SETTING STAFFING REQUIREMENTS FOR AN EMERGENCY DEPARTMENT IN EVENT OF A SURGE

by

Hengameh Vahabzadeh Sefiddarboni
Master of Applied Science

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

In the
School of Engineering Science

SIMON FRASER UNIVERSITY

Spring 2011

APPROVAL

Name:	**Hengameh Vahabzadeh**
Degree:	**Master of Applied Science**
Title of Thesis:	**SETTING STAFFING REQUIREMENTS FOR AN EMERGENCY DEPARTMENT IN EVENT OF A SURGE**

Examining Committee:

Dr. Mehrdad Moallem

Associate Professor, Engineering Science

Chair

Dr. G. Gary Wang
Associate Professor, Engineering Science

Senior Supervisor

Dr. Payman Jula
Assistant Professor, Business
Supervisor

Dr. Behraad Bahreyni
Assistant Professor, Engineering Science
Internal Examiner

Date Defended/Approved: 30 March 2011

SFU SIMON FRASER UNIVERSITY
LIBRARY

Declaration of Partial Copyright Licence

ABSTRACT

Emergency Departments often experience sudden increases in patient visits, referred as surge. Surge brings challenges to setting staffing requirement. Since the patient arrival rate is not constant over time during surge, a network of nonstationary queueing models and time-varying discrete event simulation models have been developed to model the surge in emergency department. For queueing models, to mathematically analyze this time-varying system, many approximation methods that have been proposed in literature are compared in this work in order to identify the best approach for modelling surge. Due to the lack of analytical approaches to evaluate these methods, a validated time-varying simulation model was used as the reference for comparison. In addition, a detailed discrete event simulation model was built and validated with historical data for St. Paul's hospital in BC, Canada. Both the queueing theory approach and simulation method are studied and their advantages and disadvantages are discussed for modelling surge.

Keywords: healthcare modelling, emergency department, queueing theory, discrete event simulation

ACKNOWLEDGEMENTS

I am greatly indebted to my senior supervisor Dr. G. Gary Wang, for his incessant encouragement, invaluable guidance and persistent support throughout the course of this research. I thank him for having confidence in my abilities to handle such an intricate research topic. Without his critical reviews and intellectual inputs, this thesis would not have been possible in the present form. I would also like to thank my supervisor Dr. Payman Jula for his advice and help throughout this work. I am also sincerely thankful to Dr. Behraad Bahreyni and Dr. Mehrdad Moallem for being my committee members.

Finally, I would like to offer my endless gratitude to my mother and father for their unfailing love, and also their support. I would like to thank all my wonderful friends who comforted me during the difficult times, and offered me great support and help.

TABLE OF CONTENTS

Approval ii
Abstract iii
Acknowledgements iv
Table of Contents v
List of Figures vii
List of Tables viii

1: Introduction 1
1.1 The Problem of ED overcrowding 3
1.2 Motivations 5
1.3 Objectives 6
1.4 Thesis outline 6

2: Background and literature review 8
2.1 Simulating queueing models 11
2.2 Transient queueing systems 13
2.3 Staffing queueing systems with time variant arrival rate 16
2.4 Using stationary models for non-stationary systems 18
2.4.1 Simple Stationary Approximation (SSA) 19
2.4.2 Pointwise Stationary Approximation (PSA) 22
Staffing with PSA 24
2.4.3 Segmented PSA 25
2.4.4 Infinite Server 25
2.4.5 Modified Offered Load (MOL) 28
2.4.6 The Normal Approximation 30
Square root staffing formula 31
Normal approximation refinement 32
2.4.7 Stationary Independent Period-by-Period (SIPP) Method 33
2.4.8 Busy-Hour Engineering and the Simple Peak Hour Approximation (SPHA) 34
2.5 Summary 34

3: Queueing model implementation and validation 35
3.1 Setting staffing requirements for the emergency department with time varying arrivals 35
3.2 The “Exact” method for time varying arrivals 37
3.3 Simulation approach 39
3.4 Validation of the simulation model 41

4: Comparison of the queueing approximation methods 48

4.1 Experimental design 49

4.2 Approximating probability of delay in *M/G/s* model 50

4.3 Test and Comparison Results 53

4.4 Summary 56

5: Modeling emergency department with both queueing and simulation models 58

5.1 ED at St. Paul's hospital 58

5.2 Approximating the departure process 62

5.3 Experimenting with queueing network 65

5.4 Applying the method to actual data 66

5.4.1 Fitting non-homogeneous Poisson distribution to actual data 67

R-Square Error 69

Root Mean Squared Error 70

5.4.2 Service times 71

5.5 Simulation of the ED at St. Paul's 75

5.6 Scenarios for analysis 76

5.7 Inputting queueing model outcome to detailed simulation 81

5.8 Summary 83

6: Conclusion 84

6.1 Contributions 86

6.1.1 Theoretical Contributions 86

6.1.2 Practical Contributions 87

6.2 Future work 88

Appendices 90

Reference List 92

LIST OF FIGURES

Figure 2-1 The basic queueing process 8

Figure 2-2 Arrival rate at St. Paul’s emergency department, BC, Canada. 15

Figure 2-3 Number of servers determined by the SSA method and the offered load in hours of day 20

Figure 2-4 Probability of delay with the SSA method in hours of day 21

Figure 2-5 Number of servers determined by the PSA method (upper graph) and the offered load (lower graph) versus hours of day 23

Figure 2-6 Probability of delay with PSA method in hours of day 24

Figure 2-7 Number of servers determined by the MOL method (upper graph) and the offered load (lower graph) versus hours of day 29

Figure 2-8 Probability of Delay with the MOL method in hours of day 30

Figure 3-1 Conceptual representation of simulation model 40

Figure 5-1 Flow of patients in the ED 61

Figure 5-2 Schematic representation of the ED units 62

Figure 5-3 The arrival function fitted over the actual data from the ED..............................71

LIST OF TABLES

Table 3-1 Comparing Stationary Simulation (SS) and Periodic Steady State Simulation (PSSS) for arrival rates with various cycle lengths..........................45

Table 3-2 Parameter values for computational examples..46

Table 3-3 Outcome of applying t-test to the model with six-hour long planning period..46

Table 3-4 Outcome of applying t-test to the model with four-hour long planning period..47

Table 3-5 Outcome of applying t-test to the model with two-hour long planning period..47

Table 4-1 Parameter values for computational examples..53

Table 4-2 Time average relative and maximum relative error by SCV and planning period (PP) length 55

Table 4-3 Time average relative and maximum relative error by target service level (α) and planning period (PP) length 56

Table 5-1 Comparison of the MOL solution with simulation for the Fast Track unit..........66

Table 5-2 Comparison of the MOL solution with simulation for Acute unit........................66

Table 5-3 Parameters and fitting error for the approximated arrival rate...........................69

Table 5-4 Service times mean and standard deviation..72

Table 5-5 The number of servers suggested for two-hour planning period and probability of delay estimated with MOL and simulation 73

Table 5-6 The number of servers suggested for four-hour planning period and probability of delay estimated with MOL and simulation 74

Table 5-7 The number of servers suggested for six-hour planning period and probability of delay estimated with MOL and simulation 74

Table 5-8 Results of simulation for different scenarios in terms of desired statistics (LOS, T2P, T2B)..79

Table 6-1 Comparison of the queueing and simulation models 88

1: INTRODUCTION

Almost everyone encounters queues every now and then including waiting to make a bank deposit, pay for groceries, etc. People have become used to considerable amount of waiting, but yet get frustrated by unusually long waits. From another perspective, the amount of time that people waste waiting in queues is not only a personal annoyance, but also is a major factor of quality of life and efficiency of a nation's economy. In United States, it is been estimated that people spend 37,000,000,000 hours per year waiting in queues [1]. If this time could be spent productively, it would make 20 million person-years of useful work each year!

In the healthcare section, these waiting times not only are a waste of people productive times, but also could be a source of threat to their lives. From the waiting in an emergency department to the waiting in the queue for special health services such as Magnetic Resonance Imaging (MRI), long waiting times could have undesirable impacts on people's health.

Healthcare is a large component of Canadian economy that affects every resident of this country. According to the latest OECD (Organization for Economic Co-operative and Development) total expenditure on healthcare accounted for 10.4% of the GDP in 2008, which is more than one percent higher than the average of the OECD countries [2]. High quality healthcare is one of the main factors that contribute to Canada's ranking as one of the best countries to live in, according to the United Nation's Quality of Life Survey [2].

According to a study reported in 2007, 57% of Canadians waited more than four weeks to see a specialist and 24% waited for more than four hours in the emergency department [3]. Long waiting times and shortage of medical practitioners are among the major complaints about the Canadian healthcare system.

The question is that, if we cannot take care of emergency patients on a normal day, how will we manage a large-scale disaster? When a disaster takes place, the significance of this problem is more highlighted. In the event of a surge such as a mass casualty incident, any minute of delay in serving patients can potentially cost a human's life. Disaster is referred to low probability but high impact incident that causes a large number of individuals to become ill or injured [4].

Disasters and infectious disease outbreaks over the last several years have demonstrated the importance of emergency preparedness for large-scale events affecting many people. The ability to respond effectively to events producing a massive flow of patients requires surge capacity [5]. According to Webster's Online Dictionary (2009b), surge is defined as "a sudden forceful flow" or "a sudden or abrupt strong increase." Although tools to measure "crowding" phenomenon have been developed and tested, no uniform agreement on the actual operational definition of it was attained [6]. In addition to influx (volume rate), surge is further composed of the following components: event (type, scale, and duration) and resource demand (consumption and degradation) [7]. Kelen and McCarthy proposed a recent modified definition for surge as "a sizeable increase in demand for resources compared with a baseline demand" [8].

Hospitals estimate of their average daily patient load under normal operations and prepare their resources and staffing accordingly. In case of a mass casualty incident, a healthcare facility or system may be suddenly faced with a significant increase or surge of patients. We can call this incident a surge-generating event [9].

The surge generating event that affects the operations at healthcare systems could have many forms. In terms of surge generating events, surge can be categorized as contained or population based. A contained event has a distinct geographic focus, even if the focus is very large. Such an event requires local response from hospitals, clinics and personnel within the affected region. Examples of contained events could include bombings, tornados, or hurricanes. A population-based event is not geographically defined and can spread infectiously. Such events would likely start from an initial local event but would soon spread out of the borders of a contained event. As an example, we could mention epidemic diseases [9]. Either event causes temporal high resource demands to hospitals and their Emergency Departments (ED). The ability to respond effectively to events producing a massive flow of patients requires surge capacity.

1.1 The Problem of ED overcrowding

Timely access to an emergency provider is an important factor of quality for emergency departments (ED), though many EDs are facing a number of challenges which lead to excessive waiting time and diversions. To provide a high quality healthcare to the public, healthcare issues such as overcrowding must be dealt effectively and quickly. One of the major challenges with healthcare systems is their complexity. There are many different decision making units involved with almost any individual healthcare issue. Therefore, it is very important to make informed decisions that are based on advanced methods.

One of the factors that contribute to ED overcrowding is the lack of available inpatient beds. ED patients' that get stable health conditions are usually transferred to an inpatient bed, if not sent home. In this case, the ED bed is freed up for another patient whose condition is not stable. However, because of the lack of inpatient beds in hospitals, the process of transfer may not take place. In a study done by Estey et al. [10], it is mentioned that the lack of inpatient beds is believed to be one of the main causes of ED overcrowding and long waiting times. When there is no inpatient bed available, patients end up waiting in the ED beds and consequently, the ED becomes congested. Other factors that are recognized to contribute to the ED overcrowding are

- Use of an ED bed for non-emergency cases
- Staff shortage
- Aging population and increasing patient acuity
- Increase in volume of the patients coming to ED [4]

In this work we focus on the shortage of staffing as a possible source of crowding when the patient arrival rate is increased as a result of a surge generating event. Hospitals mostly struggle to provide sufficient staffing to handle increasing demand for care. Experience shows that one of the most critical resources in the ED is usually physician. Furthermore, about 30-40% of the ED costs account for emergency physicians [4]. Therefore, determining the right number of physicians to satisfy the demand would have great impact both on the quality of the service and its costs.

1.2 Motivations

The motivation of this study in the beginning was to find out a solution for the problem of overcrowding in a local emergency department at the time of the outbreak of an infectious disease. This directed us to the definition of the surge and the fact that surge inflicts temporal challenges to the managers for determining the staffing level that could respond effectively and efficiently to the influx of patients arriving at the ED. As the surge situation is sudden and short in duration, we had to analyse it with transient models such as simulation or transient queueing models.

Both simulation and nonstationary queueing models have their own advantages and disadvantages. For example the simulation approach is easy to employ but it would take a long time and an excessive effort to construct and validate the model. In addition, examining alternative solutions with simulation requires direct experimenting and therefore, it generally requires a very long time to determine the best solution. On the other hand, queueing models are very fast and easy to build but they usually have difficulty capturing the true essence of the actual system in every aspect, or otherwise they would become so complicated that cannot be analyzed mathematically. As a compromise in this work, we built both a queueing model and a simulation model. The simulation model was validated with the result of the simple queueing model and then used in place of more complex queueing models when they could not be approached analytically.

Specifying the time dependent number of staff is done by implementing a number of approximation methods found in literature. Currently, to the best of our knowledge, there are no tools that allow us to evaluate the performance of these methods for complex queueing models. Such tool is developed in this work and is then utilized for determining the most appropriate approximation method for the specifics of our surge problem.

1.3 Objectives

The objective of this research is to develop a queueing model that represents ED at the event of surge, when the system has nonstationary arrival rate and non-exponential service time distribution. A number of approximation methods would be implemented to determine the staffing requirement of the ED according to variations in arrival rates. All these approximation methods are to be compared to identify the most suitable one for modeling surge under certain conditions.

It is also of interest in this work to use the discrete event simulation approach to model the surge in ED, and to compare both queueing models and simulation models as staffing tools.

1.4 Thesis outline

The remainder of this thesis is organized as follows: Chapter 2 provides a literature review on queueing theory and simulation and related studies about the nonstationary queueing model. The approximation methods that are developed to cope with this nonstationarity are introduced and the outcomes of implementing these methods for a challenging example are discussed. In chapter 3, a nonstationary simulation model is developed and validated with the analytical results obtained from a simpler nonstationary queueing model. In chapter 4, the validated simulation model is employed for comparison of a number of pre-introduced approximation methods for a more complex queueing model with non-exponential service time distribution. Chapter 5 describes the modeling of the ED of a local hospital by means of connecting three queueing models to form a queueing network. It also includes the outcome of applying the best approximation method found in the previous chapter, to this queueing model and presents the staffing level for each of the three nodes of the network. Finally, chapter 6 will conclude the overall research and discusses future research directions.

2: BACKGROUND AND LITERATURE REVIEW

Queueing theory is the study of waiting in various queues from the telecommunication transmissions and manufacturing to service. It uses queueing models to represent various types of the queueing systems.

The basic process for the queueing models is assumed as follows. *Customers* requiring service are generated over time by *input source*. These customers enter the *queueing system* and join a *queue*. At certain times, a member of the queue is selected for service by some rule known as the *queue discipline.* The required service is then performed for the customer by the *service mechanism*, after which the customer leaves the queueing system. Figure 2-1 illustrates a basic queueing process.

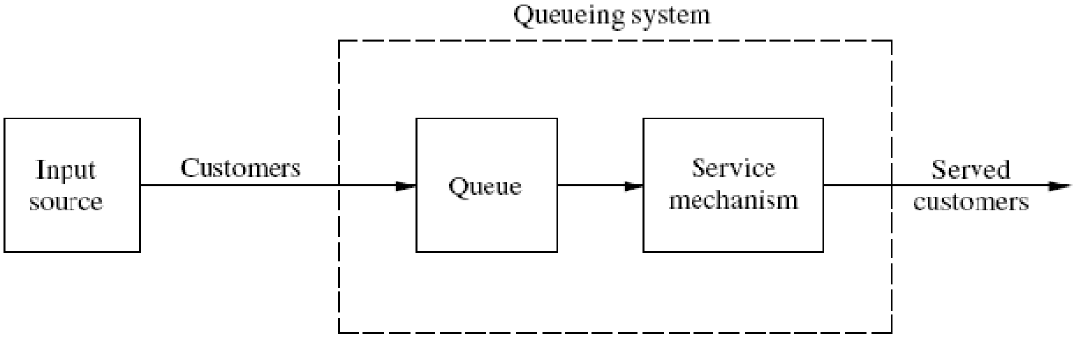

Figure 2-1 The basic queueing process [1]

Input source or the calling population is the population from which arrivals come. It could be assumed to have finite or infinite size. Since the calculations are much easier for the infinite case, this assumption is often made. Customers are generated according to a statistical pattern, which in most cases is assumed to be a *Poisson process*. There are some unusual assumptions such as balking and blocking, where the customers refuse to enter the system and is lost if the queue is too long. These assumptions must be stated if it is the case.

The queue is where the customers wait before receiving service. A queue can have an infinite or finite capacity. The finite capacity queues bring difficulty to analysis of the system, but it should be considered, if it is small enough to be filled up with some frequency.

Queue discipline refers to the order in which the members of the queue are selected for service. For instance, it could be first-come-first-served, random, according to some priority procedure, or some other order. The general assumption about the queue discipline is first-come-first-served, unless otherwise stated. The service mechanism contains one or more service facilities, each one with one or more parallel service channels, called servers [1]. Representing this information for a queueing model a specific notation is used in the form of - / - / - . Where any of the following letters that places in the first two dashed areas would denote a distribution for inter-arrival times. It is assumed that all the inter-arrival times are independent and identically distributed.

M = exponential distribution (Markovian),

D = degenerate distribution (constant times),

E_k = Erlang distribution (shape parameter = k),

G = general distribution (any arbitrary distribution allowed),

For instance, the model can be labelled as follows:

$M/M/s$

Thus in this example, the model assumes that both inter-arrival times and service times have an exponential distribution and the number of servers is s (any positive integer). Unless otherwise noted, the following standard terminology and notation will be used:

State of system = the number of customers in queueing system.

Queue length = the number of customers waiting for service to begin; equals to the state of system minus the number of customers being served.

$N(t)$ = the number of customers in queueing system at time t ($t \geq 0$).

$P_n(t)$ = the probability of exactly n customers in the queueing system at time t, given the number of customers at time 0.

s = the number of servers (parallel service channels) in the queueing system.

λ_n = the mean arrival rate (the expected number of arrivals per unit time) of new customers when n customers are in the system.

μ_n = the mean service rate for the overall system (the expected number of customers completing service per unit time) when n customers are in the system. *Note:* μ_n represents the *combined* rate at which all *busy* servers (those serving customers) achieve service completions.

When λ_n is a constant for all n, this constant is denoted by λ. When the mean service rate per busy server is a constant for all $n \geq 1$, this constant is denoted by μ. (In this case, $\mu_n = s\mu$ when $n \geq s$, that is, when all s servers are busy.) Under these circumstances, $1/\lambda$ and $1/\mu$ are the *expected inter-arrival time* and the *expected service time,* respectively. Also, $\rho = \lambda/(s\mu)$ is the utilization factor for the service facility, i.e., the expected fraction of time the individual servers are busy, because $\lambda/(s\mu)$ represents the fraction of the system's service capacity ($s\mu$) that is being *utilized* on average by arriving customers (λ) [1].

Queueing models are analytically solvable only for very specific conditions. A more general approach to gain insight about the queueing models is to simulate them on a computer.

2.1 Simulating queueing models

Simulation is a collection of methods for mimicking the behaviour of a real system, usually run on computer with appropriate software. Like most analysis methods, it involves models of a system. System here refers to a facility or process, either actual or planned, such as a manufacturing plant, a bank, and so on. These systems are usually studied to measure their performance, improve their operation, or for new system design. Managers or comptrollers of a system might also like to have an available aid for day-to-day operations. Moreover, some managers do not really care about the output of simulations; their primary goal is to understand how their system is working. Simulation could provide a great insight into what changes need to be made in a system [11].

Specialists turn to simulation when a detailed analysis is required of a complex system because mathematical or analytical modelling techniques are often not sufficient. Major weaknesses of mathematical and analytical techniques are:

1. A mathematical model of a complex system can not accurately describe the stochastic elements of the system. Randomness and nonlinearity of the discrete operations leads to inexplicit functions.

2. Since dynamic systems involve randomness that changes with time, modeling of these complex systems theoretically requires too many simplifications; therefore, the emerging models may not be valid [12]**.**

One other important use of simulation models lies in comparing and contrasting competing design alternatives without any physical costs. Many studies have been done on application of queueing models and simulation in emergency departments. For example, Green et al. [13] utilized queueing theory to adjust the staffing patterns subject to the variations in the patients' arrival rate in order to optimize the timely care of patients. They used data from 78 weeks of patient visits to a medium size ED, and examined the effect of employing a queueing model for determining the staffing requirement. However, they only considered a single *M/M/s* model and excluded other types of providers (e.g. attending physicians, residents, nurse practitioners and locations and/or types of care (e.g. multiple district and fast track area).

De Bruin et al. modeled the emergency in-patient flow of cardiac patients in university medical centre in Amsterdam using a queueing system [14]. They assumed the patient inter-arrival rate to be exponentially distributed and applied a stationary 2-D queueing system with blocking to analyze congestion in emergency care chains. They determined the optimal bed allocation over the emergency care chain, given a required service level (e.g. maximum 5% refused admissions).

Abujudeh et al. evaluated the operation of the portable X-ray machine in relation to examinations ordered by the University of Medicine and Dentistry of New Jersey, and identified bottlenecks in their system [15]. They considered a network of stationary queueing systems and estimated the arrival rate, service rate and utilization factor based on historical data. They calculated the average response time ($E[w]$) using the estimated values.

Connelly and Bair used discrete event simulation to model emergency department activities [16]. Bagust et al. also developed a stochastic simulation model to study the dynamics of bed use in accommodating emergency admissions [17]. However, to the best of our knowledge no one had yet studied the effect of surge on an emergency department with the help of queueing theory and simulation.

As we previously discussed, hospitals use the information about the daily flow of patients and determine their staffing accordingly. However, not many hospitals make capacity decisions with the help of the Operations Research (OR) model-based analyses. Statistical fluctuations in individual patient arrival times and the variability of the time needed by a provider to treat patients can have a major impact on hospital operation and capacity requirements. If the variability is ignored in the process of modeling, an unrealistic and static image of reality would appear. Such a model would not be capable of explaining the dynamics of the flow of patients [14]. To consider the variability in queueing models we turn to modeling the transient phase of systems.

2.2 Transient queueing systems

One major application of queueing theory in service systems is for determining the number of servers needed to handle the demand. A common characteristic of many service systems – ranging from telephone call centres to police patrol and hospital emergency rooms – is that the demand for service often varies greatly by time of the day. This can be seen in Figure 2-2, which depicts the hourly arrival rates to a local emergency department. Moreover when an unusual incident occurs, such as the population-based surge, the variability in arrival could become even worse, making it more difficult to determine the number of staff required for coping with the demand.

When the arrival is highly time varying, traditional queueing theory formulas do not seem to be applicable of analysing the systems. Queueing theory has mostly focused on the steady state condition, partly because the transient case is more difficult to study analytically.

When a queueing system has just begun operation, or when a sudden change occurs in the system, the state of the system (number of customers in the system) is greatly affected by its initial state and by the time that has since elapsed. The system is said to be in the transient condition.

When sufficient time has elapsed, the state of the system becomes essentially independent of the initial state and the elapsed time. The system reaches a steady-state condition, where the probability distribution of the state remains the same over time (as the steady state or stationary distribution) [1].

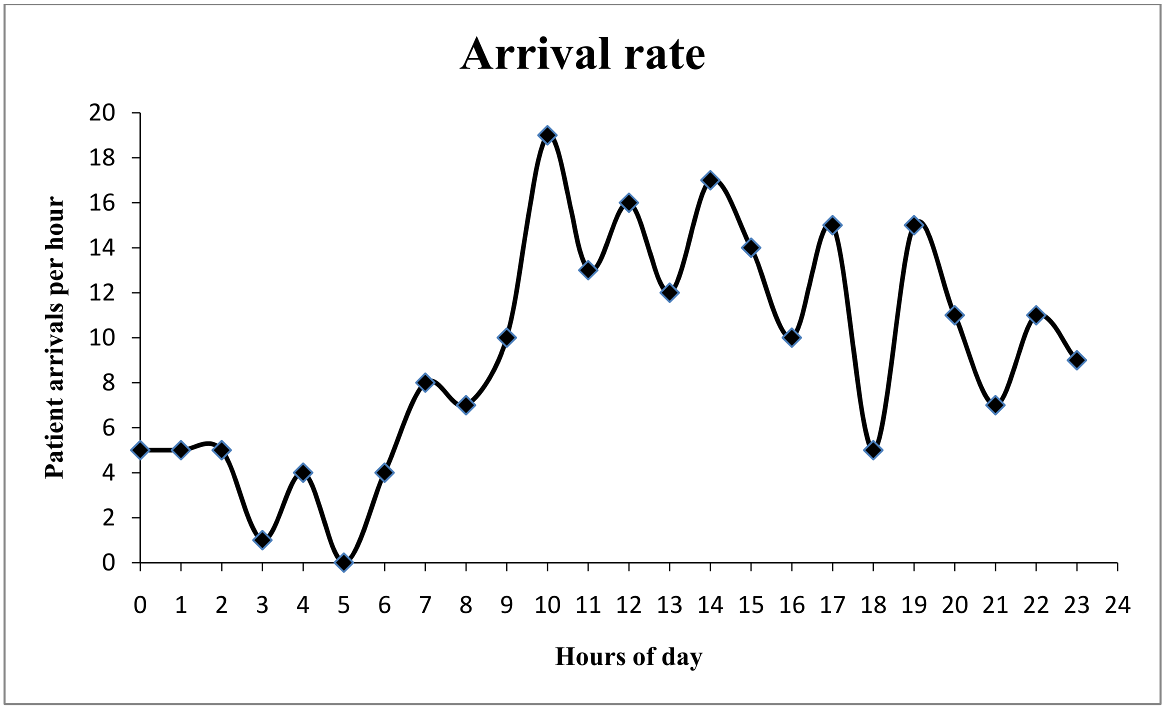

Figure 2-2 Arrival rate at St. Paul's emergency department, BC, Canada.

As mentioned before, most of the queueing models assume that the arrival process is Poisson with a rate that remains constant all the time. However, service systems usually experience cyclical demands with various cycle lengths, for example, daily, weekly, monthly and yearly cycles. Analysing the data collected from the St. Paul's Hospital emergency department in BC, Canada at the time of the H1N1 flu outbreak, we realized that the arrival rate at this time also tends to have periodic behaviour. We focus on daily cycles since the amplitude of the variations is greater on this time scale; and besides, the duration at which the ED was facing the surge was short (about one month). We assume that the facility is operating continuously for T=24 hours per day.

Different approaches are developed to cope with the time-varying demand and setting staffing requirements to satisfy the demand. Some of the approximation methods, however, assume the rate to be constant for blocks of time (e.g. two-hours), with a separate queueing model fitted for each block of time. This is called the stationary independent period-by-period (SIPP) method. One way of capturing the changes in the arrival rate is through utilizing non-homogeneous Poisson process. The first step to fit a model to the system is forecasting the arrival rate. Here we will assume that arrival rate function has been created.

2.3 Staffing queueing systems with time variant arrival rate

Setting staffing requirements is a decision that must be made in the design and management of a service system. Managers decide what the overall capacity of the system would be in a long-term planning horizon. The daily staffing decision determines the number of servers needed to work during each staffing interval over the day. After setting the staffing requirements, managers make agent-scheduling decisions, which specify the number of agents that should work on specific tasks, period by period, corresponding to the previously determined staffing levels, work rules, and legal constraints. The scheduling decision is usually addressed using an integer linear program [18].

There are a plenty of factors that could be taken in to account while developing specific staffing schedules. These include the complex scheduling constraints like employees' preferred start times, quitting times, and shift lengths; legal or policy limits on the number of consecutive hours and/or days worked; restricting the patterns of days off and on duty; providing required lunch and coffee breaks, etc. The fundamental requirement, however, is that there should be enough staff on duty at all times to meet targeted service levels.

The goal in staffing problem is to determine the specification of the staffing-requirement function *s*(*t*)— the number of servers required to be on duty as a function of time *t.* Since the changes in staffing are only allowed at certain times, e.g., once every 15 minutes, once every hour, or only once every 8 hours, we would like to determine a good staffing function subject to this constraint that allows the changes only at the end of the pre-specified *staffing intervals*.

The objective is to minimize the total number of staff hours required over the day, while meeting a desired level of service performance in each staffing interval. One of these performance constraints is *service level*: the requirement that *x%* of the demand is responded within *y* unit of time. Another performance measure, which is closely related to the former, is the delay probability, which is the probability that an arriving customer has to wait before starting service. That is the special case of the service level in which *y* = 0.

Customer abandonments is also a very important measure. Managers often place bounds on the rate of abandonment. Expected waiting time (before starting service) is also another performance measure that is commonly constrained by managers. It is also called the average speed to respond [19]. Among all these measures, the delay probability constraint is easier to compute; it is relatively more robust and insensitive to model details and has meaning independent of scale (typical the number of servers).

Simulation can also be employed to set staffing levels. In complex systems that cannot be described by analytical queueing models, simulation could be very helpful to measure performance. It is easy to evaluate the performance of a given model like $M_t/G/s_t$, for any given function of *s*(*t*), by using computer simulation. Nevertheless, determining a good function for staffing among the vast number of possibilities is challenging. As an example, consider a call centre with 100 agents, where there are 20 available staffing-change points during a day and 20 possible staffing levels at each of these staffing intervals; there would be $20^{20} \approx 10^{26}$ different staffing functions to consider. However, there are alternative analytical methods that make it possible to reduce to a small number of attractive alternatives. After determining the staffing requirements using the approximate analytical approaches, it would be helpful to simulate the system in detail to verify that the suggested staffing levels actually produce the desired performance [18].

2.4 Using stationary models for non-stationary systems

Even when the arrival rate is highly time varying, it might be possible to utilize the stationary models to specify the staffing requirements. Although it is not usually appropriate to staff to the overall average arrival rate over the entire day, surprisingly that is applicable when the arrival rate changes very rapidly (relative to the service time) [20].

Throughout this section we will consider a Markovian $M_t/M/s_t$ system with exponential service times having mean of 1 and non-homogeneous Poisson arrival process with sinusoidal arrival rate function $\lambda(t) = 30+20sin(5t)$. (A non-homogeneous Poisson process is a Poisson process that has time varying mean value, while in homogeneous Poisson process the mean value is constant). Assume that the target delay probability is 0.13, there are no abandonments, and patients would wait in a waiting room with unlimited capacity. This is a challenging example, since the arrival rate fluctuates rapidly relative to the mean service time. If we think of daily cycles, the mean service time is about 0.8 days (each cycle is $2\pi/5 \approx 1.256$).

2.4.1 Simple Stationary Approximation (SSA)

We applied the Simple Stationary Approximation (SSA) to this arrival rate, which takes an overall average on the arrival rate, and determine a single staffing level for the entire day accordingly. The average arrival rate is 30, in this example. The steady state delay probability with this constant arrival rate is 0.112 if we have 38 servers. If we reduce the number of servers by one, the delay probability would be 0.155, which violates the targeted delay.

The outcome of this experiment is shown in Figure 2-4 and Figure 2-3. As we expected, because of the rapid fluctuations of the arrival rate SSA performs fairly well and the probability of delay varies within [0 0.25] which is relatively close to the previously defined target of 0.13.

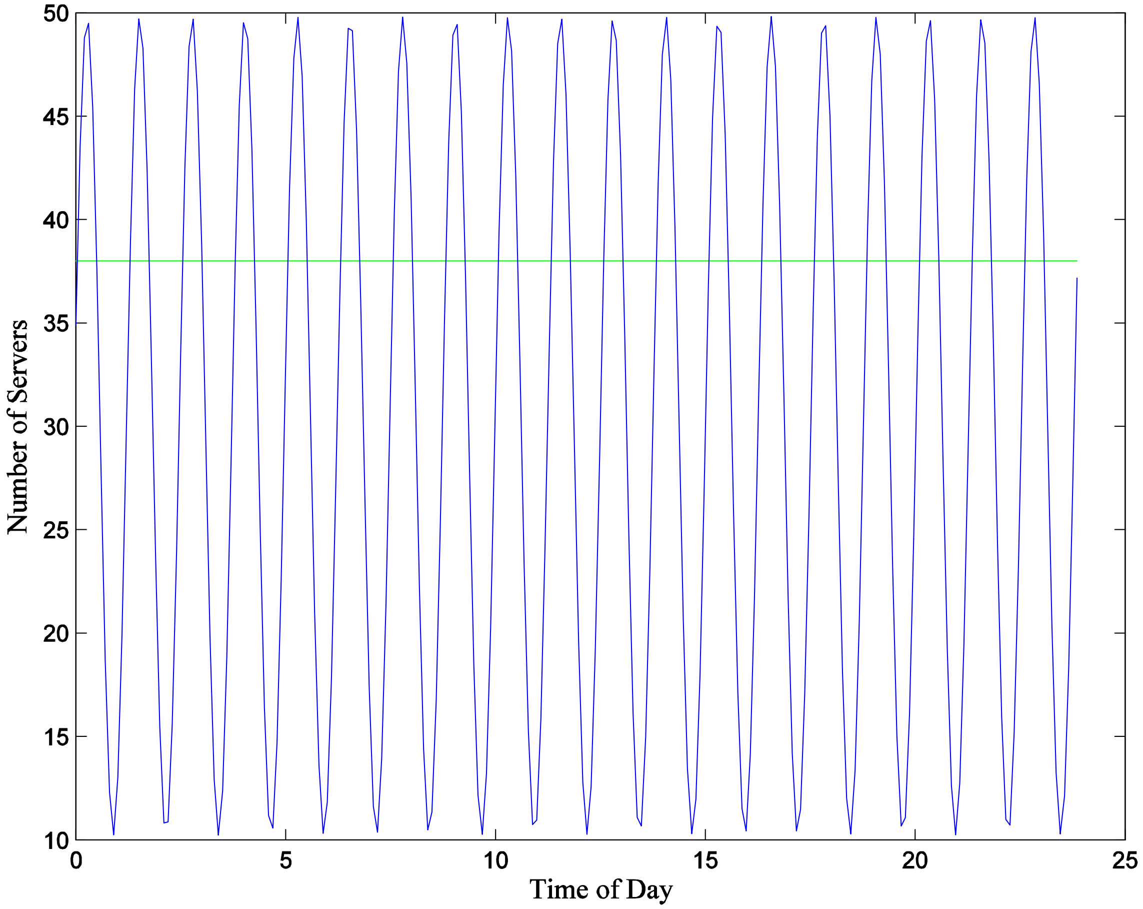

Figure 2-3 Number of servers determined by the SSA method and the offered load in hours of day

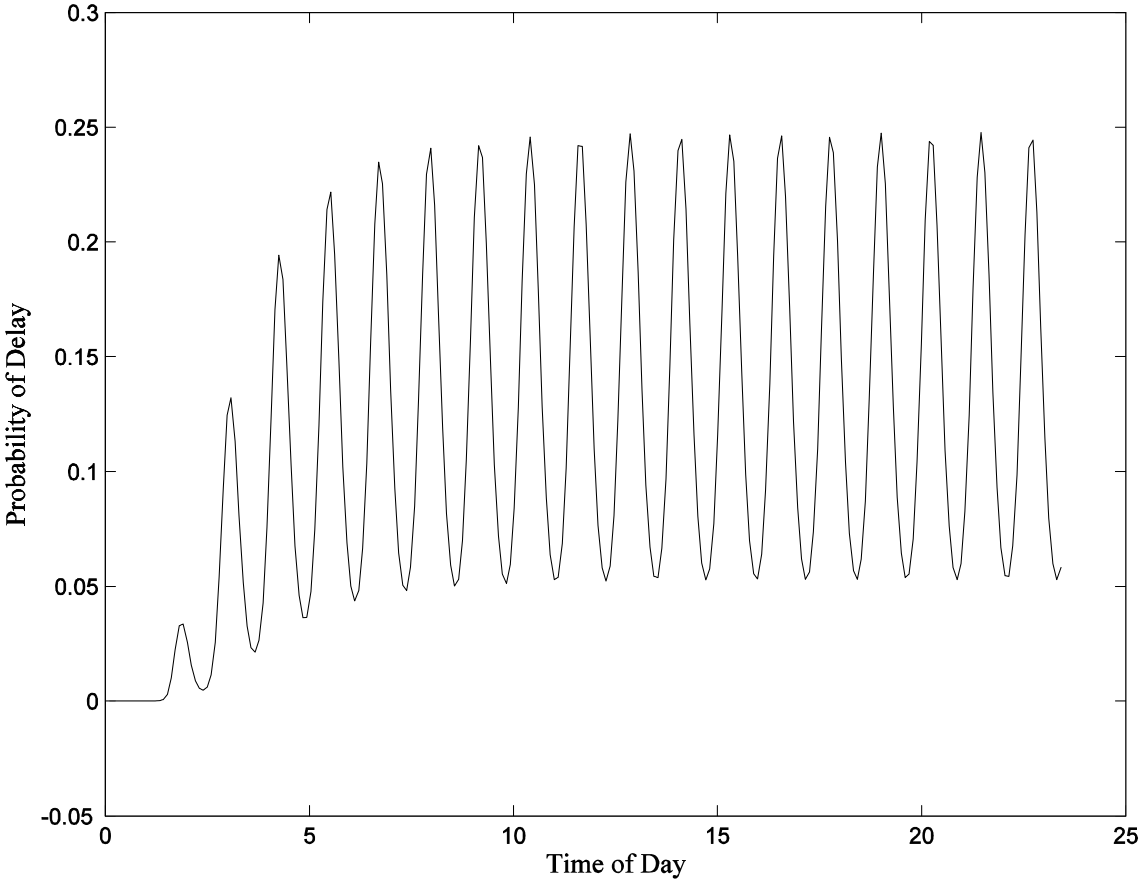

Figure 2-4 Probability of delay with the SSA method in hours of day

Generally, it is possible to use stationary models in a nonstationary manner. That is to divide time into segments and use a stationary model for each segment. This is appropriate when the service times are short (e.g. 5-10 minutes) and the quality-of-service standard is high. Under such conditions, the systems are seldom overloaded and staffing requirements follow predictable patterns. These methods are for the cases where the staffing intervals are short; however, some modifications are applicable for the longer staffing intervals [18].

Here we discuss a number of approaches that are developed for staffing nonstationary systems based on the approximations with a variant of stationary models.

2.4.2 Pointwise Stationary Approximation (PSA)

The classic case with short service times, a high quality of service standard, and short staffing intervals is addressed by an effective analytic strategy called Pointwise Stationary Approximation (PSA). *PSA provides a time-dependent description of performance based on a stationary model, using the arrival rate and other parameters that prevail at each moment in time to describe the performance at that time* [21].

However, PSA approach generates a time-dependent staffing function, which does not limit the changes to be at the boundaries of the staffing intervals. If we could staff in this fully time-dependent manner, we expect to produce a good staffing function. Nevertheless, one is usually forced to keep the staffing level constant during each staffing interval.

Even though PSA is expected to be an excellent approach for staffing, if one does not specify any constraints for the length of the staffing intervals, experimenting with the aforementioned arrival rate reveals that there exist circumstances for which PSA fails miserably. As Figure 2-5 depicts, the PSA method applied to the arrival rate function $\lambda(t)= 30+20sin(5t)$ does not provide very effective results. Since the offered load is changing between 10-50, with the target delay 0.13, PSA suggests the number of servers that varies between 15 to 50, and the delay probability oscillates nearly over the full range of 0 and 1.

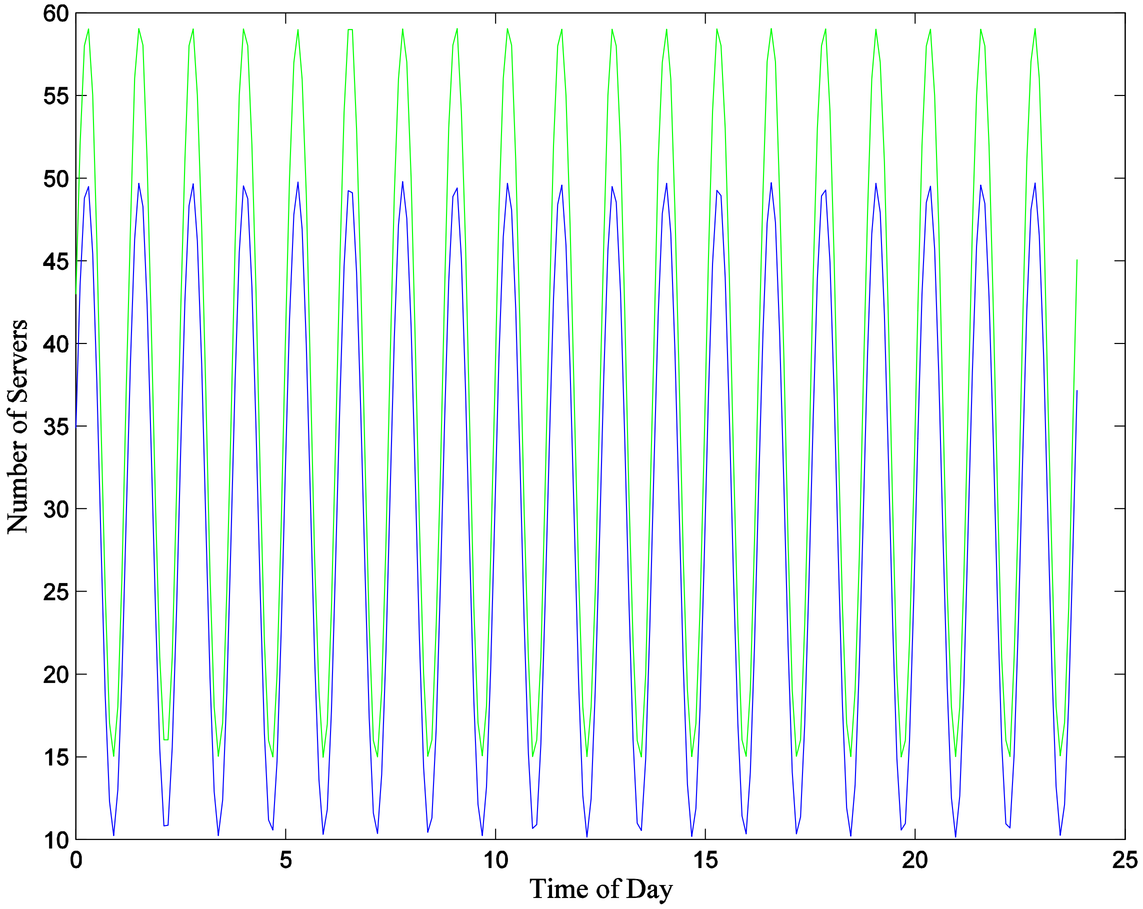

Figure 2-5 Number of servers determined by the PSA method (upper graph) and the offered load (lower graph) versus hours of day

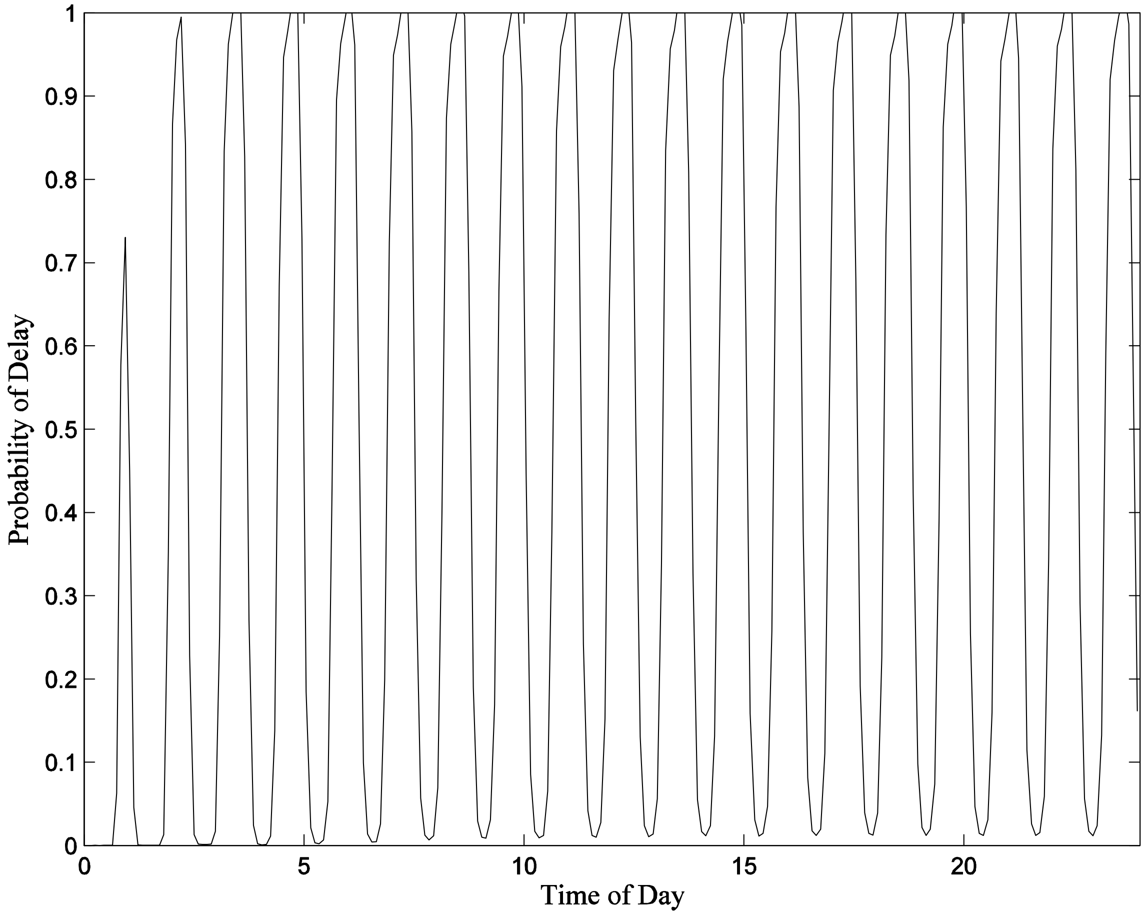

Figure 2-6 Probability of delay with PSA method in hours of day

Staffing with PSA

Let $N(t)$ be the number of customers in the system, either waiting or being served at time t. We focused on the probability of delay aiming to choose time dependent staffing level $s(t)$ such that

$$P(N(t) \geq s(t)) \leq \alpha < P(N(t) \geq s(t)\text{-}1) \text{ for all } t \tag{2-1}$$

where α is the target delay probability. This problem is challenging since the time dependent delay probability $P(N(t) \geq s(t))$ in (2-1) depends on the staffing function before time t as well as at time t.

When we apply the PSA (or use an alternative method, such as modified offered load (MOL) which we discuss later), we replace our initial $M_t/M/s_t$ model with a stationary *M/M/s* model. With PSA at time *t*, we use the limiting steady-state distribution for the model with a fixed arrival rate $\lambda(t)$. The Markovian case *M/M/s* (Erlang-*C* or delay model) is not difficult to analyze, since closed form formulas for computing all the performance measures of the system are available.

2.4.3 Segmented PSA

Some adjustments have been made to the original PSA to reflect the staffing-interval constraint. One of them is the *Segmented PSA,* which works well when the staffing intervals are short. It generates the PSA-required staffing at each time *t* and then sets the staffing to be the maximum of these staffing requirements over the staffing interval. *Segmented PSA* returns an upper bound on the required staffing. Although this approach may slightly overstaff, its results could be used as an initial policy and be evaluated and refined using simulation [18].

2.4.4 Infinite Server

Another approach for approximating transient systems is the infinite server (IS) method. The main idea behind this approach is to assume infinitely many servers are available all the time and then approximate the distribution of the number $N_\infty(t)$ of busy servers at time *t* [22]. The reason to consider the infinite server method is that it is remarkably tractable [23]. Moreover the IS method can be used to show the amount of capacity that would actually be used (and therefore is needed) if there are no capacity constraints (i.e. limited number of servers). The IS approximation is a procedure that produces effective arrivals at all times. It also reveals the role of the service time distribution in this averaging. Given an approximate distribution of $N_\infty(t)$ for each t, we try to choose $s(t)$ so that

$$\begin{aligned} &P(N_\infty(t) \geq s(t)) \leq \alpha \quad \text{and} \\ &P(N_\infty(t) \geq s(t)\text{-}1) > \alpha \quad \text{for all } t, \end{aligned} \tag{2-2}$$

for some prescribed target probability α. Therefore the infinite-server staffing function $s_\infty(t)$ is obtained by applying Eq. (2-1) with $N_\infty(t)$ instead of $N(t)$.This approximation simplifies the problem greatly because (i) the tail probability $P(N_\infty(t) \geq s(t))$ at time t depends on the staffing function $\{s(t)\text{: } t \geq 0\}$ only through its value at the single time t and (ii) the exact time-dependent distribution of $N_\infty(t)$ is known.

The first simplification is due to the fact that the distribution of the stochastic process $\{N_\infty(t)\text{: } t \geq 0\}$ is totally independent of the staffing function $\{s(t)\text{: } t \geq 0\}$. When we calculate $P(N_\infty(t) \geq s(t))$, the staffing level $s(t)$ serves just as the argument of the tail-probability function [23].

The second simplification follows from the basic properties of $M_t/G/\infty$ queues. For each t, $N_\infty(t)$ has a Poisson distribution whenever the number in the system at the initial time has a Poisson distribution (Being empty is a degenerated case of a Poisson distribution). That Poisson distribution is fully known by its mean $m_\infty(t)$ which can be expressed in terms of the arrival rate function and the service time c.d.f. G as [24]

$$m_\infty(t) = E[\int_{t-S}^{t} \lambda(u)\, du] = E[\lambda(t - S_e)]E[S] = \int_{-\infty}^{t} [1 - G(t-u)]\lambda(u) du \qquad 2\text{-}3)$$

where S_e is a random variable with stationary-excess (or residual lifetime) c.d.f. associated with the service-time c.d.f. G. A service time in process in an endless succession of service times in equilibrium will have a residual remaining lifetime distributed as S$_e$. in equilibrium, the remaining service times of the patients in service in the stationary *M/G/∞* model, conditioned on that number in service, are iid random variables, each distributed according to *S$_e$*.

$$P(S_e \leq t) \equiv \frac{1}{E[S]} \int_0^t [1 - G(u)] du, \quad t \geq 0 \tag{2-4}$$

with k-th moment

$$E[S_e^k] = \frac{E[S^{k+1}]}{(k+1)E[S]} \tag{2-5}$$

And so $[S_e] = E[S](c_s^2 + 1)/2$, where c_s^2 is the standard coefficient of variation (SCV) of the service time *S*.

After employing the IS method we can apply normal approximation to solve formula (2-2). This will be discussed further under the normal approximation section.

In a $M_t/M/s_t$ system where it is estimated by $M_t/M/\infty$ system, the expected number of servers can be calculated from a simple differential equation as follows [24].

$$E[N_\infty(t)]' = \lambda(t) - \mu E[N_\infty(t)] \tag{2-6}$$

Viewing the number of busy servers as in (2-6) motivates a different choice of distribution to approximate *P$_n$*(*t*) (probability of *n* patients present in the system at time *t*). By applying the Little's law in a stationary *M/M/s* system, we have the expected number of servers to be *λ/μ*. This approximation is used in a method known as the Modified Offered Load (MOL) that we will be explained in the following section.

2.4.5 Modified Offered Load (MOL)

Usually the infinite server model is applied as the first step in a two-step procedure for generating better approximations for the time dependent measures and the required staffing. The second step is MOL, which also assumes that the system is never overloaded. We use stationary finite server *M/G/s* model at each *t* with the "modified time dependent arrival rate.

$$\lambda_{MOL}(t) \equiv \frac{m_\infty(t)}{E[S]} \tag{2-7}$$

where as previously mentioned $m_\infty(t)$ is the infinite server mean.

The MOL approximation uses the stationary distribution for an *M/M/s* system to approximate $P_n(t)$, with the number of busy servers that is obtained from solving (2-6). According to Little's law we have $\lambda/\mu = E[N(t)]$. Solving this equation for λ gives $\lambda = E[N(t)]\mu$. So we approximate the $P_n(t)$ with the stationary distribution for an *M/M/s* system with the arrival rate $E[N(t)]\mu$, service rate μ, and $s(t)$ servers. (Or equivalently, arrival rate $E[N(t)]$, service rate 1, and $s(t)$ servers, since the stationary distribution is insensitive to multiplication of the same constant with both the arrival and service rates). The arrival rate proposed by MOL approximation can be viewed as an "effective arrival rate" $\lambda_{eff}(t)$ [25].

The MOL approximation is expected to work best when utilization is low enough that the solution obtained from Eq. (2-6**)** provides a good approximation to the number of busy servers over time. In cases where $E[N(t)] > s(t)$ the MOL approximation is expected to be poor. When utilization $\lambda(t)/(s(t)\mu)$ exceeds 100%, the MOL approximation may fail since the approximating stationary system is unstable. When this happens, we set the service level to zero [22].

Figure 2-7 and Figure 2-8 depict the result of applying the MOL method to our challenging arrival function $\lambda(t) = 30+20sin(5t)$. As it is clear from the figure, application of the MOL method confines the variation of the number of servers needed to the interval of value [33 42] and keeps the probability of delay variations within [0.05 0.33] which is better than the PSA method.

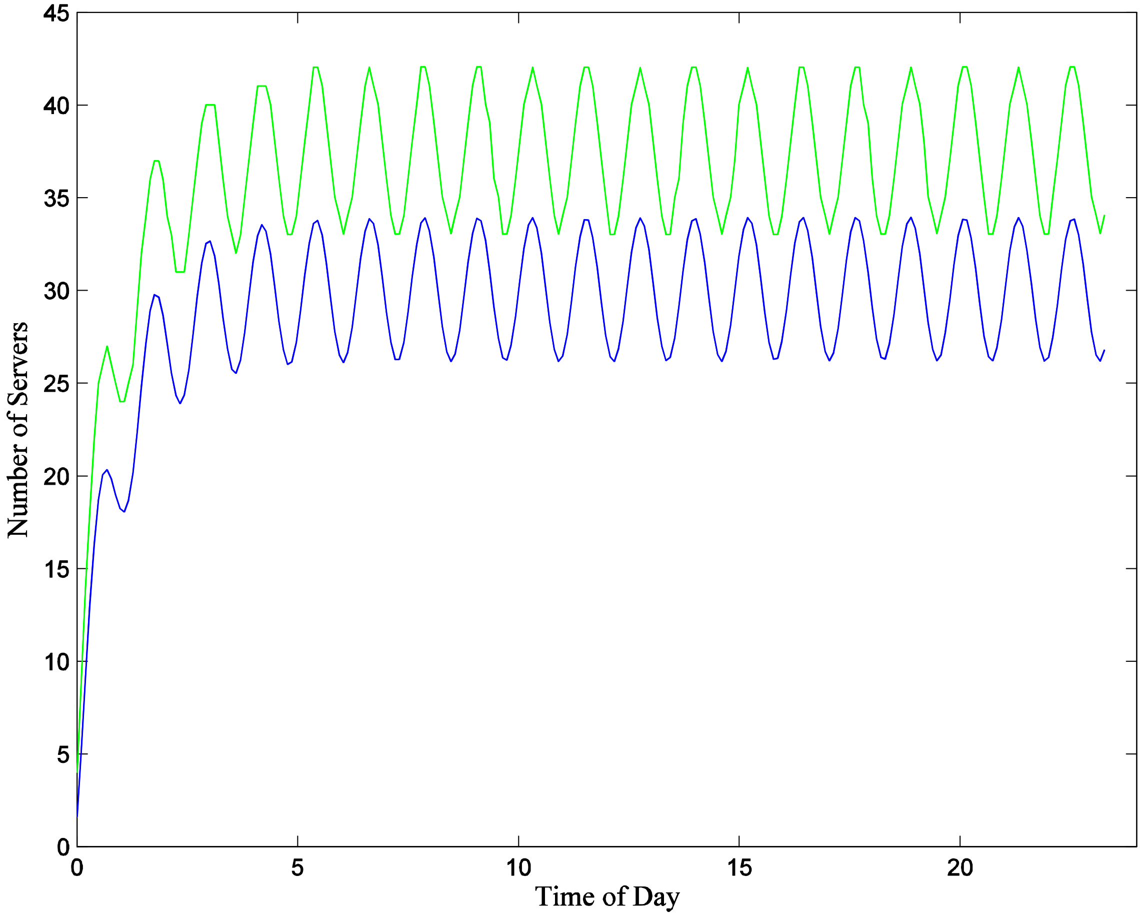

Figure 2-7 Number of servers determined by the MOL method (upper graph) and the offered load (lower graph) versus hours of day

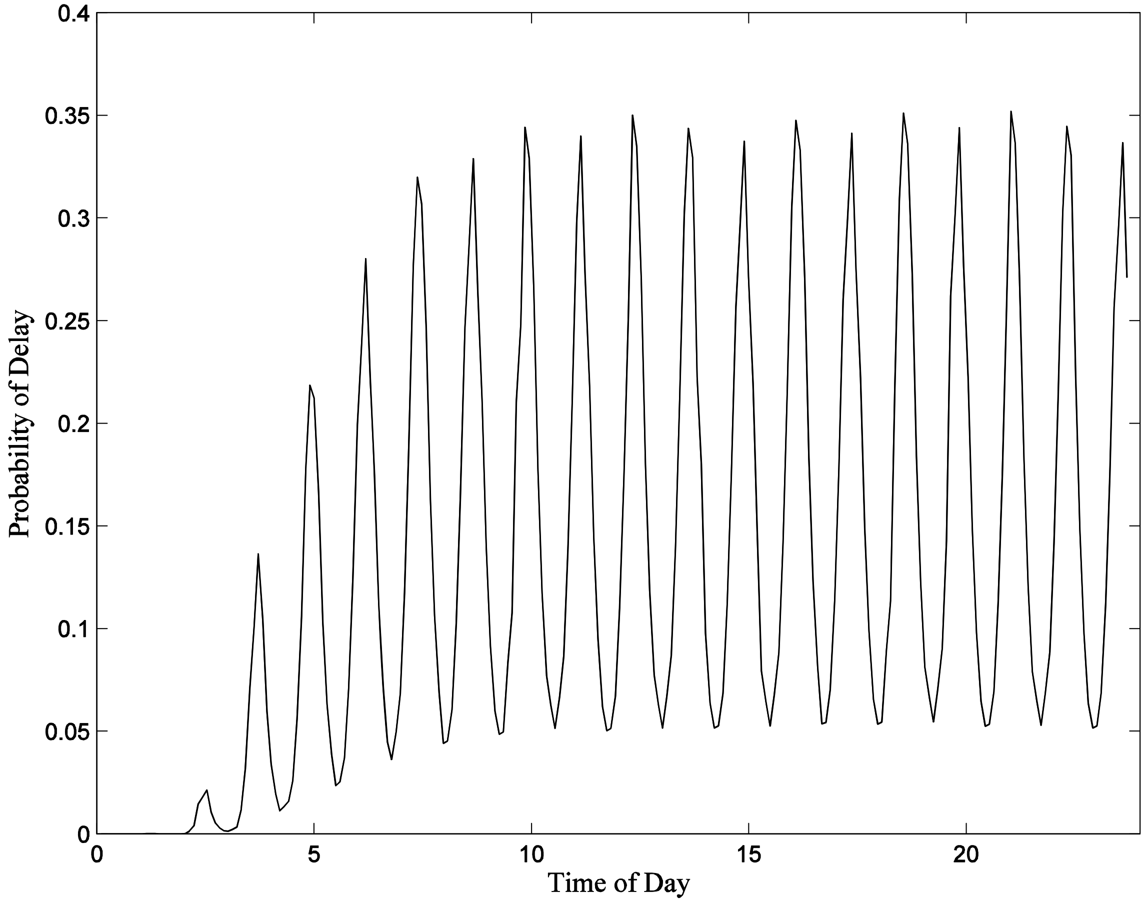

Figure 2-8 Probability of Delay with the MOL method in hours of day

2.4.6 The Normal Approximation

For determining the staffing requirements there are easier ways that do not require calculation of the steady state performance measures in the staffing interval. The fact behind this method is that when the offered load is not too small (say at least five) and the targeted quality of service is high, the number of customers in the system is approximately normally distributed. Deriving the normal approximation, we first approximate the *M/G/s* model by an infinite server *M/G/*∞ model, having the same arrival rate and the same service time distribution. The steady state number of busy servers in *M/G/*∞ model has a Poisson distribution with a mean equal to the offered load $a \equiv \lambda E[S]$, independent of the service time distribution beyond its mean. The Poisson distribution itself can be approximated by normal distribution. Since the actual distribution is Poisson, the variance equals the mean so the offered load is the only parameter in normal distribution [26], [18].

Square root staffing formula

From the normal approximation, the square root staffing formula can be obtained,

$$s(t) = a(t) + \beta\sqrt{a(\mathrm{t})}, \tag{2-8}$$

where $a(t) \equiv \lambda E[S]$, the offered load which is also the mean number of busy servers in the infinite server model, and β is a parameter that reflects the quality of service—in terms of delay congestion— the quality of service (QoS) improves as the β increases. A feasible integer staffing level is the least integer greater than or equal to $s(t)$ in (2-8) [20]. When we divide the time into segments and deal with each segment of time with a stationary queueing model the offered is considered to be constant over that interval and would be calculated using the average of the arrival rate over that time segment.

Utilizing the normal approximation, it is very easy to relate the steady state delay probability, which is denoted by α to the QoS parameter β. Letting Q denote the number of busy servers in the infinite server model, we can approximate the steady state delay probability α by

$$P(Delay) \equiv \alpha \approx P(Q \geq s) = P\left(\frac{Q-a}{\sqrt{a}} \geq \frac{s-a}{\sqrt{a}}\right) \approx 1 - \Phi(\beta), \tag{2-9}$$

where Φ is the c.d.f. of the standard normal distribution.

Although the derivation of the aforementioned methods was focused on models with a large number of servers (i.e., high offered load), they work for any number of servers [20].

Normal approximation refinement

In an actual *M/G/s* model, the steady state number of customers in the system is not exactly normally distributed. Therefore, the normal approximation may need to be refined. An effective way to do so is based on Many Server Heavy Traffic Limits (MSHTL) [27]. The idea is to let $s \rightarrow \infty$ and $\lambda \rightarrow \infty$, leaving the service time unchanged. Halfin and Whitt showed for the *M/G/s* model that the limits λ and s are related in a way that we should let $s \rightarrow \infty$ and $\lambda \rightarrow \infty$ so that

$$\frac{s-a}{\sqrt{a}} \rightarrow \beta \qquad \text{where } a \equiv \frac{\lambda}{\mu} \tag{2-10}$$

In that limit, the steady state delay probability $\alpha \equiv \alpha(\lambda,\mu,s)$ in the *M/G/s* model approaches a limit strictly between 0 and 1. This confirms that the delay probability is a good performance measure, because it tends to have meaning independent of scale. That is not true for most of the other performance measures. For instance the mean waiting time is asymptotically of order $1/\sqrt{s}$ in the limiting regime (2-10**)**.

From (2-10), we see that the MSHTL also produces a square root staffing law, which coincides with (2-8).

As a consequence of the MSHTL for the *M/G/s* model, there is a continuous increasing function mapping the QoS parameter into the limiting delay probability α, now commonly called the *Halfin-Whitt delay function*

$$P(Delay) \equiv \alpha \approx HW(\beta) \equiv [1 + \left(\frac{\beta\Phi(\beta)}{\varphi(\beta)}\right)]^{-1}, \qquad 0 < \beta < \infty, \tag{2-11}$$

where, Φ is the c.d.f. and φ is the associated probability density function (pdf) of the standard normal distribution [27].

2.4.7 Stationary Independent Period-by-Period (SIPP) Method

A very common method to maintain the staffing requirement is through using an approach that is called stationary independent period-by-period (SIPP). In this approach, first the workday or workweek is divided into "planning periods" such as shifts, hours, quarter-hours, etc. Then a series of stationary queueing models, usually *M/M/s* models, are constructed for each planning period. After that each of these models are independently solved to determine the minimum number of servers required to satisfy the target service level in that period. The staffing requirement generated by this method could be used to set the actual staffing schedules or could become the right-hand sides of the key constraints in a large optimization model that results in the actual workforce schedules. SIPP is appropriate for the classic case with short service times and short staffing intervals provided that the arrival rate function does not fluctuate too greatly over staffing intervals [28]. SIPP is also used in some of the commercial software packages developed for call centre management.

Although SIPP is widely used, it is based on some assumptions that make it not applicable in some cases. These assumptions are (1) delays in consecutive planning periods are independent of one another, (2) within each planning period the system achieves steady state; and (3) the arrival rate does not change during the planning period.

The common idea behind the segmented-PSA and SIPP approaches is that they both use a stationary independent period-by-period approach. Nonetheless, segmented PSA first specifies the staffing level at each time point, while SIPP first averages the arrival rate over the staffing interval. If the arrival-rate function does not greatly fluctuate within individual staffing intervals, these methods attain similar results [21].

2.4.8 Busy-Hour Engineering and the Simple Peak Hour Approximation (SPHA)

Another classic case is when the service times are short and the quality of service standard is high, but the staffing interval is long (e.g. 8 hours or even an entire day). One approach is to reduce the problem to one with stationary demand and determine the staffing requirement so that the satisfactory performance is obtained at all times. This is done by setting the staffing requirement to meet the peak demand during the long staffing interval. This method is also referred to as MaxSIPP in some of the papers.

In some cases, the managers may staff to meet the average performance instead of peak performance over the long staffing interval. However, this is risky, because it leads to understaffing at peak times [18].

2.5 Summary

To summarize, in this chapter we discussed the related literature about nonstationary queueing systems with time varying arrival rates. We introduced a number of approximation methods that have been developed to set staffing requirement according to these variations. We discussed that appropriate stationary models can provide effective solutions to the surge staffing problems with varying arrival rates. Therefore, to determine the staffing requirement for the $M_t/G/s_t$ model, it is sufficient to consider the staffing problem for the stationary *M/G/s* model with a specific arrival rate over a certain interval. We realized the performance of these methods is dependent of various parameters of the queueing systems such as the frequency and amplitude of changes in the arrival rate, the service rate etc. In the next chapter, we would like to develop a baseline model on which we could compare the performance of these approximation methods.

3: QUEUEING MODEL IMPLEMENTATION AND VALIDATION

In the previous chapter, we discussed different methods that are used to approximate nonstationary queueing models. However, it is not clear which one of these methods would perform best for the given surge situation. In this chapter, we build a queueing model that captures the variation in arrival rates, simulating a surge situation, and set the staffing requirement according to these variations so that a target level of service is maintained.

For evaluating the performance of such a model, an exact numerical solution was found in literature. This method, however, is limited to certain special conditions and is not applicable for many other queueing models. Therefore in this work, a simulation model with time varying arrival rate was built as a baseline, in order to compare and evaluate different queueing models for modelling the surge situation. This simulation model is validated with exact numerical solutions for 162 different model variants with a wide range of conditions. The validation not only gives us a credible simulation model, it also sheds some lights on the type and condition of queueing models.

3.1 Setting staffing requirements for the emergency department with time varying arrivals

As mentioned before, most of the queueing models assume that the arrival process is Poisson with a rate that remains constant all the time. However, service systems usually experience cyclical demands with various cycle lengths. As we mentioned in the previous chapter, different approaches have been developed to cope with the time-varying demand and setting staffing requirements to satisfy the demand. Some of these approximation methods assume the rate to be constant for blocks of time (e.g. two-hours), with a separate queueing model fitted for each block of time. This is also called the stationary independent period by period (SIPP) method. One way of capturing the changes in the arrival rate is through utilizing non-homogeneous Poisson process.

We perform the statistical tests provided in [29] on the data from the ED at the time of surge and we could validate the assumption of the ED arrival rate following non-homogeneous Poisson process. This observation motivated us to apply a couple of approximation methods to a wide range of models with periodic non-homogeneous Poisson arrivals and evaluate the performance of these methods.

In order to be able to evaluate the performance of approximation methods we need to have a baseline model that does not involve so much approximation as these methods do. Ingolfsson et al. used the analytical solution to the differential equations associated with the queueing model with time varying arrival rate ($M_t/M/s$) as a baseline for comparing other approximation methods [22]. Although this approach is a fast and convenient way of conducting this comparison, it has some shortcomings. This method, which is also referred to as “exact” method, involves approximations regarding to solving an infinite set of differential equations. Moreover, it can only be applied to the $M_t/M/s$ model. As soon as the model gets a little more complicated (e.g. the service time distribution does not follow an exponential distribution), it would become analytically intractable. Hence, the exact method would not be applicable.

3.2 The “Exact” method for time varying arrivals

In this paper, we build a simulation model for the Markovian $M_t/M/s$ system for which we have the analytical solution. The solution to this queueing model is analytically calculated through solving a set of differential equations. We would like to evaluate with simulation the time-varying probability of delay resulting from a given staffing schedule. The staffing schedule could be selected manually or by calling a staffing subroutine. Here, we choose s according to the square root formula,

$$s = a + \beta\sqrt{a}, \tag{3-1}$$

where $a \equiv \lambda E[S]$, the offered load which is also the mean number of busy servers; in the infinite server model; λ is the arrival rate; $E[S]$ is the mean service time; and β is a parameter that reflects the quality of service—in terms of delay congestion.

To obtain the time dependent probability of delay, we calculate the time dependent distribution of the number of customers present in the system, by numerically solving the Chapman-Kolmogorov forward equations (a system of ordinary differential equations), as described below, using a variant of Runge-Kutta algorithm.

$$\begin{aligned} p_0'(t) &= -\lambda(t)p_0(t) + \mu p_1(t), \\ p'_n(t) &= \lambda(t)p_{n-1}(t) + (n+1)\mu p_{n+1}(t) - (\lambda(t) + n\mu)p_n(t), \qquad 1 \le n < s, \\ p'_n(t) &= \lambda(t)p_{n-1}(t) + s\mu p_{n+1}(t) - (\lambda(t) + s(t)\mu)p_n(t), \qquad n \ge s. \end{aligned} \tag{3-2}$$

where $\lambda(t)$ is the arrival rate at time t, μ is the service rate, and $P_n(t)$ is the probability of n customers in the system at time t. This involves approximation of the infinite set of forward equations with the first K+1 equations. We choose the finite capacity K (the maximum number of patients in the system) sufficiently large so that $P_K(t)$ is negligibly small (less than some user specified value, e.g., $\varepsilon = 10^{-6}$).

The numerical integration is performed using forth- and fifth- order Runge-Kutta methods recursively. The length of the recursion interval is determined so that a user specified error is not exceeded. The integrations are initialized using the steady-state *M/M/s* solution for $\lambda(0)$ as the stationary arrival rate. We could also start the system from empty; it will again end up giving us the same results. We have utilized the *ode45* Runge-Kutta ODE solver from the Matlab ODE suite [30].

Since we have time varying arrivals, we should be very careful in definition and estimation of the performance measures. The measures should also be time-varying and should be defined for each time interval t, $t \in [0, T]$.

For computing the probability of delay, we use the formula in Eq. (3-3) that averages the instantaneous measures provided at PP*60 time segments of one minute (time step=1) gird (PP is the length of the planning period over which we are calculating the performance measure) [31]. Assuming that segment 1 begins at midnight and segments are numbered consecutively, let λ_i be the average arrival rate at the start of the segment i, so that $\overline{\lambda} = \sum_{1}^{PP*60} \lambda_i/(PP * 60)$. Let p_{ni} be the probability that n customers are in the system at the start of segment i. Then the average probability of delay in each interval is computed as

$$P_d = \sum_{i=1}^{PP*60} \lambda_i(1 - \sum_{n=0}^{s-1} p_{ni})/(PP * 60 * \bar{\lambda}) \\ =\sum_{i=1}^{PP*60} \lambda_i p_{di}\,/(PP * 60 * \bar{\lambda}) \qquad (3\text{-}3)$$

where p_{di} is the probability that all the servers are busy at the start of segment i.

3.3 Simulation approach

It is important to pay attention to the fact that a physical phenomenon, a mathematical model of that physical phenomenon, and a simulation of that mathematical model are three different things. A mathematical model, whether simulated or analyzed, may provide useful and accurate information about the physical phenomenon [32]. We selected the mathematical queueing model because of its ability to explain queueing phenomena. Here, we also developed a stochastic simulation model to reveal statistical regularity. We expect the simulation model to capture key features of the queueing model, but do not expect a perfect fit with it. In this case, we first choose a standard queueing model with "exact" solutions as our reference and use it to validate our simulation model. Later, when queueing models involve non-exponential service rates and no analytical solution is available, the validated simulation model can be used as a reference to compare different approximation methods that are based on the queueing theory. The literature on building and validating simulation models for nonstationary queueing models is scarce.

We built the simulation model in Arena^{TM}. It is shown in the conceptual model represented in Figure 3-1 that the simulation model consists of different modules. Since the arrival rate is not constant, we developed a separate module for advancing the time.

We set the time step by which the simulation time is advanced to be close to the time steps used in queueing model for evolving the vector of probabilities through the Chapman-Kolmogorov forward equations. We recorded the time in a variable and used it as the argument for the inter-arrival function.

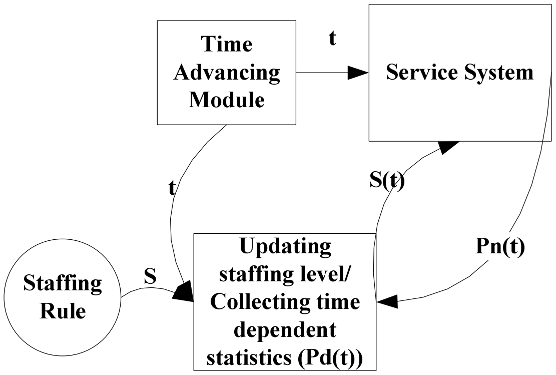
Time Advancing Module
t
Service System
t
S(t)
Staffing Rule
S
Updating staffing level/ Collecting time dependent statistics (Pd(t))
Pn(t)

Figure 3-1 Conceptual representation of simulation model

The inter-arrival time is changing according to an exponential distribution function with a time varying parameter. We used a sinusoidal function with a 24-hour cycle,

$$\lambda(t) = \bar{\lambda}(1 + RA\sin(2\pi t/24)) \tag{3-4}$$

where $\bar{\lambda}$ is the average arrival rate and $RA \in [0, 1]$ is the relative amplitude. The sinusoidal form is used for the purpose of simplicity; the method applies to general arrival rate functions estimated from data.

Another module is designed for collecting time dependent statistics, such as the probability of delay. This module is also used for changing the staffing level while the model is running. In this module at the end of the pre-specified planning period the staffing level is updated according to the staffing method that we had used. Then we calculate the probability of delay according to Eq. (3-5), which is to be explained later. This measure is recorded in a variable and compared with the time-dependent probability of delay calculated from the "exact" method.

At the end of each planning period when the number of servers changes, if the number is decreasing we would return the patient that is currently receiving service back to the queue. This is basically the underlying assumption in the queueing model as well.

3.4 Validation of the simulation model

The number of servers over time ideally is supposed to match the variation in demand. However, this match might not be perfect because of various circumstances, including (1) cost savings from reducing variability in staffing, (2) an upper limit on staffing, (3) limitations on when servers can begin and end work, and (4) lack of planning. We do not need, for the purpose of validation, the number of servers to be in perfect match with the arrival. Besides, such situations do not occur in reality very often. In addition, a scheduling algorithm that may call this method as a subroutine might need to alter the number of servers to improve a poor schedule [22].

Most of the papers that have discussed the validation of either the queueing models or simulation with each other have mostly considered the steady state queueing models, in which case the simulation is usually run for one long time. The length of run is chosen so that the system reaches the steady state condition. The warm up period is determined to leave out the statistics associated with the transient phase of the beginning of the simulation and the effects of the system initial state [33]. This approach usually works well for stationary models. Another approach for validating simulation, which is expected to be more effective for nonstationary queueing models, is based on periodic steady state condition.

In this chapter, we implemented both steady state and periodic steady state conditions and realized, for the same simulation run time both approaches would give the same result, given the cycle length of the arrival function is not too large (e.g. less than 24 hours). In our tests, we examined both approaches with a couple of arrival functions that would solely differ in the length of the cycle time and compared the probability of delay predicted by either one of the simulation approaches with the exact analytical solution. The result of this experiment is presented in Table 3-1. In this table, the first column is the probability of delay resulted from applying the exact method, and the two other columns represent, periodic steady state simulation, and stationary simulation respectively. The last two columns show the discrepancies (error percentage) of each of these simulation models with the exact method. It can be seen that in the cases where the length of period is large the PSSS approach yields very close results to the exact method, while the SS yields poor results.

We used the PASTA (Poisson Arrivals See Time Averages) property [34] to calculate the performance measure in the stationary simulation case by dividing the number of patients that have not been served immediately in a planning period by the total number of patients arrived in that period.

A slightly different approach is taken to calculate performance measures from simulation models in the periodic steady state case. According to Heyman and Whitt [35], the periodic $M_t/G/s$ reaches periodic steady state meaning that if we let $N(t)$ be the number of customers in the system at time t, there exist T>0 such that $\{N(nT+t)=N(t), n\geq 0\}$. In this approach, the model is run in shorter length for numerous times. For the periodic steady state case, for replication k, the delay probability in interval t is estimated by the fraction of patients who are not served immediately upon arrival, out of all arriving patients during the t time interval. Namely, for the *k*-th replication, the estimator is

$$\hat{\alpha}_k(t) = \frac{\sum_i \ 1\{customer\ i\ entered\ queue\ at\ interval\ t\}}{\sum_i \ 1\{customer\ i\ entered\ system\ at\ interval\ t\}} \equiv \frac{\hat{Q}_k(t)}{\hat{S}_k(t)} \tag{3-5}$$

We obtain the overall estimator $\hat{\alpha}(t)$ by averaging over all replications. This also happened to be the same as the ratio of the average of the $\hat{Q}_k(t)$ over all replications to the average of $\hat{S}_k(t)$ [23].

The process of estimating the time dependent delay probability for any given staffing function by computer simulation is subject to sampling error. This statistical sampling error decreases as we increase the number of independent replications. Therefore the error can be reduced to a certain degree at the expense of computational effort. But it will always be present for any amount of computational effort [23].

We ran the periodic simulation for various numbers of replications and realized that increasing the number of replications beyond 100 would not have a great impact on the average values of probability of delay estimated from it.

Therefore, we ran the periodic simulation model for 100 replications with the length of 8640 minutes (6 days). For the other simulation model to have the same run time, we considered one single replication of length 600 days (864000 minutes). We considered the first two days as the warm up period for the former case and 200 days (288000 minutes) as the warm up period for the latter.

Generally, proving that the model produces the same results as the original system under all circumstances require a substantially large amount of resources. Therefore, the validation exercise is confined to a limited number of scenarios, which would logically cover all the important cases, thereby would increase the degree of confidence in the model results [36].

Table 3-1 Comparing Stationary Simulation (SS) and Periodic Steady State Simulation (PSSS) for arrival rates with various cycle lengths

$\lambda(t)$	s	Pd(EX)	Pd(PSSS)	Pd(SS)	Disc (PSSS-EX)	Disc (SS-EX)
30+20sin (0.0001t)	8	0.167401	0.150693	0.611429	9.98082449	265.248117
	8	0.167663	0.150653	0.584786	10.1453511	248.786554
	8	0.167929	0.168231	0.594599	0.17983791	254.077616
	8	0.168223	0.177384	0.590023	5.44574761	250.738603
30+20sin (0.01t)	8	0.18128	0.182803	0.343326	0.83982554	89.3894853
	9	0.108605	0.105442	0.262348	2.91230218	141.561299
	9	0.128786	0.123911	0.220673	3.78478354	71.349436
	9	0.152535	0.156343	0.232829	2.49641718	52.6401882
30+20sin (0.1t)	10	0.105038	0.090473	0.125596	13.8664075	19.5718759
	12	0.103012	0.099805	0.035745	3.11357696	65.2996893
	12	0.16846	0.17143	0.035949	1.76260406	78.660422
	12	0.130202	0.140319	0.041687	7.7701222	67.9825659
30+20sin (t)	8	0.495064	0.495253	0.456162	0.03808903	7.85801609
	8	0.488087	0.458439	0.439468	6.07421855	9.96103191
	8	0.485109	0.446523	0.450723	7.95394254	7.08830297
	9	0.326246	0.319829	0.333187	1.96699189	2.12763108
30+20sin (5t)	8	0.327951	0.308794	0.289168	5.84158617	11.8260323
	9	0.200659	0.184616	0.175674	7.99476728	12.4512498
	8	0.296494	0.283732	0.29605	4.30441532	0.1499415
	8	0.305361	0.310816	0.288716	1.78640991	5.45066904

We have also applied the queueing model and simulation to a set of test problems with a wide range of conditions. Table 3-2 shows the parameters we varied and their values. The combinations of these values resulted in 162 different test problems. The number of servers was determined according to the square root formula for each planning period. We calculated probability of delay at one minute intervals, starting at time zero and ending at time 24, and we employed Eq. (3-3) for each planning period.

Table 3-2 Parameter values for computational examples

Factor	Low value	High value
Service rate (μ)	2	8
Offered load ($r=\bar{\lambda}/\mu$)	2	8
Arrival rate relative amplitude (RA)	0.1	1
Planning period (PP)	2	6
Service level target (α)	0.1	0.5
Total combinations	**162**	

Model validation is normally performed by using statistical techniques to compare the model output data with the corresponding simulation output when the simulation model is run with the “same” input parameters. Performing a t-test on the output (probability of delay in each planning period) of the queueing model and simulation with 100 replications, we could not reject the null hypothesis of the two sets of results having the same means (H_0). The t-test has been done separately for each planning period (PP) over data provided with varying system parameters.

Table 3-3 presents the t-test 95% confidence intervals (CI) and the p-values for each planning period when the length of the planning period is six hours.

Table 3-3 Outcome of applying t-test to the model with six-hour long planning period

PP	**1**	**2**	**3**	**4**
CI	[-0.0817 0.0922]	[-0.1187 0.0659]	[-0.1371 0.0910]	[-0.1347 0.0929]
H_0	Not rejected	Not rejected	Not rejected	Not rejected

Table 3-4 and Table 3-5present the 95% confidence intervals and p-values related to the queueing model with four and two hours of length respectively.

Table 3-4 Outcome of applying t-test to the model with four-hour long planning period

PP	1	2	3	4	5	6
CI	[-0.0746 0.1085	[-0.1092 0.0926	[-0.1108 0.0877	[-0.1283 0.1039	[-0.1186 0.0929	[-0.1320 0.1011
H_0	Not rejected	Not rejected	Not rejected	Not rejected	Not rejected	Not rejected

Table 3-5 Outcome of applying t-test to the model with two-hour long planning period

PP	1	2	3	4	5	6	7	8	9	10	11	12
CI	[- 0.067 0.115]	[-0.096 0.091]	[-0.109 0.089]	[-0.110 0.092]	[-0.113 0.093]	[-0.106 0.085]	[-0.127 0.102]	[-0.13 0.101]	[-0.122 0.093]	[-0.121 0.089]	[-0.127 0.093]	[-0.137 0.098]
H_0	Not reject	Not reject	Not reject	Not reject	Not reject	Not reject	Not reject	Not reject	Not reject	Not reject	Not reject	Not reject

From the results presented in the tables above, we can conclude that the simulation model is predicting the same results as the queueing model and therefore is validated.

As we discussed the simulation model is developed to replace the exact method for the cases where exact method is not applicable. Statistically analyzing the results from planned experiments performed on the queueing models with a wide range of parameters assured us that the simulation model generates the statistically equivalent results as the exact method. Now this model is ready to be used as a baseline for comparison for other queueing models. It is also found that for queueing/simulation models with time varying arrival rate, periodic steady state condition, rather than the steady state condition, is to be used to accurately reflect the situation. In the next step we would use the validated simulation model for evaluating the performance of a number of approximation methods applied to the $M_t/G/s_t$ and would determine the conditions under which either of these methods would be most suitable.

4: COMPARISON OF THE QUEUEING APPROXIMATION METHODS

One of the major problems in modeling is lack of data. Although the arrival data can be collected through patient registration process, there are no systems for collecting the “service” times, which include taking patient histories, performing physical examination, ordering and reading test results, consulting with other physicians, administering treatments and writing up reports. Moreover, these service times are very difficult to capture since they are usually discontinuous because physicians make up their treatment or discharge decisions based on the results of the tests that they have to order; and this procedure often takes time. While the tests are conducted physicians would handle other patients [18].

According to [29], service time distributions for call centers tend to be non-exponential most of the time. It has been observed that the service time distributions can be better approximated with lognormal distribution; however the variability of the lognormal distribution in this case would not be too great. Particularly, the squared coefficient of variation (SCV, variance divided by the square of the mean) would be between 1 and 2. (SCV is independent of the mean, meaning it will not change if the random variable is multiplied by a constant, therefore it is a better measure to use than variance).

Since there is not enough data to approximate a physician’s service time distribution in the ED properly, it is common to make a simplifying assumption and consider it to be exponentially distributed. Although the previous discussion was about the call centers, it might be useful if we remove this simplified assumption for the ED service times, and instead use the lognormal distribution to better reflect the reality.

In this chapter, we would study models with the lognormal service time distribution with various SCV values. We would utilize the simulation model validated in the previous chapter for comparing the performance of some of the approximation methods introduced previously. We will run 324 test problems with different parameters for each approximation method, using both the $M_t/G/s_t$ queueing model and the simulation model. Then compare the time varying average error and maximum error and rank the approximation methods according to these measures.

4.1 Experimental design

We would apply a couple of approximation methods introduced in chapter two to an $M_t/G/s_t$ model, for which there is no numerical solution available, and compare the performance of them by evaluating the probability of delay calculated by each method.

One approach to simplify the problem of dealing with complex nonstationary queueing models is to approximate these models with a stationary model that uses the average arrival rate over a planning period and determines the minimum number of servers needed in the planning period to provide a specified level of service. However, for the case of $M_t/G/s_t$ model this approach would not help the problem of intractability very much. Since the *M/G/s* model also tends to be intractable. Mathematical formulation of this model does not provide analytical insight and it is not suitable for numerical computation either [37]. Therefore, for calculating the probability of delay for this model we have to use approximations.

4.2 Approximating probability of delay in *M/G/s* model

We consider the standard *M/G/s* queueing system with $s(\geq 1)$ homogeneous servers in parallel, unlimited waiting room and with the first-come first-served rule. Customers arrive according to a Poisson process with a rate λ (>0). Their service times are i.i.d. with a c.d.f. G having finite $1/\mu$ and independent of the arrival process. For simplicity we assume $G(0)=0$, but this assumption is not essential. Let c^2 be the SCV of G and let $\rho=\lambda/s\mu$ be the traffic intensity. In addition, let $\Pi(M/G/s)$ denote the delay probability in the *M/G/s* model, assuming that the system is stable and in steady state; i.e. $\rho<1$.

It is been known that the Erlang delay formula, as in Eq. (4-1), which is for *M/M/s* queue is a good approximation for *M/G/s* queue

$$\Pi(M/M/s) = \frac{(s\rho)^s}{s!\,(1-\rho)}\left[\sum_{j=0}^{s-1}\frac{(s\rho)^j}{j!}+\frac{(s\rho)^s}{s!\,(1-\rho)}\right]^{-1}, \tag{4-1}$$

Mathematically

$$\Pi(M/G/s) \simeq \Pi(M/M/s) \tag{4-2}$$

There has been a number of other approximations proposed in literature (i.e. in [38] and in [39]) that have improved the Erlang delay approximation in (4-1). For large systems, Eq. (4-2) can be justified by the insensitive property of the *M/G/∞* queue. For any distribution we have

$$\lim_{s\to\infty} \frac{\Pi(M/G/s)}{\Pi(M/M/s)} = 1 \tag{4-3}$$

Considering this property, we are looking for an approximation for a relatively small s (e.g., $s \leq 10$) with the form

$$\Pi(M/G/s) \simeq k_G \Pi(M/M/s) \tag{4-4}$$

where $k_G \equiv k_G(s,\rho)$ denotes a correction factor. The approximation method that we employed is based on estimating the probability of delay with the mean waiting time in *M/G/s* queue and is defined ad below [37].

$$\Pi(M/G/s) \simeq \left\{(1-\rho)sI_G(s) + \rho\frac{1+c^2}{2}\right\}^{-1} R_G \Pi(M/M/s) \tag{4-5}$$

where

$$I_G(s) = \int_0^\infty \{1 - G_e(t)\}^s dt, \qquad s \geq 1. \tag{4-6}$$

$$G_e(t) = \mu \int_0^t \{1 - G(u)\} du, \qquad t \geq 0, \tag{4-7}$$

where G_e is the stationary-excess c.d.f. associated with the service time c.d.f. G i.e., and

$$R_G \equiv \frac{R_D(1+c^2)}{(2R_D - 1)J_G(s) + 1} \tag{4-8}$$

with

$$R_D \equiv \frac{1}{2}\{1 + f(s)g(\rho)h(s,\rho)\}, \tag{4-9}$$

In which

$$f(s) = \frac{(s-1)\sqrt{4+5s}-2}{16s}, \tag{4-10}$$

$$g(\rho) = \frac{1-\rho}{\rho}, \tag{4-11}$$

And the bivariate function *h*(*s*,*ρ*) is given by

$$h(s,\rho) = \xi\big(s, a(\rho)\big)\eta(b(s),\rho) \tag{4-12}$$

with

$$\xi(s,x) = \sqrt{1-\exp\left(\frac{-2x}{s-1}\right)}, \qquad x \geq 0, \tag{4-13}$$

$$\eta(y,\rho) = 1-\exp\left(\frac{-\rho y}{1-\rho}\right), \qquad y \geq 0. \tag{4-14}$$

The functions *a*(*ρ*) and *b*(*s*) are defined by

$$a(\rho) = \frac{25.6}{\{g(\rho)\eta(\beta,\rho)\}^2} \tag{4-15}$$

$$b(s) = \frac{s-1}{(s+1)f(s)\xi(s,\alpha)} \tag{4-16}$$

Respectively, where α and β are arbitrary positive constants satisfying the relation

$$\alpha\beta^2 = 25.6 \tag{4-17}$$

It is been shown in literature by numerical experiments that *α*=2.2 is an optimal value for the best performance of the approximation [37].

Eq. (4-5) is the approximation that we would use in this chapter for calculating the probability of delay in the *M/G/s* model. We would calculate the probability of delay at the end of each planning period, and compare it with the probability of delay estimated from the simulation model that is run with the same parameters.

4.3 Test and Comparison Results

Similar to the validation procedure we applied the approximation methods to a set of 324 test problems with a wide range of conditions. Table 4-1 shows the parameters that we vary and their values. We calculated errors by comparing the results with the outcome of the related simulation models, and evaluate the errors of different approximation methods according to Eq. (5-2). We calculate the time averages and maxima of the errors for each test problem and each method.

$$\frac{P_d(SIM) - P_d(APPROX)}{P_d(SIM)} \tag{4-18}$$

Table 4-1 Parameter values for computational examples

Factor	Value
Service rate (μ)	2, 4, 8
Offered load (r=$\bar{\lambda}/$μ)	2, 4, 8
Arrival rate relative amplitude (RA)	0.1, 0.5
Planning period (PP)	2, 4, 6
Service level target (α)	0.1, 0.2, 0.5
SCV	0.5, 1, 2, 4
Total combinations	**324**

Table 4-2 presents the mean and median for the time average relative and maximum relative errors calculated over all the test problems. We summarized the results for each method with different SCV values to analyse how the error would be affected by increase in the variation in service time distribution. As we expected, the error would grow when increasing the variability of service time distribution, especially when SCV jumps to four, the change in error is much more visible. Among these approximation methods, MaxSIPP shows less sensitivity to change in SCV, probably because it mitigates the effect of the variations by offering a higher (upper bound for) staff level. Computing the staff-hours, that each of these methods propose, we realize that MOL offers the least number of staff-hours. SIPP is in the second position with about an average of 0.4 staff-hours more than MOL and MaxSIPP ranks third with an average of 14.2 more staff-hours.

Analysing Table 4-2 we realize that MOL outperforms MaxSIPP and SIPP considering both lower staff-hours and higher accuracy. Nonetheless, MaxSIPP and MOL compete very closely from the accuracy aspect, especially when the planning period is long (more than two hours). We can see from the table that for the smaller planning period, the MaxSIPP performances much better than MOL.

Table 4-2 Time average relative and maximum relative error by SCV and planning period (PP) length

			SIPP		MOL		MaxSIPP	
SCV	**PP(hrs)**	**Mean/ Median**	**Max Error%**	**Time Ave Error%**	**Max Error%**	**Time Ave Error%**	**Max Error%**	**Time Ave Error%**
0.5	2	mean	46.623	22.34606	46.1386	21.82092	16.047	8.038
		median	36.4622	18.07707	39.8848	17.53242	13.165	6.105
	4	mean	38.7378	21.08358	13.8593	7.530134	16.345	9.79
		median	33.2374	19.46694	11.3894	5.506687	12.297	7.261
	6	mean	29.742	17.12368	11.4679	6.052025	17.012	12.6
		median	26.3825	15.74273	8.7812	5.232265	12.247	8.194
1	2	mean	47.5699	23.22282	47.5015	22.79289	16.243	8.313
		median	38.8264	19.1748	38.5256	18.76471	15.087	7.299
	4	mean	42.6732	22.56089	14.9166	7.826452	16.21	9.886
		median	34.7452	19.12795	12.4852	6.481024	14.831	8.12
	6	mean	32.1418	18.70206	12.0299	6.883359	16.132	12.199
		median	26.9269	16.26083	9.3777	5.919824	12.36	8.813
2	2	mean	47.6016	25.6004	45.8793	24.27017	17.561	8.711
		median	46.1436	24.10426	44.9372	21.43283	16.76	7.993
	4	mean	41.6944	25.39723	18.422	9.18414	16.674	9.226
		median	38.9898	23.18064	16.9209	9.058691	14.54	7.146
	6	mean	36.1125	22.25201	14.5588	8.545564	14.162	10.499
		median	36.7203	22.28914	11.7577	7.381624	8.939	5.465
4	2	mean	69.5943	51.94616	70.1395	51.54401	29.576	15.997
		median	73.1412	58.63171	73.1997	59.45892	27.418	14.067
	4	mean	67.6459	52.77069	32.1167	20.03433	27.623	14.609
		median	71.3281	59.36151	29.3437	19.97489	25.999	12.436
	6	mean	62.7614	52.6112	29.507	19.82894	18.672	13.273
		median	70.079	61.45038	28.6813	19.79942	13.544	9.592

Next, we would like to analyse the result of the experiments from another perspective. Table 4-3 shows the time average relative and maximum relative error by the target service level and the planning period length.

Table 4-3 Time average relative and maximum relative error by target service level (α) and planning period (PP) length

			SIPP		MOL		MaxSIPP	
α	PP(hrs)	Mean/ Median	Max Error%	Time Ave Error%	Max Error%	Time Ave Error%	Max Error%	Time Ave Error%
0.1	2	mean	66.9343	40.2998	66.1669	39.90555	12.789	6.363
		median	64.0608	35.6181	65.8937	35.82605	10.293	5.045
	4	mean	60.0345	38.9558	13.7594	7.3497	11.676	5.733
		median	57.2515	36.6889	10.8433	5.8049	9.786	4.848
	6	mean	51.1906	35.4289	13.4921	7.5106	7.722	5.185
		median	48.8189	28.2542	11.2088	6.1608	7.126	4.848
0.2	2	mean	54.8198	32.69206	53.287	32.2024	17.119	9.019
		median	53.1887	26.54291	52.1208	26.8106	15.573	7.993
	4	mean	48.0637	31.9792	17.8109	10.6893	15.43	8.3
		median	43.0697	25.1673	14.874	8.9904	14.069	7.327
	6	mean	45.9172	31.7473	17.0711	10.6827	11.864	8.046
		median	45.6423	25.9827	15.6936	8.2632	10.839	7.093
0.5	2	mean	37.2838	19.4536	38.3281	19.4419	30.117	15.797
		median	36.4353	18.507	37.0949	18.2435	27.92	13.796
	4	mean	34.9062	20.1646	28.031	15.4068	30.679	18.59
		median	33.444	19.0664	24.7844	13.68	27.595	16.013
	6	mean	23.8693	15.794	21.0521	13.22	30.756	23.578
		median	20.9631	14.137	17.6143	11.94	26.137	20.13

We can see from the table that as the target service level (α) increases the accuracy of these methods are affected differently. The SIPP and MOL provide more accurate approximations for models with lower quality of service, while MaxSIPP is better when higher quality of service is required. One noticeable trend is the effect of the length of planning period on the performance of these methods.

4.4 Summary

A time-varying simulation model developed and validated with the exact method, which is then used to compare different approximation methods for nonstationary queueing models. Statistical analysis of planned experiment results on the test problems with a wide range of parameters assured us that the model generates statistically equivalent results as the exact method. It is also found that for queueing/simulation models with a time varying arrival rate, periodic steady state condition, rather than the steady state condition, is to be used to accurately reflect the situation. Then we used the validated simulation model to evaluate the performance of a number of approximation methods applied to the $M_t/G/s_t$ and determined conditions under which either of these methods would be most suitable. We realized that the MOL method is more accurate and requires less number of servers in comparison to other methods. MaxSIPP performs better when the quality of service is very high; however it demands a higher number of staff.

5: MODELING EMERGENCY DEPARTMENT WITH BOTH QUEUEING AND SIMULATION MODELS

As we previously discussed the purpose of modelling was to represent the emergency department and help with decision-making regarding determining the number of staff when there is a sudden variation in arrival rate. In previous chapters, we built nonstationary queueing models that could capture the variation in arrival rate and used approximation methods to determine the staffing requirements in accordance with these variations. Then we calculated the performance measure resulted from employing those proposed number of servers. We then built a parallel simulation model to evaluate the accuracy of these approximation methods for the cases where there were not numerical solutions available.

In all the experiments run in the previous chapter, we considered a single $M_t/G/s_t$ queueing model. However, assuming the emergency department as a single queueing model is not realistic. Since the actual system consists of different units which require different types of staff with their specific service time distributions.

5.1 ED at St. Paul's hospital

St. Paul's Hospital is an acute care, academic, and research hospital located in downtown Vancouver. ED at St. Paul's hospital consists of different parts:

- Fast track (FT)
- Diagnostic and Treatment Unit (DTU)
- Triage and Registration
- Acute Zone

- Rapid Assessment Zone (RAZ)

The patient route in the St. Paul's hospital starts by a patient getting into the triage and registration area either by foot or in an ambulance. In this area, one registration clerk registers the patient and then the patient waits for one of the two triage nurses to do a quick interview to determine the level of severity of his/her condition. A CTAS (Canadian Triage and Acuity Scale) level is then assigned to the patient and s/he will wait until the appropriate service is available. Complete registration of the patients is usually done after triage if the patient is not of high severity level. This is done by the same registration clerk or the other one who comes to work during peak hours.

Once the CTAS level is determined for the patient, s/he will go through a special path designed for that CTAS level. Patients of the CTAS level I are usually assigned to the trauma room directly and immediately. These patients have the highest priority for handling and treatment in the ED. Once the patient is in the bed, a nurse would start the treatment and soon after, a physician will visit this patient and will order appropriate medications and/or tests for him/her.

Patients of CTAS level II are supposed to be assigned to an acute bed shortly after they are triaged. These patients may go to the Rapid Assessment Zone (RAZ) if no acute beds are available. They are then seen by a nurse and a physician.

Patients of CTAS level III belong to a heterogeneous category that includes a range of presenting complaints with a spectrum of severity. Higher acuity level III patients will mostly require a bed for care, whereas the lower acuity level III patients may not require a bed. In the current system, lower acuity III patients are supposed to be seen in the RAZ likewise by a nurse and then a doctor. Higher level III patients should be placed in an acute bed if available. If no bed is available, they should be seen in the RAZ.

The rest of the patients (levels IV and V) will go to the Fast Track area and will wait till a bed is available. In a research study it was shown that a fast track lane can help decreasing 50% of the required resources while resulting in less waiting time for all the patients on average due to providing the possibility of handling patients of CTAS levels III and IV, and therefore providing more space and resources for patients with higher acuity levels [40].

However, there are some exceptions to the flow of the patients. For example, patients who have an emergency eye problem and are of any of the five CTAS levels would go to the Fast Track area since the required equipment for their treatment is located there. Flow of patients in the hospital is shown in Figure 5-1.

In a practical project a detailed simulation model for the ED was constructed and validated with historical data of the patients' visits to the ED during the year 2009. The model was fed with the actual arrival times and other parameters of the systems were estimated from data. The performance statistics recorded by the model was compared to the reported values of the real system.

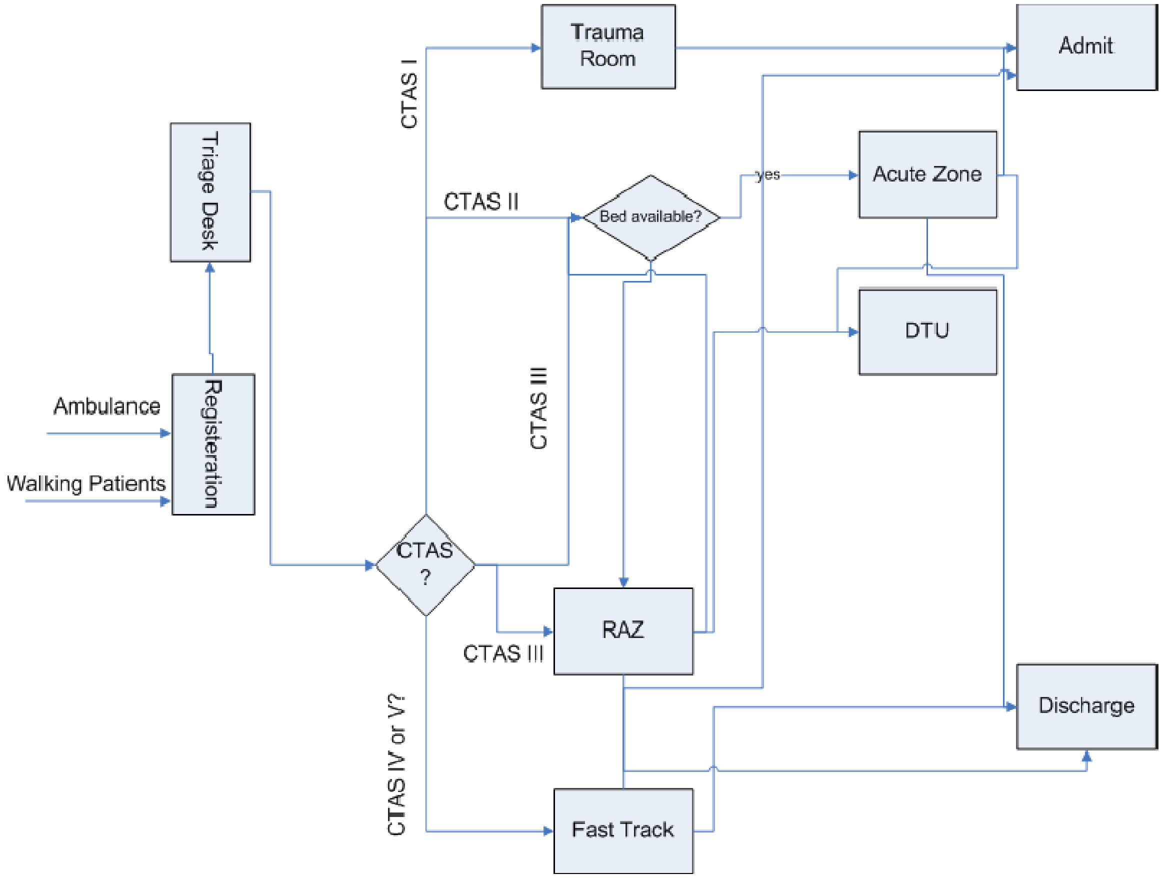

Figure 5-1 Flow of patients in the ED

Analysing ED at St. Paul's we realized that some of these different parts share the same resources most of the time, so we integrated them into one unit. Finally, we had three separate units that had their own specific resources. We named these units as Triage, Acute care and Fast Track.

In this chapter, we would like to analyse the three-node network of queueing systems for the specific case of the ED. Figure 5-2 is a schematic representation of these units. In this network, the arrival to the Acute and Fast Track is the departure from the Triage unit. Therefore, first we have to determine the departure of this unit as a function of its arrival.

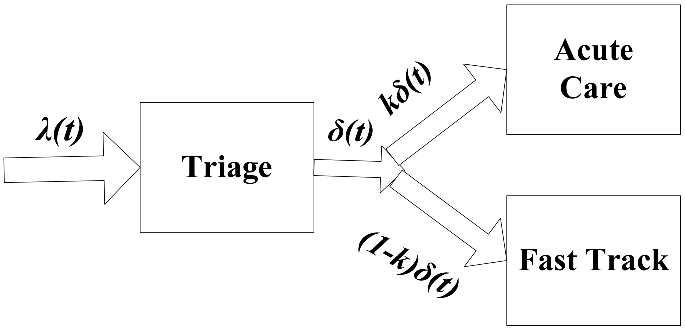

Figure 5-2 Schematic representation of the ED units

5.2 Approximating the departure process

The arrival to the first unit (Triage), as discussed before, is supposed to be a non-homogeneous Poisson process with parameter $\lambda(t)$. If we consider the service time of the first queueing model to be exponentially distributed, the outflow of that model will follow a non-homogeneous Poisson distribution with the rate $\mu s(u)$, for $u \geq t$ (assuming all servers are busy throughout the interval, which could be realistic if we set the service level low enough) [41]. The outflow of this unit then divides into two parts, each entering another queueing system. So, the arrival rate to one unit would be a non-homogeneous Poisson distribution with a parameter $k\mu s(u)$, for $u \geq t$, where k is the percentage of the outflow entering that unit. Consequently, the arrival rate to the other unit would be a non-homogeneous Poisson distribution with parameter $(1\text{-}k)\mu s(u)$, for $u \geq t$. The value for k can be estimated using historical data from ED.

However, the aforementioned results hold only if all the servers are busy all the time, which may not be always true. Therefore, we turn to another strategy to approximate the departure rate of a $M_t/G/s_t$ model.

We can approximate the outflow of the $M_t/G/s_t$ model with the outflow of a $M_t/G/\infty$ using the following theorems. In this model, where the number of customers in the system at time *t* is represented by *Q*(*t*) the following theorem can be proved. Let *S* be the generic service time random variable and let *G* be its cumulative distribution function (c.d.f.). The service times should be i.i.d and independent of the arrival process [42]. S_e is a random variable associated with stationary-excess or equilibrium-residual-lifetime c.d.f..

$$G_e(t) \equiv P(S_e \leq t) \equiv \frac{1}{E[S]} \int_0^t G^c(u)du, \qquad t \geq 0, \tag{5-1}$$

where $G^c(t)$= 1-*G*(*t*),

Theorem 1. [42] For each *t*, *Q*(*t*) has a Poisson distribution with mean

$$m(t) = E[\int_{t-S}^{t} \lambda(u)\, du] = E[\lambda(t - S_e)]E[S]. \tag{5-2}$$

The departure process is a Poisson process with time dependent rate function δ, where

$$\delta(t) = E[\lambda(t\text{-}S)] \tag{5-3}$$

For each *t*, *Q*(*t*) is independent of the departure process in the interval (-∞,*t*]. The departure process is directly associated with the derivative of *m*(*t*) in the theorem described below.

Theorem 2. [42] If the departure function δ in Eq. (5-3) is integrable in a neighbourhood of *t*, then the mean function *m* in Eq. (5-2) is absolutely continuous with respect to a *Lebesgue* measure in a neighbourhood of *t*, with density

$$m'(t) = \lambda(t) - \delta(t) \tag{5-4}$$

If we revisit Eq. (2-6) from chapter two

$$E[Q(t)]' = \lambda(t) - \mu E[Q(t)] \tag{5-5}$$

we can conclude that if the service time of the queueing system is exponential, the departure process can be estimated with a non-homogeneous Poisson process with rate $\mu m(t)$. In other words, as a generalization to what was discussed earlier, the departure process is always a non-homogeneous Poisson process with rate $m(t)/E[S]$, if the service times are exponential with mean $E[S]=1/\mu$.

If the service time is not exponential, we cannot draw the last conclusion. We are left with Eq. (5-3) which is complicated since the time lag of the arrival rate S appears inside the arrival function $\lambda(t)$, inside the expectation $E(.)$. If the expectation could be moved inside, we could produce a deterministic time lag $E[S]$ and $\delta(t)$ could be expressed more generally in terms of the moment of S. This would have been possible if the arrival rate function $\lambda(t)$ were a polynomial. Although the arrival rate function will not usually be polynomial, it can be approximated by polynomials in the neighbourhood of individual arguments, using Taylor-series approximations [43].

Consider we are interested in performance in some time t. we can approximate the arrival rate function, $\lambda(t)$, in an interval before time t using a first order Taylor-series approximation centered at t

$$\lambda(t-u) \approx \lambda(t) - \lambda^{(1)}(t)u \qquad\qquad for\ u \geq 0 \tag{5-6}$$

where $\lambda^{(k)}(t)$ is the kth derivative of $\lambda(t)$ evaluated at time t. Applying Eq. (5-6) to Eq. (5-3) we have

$$\delta(t) \approx \lambda(t - E[S]), \tag{5-7}$$

which shows that $\delta(t)$ is approximately the PSA arrival rate modified by the deterministic time lag $E[S]$.

5.3 Experimenting with queueing network

In chapter three, we validated a single queueing model with the exact method. Here, we build a queueing network by connecting three of those single queueing models together with minor changes in their service times. To make sure that this network of the validated models is still valid we compare its results with the analytical solution obtained from the exact method, inputting a challenging arrival rate ($\lambda(t)=30 + 20sin(5t)$).

Since the MOL method was proved to be better than the other methods in the last chapter for surge, we apply it to the units in the second layer of the network (Acute and Fast Track). We assume the arrival rates to these units are estimated from Eq. (5-3), and are divided into two branches with the coefficient k =0.4. We also extend the single simulation model to a three-node one and let it run with same parameters as the queueing network. Since we would like to compare the simulation outcomes with the results of the exact method we have to assume that all the service times are exponential.

As it is clear from Table 5-1 and Table 5-2, the simulation model output, except for the first planning period, are close to the results obtained by the exact method. The large difference in the first period is due to the fact that there is a time lag in the arrival to the second layer models as a result of the delay of service in the Triage unit preceding them in the network. This causes the Fast Track and Acute unit to be empty during the time that patients are being served at Triage in the beginning of each day. This time lag which is generated at the start of each day has not been included in the queueing model and that is the reason why the errors in the first planning period are substantially greater than other periods.

Table 5-1 Comparison of the MOL solution with simulation for the Fast Track unit

PP	No. Servers	Pd (MOL)	Pd(Simulation)	Error %
1	8	0.127811	0.284289	122.4292
2	9	0.183496	0.176322	3.909622
3	10	0.100265	0.093006	7.239814
4	9	0.162365	0.145007	10.69073
5	10	0.105551	0.102717	2.684958
6	9	0.159756	0.163029	2.048749
7	9	0.187418	0.170082	9.249912
8	10	0.098375	0.10376	5.473952
9	9	0.165752	0.141202	14.81128
10	10	0.104769	0.110627	5.591349
11	9	0.158454	0.146426	7.590847
12	9	0.189386	0.172901	8.704445

Table 5-2 Comparison of the MOL solution with simulation for Acute unit

PP	No. Servers	Pd (MOL)	Pd(Simulation)	Error %
1	8	0.087064	0.336088	286.0241
2	10	0.118574	0.126574	6.746842
3	10	0.132206	0.120488	8.86344
4	10	0.131164	0.087326	33.42228
5	10	0.138698	0.106619	23.12867
6	10	0.128311	0.140796	9.730265
7	10	0.138797	0.13857	0.163548
8	10	0.132531	0.124009	6.430194
9	10	0.132908	0.099382	25.22497
10	10	0.137905	0.133627	3.102136
11	10	0.128276	0.120276	6.236552
12	10	0.139362	0.139532	0.121984

5.4 Applying the method to actual data

As mentioned before, the purpose of modelling the ED was to develop a tool that could assist decision making in the event of surge, when managers encounter difficulty setting the staffing requirements. We built a queueing network of three $M_t/G/s_t$ models that represented the actual system; then selected the best approximation method for surge to determine the staff level and predicted the resulting probability of delay. Now, we would apply the chosen method to the queueing network whose arrival rate is estimated from historical data of the actual ED.

5.4.1 Fitting non-homogeneous Poisson distribution to actual data

We first have to make sure, if the data from the ED would actually fit into the assumption of non-homogeneous Poisson process. For determining whether the arrivals of the considered process form a non-homogeneous Poisson process, we have to build a test for this null hypothesis. The first step in constructing this test is to break up the duration of the day into relatively short blocks of time; short enough so that the arrival rate does not change significantly within a block. The length of the time blocks does not have to be equal; however, for convenience it is assumed so. Let T_{ij} denote the j-th ordered arrival time in the i-th block, i=1, … I. Thus $T_{i1} \leq \ldots \leq T_{iJ(i)}$, where $J(i)$ denotes the total number of arrivals in the i-th block. Then define T_{i0}=0 and

$$R_{ij} = (J(i) + 1 - j)\left(-log\left(\frac{L - T_{ij}}{L - T_{i,j-1}}\right)\right), \qquad j = 1, \ldots, J(i) \qquad (5\text{-}8)$$

Under the formal null hypothesis that the arrival rate is constant within each given time interval, the $\{R_{ij}\}$ will be independent standard exponential variables, as discussed below.

Let U_{ij} denote the j-th (unordered) arrival time in the i-th block. Then as we have assumed the arrival rate within this block to be constant Poisson, conditionally on $j(i)$, the unordered arrival times are independent and uniformly distributed, that is, $U_{ij} \sim$ U(0, L). Note that $T_{ij}= U_{i(j)}$. It follows that $\frac{L-T_{ij}}{L-T_{i,j-1}}$ are independent $beta(J(i)+1-j,1)$ variables. By a standard change of variables the conditional exponentiality of the R_{ij} given the value of

$J(i)$ is obtained. (Equivalently we may base the test on the variables $R^{*}_{ij} = j(-log\frac{T_{ij}}{T_{i,j+1}})$, where j=1,…, $J(i)$ and $T_{i,J(i)+1}$=L. under the null hypothesis, these will also be independent standard exponential variables [44].

The null hypothesis does not imply that the arrival rate of different intervals should be equal or have any other prescribed relationship. Any customary test for the exponential distribution can be applied to test the null hypothesis. We use the Kolmogorov-Smirnov test, which is available in Matlab statistical toolbox. Moreover, exponential Q-Q plots can be very useful in determining goodness of fit to the exponential distribution [29].

We considered one-hour intervals, and could not reject the null hypothesis. This implies that the arrival rate follows non-homogeneous Poisson distribution. After confirming the assumption of non-homogeneous Poisson arrival distribution, we can estimate the arrival rate with a periodic function. We extracted the hourly arrival rates to the ED at the time of the outbreak of the H1N1 flu in fall 2009 from the data provided to us and utilized Matlab Curve Fitting tool to fit a sum of Sine functions to it. Figure 5-3 depicts our data and the curve that is fit to it. The mathematical representation of this function is as follows:

$$\lambda(t) = a_1 \sin(b_1 x+c_1) + a_2 \sin(b_2 x+c_2) \tag{5-9}$$

Table 5-3 presents the value for the parameters in arrival rate and the R-square and RMSE error for this fit.

Table 5-3 Parameters and fitting error for the approximated arrival rate

a_1	13.04
b_1	0.06042
c_1	0.317
a_2	4.173
b_2	0.3988
c_2	2.918
R-Square	0.7072
RMSE	3.152

R-Square Error

R-square is a statistic measure that shows how successful the fit is in explaining the variation of the data. In other words, R-square is the square of the correlation between the response values and the predicted response values. R-square is defined as the ratio of the sum of squares of the regression (*SSR*) and the total sum of squares (*SST*). *SSR* is defined as

$$SSR = \sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y})^2 \qquad (5\text{-}10)$$

SST is also called the sum of squares about the mean, and is defined as

$$SST = \sum_{i=1}^{n} w_i(y_i - \bar{y})^2 \qquad (5\text{-}11)$$

where *SST* = *SSR* + *SSE*, and *SSE* is sum of squares due to error calculated as below

$$SSE = \sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2 \qquad (5\text{-}12)$$

Given these definitions, R-square is expressed as

$$R - square = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad (5\text{-}13)$$

R-square can take on any value between 0 and 1, with a value closer to 1 indicating that a greater proportion of variance is accounted for by the model.

Root Mean Squared Error

This statistic is also known as the fit standard error and the standard error of the regression. It is an estimate of the standard deviation of the random component in the data, and is defined as

$$RMSE = \sqrt{MSE} \qquad (5\text{-}14)$$

where *MSE* is the mean square error or the residual mean square

$$MSE = \frac{SSE}{v} \qquad (5\text{-}15)$$

Just as with *SSE*, an *MSE* value closer to zero indicates a fit that is more useful for prediction. The residual degrees of freedom is defined as the number of response values *n* minus the number of fitted coefficients *m* estimated from the response values (v= *n*-*m*) [45]. Based on the values for these error measures, we concluded that this fit is appropriate for our data.

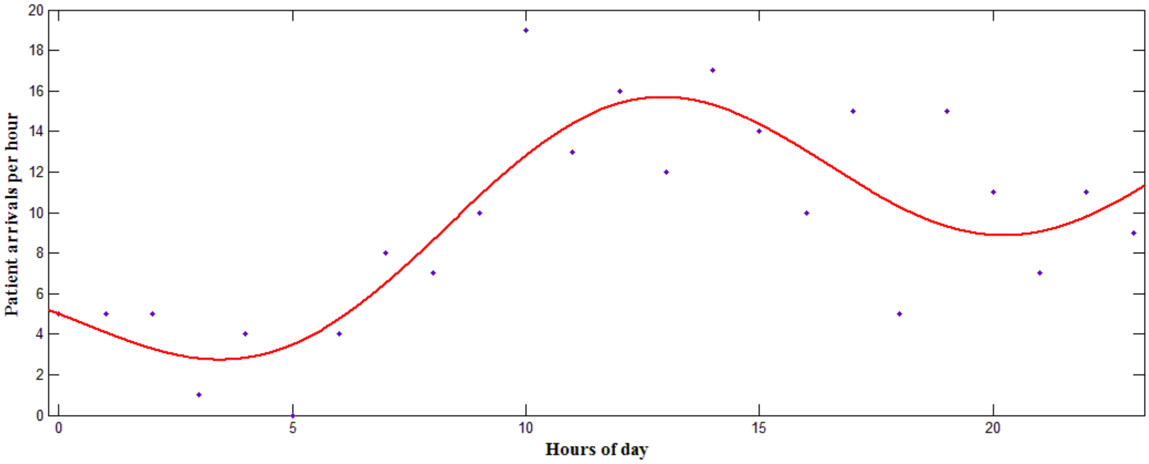

Figure 5-3 The arrival function fitted over the actual data from the ED

5.4.2 Service times

From the data, we also estimated the percentage of patients that would enter the Acute unit to be k = 0.42. That results in the arrival rate to this unit to be $k\lambda(t)$ and the arrival to the fast track to be $(1\text{-}k)\lambda(t)$. As mentioned before, determining service times for different units in the ED is a challenging problem. Here we consider some estimation that have been recommended by the experts at ED. Table 5-4 shows the mean service times for each unit. In queueing network, all the units are assumed to have general distribution and are considered to have SCV=2.

Table 5-4 Service times mean and standard deviation

Unit	**Service Time Mean (min)**	**Service Time Standard Deviation (min)**
Triage	10	7.07
Fast Track	20	14.42
Acute	30	21.21

In the next step, we input the estimated arrival function to our queueing network, apply the MOL method to each unit and determine the staffing level for each unit. We compare suggested solutions for two strategies Quality-driven and Quality and Efficiency driven (Rationalized) with setting the target probability of delay to α=0.1 and 0.5 respectively. We also determined the staffing with different planning periods. Tables below summarize the outcome of these experiments.

Table 5-5 The number of servers suggested for two-hour planning period and probability of delay estimated with MOL and simulation

PP	α	Acute				Fast Track				Triage			
		s	Pd(MOL)	Pd(Sim)	Err.	s	Pd(MOL)	Pd(Sim)	Err.	s	Pd(MOL)	Pd(Sim)	Err.
1	0.1	3	0.101	0.270	62.5	3	0.069	0.183	62.3	2	0.180	0.286	37.1
2		2	0.176	0.136	29.7	2	0.131	0.091	44.6	1	0.700	0.341	105
3		2	0.124	0.106	17.8	2	0.118	0.065	81.5	1	0.760	0.381	99.5
4		3	0.077	0.083	7.7	3	0.086	0.096	10.5	2	0.354	0.322	9.8
5		4	0.114	0.122	6.9	4	0.118	0.113	4.5	3	0.325	0.288	12.7
6		5	0.150	0.148	1	5	0.134	0.119	13.2	4	0.265	0.238	11.2
7		6	0.121	0.113	7.1	6	0.092	0.083	10.2	4	0.353	0.340	3.8
8		6	0.124	0.113	9.8	6	0.083	0.075	9.5	4	0.291	0.252	15.4
9		5	0.177	0.125	41.9	5	0.115	0.091	26.4	3	0.441	0.391	12.9
10		5	0.077	0.059	29.8	4	0.153	0.145	5	3	0.267	0.262	1.9
11		4	0.138	0.113	22.5	4	0.107	0.090	19.2	3	0.236	0.224	4.9
12		4	0.177	0.187	5.3	4	0.156	0.127	23.2	3	0.354	0.328	8
1	0.5	2	0.380	0.570	33.3	1	0.967	0.895	8	1	0.861	0.827	4.1
2		1	0.857	0.614	39.4	1	0.792	0.731	8.4	1	0.700	0.514	36.3
3		1	0.781	0.614	27.2	1	0.770	0.673	14.5	1	0.760	0.401	89.6
4		1	0.983	0.761	29.3	2	0.338	0.616	45.1	2	0.354	0.351	0.8
5		2	0.836	0.736	13.6	2	0.849	0.672	26.3	2	0.815	0.649	25.6
6		3	0.791	0.738	7.1	3	0.744	0.695	7.1	3	0.626	0.585	7
7		4	0.603	0.627	3.8	3	0.996	0.770	29.3	3	0.768	0.685	12.2
8		4	0.611	0.585	4.5	3	0.950	0.814	16.7	3	0.670	0.659	1.6
9		3	0.867	0.609	42.3	3	0.683	0.719	4.9	2	0.988	0.804	22.9
10		3	0.544	0.561	2.9	2	0.956	0.807	18.5	2	0.721	0.737	2.2
11		2	0.914	0.737	23.9	2	0.811	0.811	0	2	0.665	0.679	2.1
12		3	0.467	0.594	21.4	2	0.967	0.821	17.7	2	0.861	0.744	15.7

Table 5-6 The number of servers suggested for four-hour planning period and probability of delay estimated with MOL and simulation

PP	α	Acute				Fast Track				Triage			
		s	Pd(MOL)	Pd(Sim)	Err.	s	Pd(MOL)	Pd(Sim)	Err.	s	Pd(MOL)	Pd(Sim)	Err.
1	0.1	3	0.062	0.192	67.6	2	0.202	0.325	37.7	2	0.127	0.217	41.5
2		2	0.197	0.227	13.5	2	0.209	0.211	0.9	2	0.217	0.240	9.7
3		5	0.079	0.114	30.6	5	0.075	0.092	18.5	4	0.178	0.196	9.1
4		6	0.127	0.117	8.5	6	0.089	0.082	8.8	5	0.127	0.129	1.1
5		5	0.120	0.108	10.9	5	0.077	0.085	9.6	4	0.119	0.131	9.4
6		4	0.151	0.139	8.6	4	0.127	0.133	4.2	4	0.093	0.101	7.9
1	0.5	1	0.950	0.899	5.7	1	0.200	0.889	77.5	1	0.785	0.786	0.1
2		1	0.880	0.850	3.6	1	0.894	0.800	11.7	1	0.902	0.739	22
3		3	0.553	0.688	19.6	3	0.537	0.644	16.6	3	0.469	0.505	7.1
4		4	0.621	0.622	0.2	3	0.983	0.795	23.7	3	0.722	0.676	6.9
5		3	0.699	0.654	6.7	3	0.544	0.632	14	2	0.852	0.789	8
6		2	0.951	0.787	20.9	2	0.879	0.756	16.3	2	0.759	0.701	8.3

Table 5-7 The number of servers suggested for six-hour planning period and probability of delay estimated with MOL and simulation

PP	α	Acute				Fast Track				Triage			
		s	Pd(MOL)	Pd(Sim)	Err.	s	Pd(MOL)	Pd(Sim)	Err.	s	Pd(MOL)	Pd(Sim)	Err.
1	0.1	2	0.205	0.352	41.9	2	0.165	0.298	44.8	2	0.119	0.188	36.7
2		4	0.118	0.186	36.5	4	0.118	0.184	35.6	4	0.103	0.159	35.1
3		6	0.108	0.114	5.5	5	0.185	0.196	5.6	5	0.099	0.110	10.5
4		4	0.164	0.178	7.6	4	0.131	0.131	0.1	4	0.088	0.096	8.2
1	0.5	1	0.889	0.911	2.4	1	0.842	0.918	8.3	1	0.771	0.722	6.8
2		2	0.850	0.784	8.3	2	0.850	0.765	11.1	2	0.798	0.709	12.7
3		4	0.562	0.671	16.3	3	0.890	0.873	1.9	3	0.627	0.710	11.8
4		2	0.988	0.804	22.9	2	0.891	0.876	1.7	2	0.740	0.670	10.5

Form the tables we can see, this method works better for the Quality driven strategy, since the probability of delay estimated with queueing model and its parallel simulation model is closer to the predefined target probability of delay (α=0.1).

Moreover, the outcome of the queueing model is also more similar to the simulation for (α=0.1). It can be concluded that these methods work better with lighter traffic. However they would still work for heavier traffic.

5.5 Simulation of the ED at St. Paul's

As we discussed earlier in this chapter, the clinical procedures at St. Paul's emergency department has been simulated with Arena^{TM}. The patient flow for the emergency department of St. Paul's hospital is modelled according to the patients' data in 2009 (including arrival times, lengths of stay in different units, and information pieces such as the CTAS levels, dispositions, etc.). The probability distributions of lengths of stay at different units of ED, percentage of patients routing to different units and many other parameters have been identified from historical data and fed into the model. The simulation model was built based on the current number of resources and current procedural policy (e.g., sequence of the units through which a patient might be handled and treated during his/her stay in the ED). The model was verified and validated based on 171,000 records of patients' visits. For the cases where there were not enough data available, for example for the service time distributions the best guess of the expert physicians at the hospital was incorporated in the model.

This model was then utilized to run scenarios to analyse the effect of making changes in the system. These scenarios included adding single or multiple resources to different units, considering the effects of closing some of the units, and so on. These impacts were studied as the effects on the overall performance of ED in terms of "time to provider (T2P)", and "length of stay (LOS)".

5.6 Scenarios for analysis

Here, a number of potential options (questions) that have been studied with the help of the ED simulation model are presented

- **Dynamic resource allocation**: Different parts of the ED are designed for specific types of patients and are handled by limited resources. Given resource costs, one question is which unit to keep open or closed daily. In other words, for example "Is opening RAZ more helpful during the peak hours than other measures?", "What are the conditions (i.e. times, number of resources, utilization policies, etc.) under which each unit can be more helpful and of better performance in terms of service times?"
- **Bottleneck Analysis**: What are the major factors that contribute to the problems of ED at St. Paul's, such as overcrowding, poor throughput of patients, incompatibility of required resources (such as lack of enough nurses for the existing beds), etc.? Which sub-cycles contribute most to poor flow/overcrowding (in absolute and relative terms)? Where the target quality improvement interventions can be set to be cost-effective and to have the largest impact on improving the flow?
- **Demand/Capacity**: What is the impact of hiring new physicians/nurses or adding extra stretchers?

For the scenarios described above, results and important statistics are shown in Table 5-8 in detail; the findings for each scenario are as follows:

- Adding one doctor: As the results show, for adding a doctor (the word "doctor" refers to the doctors who visit patients in all units except FT, which is assigned separate doctors, known as "FT doctors" in the model), the best shift to make this assignment is from midnight to 7:00am. This change in number of doctors however, will result in only fifteen minutes of reduction in waiting time to provider, and consequently in length of stay for the patients on average.
- Adding two doctors: Since the number of doctors during the night shift and afternoon is less than that of other times, a scenario was to add one doctor to each of those shifts to see how the performance of ED might improve. The results indicate that this alternative performs similar to adding one doctor during the night shift, and therefore it can be concluded that adding a doctor to the afternoon shift (from 14:00 to 18:00 typically) is not of any help for reducing waiting times and such investment is of no benefit to ED performance.
- Adding beds and nurses: The number of beds in some units can be changed with the number of nurses because each nurse is supposed to handle a particular number of beds in some units. For example in Acute, each nurse is assigned to four beds while in the Trauma room each nurse takes care of exactly one bed. Therefore changing beds and nurses requires different arrangements for different units. As Table 5-8 shows, all units were considered for simultaneous increasing of beds and nurses. According to the results, the only effective resource increase in terms of beds and nurses is for RAZ, which shows a slight improvement (only five minutes for Time to Get a Bed).

- It can be concluded that it is not worth to consider merely adding beds along with nurses to any of the units. However, adding beds and nurses might result in remarkable results if it is accompanied by adding other types of resources or adding resources to other units as well.

- Adding combinatorial resources: When considering making a variety of resource changes in more than one unit (i.e., combinatorial resources), we obtain the results presented in the forth scenario of Table 5-8. Since it shows some improvement subject to adding Trauma nurse(s) along with some other changes, we decided to perform a detailed analysis based upon the Trauma resources. For this purpose, the scenario of "adding 1 Trauma nurse and 1 bed" was also considered.

Table 5-8 Results of simulation for different scenarios in terms of desired statistics (LOS, T2P, T2B)

Times Scenarios			Length of Stay (LOS)		Time to Provider (T2P)	Time to Get a Bed (T2B)	Greatest amount of decrease in average waiting time
			Admitted	Discharged			
Base case (current system)			641	280	45	17	
Adding a doctor	8:00- 15:00		640	276	43	16	Not significant
	24:00- 7:00		**624**	**265**	**31**	13	15 min (LOS, T2P)
	15:00- 23:00		**629**	**270**	**38**	14	10 min (LOS, T2P)
Adding 2 doctors	24:00- 7:00 14:00- 18:00		**625**	272	**34**	17	15 min (LOS, T2P)
Adding beds and nurses	Acute	4 beds 1 nurse	640	278	44	17	Not significant
	DTU	2 beds 1 nurse	639	275	43	15	Not significant
	FT	2 beds 1 nurse	643	280	45	16	Not significant
	RAZ	2 beds 1 nurse	643	274	**41**	**12**	5 min (T2P, T2B)
Adding resources in combination	Acute: 4 beds RAZ: 3 Care Spaces		638	278	44	**12**	5 min (T2B)
	Trauma: 2 beds Acute: 4 beds		638	279	**37** (Trauma)	15	9 min (T2P)
	1 Trauma nurse 1 doctor: 24:00-8:00 2 nurses		**620**	**272**	**30**	**12**	20 min (LOS) 15 min (T2P) 5 min (T2B)
	2 Trauma nurses 2 Acute3 nurses 2 doctors: 24:00-7:00 16:00-23:00		**610**	**269**	**32**	15	30 min (LOS) 13 min (T2P)
	1 FT doctor: 24:00-7:00 1 nurse to all units 1 bed to all units 2 doctors: 24:00-7:00 16:00-23:00		**615**	**263**	**30**	**13**	25 minutes (LOS) 15 minutes (T2P) 4 minutes (T2B)
Closing RAZ			635	*370*	*53*	21	Adding: 100 min to LOS (discharged) 8 min to T2P
Closing RAZ during nights			632	*323*	46	17	Adding 40 min to LOS (discharged)
Closing DTU			*1142*	*940*	*297*	*189*	Resulting in significant lack of resources and increasing in LOS, T2P, T2B
Decreasing DTU resources	1 nurse 2 beds		650	285	55	17	No significant increase

* The bold face numbers show decrease in statistic values (favored effect of scenario), and the italic grey highlighted values show increase in the statistic value (adverse effect of scenario)

Various types of resource increases were studied among which, adding resources of all types in a combination, offers the best results. This fact implies that in case of increased resource allocation to more than one unit, the service quality will improve more noticeably than any other improvement on each individual resource. Since adding resources will increase the operational costs at ED, a tradeoff would be made between the increased cost and the waiting time reduction. Decision about the desired resource increase and the units receiving those resources will depend on targets and constraints of the authorities and more precise decisions will depend on simultaneously considering both sides (i.e., benefits of increasing resources, and costs accompanying the resource increase).

Results of this analysis showed that while adding resources to some of the units in ED might not make significant improvement in decreasing the waiting times, closing them can remarkably add to waiting times and therefore revealed that their operation is vital in current layout and situation of the ED.

For identifying the bottleneck units, it can be concluded that no single unit or resource might be a bottleneck per se, and all of the units and resource types seem to play an important role in serving the patients. However, precise studies might be able to reveal the exact differences between different resources. This stage is possible to be evaluated only if more reliable and precise data is available.

5.7 Inputting queueing model outcome to detailed simulation

In an attempt to test our solution on the real system we applied the number of servers suggested by the methods we employed to the simulation model of the ED at St. Paul's. We input the arrival rate function estimated from data to this model and changed the service time distributions to match that of the queueing network model. Our expectation was a dramatic decrease in the patients waiting times as a result of employing, for instance, the quality driven strategy in Table 5-6 which had almost twice staff hours as the original schedule of the ED. Unfortunately, the ED simulation model did not show significant sensitivity to the alterations that we did (The output of this experiment can be found in the appendix).

We analysed the model to find out possible reasons that could have a potential role in this case, and found the following issues:

1. The simulation model constructed captures many details of the system (which queueing model either does not or considers them with extra simplifying assumptions). Namely some of the features that are not exactly modelled in the queueing model are:

    a. **Different types of staff,** such as trauma nurses, acute nurses, fast track nurses, registration clerks etc, each with their own specific service times for any of the units. In the queueing model only three types of resources are considered (triage nurse, acute doctors, and fast track doctors)

    b. **Priority of the patients**, with respect to the severity of their condition would receive priority code, CTAS, (meaning that the patients of lower CTAS levels would get through the queue faster and be served sooner). In the queueing model, in contrast, all the queues are assumed to be treated in the first-come-first-served discipline.

c. **Abandonments**, patients get intolerant with long wait times and leave without being seen by a doctor whereas in queueing model it is assumed that patients would wait until they receive service.

d. **Patients transferred within the units** due to bed unavailability or to complete their treatment services (for example as we described before, if there are no beds available in the acute area patients would be transferred to RAZ). No such flexibility is considered in the queueing model, besides all patients in the trauma room, acute area, and RAZ are considered to be in one integrated unit since they received service from the same staff members.

2. The validation and scenario testing for the simulation model were both done by inputting the actual arrival data from the real ED over the entire one-year period. The actual data is completely deterministic and does not include any randomness. On other the hand, the queueing model was fed with a stochastic arrival rate having time-varying parameter. Moreover, the queueing model was built for a specific time of the year and focused on the variation of the arrival within a two-week period of the surge caused by H1N1 flu. Although we input the time varying arrival function into the simulation model when we were to check our proposed solutions, we could see from the results that the model was not sensitive to the changes in staffing requirement. It could be partly because the model was validated for a deterministic arrival rate and by adding the randomness generated from employing a time varying distribution function it does not show the expected output. We have not examined the effect of randomness of the arrival rate on the results of the model and have not validated it under such conditions.

3. Given the argument in number 2, the scenario testing done on the validated model with deterministic arrival also does not reflect much sensitivity to changes in merely a single resource capacity (e.g. adding one or two doctors in an eight-hour shift). According to the results presented in Table 5-8, multiple resource changes of different types would best affect the output of the simulation model. This could be because no single resource can be identified as a bottleneck. Hence, the system would not stand to benefit from increasing one of them without changing the others. That could justify the unexpected results we had received from the changing just the number of physicians in the system.

5.8 Summary

This chapter discussed the construction of a network of three nodes that modelled the ED. We estimated the arrival rate to each node from data and with the help of the queueing theory rules. Then we applied the proper approximation method to this model and determined the number of servers in various planning periods with different lengths. We applied this solution to the detailed simulation of the ED as well as the three-node simulation. The output of latter was similar to the queueing network. However, we failed to confirm it with the detailed simulation model due to a number of reasons that we discussed in this chapter. Nonetheless, the queueing model solution can be utilized by the managers of the ED as estimation for the number of staff required under the surge circumstance. It also can be used as an input to a scheduling routine to determine appropriate shifts for the staff at ED.

6: CONCLUSION

A major challenge in emergency departments is to determine the number of servers required to satisfy the demand, especially when a surge such as a mass casualty incident occurs. Queueing theory has long been employed to assist decision making regarding setting the staffing requirement to cope with the variation in the arrival rate. Specifically in the cases with highly varying arrival rate, often nonstationary queueing models would be developed. Since some of the nonstationary queueing models built to represent these complex systems are too complicated and are intractable mathematically, other approaches such as simulation and approximation methods are applied to analyse them.

In this work, we constructed a queueing model that captured the variation in arrival rates, simulating a surge situation, and set the staffing requirement according to these variations so that a target level of service was maintained. For evaluating the performance of such a model, an exact numerical solution was found in literature. This method, however, was limited to certain special conditions and was not applicable for many other queueing models. Therefore, we developed a simulation model with time varying arrival rate as a baseline, in order to compare and evaluate different queueing models for modeling the surge situation. This simulation model was validated with exact numerical solutions for 162 different model variants with a wide range of conditions. The validation not only gave us a credible simulation model, it also shed some lights on the type and condition of queueing models.

Then we utilized the validated simulation model for comparing the performance of approximation methods. We ran the 324 scenarios for each approximation method both on simulation and queueing model and identified the one that better suggested less number of staff and predicted more accurate estimates of the performance measure.

At the end, a three-node network of queueing models was constructed. It was shown that the result of the queueing network confirmed the related simulation model output. The best approximation method chosen previously was applied to this network inputting the actual arrival rate estimated from data of the real ED at the event of surge. We omitted the simplifying assumption of exponentiality of the service time distributions and used a general distribution with specified mean and variance. The queueing network then predicted the staff level required for each node (unit) and the performance resulted from employing them.

As a part of a practical project, a detailed discrete event simulation model of the ED at hospital had been developed and validated based on the historical data from the actual patients' visits. As the final step, we fed the staff number suggested by the queueing network to this simulation model. However, we failed to confirm the effect of these changes on this simulation model due to following reason.

- Incompatibility of the queueing model with the simulation model in terms of the incorporated level of details from the actual system
- Insensitivity of the simulation model to the single resource changes

- Determinism built in the simulation model as a result of employing the exact actual arrival times of the patients from the historical data, as opposed to the randomness of the queueing model caused by its time varying Poisson arrival rate

6.1 Contributions

The main contributions of this thesis can be categorized from theoretical and practical perspectives as below.

6.1.1 Theoretical Contributions

- It is the first time that the nonstationary queueing models have been utilized to analyse the impact of a surge generating event on an emergency department.
- The three-node network queueing model proposed in this work captures more details of the real system compared to the previous single nonstationary queueing models developed in literature for this purpose.
- The previous models assumed the ED with exponential service times. Whilst, in this work we considered a network of queueing models with general service time distributions.
- A time-varying simulation model was developed to replace the exact method (when it was not applicable) and utilized for comparing the approximation methods on more complex queueing models, based on which the best approximation method has been identified for surge modeling.

6.1.2 Practical Contributions

In addition to theoretical contributions, this work also bears practical merits, listed as follows:

- The detailed simulation model developed for ED of St. Pauls' Hospital has been employed for studying the effects of changes in the ED. This model has been applied by the managers of the ED to run many scenarios to answer what if questions. In addition, many alternative solutions were proposed for reducing patients waiting time based on the output of the simulation model for ED of St. Paul's hospital.
- Both queueing model and simulation have been applied to model surge in support of staff planning. The pros and cons of both methods are compared, which should be of help for future research and practice. Table 6-1 below presents a comparison between these models.

Table 6-1 Comparison of the queueing and simulation models

	Pros	**Cons**
Queueing model	- Easy to construct - Fast to test alternatives - Provides good estimates	- Difficult to solve analytically - Difficult to understand and grasp for practitioners - Assumptions are hard to validate - Very abstract; hard to capture details and flexibilities in the actual system
Simulation model	- Captures many details and flexibilities of the actual system - Versatile and can model virtually any system - Easy to analyse the output	- Time consuming to construct the model - Difficult to validate the model - Time consuming for experimenting with alternative solutions - Hard to define various measures of performance

6.2 Future work

Although we tried to capture more details of the system in our queueing model than the previous models in literature, there is a still a long way before we could fully represent all the important features of the actual system. The next step could be to incorporate more details in to this queueing model to better reflect reality. Of the most important features that could have a significant impact if employed, we can name,

- Abandonment

- Multiple servers type in each unit
- Priority in service

In addition, in order to be able to test the proposed solution of the queueing model we could modify the detailed simulation model and validate it with stochastic arrival rate. We could also record time varying performance measure in the simulation model rather than work with the averages over the total simulation time. This would help to study the effect of changes in staffing better.

APPENDICES

Table A- 1 Sample of the experiment done for comparing different approximation methods

μ	α	RA	λ	SCV	s1	s2	s3	s4	S-Pd 1	S-Pd 2	S-Pd 3	S-Pd 4	Q-Pd 1	Q-Pd 2	Q-Pd 3	Q-Pd 4	Time Ave Err	Max Err
2	0.1	0.4	4	1	4	4	4	4	0.189	0.199	0.144	0.141	0.201	0.201	0.148	0.148	0.038	0.063
		0.8	8	1	7	7	7	7	0.149	0.167	0.103	0.089	0.181	0.181	0.115	0.115	0.177	0.288
		1.6	16	1	13	13	11	11	0.102	0.121	0.186	0.160	0.130	0.130	0.189	0.189	0.138	0.279
4		0.8	8	2	4	4	4	4	0.221	0.231	0.201	0.178	0.201	0.201	0.148	0.148	0.163	0.264
		1.6	16	2	7	7	7	7	0.164	0.171	0.136	0.128	0.181	0.181	0.115	0.115	0.104	0.156
		3.2	32	2	13	13	11	11	0.118	0.131	0.208	0.187	0.130	0.130	0.189	0.189	0.051	0.098
8		1.6	16	4	4	4	4	4	0.211	0.223	0.188	0.184	0.201	0.201	0.148	0.148	0.138	0.213
		3.2	32	4	7	7	7	7	0.194	0.198	0.166	0.155	0.181	0.181	0.115	0.115	0.180	0.308
		6.4	64	4	13	13	11	11	0.144	0.148	0.241	0.235	0.130	0.130	0.189	0.189	0.157	0.215
2	0.2	0.4	4	1	4	4	4	4	0.189	0.199	0.144	0.141	0.201	0.201	0.148	0.148	0.038	0.063
		0.8	8	1	6	6	6	6	0.304	0.351	0.240	0.214	0.338	0.338	0.226	0.226	0.066	0.112
		1.6	16	1	11	11	10	10	0.293	0.343	0.338	0.288	0.338	0.338	0.311	0.311	0.082	0.153
4		0.8	8	2	4	4	4	4	0.221	0.231	0.201	0.178	0.201	0.201	0.148	0.148	0.163	0.264
		1.6	16	2	6	6	6	6	0.321	0.362	0.281	0.252	0.338	0.338	0.226	0.226	0.104	0.195

Various scenarios that are run for each approximation method are presented in the first five columns, the number of servers in each planning period and the probability of delay estimated with simulation and queueing model are presented in the following columns respectively. The last two columns represent the time average and maximum error .

Table A- 2 A sample of the data collected from detailed simulation model before and after applying number of staff suggested from queueing model. Times are in hours.

Before applying suggested staffing					After applying suggested staffing				
LOS	Triage	Fast Track	RAZ	Acute	LOS	Triage	Fast Track	RAZ	Acute
0.5440	0.2402	0.3382	0.0016	0.2252	0.5468	0.2402	0.3382	0.0016	0.0033
1.3421	0.3382	0.3209	1.2565	3.8897	1.3421	0.3382	0.3246	1.2565	0.0050
1.4387	0.3209	0.2402	1.6577	0.9867	1.8282	0.3209	0.2447	0.0106	0.0664
2.2351	0.0016	0.2645	0.0051	0.3306	1.9433	0.0016	0.2483	0.0151	0.9410
2.0612	0.2645	0.2839	0.0027	0.3436	2.2124	0.2645	0.2738	0.0046	0.3486
2.4960	1.2565	0.2141	0.0013	0.8223	0.6430	1.2565	0.2566	0.0014	0.7343
2.3452	0.2839	0.2150	0.0029	0.8415	0.9022	0.2839	0.3196	0.0024	0.8356
2.2243	0.2150	0.3301	0.0013	1.1117	2.4517	0.2150	0.2331	0.0001	1.1786
2.6480	0.2141	0.2411	0.9183	0.4984	0.5814	0.2141	0.2357	0.9193	0.4019
0.2958	0.2529	0.2388	0.2159	0.2700	4.9629	0.2529	0.2480	0.1934	0.2711
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
1.4372	0.3392	0.3392	1.0700	0.8707	1.4307	0.3420	0.3421	0.8195	0.4578

LOS (Length of stay of the patients in the ED), other columns are the time that it takes a patients to see a provider in the named unit. The last row is the average of the entire column which is not presented here.

REFERENCE LIST

[1] F. S. Hillier, G. J. Lieberman, and G. J. Liberman, *Introduction to operations research*. McGraw-Hill New York, 1990.

[2] *OECD Health Data 2010. How Does Canada Compare.* .

[3] K. Davis., *Mirror on the wall: An international update on the comparative performance of american health care*. Commonwealth Fund, 2007.

[4] Committee on the Future of Emergency Care in the United States Health System, *Hospital-Based Emergency Care: At the Breaking Point*. Washington, D.C.: The National Academies Press, 2007.

[5] D. F. Barbisch and K. L. Koenig, “Understanding surge capacity: essential elements,” *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*, vol. 13, no. 11, pp. 1098-1102, Nov. 2006.

[6] C. H. Schultz, K. L. Koenig, and E. K. Noji, “A medical disaster response to reduce immediate mortality after an earthquake,” *The New England Journal of Medicine*, vol. 334, no. 7, pp. 438-444, Feb. 1996.

[7] J. L. Hick et al., “Health care facility and community strategies for patient care surge capacity,” *Annals of Emergency Medicine*, vol. 44, no. 3, pp. 253-261, Sep. 2004.

[8] G. Kelen and M. Mccarthy, “The science of surge,” *Acad Emerg Med*, vol. 13, no. 11, pp. 1089-94, 2006.

[9] C. J. Bonnett et al., “Surge capacity: a proposed conceptual framework,” *The American Journal of Emergency Medicine*, vol. 25, no. 3, pp. 297-306, Mar. 2007.

[10] A. Estey, K. Ness, L. Saunders, A. Alibhai, and R. Bear, “Understanding the causes of overcrowding in emergency departments in the capital health region in alberta: A focus group study,” *Canadian Journal of Emergency Medicine*, vol. 5, no. 2, pp. 81-94, 2003.

[11] Kelton. D. Sadowski. R. P. Sturrock. D. T. , *Simulation with Arena*, 4th ed. Mc Graw Hill, 2007.

[12] Q. Wang and C. Chatwin, “Key issues and developments in modelling and simulation-based methodologies for manufacturing systems analysis, design and performance evaluation,” *The International Journal of Advanced Manufacturing Technology*, vol. 25, no. 11, pp. 1254-1265, 2004.

[13] L. V. Green, J. Soares, J. F. Giglio, and R. A. Green, “Using queueing theory to increase the effectiveness of emergency department provider staffing,” *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*, vol. 13, no. 1, pp. 61-68, Jan. 2006.

[14] A. M. de Bruin, G. M. Koole, and M. C. Visser, “Bottleneck analysis of emergency cardiac in-patient flow in a university setting: an application of queueing theory,” *Clinical and Investigative Medicine. Médecine Clinique Et Experimentale*, vol. 28, no. 6, pp. 316-317, Dec. 2005.

[15] H. Abujudeh, B. Vuong, and S. R. Baker, “Quality and operations of portable X-ray examination procedures in the emergency room: queuing theory at work,” *Emergency Radiology*, vol. 11, no. 5, pp. 262-266, Jul. 2005.

[16] L. G. Connelly, “Discrete Event Simulation of Emergency Department Activity: A Platform for System-level Operations Research,” *Academic Emergency Medicine*, vol. 11, no. 11, pp. 1177-1185, 2004.

[17] A. Bagust, M. Place, and J. W. Posnett, “Dynamics of bed use in accommodating emergency admissions: stochastic simulation model,” *BMJ*, vol. 319, no. 7203, pp. 155 -158, Jul. 1999.

[18] L. V. Green, P. J. Kolesar, and W. Whitt, “Coping with time-varying demand when setting staffing requirements for a service system,” *Production and Operations Management*, vol. 16, no. 1, pp. 13–39, 2007.

[19] B. Cleveland and J. Mayben, *Call center management on fast forward: succeeding in today's dynamic inbound environment*. Call Center Pr, 1997.

[20] O. B. Jennings, A. Mandelbaum, W. A. Massey, and W. Whitt, “Server staffing to meet time-varying demand,” *Management Science*, vol. 42, no. 10, pp. 1383–1394, 1996.

[21] L. Green and P. Kolesar, “The pointwise stationary approximation for queues with nonstationary arrivals,” *Management Science*, pp. 84–97, 1991.

[22] A. Ingolfsson, E. Akhmetshina, S. Budge, Y. Li, and X. Wu, “A Survey and Experimental Comparison of Service-Level-Approximation Methods for Nonstationary M(t)/M/s(t) Queueing Systems with Exhaustive Discipline,” *INFORMS Journal On Computing*, vol. 19, no. 2, pp. 201-214, Jan. 2007.

[23] Z. Feldman, A. Mandelbaum, W. A. Massey, and W. Whitt, “Staffing of Time-Varying Queues to Achieve Time-Stable Performance,” *Management Science*, vol. 54, no. 2, pp. 324-338, Feb. 2008.

[24] S. G. Eick, W. A. Massey, and W. Whitt, “Mt/G/{infty} Queues with Sinusoidal Arrival Rates,” *Management Science*, vol. 39, no. 2, pp. 241-252, Feb. 1993.

[25] D. Y. Sze, “A Queueing Model for Telephone Operator Staffing,” *Operations Research*, vol. 32, no. 2, pp. 229-249, Mar. 1984.

[26] W. Whitt, “What You Should Know About Queueing Models To Set Staffing Requirements in Service Systems,” 2007.

[27] S. Halfin and W. Whitt, “Heavy-Traffic Limits for Queues with Many Exponential Servers,” *Operations Research*, vol. 29, no. 3, pp. 567-588, May. 1981.

[28] L. V. Green, P. J. Kolesar, and J. Soares, “Improving the SIPP approach for staffing service systems that have cyclic demands,” *Operations Research*, pp. 549–564, 2001.

[29] Lawrence Brown et al., *Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective*. Wharton School Center for Financial Institutions, University of Pennsylvania.

[30] L. F. Shampine and M. W. Reichelt, “The MATLAB ODE Suite,” *SIAM Journal on Scientific Computing*, vol. 18, no. 1, pp. 1-22, Jan. 1997.

[31] L. Green, P. Kolesar, and A. Svoronos, “Some effects of nonstationarity on multiserver Markovian queueing systems,” *Operations Research*, vol. 39, no. 3, pp. 502–511, 1991.

[32] W. Whitt, *Stochastic-process limits: an introduction to stochastic-process limits and*

their application to queues. Springer, 2002.

[33] W. Whitt, “Planning Queueing Simulation,” *Management Science*, vol. 35, no. 11, pp. 1341-1366, 1989.

[34] R. W. Wolff, “Poisson Arrivals See Time Averages,” *Operations Research*, vol. 30, no. 2, pp. 223-231, Mar. 1982.

[35] D. P. Heyman and W. Whitt, “The Asymptotic Behavior of Queues with Time-Varying Arrival Rates,” *Journal of Applied Probability*, vol. 21, no. 1, pp. 143-156, Mar. 1984.

[36] R. K. Jain, *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. Wiley, 1991.

[37] T. Kimura, “Approximations for the delay probability in the M/G/s queue,” *Mathematical and Computer Modelling*, vol. 22, no. 10, pp. 157-165, Nov. .

[38] H. C. Tijms, M. H. V. Hoorn, and A. Federgruen, “Approximations for the Steady-State Probabilities in the M/G/c Queue,” *Advances in Applied Probability*, vol. 13, no. 1, pp. 186-206, Mar. 1981.

[39] P. Hokstad, “Approximations for the M/G/m Queue,” *Operations Research*, vol. 26, no. 3, pp. 510-523, May. 1978.

[40] R. P. Kittel and A. Palin, “Mercy Hospital: Simulation Techniques For ER Processes,” *Industrial Engineering*, vol. 24, no. 2, pp. 35-37, 1992.

[41] L. V. Green and J. Soares, “Computing time-dependent waiting time probabilities in M (t)/M/s(t) queuing systems,” *Manufacturing & Service Operations Management*, vol. 9, no. 1, pp. 54–61, 2007.

[42] S. G. Eick, W. A. Massey, and W. Whitt, “The Physics of the $M_t/G/\infty$ Queue,” *Operations Research*, vol. 41, no. 4, pp. 731-742, 1993.

[43] W. A. Massey and W. Whitt, “Peak congestion in multi-server service systems with slowly varying arrival rates,” 1995.

[44] Lawrence Brown et al., *Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective*. Wharton School Center for Financial Institutions, University of Pennsylvania.

[45] “Residual Analysis,” MathWorks, Source: on WWW at http://www.mathworks.com/help/toolbox/curvefit/bq_5ka6-1_1.html, visited on March 10, 2011.