFS comparator queue.

During this initial period $\mathcal{W}$, the process of later arrivals that will gain accreditation (and therefore be served ahead of the tagged class-2 customer) constitutes a Poisson process at rate $\lambda_1(1-b)$, as established in Stanford *et al.* (2014) Lemma 4.2. While some of these later class-1 arrivals may enter service according to the APQ service discipline ahead of some of the class-2 customers already present at the arrival instant of the tagged customer, the actual order of service does not matter, so long as all such work is completed prior to the tagged customer's entry into service.

Thus we can write $\mathcal{W}_2 = \mathcal{W} + Y$, where Y represents the total of all of the service times for class-1 customers that gain accreditation relative to the tagged customer prior to its entry into service. We are able to characterize Y in terms of $\mathcal{W}$ as follows. We do so by resorting to the following discipline, which parallels the rearrangement of service times in the derivation by Conway *et al.* (1967) of the implicit transform equation for the duration of a busy period in an M/G/1 queue.

Our rearrangement places all of the N class-1 customers who gain accreditation during $\mathcal{W}$ in a special queue. Upon completion of $\mathcal{W}$, the first (if any) of these N customers is selected for service. The server then selects newly-accredited class-1 customers in the main queue until none remain, at which point the next (if any) of the class-1 customers in the special queue is selected, and so on. We observe that, by construction, each such "sub-busy period" comprising the service of one customer from the special queue and the subsequent accredited arrivals to the main queue is identically distributed to the busy period η in an M/G/1 queue with arrivals at rate $\lambda_1(1-b)$ and whose service times are drawn from the class-1 service time distribution.

In this way, Y can be expressed according to (7.3), with N representing the number of Poisson events at rate $\lambda_1(1-b)$ to occur during $\mathcal{W}$, and the ith such accrediting customer, $i = 1, 2, \ldots, N$, adding an i.i.d. busy period duration η_i from such an M/G/1 queue. □

Chapter 8

Conclusions

In this thesis, I worked with both theoretical and practical problems related to a heath care service system (emergency department) where patients are served based on their acuity (urgency of need). Among four chapters of novel material, two chapters (4 and 5) present empirical work involving ED arrival and service processes based upon two years of data from a particular hospital in Ontario. The remaining two chapters (6 and 7) dealt with theoretical work related to a priority queueing system, where the former determined the waiting time distribution for high and low acuity patients, while the latter pertained to minimizing the expected excess waiting time for high and low acuity patients beyond their stipulated waiting time limits.

In chapter 4, we investigated both regression and time-series based forecasting methods to identify an appropriate forecasting model to accurately predict ED arrivals in short term. Based on the type of data, we fitted and compared seasonal ARIMA, GLM, harmonic regression, and GLARMA models. Comparing performance based on different accuracy measures, we observed that the GLARMA model produced better forecast accuracy than others. Additionally, we employed the rolling horizon approach, frequently used in operations management literature, to validate our proposed (GLARMA) forecasting model.

In chapter 5, we used statistical models to investigate the effect of workload on ED productivity. GLM and Cox-PH models were used to model the effect of covariates on the count and time-to-event type responses, respectively. An increase in workload is found to be associated with an increase in the number of patients discharged and the waiting time, and a decrease in the patient LOS and service times. However, there is a fatigue effect associated with how long a server is continuously working at a faster rate. A continuous increase in workload for several hours is associated with a decrease in productivity.

In chapter 6, we presented a multi-server APQ model for an arbitrary number of customer classes, where the service time distributions for each classes of customer follows an exponential distribution with common mean. Numerical investigations showed how to ascertain whether a given set of priority accumulation rates will enable compliance with a given set of KPIs for a given traffic pattern of patient arrival and service rates. An optimization problem related to the model under study was also developed to select the best priority accumulation rate for which we can get the maximum possible utilization that complies with the KPIs.

In chapter 7, we developed an optimization model to minimize the expected excess waiting time for the high and low acuity patients. Numerous numerical investigations lead us to discover a “Rule of Thumb” which can be easily implemented by practitioners without knowing the details about the mathematical model behind the mechanism. In order to minimize expected excess waiting time, patients should accumulate credit in inverse proportion to the ratio of the time limits for their respective classes. Based upon the Rule of Thumb, the patient who has already waited the largest proportion of their permissible waiting time limit should be selected for service. The Rule of Thumb will approximately minimize the expected excess waiting time for high and low acuity patients, respectively.

8.1 Future Research Directions

Several extensions to our work are possible. We highlighted below some of the obvious extensions that can be considered as future research.

In chapter 4, we proposed GLARMA model to accurately forecast future ED arrivals based on data from an Ontario hospital. Since ED varies based on the patient population, severity of illness, and so on, an obvious extension would be to study multiple EDs to obtain more generalizable result. From a modeling perspective, one may use more complicated models such as fractionally integrated auto-regressive moving average (ARFIMA) model or generalized estimating equations (GEE) to see whether such models produce better forecast.

In chapter 5, we studied the effect of workload on the ED service process. For count responses, one may use distributions other than Poisson or Negative Binomial model if the number discharged follows a different distribution. For time-to-event responses, one may use a parametric survival models (AFT) as an alternative to Cox-proportional hazard model. Both the AFT and Cox-PH model produce similar results for few parametric survival distributions (e.g., exponential, Weibull). However, they may produce completely different results for other survival distributions. Finally, one could contemplate a more extensive statistical model which includes the impact of interaction between covariates on the service quality measures.

In chapter 6, we derived the waiting time distributions for a multi-class multi-server APQ model where service times are selected from a common exponential distribution. An obvious extension of our current model is to the situation where the rate of service differs for different classes of patients. Another possible extension would be to consider non-exponential service time distributions. Although some results might be possible in the case of a small number of servers for small-order Erlang distributions, it effectively leads to intractable results in the general multi-server context.

In chapter 7, we studied an optimization model that minimized total expected excess waiting time for a queueing system that operates under APQ. There may well be appropriate settings for this model involving certain non-work-conserving queues, such as call centres with fully and partially trained agents. Customers with standard requirements could be seen by any agent, while customers with specialized needs (typically, these customers would be of higher priority) could only be served by a subset of the agents. The optimization problem in such a setting would rely upon the availability of the waiting time LSTs for the different classes of customer.

Bibliography

Addis, B., Carello, G., Grosso, A. & Tanfani, E. (2014). A rolling horizon framework for the operating rooms planning under uncertain surgery duration. *HAL Archives (ID 00936085)* .

Akan, M., Ata, B. s. & Olsen, T. (2012). Congestion-based lead-time quotation for heterogenous customers with convex-concave delay costs: Optimality of a cost-balancing policy based on convex hull functions. *Operations research* **60**(6), 1505–1519.

Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y. N., Tseytlin, Y. & Yom-Tov, G. B. (2011). Patient flow in hospitals: A data-based queueing-science perspective. *Submitted to Stochastic Systems* **20**.

Arnett, G., Hadorn, D. & the Steering Committee of the Western Canada Waiting List Project. (2003). Developing priority criteria for hip and knee replacement surgery: Results from the Western Canada Waiting List Project. *Canadian Journal of Surgery* **46**(4), 290–296.

ATS (2000). The Australasian Triage Scale. From the website. `http://www.acem.org.au/media/policies_and_guidelines/P06_Aust_Triage_Scale_Nov_2000.eps`.

Au-Yeung, S., Harder, U., McCoy, E. & Knottenbelt, W. (2009). Predicting patient arrivals to an accident and emergency department. *Emergency medicine journal* **26**(4), 241–244.

Babu, G. J. (2012). *Modern Statistical Methods for Astronomy*. Cambridge University Press Textbooks.

BARTEL, A. P., CHAN, C. W. & KIM, S.-H. H. (2014). Should hospitals keep their patients longer? the role of inpatient and outpatient care in reducing readmissions. Tech. rep., National Bureau of Economic Research.

BATT, R. J. & TERWIESCH, C. (2012). Doctors under load: An empirical study of state-dependent service times in emergency care. Tech. rep., Working Paper, The Wharton School. 1.

BERGS, J., HEERINCKX, P. & VERELST, S. (2014). Knowing what to expect, forecasting monthly emergency department visits: A time-series analysis. *International emergency nursing* **22**(2), 112–115.

BERRY JAEKER, J. & TUCKER, A. L. (2013). An empirical study of the spillover effects of workload on patient length of stay. *Harvard Business School Working Paper* .

BERRY JAEKER, J., TUCKER, A. L. & FIELD, S. (2012). Hurry up and wait: Differential impacts of congestion, bottleneck pressure, and predictability on patient length of stay. *Harvard Business School Working Paper* .

BERRY JAEKER, J., TUCKER, A. L. & LEE, M. H. (2013). Increased speed equals increased wait: The impact of a reduction in emergency department ultrasound order processing time. *Harvard Business School Working Paper* .

BOND, K., OSPINA, M., BLITZ, S., AFILALO, M., CAMPBELL, S., BULLARD, M., INNES, G., HOLROYD, B., CURRY, G., SCHULL, M. *et al.* (2006). Frequency, determinants and impact of overcrowding in emergency departments in Canada: a national survey. *Healthcare quarterly (Toronto, Ont.)* **10**(4), 32–40.

BOUDREAU, J., HOPP, W., MCCLAIN, J. O. & THOMAS, L. J. (2003). On the interface between operations and human resources management. *Manufacturing & Service Operations Management* **5**(3), 179–202.

Box, G. E. P. & Jenkins, G. M. (1970). *Time Series Analysis, Forecasting and Control*. Holden-Day, San Francisco.

Boyle, J., Jessup, M., Crilly, J., Green, D., Lind, J., Wallis, M., Miller, P. & Fitzgerald, G. (2012). Predicting emergency department admissions. *Emergency Medicine Journal* **29**(5), 358–365.

Boyle, J., Le Padellec, R. & Ireland, D. (2010). Statewide validation of a patient admissions prediction tool. In: *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*. IEEE.

Brigham, E. & Morrow, R. (1967). The fast Fourier transform. *Spectrum, IEEE* **4**(12), 63–70.

Brockwell, P. J. & Davis, R. A. (2002). *Introduction to time series and forecasting*, vol. 1. Taylor & Francis.

Brockwell, P. J. & Davis, R. A. (2009). *Time series: theory and methods*. Springer Science & Business Media.

Canadian Institute for Health Information (2014). Patients admitted to hospital spend almost 5 times longer in ED than non-admitted. `http://www.cihi.ca/CIHI-ext-portal/internet/en/Document/types+of+care/hospital+care/emergency+care/RELEASE_07102014`. Accessed: 2015-01-06.

Çelik, S. & Maglaras, C. (2008). Dynamic pricing and lead-time quotation for a multiclass make-to-order queue. *Management Science* **54**(6), 1132–1146.

Champion, R., Kinsman, L. D., Lee, G. A., Masman, K. A., May, E. A., Mills, T. M., Taylor, M. D., Thomas, P. R. & Williams, R. J. (2007). Forecasting emergency department presentations. *Australian Health Review* **31**(1), 83–90.

Chan, C. W., Farias, V. F., Bambos, N. & Escobar, G. J. (2011). Maximizing throughput of hospital intensive care units with patient readmissions. Tech. rep., Working Paper, Columbia Business School.

Chan, C. W., Farias, V. F. & Escobar, G. (2013). The impact of delays on service times in the intensive care unit. *Working Paper. 2, 2014* .

Chand, S., Hsu, V. N. & Sethi, S. (2002). Forecast, solution, and rolling horizons in operations management problems: a classified bibliography. *Manufacturing & Service Operations Management* **4**(1), 25–43.

Chatfield, C. (2013). *The analysis of time series: an introduction*. CRC press.

Chen, C. F., Ho, W. H., Chou, H. Y., Yang, S. M., Chen, I. T. & Shi, H. Y. (2011). Long-term prediction of emergency department revenue and visitor volume using autoregressive integrated moving average model. *Computational and mathematical methods in medicine* **2011**.

Conway, R., Maxwell, W. & Miller, L. (1967). *Theory of Scheduling*. Addison-Wesley.

Côté, M. J., Smith, M. A., Eitel, D. R. & Akçali, E. (2013). Forecasting emergency department arrivals: A tutorial for emergency department directors. *Hospital topics* **91**(1), 9–19.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* , 187–220.

CTAS (2005). The Canadian Triage and Acuity Scale. From the website. `http://www.calgaryhealthregion.ca/policy/docs/1451/Admission_over_capacity_AppendixA.eps`.

Czado, C., Gneiting, T. & Held, L. (2009). Predictive model assessment for count data. *Biometrics* **65**(4), 1254–1261.

Davies, J. & Little, L. (2012). The Taming of the Queue: Looking Back, Going Forward: Reframing Timely Access as Part of Health System Transformation.

From the website. `http://www.cfhi-fcass.ca/sf-docs/default-source/taming-of-the-queue-english/TQ2012-FinalReport-EN.pdf?sfvrsn=0` [Accessed: July 30, 2014].

Davis, R. A., Dunsmuir, W. & Wang, Y. (1999). Modeling time series of count data. *STATISTICS TEXTBOOKS AND MONOGRAPHS* **158**, 63–114.

Davis, R. A., Dunsmuir, W. T. & Streett, S. B. (2003). Observation-driven models for Poisson counts. *Biometrika* **90**(4), 777–790.

Derlet, R. W. & Richards, J. R. (2000). Overcrowding in the nation's emergency departments: complex causes and disturbing effects. *Annals of Emergency Medicine* **35**(1), 63–68.

Dobson, A. J. & Barnett, A. (2011). *An introduction to generalized linear models*. CRC press.

Dodd, G. (2011). Key performance indicator (kpi) report. *South West London and St George's, Mental Health NHS Trust* **Quarter 1**, 1–19.

Dodge, Y. (2006). *The Oxford dictionary of statistical terms*. Oxford University Press.

Dreyer, J. F., McLeod, S. L., Anderson, C. K., Carter, M. W. & Zaric, G. S. (2009). Physician workload and the canadian emergency department triage and acuity scale: the predictors of workload in the emergency room (power) study. *CJEM* **11**(04), 321–329.

Dunsmuir, W. T. & Scott, D. J. (2015). The glarma Package for Observation Driven Time Series Regression of Counts. *Journal of Statistical Software* **67**(1), 1–36.

Ekström, A., Kurland, L., Farrokhnia, N., Castrén, M. & Nordberg, M. (2015). Forecasting emergency department visits using internet data. *Annals of Emergency Medicine* **65**(4), 436–442.

Faraway, J. J. (2005). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press.

GAVER, D. (1966). Observing stochastic processes and approximate transform inversion. *Operations Research* **14**.

GRANGER, C. W. & JOYEUX, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of time series analysis* **1**(1), 15–29.

GREEN, L. V. & NGUYEN, V. (2001). Strategies for cutting hospital beds: the impact on patient service. *Health Services Research* **36**(2), 421.

GREEN, L. V., SAVIN, S. & SAVVA, N. (2013). Nursevendor problem: Personnel staffing in the presence of endogenous absenteeism. *Management Science* **59**(10).

GREENE, W. H. (2011). *Econometric analysis*. Prentice Hall.

HAY, A., VALENTIN, E. & BIJLSMA, R. (2006). Modeling emergency care in hospitals: A paradox - the patient should not drive the process. *Proc. 2006 Winter Simulation Conference* .

HOLLEMAN, D. R., BOWLING, R. L. & GATHY, C. (1996). Predicting daily visits to a walk-in clinic and emergency department using calendar and weather data. *Journal of General Internal Medicine* **11**(4), 237–239.

HOOT, N. R., LEBLANC, L. J., JONES, I., LEVIN, S. R., ZHOU, C., GADD, C. S. & ARONSKY, D. (2008). Forecasting emergency department crowding: a discrete event simulation. *Annals of Emergency Medicine* **52**(2), 116–125.

HOOT, N. R., LEBLANC, L. J., JONES, I., LEVIN, S. R., ZHOU, C., GADD, C. S. & ARONSKY, D. (2009). Forecasting emergency department crowding: a prospective, real-time evaluation. *Journal of the American Medical Informatics Association* **16**(3), 338–345.

JONES, S. S., EVANS, R. S., ALLEN, T. L., THOMAS, A., HAUG, P. J., WELCH, S. J. & SNOW, G. L. (2009). A multivariate time series approach to modeling and forecasting demand in the emergency department. *Journal of Biomedical Informatics* **42**(1), 123–139.

Jouini, O., Dallery, Y. & Nait-Abdallah, R. (2008). Analysis of the impact of team-based organizations in call center management. *Management Science* **54**(2), 400–414.

Kadri, F., Harrou, F., Chaabane, S. & Tahon, C. (2014). Time series modelling and forecasting of emergency department overcrowding. *Journal of Medical Systems* **38**(9), 1–20.

Kalbfleisch, J. D. & Prentice, R. L. (2011). *The statistical analysis of failure time data*, vol. 360. John Wiley & Sons.

KC, D. S. (2013). Does multitasking improve performance? evidence from the emergency department. *Manufacturing & Service Operations Management* **16**(2), 168–183.

KC, D. S. & Terwiesch, C. (2009). Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* **55**(9), 1486–1498.

KC, D. S. & Terwiesch, C. (2012). An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management* **14**(1), 50–65.

Keskinocak, P., Ravi, R. & Tayur, S. (2001). Scheduling and reliable lead-time quotation for orders with availability intervals and lead-time sensitive revenues. *Management Science* **47**(2), 264–279.

Kleinrock, L. (1964). A delay dependent queue discipline. *Naval Research Logistics Quarterly* **11**, 329–341.

Kuntz, L., Mennicken, R. & Scholtes, S. (2014). Stress on the ward: Evidence of safety tipping points in hospitals. *Management Science* **61**(4), 754–771.

Liang, K. Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**(1), 13–22.

Lucas, R., Farley, H., Twanmoh, J., Urumov, A., Olsen, N., Evans, B. & Kabiri, H. (2009). Emergency department patient flow: the influence of hospital census variables on emergency department length of stay. *Academic Emergency Medicine* **16**(7), 597–602.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting* **1**(2), 111–153.

Makridakis, S., Wheelwright, S. C. & Hyndman, R. J. (2008). *Forecasting methods and applications*. John Wiley & Sons.

Marcilio, I., Hajat, S. & Gouveia, N. (2013). Forecasting daily emergency department visits using calendar variables and ambient temperature readings. *Academic Emergency Medicine* **20**(8), 769–777.

McCarthy, M. L., Zeger, S. L., Ding, R., Aronsky, D., Hoot, N. R. & Kelen, G. D. (2008). The challenge of predicting demand for emergency department services. *Academic Emergency Medicine* **15**(4), 337–346.

McCarthy, M. L., Zeger, S. L., Ding, R., Levin, S. R., Desmond, J. S., Lee, J. & Aronsky, D. (2009). Crowding delays treatment and lengthens emergency department length of stay, even among high-acuity patients. *Annals of Emergency Medicine* **54**(4), 492–503.

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models (Second edition)*. London: Chapman & Hall.

Milner, P. (1988). Forecasting the demand on accident and emergency departments in health districts in the trent region. *Statistics in Medicine* **7**(10), 1061–1072.

Milner, P. (1997). Ten-year follow-up of arima forecasts of attendances at accident and emergency departments in the trent region. *Statistics in Medicine* **16**(18), 2117–2125.

Ministry of Health and Long-Term Care (2010). 2010 Annual Report of the Office of the Auditor General of Ontario. `http://www.auditor.on.ca/en/reports_en/en10/305en10.pdf`. Accessed: 2015-01-06.

NHS-Leeds (2013). NHS Leeds West Clinical Commissioning Group Quality & Performance Report. From the website. `http://www.leedswestccg.nhs.uk/downloads/about%20us/gb%20may%202013/68%20-%20performance%20report%20leeds%20west%20april%20minus%20deep%20dive.pdf` [Accessed: July 30, 2014].

NHS-Stockport (2014). ACCIDENT & EMERGENCY CLINICAL QUALITY INDICATORS. From the website. `https://www.stockport.nhs.uk/webdocs/AE_Clinical_Quality_Indicators_201403.pdf` [Accessed: July 30, 2014].

Pines, J. M., Hilton, J. A., Weber, E. J., Alkemade, A. J., Al Shabanah, H., Anderson, P. D., Bernhard, M., Bertini, A., Gries, A., Ferrandiz, S. *et al.* (2011). International perspectives on emergency department crowding. *Academic Emergency Medicine* **18**(12), 1358–1370.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Reis, B. Y. & Mandl, K. D. (2003). Time series modeling for syndromic surveillance. *BMC Medical Informatics and Decision Making* **3**(1), 2.

Rohleder, T. R. & Klassen, K. J. (2002). Rolling horizon appointment scheduling: a simulation study. *Health Care Management Science* **5**(3), 201–209.

Rotstein, Z., Wilf-Miron, R., Lavi, B., Shahar, A., Gabbay, U. & Noy, S. (1997). The dynamics of patient visits to a public hospital ED: a statistical model. *The American Journal of Emergency Medicine* **15**(6), 596–599.

Rowe, B. H., Bond, K., Ospina, M. B., Blitz, S., Afilalo, M., Campbell, S. G. & Schull, M. (2006). Frequency, determinants, and impact of overcrowding in emergency departments in Canada: a national survey of emergency department directors. *Academic Emergency Medicine* **13**(5 Supplement 1), S27.

Schull, M. J., Kiss, A. & Szalai, J.-P. (2007). The effect of low-complexity patients on emergency department waiting times. *Annals of Emergency Medicine* **49**(3), 257–264.

Schultz, K. L., Juran, D. C., Boudreau, J. W., McClain, J. O. & Thomas, L. J. (1998). Modeling and worker motivation in JIT production systems. *Management Science* **44**(12-part-1), 1595–1607.

Schweigler, L. M., Desmond, J. S., McCarthy, M. L., Bukowski, K. J., Ionides, E. L. & Younger, J. G. (2009). Forecasting models of emergency department crowding. *Academic Emergency Medicine* **16**(4), 301–308.

Sethi, S. & Sorger, G. (1991). A theory of rolling horizon decision making. *Annals of Operations Research* **29**(1), 387–415.

Sethi, S. P., Yan, H. & Zhang, H. (2006). *Inventory and Supply Chain Management with Forecast Updates*. Springer Science & Business Media.

Sharif, A. B., Stanford, D. A., Taylor, P. & Ziedins, I. (2014). A multi-class multi-server accumulating priority queue with application to health care. *Operations Research for Health Care* **3**(2), 73–79.

Shumway, R. H. & Stoffer, D. S. (2010). *Time series analysis and its applications: with R examples*. Springer Science & Business Media.

Song, H., Tucker, A. L. & Murrell, K. L. (2015). The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science* .

Stanford, D. A., Taylor, P. & Ziedins, I. (2013). Waiting time distributions in the accumulating priority queue. *Accepted for publication in Queueing Systems: Theory and Applications, October 21,2013, 40 pages.* .

Stanford, D. A., Taylor, P. & Ziedins, I. (2014). Waiting time distributions in the accumulating priority queue. *Queueing Systems* **77**(3), 297–330.

STASINOPOULOS, D. M. & RIGBY, R. A. (2007). Generalized additive models for location scale and shape (gamlss) in R. *Journal of Statistical Software* **23**(7), 1–46.

STEHFEST, H. (1970). Numerical inversion of Laplace transforms. *Communications of the ACM* **13**.

SUN, Y., HENG, B. H., SEOW, Y. T. & SEOW, E. (2009). Forecasting daily attendances at an emergency department to aid resource planning. *BMC Emergency Medicine* **9**(1), 1.

TAN, T. F. & NETESSINE, S. (2014). When does the devil make work? an empirical study of the impact of workload on worker productivity. *Management Science* **60**(6), 1574–1593.

TANDBERG, D. & QUALLS, C. (1994). Time series forecasts of emergency department patient volume, length of stay, and acuity. *Annals of Emergency Medicine* **23**(2), 299–306.

THEGUARDIAN (2011). NHS waiting lists: how long are patients waiting? From the website. `http://www.theguardian.com/news/datablog/2011/jul/11/nhs-waiting-lists-data` [Accessed: July 30, 2014].

WARGON, M., GUIDET, B., HOANG, T. & HEJBLUM, G. (2009). A systematic review of models for forecasting the number of emergency department visits. *Emergency Medicine Journal* **26**(6), 395–399.

WOLFF, R. W. (1982). Poisson arrivals see time averages. *Operations Research* **30**(2), 223–231.

WOLFF, R. W. (1989). *Stochastic modeling and the theory of queues.* Industrial and Systems Engineering, Prentice Hall Inc., Upper Saddle River, NJ.

XIONG, Y., MURDOCH, D. J. & STANFORD, D. A. (2013). Perfect and nearly perfect sampling of work-conserving queues. *Submitted for publication, October 2013, 20 pages.* .

ZHANG, D. & VOSSEN, T. W. (2013). An approximate dynamic programming approach to a rolling-horizon appointment scheduling problem. Tech. rep., Working paper, Leeds School of Business, University of Colorado at Boulder.

Curriculum Vitae

Name:	Azaz Sharif
Post-Secondary Education and Degrees:	University of Western Ontario London, ON 2008 - 2010 M.Sc. (Biostatistics)
	University of Western Ontario London, ON 2010 - 2015 Ph.D. (Biostatistics)
Honours and Awards:	WGRS 2008-2014
Related Work Experience:	Sessional Lecturer The University of Western Ontario Fall'2014, Fall'2015 - todate
	Research and Teaching Assistant The University of Western Ontario 2008 - 2015

Publications:

Sharif, A. B., Stanford, D. A., Taylor, P., Ziedins, I. (2014). A multi-class multi-server accumulating priority queue with application to health care. Operations Research for Health Care, 3(2), 73-79.

Li, N., Stanford, D. A., Sharif, A. B., Caron, R. J. (2015). Optimization of queues operating under waiting time limits. Operations Research, (under review).