The Offered-Load Process: Modeling, Inference and Applications

Research Thesis

Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science in Statistics

Michael Reich

Submitted to the Senate of the
Technion - Israel Institute of Technology

Sivan, 5772 Haifa June, 2012

Acknowledgements

This research thesis was written under the supervision of Professor Avisai Mandelbaum in the Faculty of Industrial Engineering and Managemen. I would like to thank him for close guidance through my research and for his goodwill and understanding.

An essential part of this work was done under the co-supervision of Professor Ya'acov Ritov. I would like to thank him for sharing with me his vast experience, and for the invaluable guidance and support throughout every stage of this work.

I would like to thank Professor Paul Feigin, Professor Haya Kaspi and Professor Malka Gorfine-Orgad for many helpful discussions.

Also I would like to thank my dear family for supporting and encouraging me all the way through. Special thanks to my beloved Rita, who strengthened me with her love, and to my son Itai.

Finally, the generous financial help of the Technion Graduate School is gratefully acknowledged.

Contents

List of Tables

List of Figures

Abstract

A standard assumption in queueing models is that the service time of customers and their patience are independent. Practice shows that this assumption is often violated, as one expects longer service times for customers who waited longer or, alternatively, higher patience for customers expecting longer services.

We introduce a model for the relationship between the service time and the patience of a customer, including a statistical test for the existence of such dependency. We show that in the presence of a relationship, various performance measures are influenced, and that the classical estimation procedures of service parameters yield biased estimators.

The above mentioned dependence affects the standard way of calculating offered-load. Specifically, the offered-load of a system must account for the work that would have been required by customers who abandon prior to service. We thus carefully define the offered-load process and the offered-load function, and then derive a method for estimating and predicting them, in service systems where service-times and patience are dependent. Finally, we discuss the effect of this dependence on the square-root staffing rule.

List of Symbols

λ	Arrival rate function
A	Arrival process
μ	Service rate
n	Number of agents
S	Service time
f_S	Service time density
G	Service time distribution
G^c	Survival function of service time ($G^c = 1 - G$)
$P\{Ab\}$	Abandonment probability
W	Waiting time
V	Offered waiting time
τ	Patience time
f_τ	Patience time density
h_τ	Hazard rate of patience time
L	Offered-Load process
R	Offered-Load function

E	Expectation
Var or σ^2	Variance
$\mathbb{1}_{\{\}}$	Event indicator
$\sim$	Distributed as
Φ	The standard normal *distribution* function
ϕ	The standard normal *density* function
h	The hazard rate of the standard normal distribution
$f(x) \to a$	$\lim_{\lambda \to \infty} f(x) = a$

List of Acronyms

ANOVA	Analysis Of VAriance
cdf	cumulative distribution function
CV	Coefficient of Variation
ED	Efficiency Driven
FCFS	First Come First Served
iff	if and only if
iid	independent and identically-distributed
ISA	Iterative Staffing Algorithm
MLE	Maximum Likelihood Estimator
PASTA	Poisson Arrivals See Time Averages
pdf	probability density function
PSA	Pointwise Stationary Approximation
QD	Quality Driven
QED	Quality and Efficiency Driven
QoS	Quality of Service
wp	with probability

Chapter 1

Introduction

During the last century and the beginning of the current one, the service sector has grown significantly and now accounts for approximately 70% of the national income in the United States. The service sector covers a wide spectrum of activities, e.g. education, professional services, financial services and government services. In this thesis we focus on telephone call centers and on the health care system.

Call centers are very commonly used by companies and organizations for managing their customer relationships. This covers both the public and private sector. For some of the companies, such as banks and cellular operators, their call centers are the main channel for maintaining contact with their customers. In general, call centers are becoming a vital part of the service-driven society nowadays. As a result call centers have also become an object for academic research.

For the analysis of call centers operations, queueing theory and statistics are being used. The problems that call centers operators are dealing with, are often related to statistical characteristics of the calls arrival and handling processes. Very often, about 70% of the operations costs are devoted to human resources. As a result, forecasting the calls arrival rate, understanding the handling process, setting the right service performance measures and staffing levels, are central problems at all call centers. These are issues that any call center manager deals with on a daily basis.

Another important area of the service sector is the health care system. Hospital managers are increasingly aware of the need to use their resources as efficiently as possible

in order to continue to assure their institutions' survival and prosperity. Moreover, there is consistent pressure, coming from patients, to increase the quality of care by technical improvements and medical innovations and to shorten the sojourn time in the hospital. Green [14] describes the general background and issues involved in hospital capacity planning and provides examples of how Operations Research models can be used to provide important insights into operational strategies and practice.

Satisfactory customer service can be defined in many ways, based on various performance measures. Focusing on operational measures, a customer enjoys satisfactory service if his delay in queue is at most τ seconds [25]. For an emergency medical service system, the main performance measure is the fraction of calls that are being held within some time standard. The response time is typically considered to begin at the instant the call was made and end when an ambulance reaches the call address. In North America, a typical target is to reach 90% of the most urgent calls within 9 minutes [16]. The measures of quality are not absolute. They differ among service system in accordance with the system's goals, its environment and the services it offers.

The staffing problem is a key for providing satisfactory service. Managers of call centers must decide on how many agents to hire and how many to login at any time in order to ensure satisfactory customer service. In emergency medical services and in hospitals the system is even more complicated. Staffing consists of scheduling decisions for ambulances and their crews, for doctors and nurses and it must also account for the number of inpatient beds.

Service environments are typically very complex. Some of the complexity factors are, for example: changes of the environmental parameters (arrival/service rates), uncertainty of these parameters due to either random variation over time or lack of information, variety of services with different requirements (i.e. time and skills). Naturally, staffing decisions must account for all this complexity, yet the challenge is to develop staffing rules that are simple and insightful enough for implementation. For example, the "square-root staffing" rule is such a rule, as will be surveyed below.

Mathematical models have been developed to model the complexity of the call center environment. Their strength is their simplicity and the theoretical insights they provide. On the other hand, their modeling scope is limited, and an analytical knowledge is needed in order to apply them efficiently. These weaknesses could be the reason why such models

are not being used as often as expected. The most commonly-used models are the $M/M/n$ queue (Erlang-C) and $M/M/n+M$ queue (Erlang-A), where the first M denotes that the arrivals to the queue are according to a time-homogenous Poisson process, the second M stands for an exponentially distributed service-time, the n represents the staffing level and the last M in the Erlang-A model denotes that the patience of customers is modeled by an exponential distribution (In Erlang-C, customers are assumed to have infinite patience). A more realistic call-center model is the $M_t/GI/n_t+GI$ queue. In this model, the arrival process need not be homogenous over time (denoted by the subscript t) and the service time and patience time are allowed to be drawn independently from a general distribution (denoted by GI). These models are described in Chapter 2.

A standard assumption in service system modeling is that the service-time and the patience of customers are independent. Practice suggests that this assumption is often violated. For example, customers who expect longer service times are likely to be willing to wait longer before abandoning the queue. In this research, we model queues where the patience and the service time are *dependent*, and explore the influence of such dependency on staffing rules and other associated performance measures.
We shall also propose methods to estimate and predict the offered-load process and the offered-load function in the presented queueing models, focusing on time-dependent queues and queues that face a dependency between patience and service time.

Computer simulation presents an alternative to mathematical analysis of queueing models, such as those considered above. When created and used properly, simulation models can handle almost any model complexity and take into account the very small details. But there are limitations as well - the development and maintenance of simulation models are expensive, and it can take hours to run a simulation even with todays computing capabilities. A research trend in which mathematical models and simulation have been combined has recently emerged, in order to overcome the weaknesses and utilize the advantages of the two methods. For examples of models where the patience and the service time are dependent, that are not tractable mathematically, see Chapters 6 and 7.

1.1 Contribution and Structure of the Thesis

In Chapter 2 we provide the theoretical background and survey some related literature. Section 2.1 introduces the classical Erlang-C and Erlang-A models, and Section 2.1.3

summarizes relevant asymptotic results of the $M|M|n+G$ queue. These results will be used later in our theoretical analysis. In Section 2.2 we provide a literature review.

We begin our research in Chapter 3, with the definition of two central terms: these are associated with the analysis of the required amount of resources at any time, to satisfy a predetermined service level. The first term is the **Offered-Load Process**, which is a stochastic process, representing the amount of work that is being processed in a service system at any time $t \geq 0$, in an ideal world, where any customer/task enters service right upon arrival to the system (this can be achieved by employing ample, or infinitely many resources). The second term is the **Offered-Load Function**, which is simply a function of time $t \geq 0$, representing the average of the offered-load process at time t. Both the offered-load process and the offered-load function depend on the arrival process of tasks to the resource, and on the corresponding service times.

In Section 3.1 we introduce the $M_t/GI/n_t+GI$ queue, and summarize mathematical results of the offered-load process and the offered-load function that are related to this queue. We review interesting insights through the achieved mathematical expressions. We also describe known methods and rules for staffing the $M_t/GI/n_t+GI$ queue.

In Chapter 4, we derive a model for the relationship between patience and service-time. One elementary assumption of the $M_t/GI/N_t+GI$ queue is that the queue's building blocks are independent, in particular service times and patience. Experience shows that this assumption is often violated when analyzing empirical data of real service systems, especially in those coping with human customers. This is demonstrated in Figure 7.1 in Chapter 7, which describes, for a retail service line in a call center of a large North-American commercial bank, the mean service-times of served customers, as a function of their waiting-times. One observes in this figure that the mean service time is changing over different values of waiting time.

In common analysis of service systems, the service time distribution is inferred through the distribution of only served customers. From biased sampling, it is obvious that, due to abandonment, the number of served customers with longer patience is expected to be higher than the number of served customers with shorter patience. Now, assume that the patience and the service time are positively correlated. In this case, the regular method for estimating the average service time yields an upward biased estimator, since one also tends to observe more longer service times.

We denote an $M_t/GI/N_t+GI$ with service times that depend on patience by $M_t/GI^*/N_t+$

GI: the * sign in the service time component stands for the fact that the service time distribution is dependent on patience. Our proposed model connects between the mean service-time of served customers and the overall mean service time, as a function of the patience of customers, as follows:
Denote,

- S - Service time.
- τ - (Im)patience time.
- W - Waiting time.
- $g(w) = E(S|\tau > W, W = w)$ - Mean service time of served customers who waited exactly w time units.
- h_τ - Hazard rate function of the patience.

Then, we show that the mean service time of customers with patience w is given by

$$E(S|\tau = w) = g(w) - \frac{g'(w)}{h_\tau(w)},$$

where $g'(w)$ is the derivative of $g(w)$ with respect to w.

In section 4.2 we develop a statistical test to examine if the patience and the service-time can be assumed to be independent, applying a permutation test.

In Chapter 5.1 we introduce methods to estimate the offered-load process and the offered-load function, for a variety of queueing models that are described throughout this work. Special attention is given to models where customers abandon, especially in the $M_t/GI/N_t+GI$ and $M_t/GI^*/N_t + GI$ queues. We also consider systems where the service-time distribution of customers may differ according to their arrival times.

In Section 5.2 we review work that is related to the prediction of the offered-load process. The prediction can be performed off-line, for example, one day in advance (interday), or it can be updated dynamically during the day, exploiting up-to-date system information (intraday). Most research in the field of predicting the offered-load is focused on predicting the arrival process. Then, approximations of the offered-load are made, based on the distribution of the service time. In service systems that face high variability in the demand, these approximations sometimes perform poorly. Based on work of Y. Goldberg

[32], we recommend to predict the offered-load directly. In this approach, the input for the prediction model shall be the offered-load series itself, instead of the arrivals (see [32] for the details).

In Chapter 6 we show that various performance measures of service systems are influenced by the relationship between patience and service time, since *the service time distribution that the system faces may vary as the fraction of abandoning customers is changing, which in turn is influenced by the staffing level*. In particular, for a fixed staffing level, the quality of service (e.g. the probability of waiting) is affected by that relationship, or alternatively, the staffing rule, set to achieve a specific quality of service, must be modified.

In Chapter 7 we apply our theoretical results to actual data from a call center - a retail banking service in a large North-American commercial bank. We also investigate some properties of our derived models through simulated data.

We conclude in Chapter 8 with several suggestions for future research.

Chapter 2

Theoretical Background and Literature Review

2.1 Markovian Birth & Death Queues

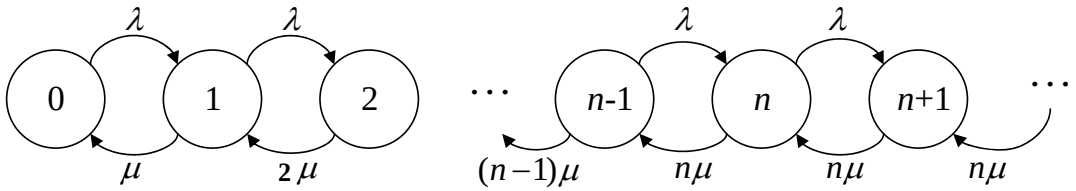
In this section, the two most common models that are used for call centers modeling and staffing are presented. The first model is Erlang-C ($M/M/n$), developed around 1910 by Erlang [3]. Its main deficiency is that it ignores customers' impatience. Impatience leads to the phenomenon of customers' abandonment, and, already around 1940, Palm [6] developed Erlang-A ($M/M/n+M$), which assumes exponential patience times, in order to capture it.
In Section 2.1.3 we present the more general $M/M/n+G$ model that allows an arbitrary distribution of patience times G (rather than the exponential one in $M/M/n+M$). $M/M/n+G$ will be the central model in our research and we shall use it to motivate four operating regimes for medium-to-large call centers: one which emphasizes service quality, another that focuses on operational efficiency, a third that carefully balances these two goals of quality and efficiency, and a fourth that is a refinement of the second by the third.

2.1.1 Erlang-C

The classical $M/M/n$ queueing model is characterized by Poisson arrivals at rate λ, iid exponential service times with an expected duration $1/\mu$, and n servers working independently in parallel. One can view this model as a simple Birth-and-Death process where λ

is the constant birth rate and μ is the constant death rate. A Markovian-state description of the process is the total number of customers in the system, either served or queued. The corresponding transition rate diagram is then the following:

Figure 2.1: Erlang-C as a Birth & Death Process.



Erlang-C is ergodic if and only if its traffic intensity $\rho_n = \dfrac{\lambda}{n\mu} < 1$; ρ_n is then the servers' utilization, namely the long-run fraction of time when a server is busy.
The stationary/limit distribution is defined as

$$\pi_j \stackrel{\triangle}{=} \lim_{t\to\infty} P\{L(t) = j\}, \quad j \geq 0,$$

where $L(t)$ is the system state at time t, namely, the total number of customers in the system.
The probability that, in steady-state, all the servers are busy is given by $\sum_{j\geq n} \pi_j$, the stationary probability of being in one of the states $\{n, n+1, \ldots\}$. This probability is sometimes referred to as the *Erlang-C formula*. It is denoted $E_{2,n}$ and is given by

$$E_{2,n} = \sum_{j\geq n} \pi_j = \frac{(\lambda/\mu)^n}{n!(1-\rho)} \cdot \left[\sum_{j=0}^{n-1} \frac{(\lambda/\mu)^j}{j!} + \frac{(\lambda/\mu)^n}{n!(1-\rho)}\right]^{-1}. \tag{2.1}$$

The Poisson distribution of arrivals has an important and useful consequence, known as PASTA (Poisson Arrivals See Time Averages) [31]: it implies that the probability $E_{2,N}$ is in fact also the probability that a customer is delayed in the queue (as opposed to being served immediately upon arrival).
Using formula (2.1) and other Erlang-C formulae, one can calculate staffing levels n for any desired service level, given the arrival rate and average service time. This can be easily done even with a spreadsheet. However, such a solution does not provide any insight on the dependence of n on model parameters; for example, how should n change if the load was doubled. Such insight comes out of a staffing rule that goes back to as early as Erlang

[3], where he derived it via marginal analysis of the benefit of adding a server. (Erlang indicated that the rule had been used in practice since 1913.) This is the *square-root safety-staffing rule*, which we now describe.

Let $R = \lambda/\mu$ denote the average *offered load*, which is defined as the amount of work, measured in time-units of service, that arrives to the system per unit of time. Then the square root safety-staffing rule states the following: for moderate to large values of R, the appropriate staffing level is of the form

$$n = R + \beta\sqrt{R} + o(\sqrt{R}), \tag{2.2}$$

where β is a *positive* constant that depends on the desired level of service; β will be referred to as the *Quality-of-Service* (QoS) parameter: the larger the value of β, the higher is the service quality. The second term on the right side of (2.2) is the excess (safety) capacity, beyond the nominal requirement R, which is needed in order to achieve an accepted service level under stochastic uncertainty.

The form of (2.2) carries with it a very important insight. Let $\Delta = \beta\sqrt{R}$ denote the safety staffing level (above the minimum $R = \lambda/\mu$.) Then, if β is fixed, an m-fold increase in the offered load R requires that the safety staffing Δ increases by only $\sqrt{m}$-fold, which constitutes significant economies of scale.

Theorem 2.1 (Halfin and Whitt [22]) *Consider a sequence of $M/M/n$ queues, indexed by $n = 1, 2 \ldots$. As the number of servers n grows to infinity, the square-root safety-staffing rule applies asymptotically if and only if the delay probability converges to a constant α $(0 < \alpha < 1)$, in which case the relation between α and β is given by the* ***Halfin-Whitt function***

$$\alpha = \left[1 + \frac{\beta}{h(-\beta)}\right]^{-1}. \tag{2.3}$$

Here $h(x) = \phi(x)/(1-\Phi(x))$ is the hazard rate of the standard normal distribution N(0,1).

Remark 2.1 *Exact performance measures and queue characteristics should be indexed by n or λ. As a rule, we omit this indexing. All asymptotic results in this chapter are valid given $n \to \infty$ (or, equivalently, $\lambda \to \infty$).*

Note that (2.2) applies if and only if $\sqrt{n}(1-\rho)$ converges to $\beta > 0$. Indeed, formally the

Halfin-Whitt Theorem states:

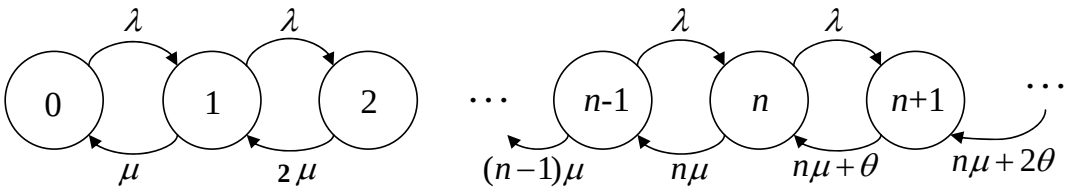
$$\begin{aligned}&\text{As } n \uparrow \infty, \quad P\{W_q > 0\} = E_{2,n} \to \alpha \qquad (0 < \alpha < 1)\\ &\text{iff } \sqrt{n}(1-\rho) \to \beta \qquad\qquad (0 < \beta < \infty)\\ &(\text{equivalently} \quad n \approx R + \beta\sqrt{R}).\end{aligned}$$

Here, W_q denotes the actual waiting time (in steady state) and $P\{W_q > 0\}$ denotes the delay probability.
In practice, this rule makes the life of a call-center manager easier: he or she can actually specify the desired delay probability α and achieve it by following the square-root safety staffing rule (2.2), simply by choosing the corresponding β, via Halfin-Whitt function (2.3).

2.1.2 Erlang-A

Trying to make the $M/M/n$ model more realistic and useful, the following assumption is added: each customer has limited patience, that is, as the waiting time in the queue grows the customer may abandon. We assume that customers patience times τ are iid $\exp(\theta)$, across customers. If a customer did not receive service till τ then abandonment occurs.
This model is referred to as Erlang-A ('A' for Abandonment) and denoted by $M/M/n{+}M$. It is also a Birth and Death process, and its transition rate diagram is depicted below. Let $P\{Ab\}$ denote the probability to abandon. The Erlang-A analogue of Theorem 2.1

Figure 2.2: Erlang-A as a Birth & Death Process.



was proved by Garnett, Mandelbaum and Massey [18], and is formulated as follows:

Theorem 2.2 (Garnett, Mandelbaum and Massey [18]) *Consider a sequence of $M/M/n+M$ queues, indexed by $n = 1, 2 \ldots$. As the number of servers n grows to infinity, the square-root safety-staffing rule (2.2) applies asymptotically if and only if the delay probability converges to a constant α, $0 < \alpha < 1$, in which case the relation between α and β is given by the* ***Garnett function***

$$\alpha = \left[1 + \sqrt{\frac{\theta}{\mu}} \cdot \frac{h(\beta\sqrt{\mu/\theta})}{h(-\beta)} \right]^{-1}, \quad -\infty < \beta < \infty. \tag{2.4}$$

Moreover, the above conditions apply if and only if $\sqrt{n}P\{Ab\}$ converges to some positive constant γ that is given by

$$\gamma = \alpha\beta \left[\frac{h(\beta\sqrt{\mu/\theta})}{\beta\sqrt{\mu/\theta}} - 1 \right].$$

Formally, Theorem 2.2 reads:

$$\begin{aligned} &\text{As } n \uparrow \infty, \quad P\{W_q > 0\} \to \alpha \qquad (0 < \alpha < 1) \\ &\text{iff } \sqrt{n}(1-\rho) \to \beta \qquad\quad (-\infty < \beta < \infty) \\ &(\text{equivalently} \quad n \approx R + \beta\sqrt{R}) \\ &\text{iff } \sqrt{n}P\{Ab\} \to \gamma \qquad\quad (0 < \gamma < \infty). \end{aligned}$$

An important feature of Erlang-A is that, unlike Erlang-C, it is always stable whenever the abandonment rate θ is positive.
Theorem 2.2 demonstrates that the square-root staffing rule prevails for Erlang-A as well. The QoS parameter β now depends on both the abandonment rate θ and the delay probability α. It is significant that here β may take also negative values (since Erlang-A is always stable), in which case the system exhibits *diseconomies* of scale [26].

2.1.3 Asymptotic Operational Regimes: QED, QD and ED

Each organization has its own preferences in everyday functioning. Some try to get the most from the available resources, while others see customers' satisfaction as the most important target. Depending on organizational preferences, three different operational regimes arise:

1. Efficiency driven (ED).

2. Quality driven (QD).

3. Quality-Efficiency driven (QED).

A fourth regime turns out useful as well: introduced in [1], it is a QED refinement of the ED regime.

For moderate to large call centers, these regimes can be formally characterized by relating the number of servers to the offered load.

1. Efficiency Driven (ED) Regime

The efficiency driven regime is characterized by very high servers utilization ($\sim 100\%$) and relatively high abandonment rate (over 10%). In the ED regime, the offered-load $R = \lambda/\mu$ is noticeably larger than the number of agents N. This means that the number-in-system would explode unless abandonment take place. The formal characterization of the ED regime is in terms of the following relationship between n and R :

$$n \;=\; R(1-\gamma) + o(R), \qquad 0 < \gamma < 1\,, \tag{2.5}$$

where γ is a QoS parameter: a larger value of γ implies longer waiting times and more abandonment.

2. Quality Driven (QD) Regime

In the quality driven regime, the emphasis is given to customers' service quality. This regime can be characterized by relatively low server utilization (for large call centers below 90%, and for smaller ones around 80% and perhaps less), which leads to very low abandonment rates. The formal characterization of the QD regime is in terms of the following relationship between n and R :

$$n \;=\; R(1+\delta) + o(R), \qquad \delta > 0. \tag{2.6}$$

3. Quality and Efficiency Driven (QED) Regime

The QED regime is the most relevant for call centers operation. It combines relatively high utilization levels of servers (around 95%) and low abandonment rates (1%-3%). For more information and motivation regarding the QED regime, see [17].

Let $R = \lambda/\mu$ denote the Offered load. The number of servers in this regime is given by the square-root staffing rule:

$$n = R + \beta\sqrt{R} + o(\sqrt{R}), \quad -\infty < \beta < \infty\,, \tag{2.7}$$

where β is the QoS parameter. Note that in Erlang-A β can take negative values, while in the case of Erlang-C it is restricted to be positive in order to ensure stability.

4. ED+QED Regime

As described in [1], this operational regime combines the staffing rules of the ED and QED regimes. It arose from the need to accommodate the constraint $P(W > T) \leq \alpha$, assuming that T is of the order of the service time and α is not too close to 0 or 1. In this regime, the number of servers is characterized by the following formula:

$$N \approx R\,(1-\varepsilon) + \beta\,\sqrt{R} + o(\sqrt{R}), \quad 0 < \varepsilon < 1,\ -\infty < \beta < \infty\,. \tag{2.8}$$

One observes that ED+QED staffing amounts to QED fine-tuning of ED staffing.

2.2 Literature Review

Most service systems face a time-varying environment, such as a demand for service that varies over time or a service time distribution which may differ at different time periods (e.g. due to different types of customers over different day hours). Eick, Massey and Whitt, review in [21] the properties of the $M_t/G/\infty$ queue. Some of these results are given in Chapter 3.

Feldman et al. [35] consider the staffing problem under time-varying demand ($M_t/G/n_t+G$ systems). They demonstrate, via simulations, that in order to achieve time stable operational performance in the face of general arriving rates, it is enough to target for a stable delay probability which, in fact, produces stable performance of several other operational measures. More details on [35] are given in Chapter 6.

Here we survey several papers that cope with the prediction of the demand for service in time-varying service systems.

"Dynamic Staffing in a Telephone Call Center Aiming to Immediately Answer All Calls" by Whitt W.[29]

All the queueing models, that are described previously in this chapter, deal with a long term planning managing of service systems, based on the steady-state performance of the system. The main idea in dynamic staffing is to dynamically use all available information up to the current time in order to more accurately predict the demand in a relatively short time t in the future. In this paper, Whitt is aiming to achieve the goal of immediately answering all calls. Thus, in the analysis, he assume that all calls are indeed immediately answered. Of course, actually answering all calls immediately may be an unrealistic objective, but it is often possible to come very close to this goal by dynamically staffing based on recently updated information. The arrival process is assumed to be a nonhomogeneous Poisson process, and the call service times are assumed i.i.d..

The future demand at time t is divided into two different components: those who are currently in system and will stay at least until time t (denoted by $C(t)$), and those who will come before time t and will stay after it (denoted by $N(t)$). Under the assumptions of the model, it is reasonable to regard these future demand components as independent. Thus, each component is analyzed separately.

Whitt shows, that for a relatively large of number of calls, each of the components is approximately normally distributed. He suggests a method to describe the mean and variance of each component. Then, the overall mean and variance of the total future staffing requirement will be the sums of the components' means and variances. Let the total future demand be the sum of the two components: $D(t) = C(t) + N(t)$, $t \geq 0$. Whitt offers to let the required number of servers at time t be
$s(t) = \lceil ED(t) + Z_\alpha \sqrt{VarD(t)} + 0.5 \rceil$, where $P(N(0,1)) > Z_\alpha = \alpha$ and $\lceil x \rceil$ is the least integer greater than x.

As the time interval between staffing changes decreases (and respectively the lead time for the prediction is smaller), with respect to the average service time, then the significance of current calls in the total future demand increases. From the other point of view, as the average service time increases, the current number of calls in the system becomes more significant for estimating future demand. The current calls in progress may also provide

a useful information for predicting their remaining service times (according to their class and service time distribution).

"Bayesian Forecasting of an Inhomogeneous Poisson Process With Applications to Call Center Data" by Weinberg J., Brown L.D. and Stroud J.R.[12]
Weinberg et al. propose a multiplicative effects model for estimating and forecasting Poisson arrival rates for short intra-day time intervals, with a 1-day lead time. In their setting, the call arrival rate for a given time interval of a particular day of the week is modeled as the product of the forecasted volume for that day of the week and the proportion of calls that arrive in that time interval plus a random error.

The authors assume that the daily patterns behave according to an autoregressive model. To estimate the autoregressive model's parameters, they used Bayesian techniques, proposing a set of prior distributions, and using a Monte Carlo Markov Chain (MCMC) simulation to estimate the parameters of the posterior distribution. They also suggest an efficient method to adopt new data that arrive throughout the day in order to update the continuation of the current day forecast.

Shen and Huang develop in [11] another autoregressive statistical model for forecasting call volumes for each interval of a given day, using a singular value decomposition to achieve a substantial dimensionality reduction. They also provide an extension of their model to account for intra-day forecast updating.

"Forecasting Demand for a Telephone Call Center: Analysis of Desired versus Attainable Precision" by Aldor-Noiman S.[19]
Aldor-Noiman suggests a method for prediction of the offered-load in a call center, by predicting the arrival process and the average service time separately. The work focuses on predicting the number of calls that reach the call center based on mixed models approach, which enables incorporating exogenous variables such as billing cycles indicators. In this work, an auto-regressive process is also considered for modeling an intra-day and inter-day correlations in the arrival process.

The results of the mixed model are compared to other models, such as the Bayesian model developed by Weinberg et al. in [12]. It was shown that the results have a similar level of precision though they are based on a significant lower number of learning data. The results of this thesis were published in [20].

Chapter 3

The Offered-Load

An important aspect of this work deals with the analysis of the required amount of resources at any time, to satisfy the demand for work in a service system, at a predetermined service level. In order to model this problem, one must first understand the structure and the components of the analyzed system: the available type of resources, their skills, the number or amount of available resources of each type, the tasks that reach the system, the sequence of tasks performed by the resources, and the operational information associated with each task and resource.

We refer here to the following definitions:

1. **Resource-k Offered-Load Process** - A stochastic process, representing the amount of work being processed by resource k at time t, under the assumptions of infinitely many resources of type k, and that a task that reaches resource k, enters immediately to process.

2. **Resource-k Offered-Load Function** - A function of time $t \geq 0$, representing the average of Resource-k Offered-Load Process at time t.

As shown later in this chapter, these last two definitions play a central role in the world of analyzing, understanding and staffing of service systems.

It is obvious that both resource-k's offered-load process and function depend on the arrival process of tasks to resource k, and on their corresponding service times. It is important to notice that, given a service network, the arrival process to resource k and the service

times of the tasks are likely to depend on the performance of other resources and the flow of other tasks in the system. For example, consider a simple case where customers reach a servers pool according to a homogenous Poisson arrival process, with a rate λ. Each customer has an exponential service time with rate μ, and after completing service, the customer leaves the system with probability p or, alternatively seeks for an additional independent service, again exponentially distributed with rate μ, from the same servers pool with probability $1-p$, as demonstrated in Figure 3. Then, this procedure repeats, independently on the past. In this example, the *effective* arrival rate to the servers pool is actually λ/p, since $\lambda_{eff} = \lambda + \lambda \cdot (1-p) + \lambda \cdot (1-p)^2 + \ldots$.

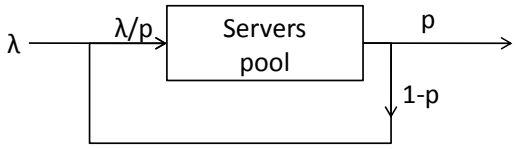


Figure 3.1: The arrival rate of new customers is λ. Each customer asks for additional service with probability $1-p$. The effective arrival rate to the servers pool is λ/p.

In systems that involve a complex service network, or parameters that change over time, this analysis is expected to be much more complicated. In such systems, we propose that the calculation of the offered-load, at station k, is to be performed by allocating the number of resources in the explored resources pools to be infinity, while keeping the rest of the system fixed.

3.1 The $M_t/GI/N_t + GI$ Queue

In this section, we introduce the offered-load process and the offered-load function of the $M_t/GI/N_t + GI$ queue. Here M_t indicates that the arrival process is assumed to be non-homogenous Poisson with arrival rate $\lambda(t)\,, t \geq 0$. The first GI indicates that the service times are iid with cdf G. The N_t notation indicates that the number of servers (possibly infinity) can be varied over time. Finally, the last GI indicates that each customer has patience with general cdf, iid across customers, independent of the system's state. This $M_t/GI/N_t + GI$ model is likely to be much more appropriate, than the stationary queues discussed in Chapter 2, for approximating real life service system.

In stationary models, we have shown in Chapter 2 that the staffing level is determined by the offered-load R, which is defined as the arrival rate divided by the service rate. Considering the $M/GI/\infty$ model in steady state, this is exactly the average number of customers in service at any time t, which is consistent with the definition of the offered-load at the beginning of this chapter. However, while the $M_t/GI/N_t + GI$ model is mathematically intractable, our goal is still to determine the required staffing level at time t, in a time varying manner, in order to achieve a specific service performance.

A common approach to handle time varying arrival rates is to use a *Pointwise Stationary Approximation* (PSA). This method provides a time dependent description of performance based on the steady-state behavior of a corresponding stationary model, using the parameters that prevail at the time at which we carry out the analysis. For example, we can approximate the performance of $M_t/GI/N_t + GI$ queue at time t by the steady-state performance in the associated stationary $M/GI/s + GI$ queue, with the same service time and patience distributions, but with a constant arrival rate and number of servers equal to the values $\lambda(t)$ and N_t respectively. Clearly, a steady-state assumption could be problematic when the arrival rate changes over time, especially during time periods when the variation of $\lambda(t)$ is large. Indeed, in practice, PSA often seems to be inappropriate, as demonstrated in [36].

It follows from [36] that a first step to solve the staffing problem for time-varying systems is to understand better the offered-load at time t. To this end, we introduce the $M_t/GI/\infty$ model. This model differs from the previous one by having infinitely many servers, which means that each customer who joins the system is facing a ready-to-answer server and does not need to wait. Consequently, there is no need to consider the patience of customers in this model. In our study, we use the $M_t/GI/\infty$ model for several reasons: First, it is remarkably tractable, as will be shown later. Second, one can use the $M_t/GI/\infty$ model to approximate a system that aims at immediately answering all calls or to analyze the level of required capacity. Moreover, we can use this model to get an upper bound on the performance that could be achieved if the staffing level was as high as needed. But most importantly, as explained momentarily, analyzing $M_t/GI/\infty$ yields the "right" definition of the offered-load for $M_t/GI/N_t + GI$.

Applying the above assumptions to $M_t/GI/\infty$, the number of arrivals in the time interval $[a, b]$ has a Poisson distribution with mean $\int_a^b \lambda(u)du$. In the special case where $\lambda(t) = \lambda$, the steady-state number of busy servers has a Poisson distribution with mean $\lambda \cdot E[S]$, independently of the service time distribution beyond its mean (insensitivity). Immediately

we show that, under a time-varying rate $\lambda(t)$, the number of busy servers at time t, $L(t)$, also has a Poisson distribution with a time-varying mean, $R(t) = E[L(t)]$. The measure $R(t)$ is again interpreted as the offered-load at time t, and will be discussed extensively in the next section.

3.1.1 The Offered-Load of $\mathbf{M_t/GI/N_t + GI}$

For the $M_t/GI/N_t + GI$ queue, the *offered-load* $R = \{R(t),\, t \geq 0\}$ is given by $R(t) = E[L(t)]$, where $L(t)$ is the number of customers in service (number of busy servers) at time t, in a corresponding $M_t/GI/\infty$ queue (same arrivals and services). The stochastic process $L = \{L(t),\, t \geq 0\}$ is by definition the *offered-load* process.

In order to describe the offered-load process formally, we introduce the following notations:

- $0 = t_0 < t_1 < t_2 < \ldots$ - Increasing sequence of times, where t_i represents the arrival time of the i^{th} customer.
- S_i - Service time of the i^{th} customer. $S_i \in (0, \infty)$, $i = 1, 2, \ldots$.
- $A = \{A(t),\, t \geq 0\}$ - Arrival process. We define $A(0) = 0$. Then $A(t) = max\{n; t_n \leq t\}$. Recall that in the $M_t/GI/N_t + GI$ queue, we assume that the arrival process is a non-homogenous Poisson process.

We also assume that the arrival process after time t, and the service times of the corresponding customers, are independent of the system's state at time t (i.e. not influenced by any information on the customers who are currently in service). From the definition of $L(t)$, it becomes the count of all customers who arrived before time t and that, under the assumption that they were directed to service immediately upon their arrival, are still in service at time t. Thus,

$$L(t) = \int_0^t \mathbb{1}_{\{S_u > t - t_u\}} dA(u) = \sum_{i=1}^{A(t)} \mathbb{1}_{\{S_i > t - t_i\}} \tag{3.1}$$

The following theorem provides three representations for $R(t)$, in terms of the service time distribution. They are proved in [27], and we re-derive them here:

Theorem 3.1 *For any time t, $L(t)$ has a Poisson distribution with mean*

$$R(t) = E[L(t)] = E[\lambda(t - S_e)] \cdot E[S] = E\left[\int_{t-S}^{t} \lambda(u)du\right] = \int_{-\infty}^{t} \lambda(u) \cdot [1 - G(t-u)]du, \tag{3.2}$$

where
S is a generic service time;
S_e is a generic excess service time, namely with the following cdf:

$$G_e(t) = P(S_e \leq t) = \frac{1}{E(S)} \int_0^t [1 - G(u)]du, \qquad t \geq 0. \tag{3.3}$$

Proof: In order to prove the theorem, we shall prove the following lemmas, each holding for any t:

Lemma 3.1 *: $L(t)$ is Poisson distributed with mean*

$$R(t) = \int_{-\infty}^{t} \lambda(u) \cdot [1 - G(t-u)]du\,.$$

Lemma 3.2 *:* $\int_{-\infty}^{t} \lambda(u) \cdot [1 - G(t-u)]du = E\left[\int_{t-S}^{t} \lambda(u)du\right].$

Lemma 3.3 *:* $\int_{-\infty}^{t} \lambda(u) \cdot [1 - G(t-u)]du = E[\lambda(t - S_e)] \cdot E[S]\,.$

The combination of these three lemmas establishes Theorem 3.1.

Proof of Lemma 3.1: Let (x, y) be a coordinate that represent an arrival at time x with service time y. Prekpoa showed in [2] that the number of points in any finite collection of disjoint rectangles, on the two dimensional system of arrival time against service time, are independent Poisson variables (In other words, the collection of such points $(x, y) \in \mathbb{R}^2_+$ is a Poisson Point Process). The offered-load process at time t, $L(t)$, which is the set of all points with $x \leq t$ and $x + y \geq t$, is Poisson distributed with mean

$$R(t) = E[L(t)] = \int_{u=-\infty}^{t} \lambda(u) \cdot P(u + S > t)du = \int_{u=-\infty}^{t} \lambda(u) \cdot P(S > t - u)du =$$
$$= \int_{u=-\infty}^{t} \lambda(u) \cdot [1 - G(t-u)]du\,.$$

■

Proof of Lemma 3.2:

$$\int_{u=-\infty}^{t} \lambda(u) \cdot [1 - G(t-u)]du = \int_{u=-\infty}^{t} \lambda(u) \int_{s=t-u}^{\infty} dG(s)\, du = \\ = \int_{s=0}^{\infty} \left[\int_{u=t-s}^{t} \lambda(u)du \right] dG(s) = E\left[\int_{t-S}^{t} \lambda(u)du \right],$$

where the above order-interchange of the integration is justified by the non-negativity of $\lambda(t)$. ■

Proof of Lemma 3.3

$$\begin{aligned} \int_{u=-\infty}^{t} \lambda(u) \cdot [1 - G(t-u)]du &= \int_{x=0}^{\infty} \lambda(t-x) \cdot [1 - G(x)]dx = \\ &= E(S) \cdot \int_{x=0}^{\infty} \lambda(t-x) \cdot \frac{1}{E(S)} \cdot [1 - G(x)]dx = \\ &= E(S) \cdot \int_{x=0}^{\infty} \lambda(t-x) \cdot dG_e(x) = \\ &= E[\lambda(t - S_e)] \cdot E[S]\,. \end{aligned}$$

■

This completes the proof of Theorem 3.1

From the result of Theorem 3.1, one deduces that when $\lambda(t) \equiv \lambda$, the expression for $R(t)$ becomes the offered-load of a homogeneous arrival rate $\lambda \cdot E[S]$. Hence, the expression of $R(t)$ is very similar to that of the homogeneous arrival case, except for the random time lag in $\lambda(t)$. In addition, if $\lambda(t)$ changes very little before t, (in comparison to the average service time) then $R(t) \approx \lambda \cdot E[S]$. This is actually consistent with the logic underlying PSA.

3.1.2 Additional Insights

We now present some further insight on the offered-load:

1. **Another representation:** Consider the second representation, $R(t) = E\left[\int_{t-S}^{t} \lambda(u)du\right]$. It can be also written as

$$R(t) = E\Big[A(t) - A(t-S)\Big],$$

namely, the expected value of the difference between the number of arrivals until time t and the number of arrivals till a random service time before time t.
Moreover, if the service time S is deterministic, taking the value D, then it is obvious that the number of customer, being served, at time t, is $A(t) - A(t-D)$, which is actually the number of arrivals between times $t-D$ and t. Since the arrivals follow a non-homogenous Poisson process with parameter $\lambda(t)$, in this deterministic service time case, $L(t) = A(t) - A(t-D)$ is Poisson distributed, with mean $\int_{t-D}^{t} \lambda(u)du$.

2. **Biased sampling and discrete approximation:** Here we motivate the first representation of Theorem 3.1 through a discrete approximation approach. Suppose that the service time is deterministic S=D. Then S_e is uniformly distributed on the interval $[0, D]$. In this case

$$\begin{aligned} E[\lambda(t-S_e)] \cdot E[S] = D \cdot E[\lambda(t-S_e)] = D \cdot \int_0^D \frac{1}{D} \cdot \lambda(t-x)dx = \int_0^D \lambda(t-x)dx = \\ = E\left[\int_{t-D}^{t} \lambda(u)du\right]. \end{aligned} \tag{3.4}$$

For a general S, we use the fact that any random variable can be approximated as close as required by a discrete (finite-support) random variable. We thus assume that $S = D_i$ with probability p_i. From biased sampling we get

$$S_e \overset{d}{=} Uni(0, D_i) \ w.p. \ \frac{p_i \cdot D_i}{E[S]}.$$

If we denote $U_i = Uni(0, D_i)$, then:

$$\begin{aligned} E\left[\int_{t-S}^{t} \lambda(u)du\right] = \sum_i p_i \cdot \int_{t-D_i}^{t} \lambda(u)du \overset{(3.4)}{=} \sum_i p_i \cdot D_i \cdot E[\lambda(t-U_i)] = \\ = E[S] \cdot \sum_i \frac{p_i \cdot D_i}{E[S]} E[\lambda(t-U_i)] = E[\lambda(t-S_e)] \cdot E[S]. \end{aligned} \tag{3.5}$$

3. **Alternative Poisson process view:** The third representation claims that $R(t) = \int_{-\infty}^{t} \lambda(u) \cdot [1 - G(t-u)]du$. This last expression raises an alternative way to approach the offered-load process, $L(t)$. Let $\{A(s,t) = \sum_{i=1}^{A(s)} \mathbb{1}_{\{S_i > t - t_i\}}, 0 \leq s \leq t\}$ be a two parameter stochastic process, in which $A(s,t)$ represents the number of arrivals until time s, which are still in service at time t. Here again, assume that the arrivals form a non-homogenous Poisson process with arrival rate $\lambda(u)$. Notice the $L(t) = A(t,t)$.
   It is easy to see that for any pre-specified time t, $A(s,t)$ is a Poisson process in s, with rate $\lambda(u) \cdot [1 - G(t-u)]$. Therefore, the expected number of customers at time t, who arrived before time s, is $E(A(s,t)) = \int_{-\infty}^{s} \lambda(u) \cdot [1 - G(t-u)]du$, and the expected number of customers at time t, who arrived during the time interval $[s_1, s_2]$, is $\int_{s_1}^{s_2} \lambda(u) \cdot [1 - G(t-u)]du$. *This approach will be found useful in Chapter 5.1, for estimation of the offered-load, in the presence of abandonments.*

4. **A generalization of Theorem 3.1 to the $\mathbf{M_t/GI_t/N_t + GI}$ queue:**
   The $M_t/GI_t/N_t+GI$ queue is similar to $M_t/GI/N_t+GI$, apart for that it allows the service time distribution to change as a function of the arrival time of the customer. In other words, the service time distribution of a customer is determined by his arrival time, independently of other customers or of the system's state. Following the proof of Lemma 3.1, one can immediately establish that $L(t)$ is now Poisson distributed with mean

$$R(t) = E[L(t)] = \int_{u=-\infty}^{t} \big(\lambda(u)[1 - G_u(t-u)]\big)du\,, \tag{3.6}$$

   where $G_u(t)$ is the cdf of the service time of a customer who arrived at time u.

5. **A generalization of Little's Law:** Formula 3.6 can be interpreted as a generalization of Little's Law. Notice that in stationary systems, the steady-state version of this equation follows the well known Little's Law, since

$$E(L) = \lim_{t\to\infty} E(L(t)) = \lim_{t\to\infty} \int_{u=-\infty}^{t} \lambda \cdot [1 - G(t-u)]du = \\ = \lambda \cdot \int_{u=0}^{\infty} [1 - G(u)]du = \lambda \cdot E(S)\,.$$

   Moreover, it is shown in [7] that Equation 3.6 holds for much weaker assumptions then those of the $M_t/GI_t/N_t + GI$ queue. Indeed, it is generalized to any arrival processes that has an arrival rate and for cases where the sojourn times in the system

(not necessarily service times) of customers might be dependent. An example for an application of this property is taking into account not only the service period of a customer, but the entire sojourn time within the system. Then, L(t) is regarded as the number of customers in the entire analyzed system and G_t is the sojourn time distribution of a customer who arrived to the system at time t. Here, of course, $L(t)$ need not be, and typically is not, Poisson distributed.
An extensive discussion on this time-varying Little's Law is given in [7] and [4].

6. **Taylor-Series Approximations:** We now consider the first representation in (3.2): $R(t) = E[\lambda(t - S_e)] \cdot E[S]$. The time lag S_e appears inside the arrival rate function $\lambda(t)$, which appears inside the expectation. In case that $\lambda(t)$ is non-linear, the calculation of this expression might be complicated. If $\lambda(t)$ is polynomial, then one can express $R(t)$ directly in terms of moments of S_e (See Theorem 10 in [27]). However, in many cases, the arrival rate function will not be polynomial. In this case, if $\lambda(t)$ is smooth, one can approximate it by a Taylor series. In this manner, a first order approximation for the arrival rate function in a time interval before t will be $\lambda(t-u) \approx \lambda(t) - \lambda^{(1)}(t) \cdot u$, where $\lambda^{(k)}(t)$ denotes the k^{th} derivative of $\lambda(t)$ evaluated at time t. The second order Taylor series approximation is $\lambda(t-u) \approx \lambda(t) - \lambda^{(1)}(t) \cdot u + \lambda^{(2)}(t) \cdot \frac{u^2}{2}$.
From the approximations above, one obtains, as shown in [27], the first and second order approximations for $R(t)$ respectively:

   (a) $R(t) \approx \lambda(t - E[S_e]) \cdot E[S]$;

   (b) $R(t) \approx \lambda(t - E[S_e]) \cdot E[S] + \frac{\lambda^{(2)}(t)}{2} \cdot Var(S_e) \cdot E(S)$.

   The first order approximation for $R(t)$ is similar to PSA, but with a backward time shift of $E(S_e)$. This is often referred to as *lagged*-PSA [30]. From the second order approximation, one notices that there is also a space shift by $\frac{\lambda^{(2)}(t)}{2} \cdot Var(S_e) \cdot E(S)$. Since $\lambda^{(2)}(t)$ is negative at a local peak of $\lambda(t)$, and $Var(S_e) \cdot E(S)$ is always positive, we conclude that the offered-load at times of peak demand are in fact shifted down. Similarly, in demand's trough points, the offered-load is shifted up.
   Intuitively, the time shift can be considered as a "memory" of the offered-load, generated by customers' service times. In other words, the customers who are in the system at time t, arrived before time t and will stay after it (avoiding events of probability zero). Consequently, the changes in trend of the offered-load are lagged after the changes in the arrival rate, and longer service times imply longer lags. Notice

that if the arrival rate at time t is linear, the space shift expression is eliminated, if it is convex at time t, the space shift is positive, and if it is concave, then the shift space is negative.

3.1.3 The Gap/Lag - Examples

A good example for the effect of the time-gap S_e, described in Theorem 3.1, appears in [15]. In Section 6 of that thesis, empirical examples were presented to compare between several staffing methods. It has been shown that staffing according to PSA leads to over-staffing in periods when the offered-load increases, and results in under-staffing in periods when it decreases. Another method that was applied was the *lagged PSA* which is described in the next subsection. The approximation to $R(t)$ under this latter method is taken to be the PSA approximation with time lag of one mean service time, $\lambda(t - E[S]) \cdot E[S]$. This method performed better then PSA, but still encountered lack of stability in service performance. In general, PSA performs quite well for fast service rates (fast relative to variations of the arrival rate function). If service rates are slower, then the lagged PSA is more appropriate.

Another example, which we believe is related to the described phenomenon of the lag of the offered-load behind the arrival rate, is introduced in [8]. This paper describes a study, based on data from two Massachusetts hospitals, on the diversion of ambulances from emergency departments due to congestion. Specifically, a diversion to another emergency department is made when the emergency department is fully occupied, and thus is not able to accept new patients. We focus on the results of the first hospital described in this paper (which is referred to in the study as "Hospital A"). Observations of the number of arrivals to the emergency department of this hospital and the number of diversions from it, per hour, were collected over period of 42 days. Then, an average of these measures by hour of the day was made over the 42 days. By this procedure, two series of 24 records were constructed: the average number of arrivals to the emergency department per hour of the day and average number of diversions from the emergency department per hour of the day.

The average service time was found to be approximately 6 hours and therefore the writers of the paper decided to build 6 more series, based on the series of the arrivals. The first series was constructed by the addition of the arrivals of each hour to the arrivals of the previous hour. The second series was constructed by the addition of the arrivals of the

first series to the arrivals of two previous hours, and so on. In this manner, each point in the last series is actually the sum of the arrivals in the 6 hours prior to it.

The interesting insight emerging from this analysis is that even in the sixth series, where the arrivals of last 6 hours (which is the mean service time) are taken into account, there is a lag of the average number of diversions behind this series. This is another example of the fact that taking into account only the mean service time might not be good enough for the analysis of the current congestion in the system. This can be explained by the variance of the service time. Consider the relation $R(t) = E\left[\int_{t-S}^{t} \lambda(u)du\right]$ from Theorem 3.1, and compare it to $\int_{t-E(S)}^{t} \lambda(u)du$, which is what was used in [8]. If the service time was deterministic with the value of 6 hours, then under the described procedure, the cumulative arrival rate of the last 6 hour will be exactly the time-dependant offered-load. In other cases, however (depending on the distribution of S and the function $\lambda(t)$), this is totally different from $R(t)$.

3.1.4 Staffing the $M_t/GI/N_t + GI$ Queue

The staffing problem is to determine the minimal required number of agents subject to pre-specified satisfactory quality of service constraints. In the $M_t/GI/N_t + GI$ model, one allows changing the number of servers as function of time. In practice, the staffing level cannot be changed continuously. We regard the period during which the staffing level must be kept constant as the *staffing interval*. In the analysis of the arrivals to call centers, we consider two sources of variability: *predictable variability* - changes over time of the expected number of arrivals and *stochastic variability* - random fluctuations around this expected number.

We already mentioned PSA, which treats the system at time t as if it is in a steady state, with the arrival rate of time t. However, this approximation is designed to cope only with the stochastic variability. Hence, whenever the predicted variability becomes significant or the service times are not relatively short, PSA tends to be inappropriate. The first component of Theorem 3.1 shows that PSA is correct except for a random time lag S_e (the stationary-excess service time). A simple refinement for PSA, that relies on the time shift from the Taylor-series approximation, is the lagged PSA. In lagged PSA we approximate the offered-load at time t, $R(t)$, by $\lambda(t - E[S_e]) \cdot E[S]$.

Another approximation for the performance of the $M_t/GI/N_t + GI$ model is the MOL approximation. The approximation applies the performance in an associated stationary $M/GI/s + GI$ model. In the MOL approximation, we approximate the time-varying performance at time t by a *stationary* model where we take the offered-load to be $R(t)$. Since the stationary offered-load is $\lambda \cdot E(S)$, we use at time t the model with homogenous Poisson arrival process, namely $M/GI/s+GI$, with arrival rate $\lambda_{MOL}(t) = \dfrac{R(t)}{E(S)}$. Thus, the MOL method suggests staffing in the $M_t/GI/N_t + GI$ model by the square-root formula $N_t = R(t) + \beta \cdot \sqrt{R(t)}$, where β is the related service quality parameter of the stationary model as given by the Garnett function (2.4). It is important to note that the validity of that refinement depends on the assumption of nonhomogeneous Poisson arrival process. The MOL approximation is supported by the work in [36].

The authors of [36] introduced an effective staffing algorithm, called the *Iterative Staffing Algorithm* (ISA). ISA is designed to deal with both predictable and stochastic variability to achieving stable performance over the day. It is shown in [36] that in order to achieve time stable performance in the face of time-varying arriving rates it is enough to target a stable delay probability, which then yields stable performance of several other operational measures. More specifically, it is shown in [36] that for the time-varying $M_t/GI/N_t + GI$ model, it is actually possible to reduce the hard problem of time-varying staffing to an associated stationary staffing problem. Indeed, a time-varying square-root staffing, i.e. $N(t) = R(t) + \beta \cdot \sqrt{R(t)}$, applies. This staffing level gives rise to a remarkable time-stable performance in which the delay probability is constantly α. In the case of $M_t/M/N_t + M$, the relation between α and β in the time-varying model is in fact the Garnett function (2.4), derived for the stationary Erlang-A model.

Chapter 4

The Relationship Between Service Time and Patience

In Chapter 3, we described the offered-load process and the offered-load function through the effective arrival process and the service times. Consequently, it is important to understand these two components when analyzing the offered-load. One of the assumptions of the $M_t/GI/N_t+GI$ queue is that the service time and the patience time are independent. Experience shows, that this assumption is often violated when analyzing empirical data of real service systems, particularly in those coping with human customers. For example, it is reasonable to believe that customers who expect to have longer service times are willing to wait more before getting service.

We denote the model $M_t/GI/N_t+GI$ with service time which is dependent on the patience by $M_t/GI^*/N_t+GI$. The * sign in the service time component stands for the fact that the service time distribution is dependent on patience. In this section, we propose a model for the relationship between patience and service time.

We start with defining the following notation:

- S - Service time of a customer.
- τ - (Im)patience of a customer. i.e. the time the customer is willing to wait before abandoning.

- V - Virtual-waiting-time (or offered-waiting-time), namely the time a customer is required to wait before entering service. Alternatively, this is the time a customer with an infinite patience would have waited, until entering service.
- W - Waiting time of a customer. The waiting time is defined as the minimum between V and τ, which is the actual time a customer waits, for both served and abandoning customers.

We assume that each customer has an associated pair of random variables τ and S as above, that are characterized by the customer, whereas the random variable V is independent of the customer (as in telephone services, where the customer does not observe the queue) and is determined only by system conditions.

The following properties naturally follow from the above discussion:

- The pair (τ, S) is independent of V.
- For those who abandon, we observe V censored by their (im)patience. For those who are served, we observe V.
- W is observable for all customers.
- We observe S only for customers who have $\tau > V$. In this case, it also holds that $\tau > W$.

In common analyses of service systems, the service time distribution is measured through the distribution of all served customers. However, according to the last bullet point, this is in fact the projection of the service time over the probability space where the patience is greater than the required waiting time. Consequently, these methods take into account only part of the customers population, which may lead to misleading results. For instance, the prevalent estimate for the mean service time is actually $E(S|\tau > W)$. From biased sampling, it is obvious that the number of served customers with longer patience times is expected to be higher than the number of served customers with shorter patience. Now, assume that patience and service time are positively correlated. In this case, the average service time of all served customers is shifted upward, which yields a biased estimation of $E(S)$, since one also tends to observe more longer service times.
In this sense, one might be interested in understanding the relationship between patience

and service time, and also in the unconditional distribution of the service time (which appears in the representations of the offered load in Theorem 3.1).

In order to explore the relationship between patience and service time, we consider the conditional expectation

$$E(S|\tau > W, W = w),\ w > 0\,, \tag{4.1}$$

which is a natural object of study, since it can be estimated from the operational data for any $w > 0$ ($S|\tau > W, W = w$ is observable), and it also incorporates information on the patience.

4.1 The Model

We propose a non-linear regression model to relate service time and waiting:

$$S_i = g(W_i) + \varepsilon_i\,, \tag{4.2}$$

where $g(w) = E(S|\tau > W = w)$ is the mean service time of those who waited exactly w units of time and were served; S_i and W_i are the service and waiting times of customer i respectively.

Lemma 4.1 *A simpler expression for g is given by:*

$$g(w) = E(S|\tau > w)\,. \tag{4.3}$$

Proof: Notice that $g(w) = E(S|\tau > W = w) = E(S|\tau > w, W = w)$, and that the events $\{\tau > w, W = w\}$ and $\{\tau > w, V = w\}$ are identical. Hence, we get

$$g(w) = E(S|\tau > w, V = w) = E(S|\tau > w)\,,$$

where the last equality holds because of the independence of (τ, S) and V. ■

At this stage, we evaluate an expression for the expected service time of those who have patience $\tau = w$, $E(S|\tau = w)$. Here, we assume that both $E(S|\tau = w)$ and $f_\tau(w)$ are continuous with respect to w. We start with a definition of $f_{S|\tau>w}$, the density function

of the service time, given that the patience is greater than w time units:

$$f_{S|\tau>w}(s) \equiv \frac{1}{P(\tau > w)} \int_{u=0}^{\infty} f_{S,\tau}(s,u)\, \mathbb{1}_{\{u>w\}}\, du = \frac{1}{P(\tau > w)} \int_{u=w}^{\infty} f_{S,\tau}(s,u)\, du$$

Then,

$$\begin{aligned} g(w) = E(S|\tau > w) &= \int_{s=0}^{\infty} s\, f_{S|\tau>w}(s) ds = \\ &= \frac{\int_{s=0}^{\infty} s \int_{u=w}^{\infty} f_{S,\tau}(s,u)\, du\, ds}{P(\tau > w)} = \frac{\int_{s=0}^{\infty} s \int_{u=w}^{\infty} f_{S|\tau}(s|u)\, f_{\tau}(u)\, du\, ds}{\int_{u=w}^{\infty} f_{\tau}(u)\, du} \\ &= \frac{\int_{u=w}^{\infty} \left(\int_{s=0}^{\infty} s\, f_{S|\tau}(s|u)\, ds\right) f_{\tau}(u)\, du}{\int_{u=w}^{\infty} f_{\tau}(u)\, du} \\ &= \frac{\int_{u=w}^{\infty} f_{\tau}(u)\, E(S|\tau = u)\, du}{\int_{u=w}^{\infty} f_{\tau}(u)\, du}\,. \end{aligned}$$

We arrive at

$$g(w) = \frac{\int_{u=w}^{\infty} f_{\tau}(u)\, E(S|\tau = u)\, du}{\int_{u=w}^{\infty} f_{\tau}(u)\, du}\,. \tag{4.4}$$

By multiplying both sides by $\int_{u=w}^{\infty} f_{\tau}(u)\, du$, we get

$$g(w) \int_{u=w}^{\infty} f_{\tau}(u)\, du = \int_{u=w}^{\infty} f_{\tau}(u)\, E(S|\tau = u)\, du\,.$$

Differentiating both side of the last equation with respect to w yields

$$g'(w) \int_{u=w}^{\infty} f_{\tau}(u)\, du - g(w)\, f_{\tau}(w) = -f_{\tau}(w)\, E(S|\tau = w)\,.$$

We conclude that

$$E(S|\tau = w) = g(w) - g'(w)\, \frac{\int_{u=w}^{\infty} f_{\tau}(u)\, du}{f_{\tau}(w)} = g(w) - \frac{g'(w)}{h_{\tau}(w)}\,, \tag{4.5}$$

where h_{τ} is the hazard rate function of τ.

Examining equation (4.5) reveals that in order to estimate $E(S|\tau = w)$, it suffices to estimate the following expressions:

1. The hazard rate function $h_{\tau}(w)$.

2. The function $g(w)$ and its derivative $g'(w)$.

We thus intend to carry out an estimation of the two expressions above and deduce $E(S|\tau = w),\ w > 0$.

We present here several examples in order to explore the above results:

Example 1 - A Monotone function example.
Assume that τ is exponentially distributed, with mean $1/\theta$. Let the conditional expectation of the service time given the patience of a customer be of the form $E(S|\tau = w) = a\,(b - e^{-w\,(\beta-\theta)}),\ \beta > \theta$. *Consider two cases:*

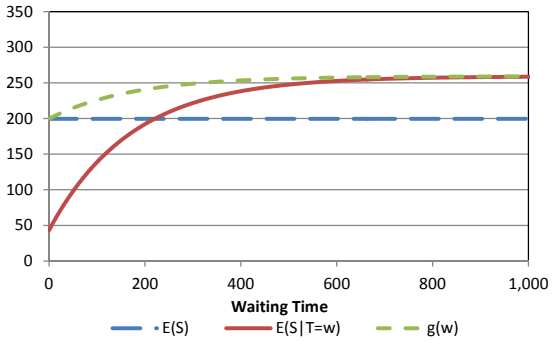
1. $a > 0$ *and* $b > 1$.

2. $a < 0$ *and* $b < 0$.

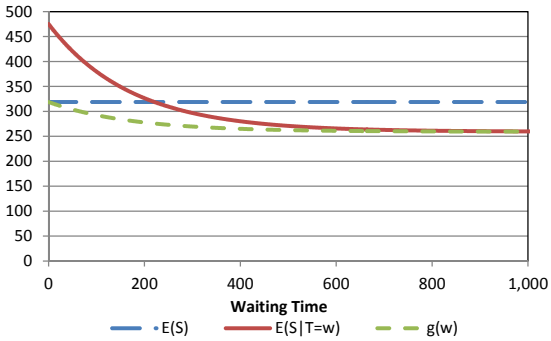
For both cases $E(S|\tau = w) \in (0, \infty)$, for any $w > 0$. It can be easily shown that $E(S) = a\,(b - \frac{\theta}{\beta})$, and by Formula (4.4), we get that $g(w) = a\,(b - \frac{\theta}{\beta}\,e^{-w\,(\beta-\theta)})$. It is clear that in the first case, $g(w) \geq E(S)$ for any $w > 0$, whereas in the second case, $g(w) \leq E(S)$ for any $w > 0$. Moreover, for these two cases, $g(w)$ equals $E(S)$ when $w = 0$. Consequently, estimating $E(S)$, taking into account only the observable records of the service times (for which $\tau_i > W_i = w_i$), may be misleading. For example, in the first case, the estimate of $E(S|\tau > W)$ is expected to be higher then $E(S)$, as shown in Figure 5.1(a).

Example 2 - A Sinusoidal function example.
Assume that τ is exponentially distributed, with mean $1/\theta$. Let the conditional expectation of the service time given the patience of a customer be of the form $E(S|\tau = w) = a + b\,sin(c\,w + d)$, *$a, b, c, d$ are all positive and $a > b$. Here one can calculate that:*
$E(S) = a + \frac{b\,\theta}{\theta^2 + c^2}\,[\theta\,\sin(d) + c\,\cos(d)]$ *and*
$g(w) = a + \frac{b\,\theta}{\theta^2 + c^2}\,[\theta\,\sin(c\,w + d) + c\,\cos(c\,w + d)]$
Here, the relationship between $g(w)$ and $E(S)$ is not so clear, but again we observe that $g(w)$ equals $E(S)$ when $w = 0$. In addition, looking at Figures 5.1(c) and 5.1(d) reveals that, in this case, $g(w) = E(S|\tau = w)$ for w's in which $g'(w) = 0$, $g(w) > E(S|\tau = w)$ when $g'(w) > 0$ and $g(w) < E(S|\tau = w)$ when $g'(w) < 0$.

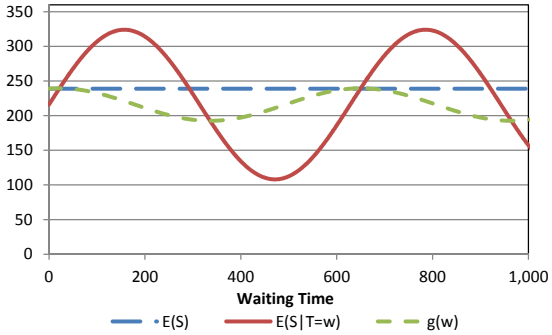
(a) $E(S|\tau = w) = 216\,(1.2 - e^{-w\,\frac{21}{3600}})$



(b) $E(S|\tau = w) = -216\,(-1.2 - e^{-w\,\frac{21}{3600}})$



(c) $E(S|\tau = w) = 216 + 108\,\sin(\frac{36}{3600}\,w)$



(d) $E(S|\tau = w) = 216 + 108\,\sin(\frac{36}{3600}\,w + 2)$

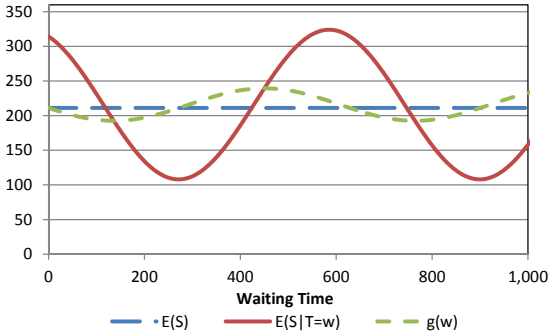


Figure 4.1: A comparison between $E(S)$, $E(S|\tau = w)$ and $g(w)$. In all figures $\tau \sim exp(\frac{8}{3600})$. Figures (a) and (b) describe the model $E(S|\tau = w) = a\,(b - e^{-w\,(\frac{29}{3600} - \frac{8}{3600})})$ - figure (a) deals with the case where $a > 0, b > 1$, whereas figure (b) is an example for $a < 0, b < 0$. Figures (c) and (d) present the model $E(S|\tau = w) = a + b\,\sin(cw + d)$.

Based on these results we deduce the following propositions:

Proposition 1 *The expectation of the service time equals the expectation of the service time given non-waiting:*

$$E(S) = E(S|W = 0)\,. \tag{4.6}$$

Proof: Using Equation (4.5),

$$E(S) = \int_{w=0}^{\infty} E(S|\tau = w)\, f_\tau(w)\, dw = \int_{w=0}^{\infty} \left[g(w) - \frac{g'(w)}{h_\tau(w)} \right] f_\tau(w)\, dw =$$
$$= \int_{w=0}^{\infty} \left[g(w)\, f_\tau(w) - g'(w)\,(1 - F_\tau(w)) \right] dw = -g(w)\,(1 - F_\tau(w))|_{w=0}^{\infty} =$$
$$= g(0) = E(S|\tau > W = 0) = E(S|W = 0)\,. \qquad \blacksquare$$

The result of Proposition (1) is also supported by the fact that, for non-waiting customers, we do not observe any information on their patience.

Proposition 2 *For a fixed w, $w \geq 0$*

1. $g(w) > E(S|\tau = w) \Leftrightarrow g'(w) > 0.$
2. $g(w) < E(S|\tau = w) \Leftrightarrow g'(w) < 0.$
3. $g(w) = E(S|\tau = w) \Leftrightarrow g'(w) = 0.$

Proof: By Equation 4.5 we get $g'(w) = (g(w) - E(S|\tau = w))\, h_\tau(w)$. Since $h_\tau(w) \geq 0$ for $w \geq 0$, the sign of $g'(w)$ is equal to the sign of $g(w) - E(S|\tau = w)$. $\blacksquare$

Corollary 4.1 *Let $E(S|\tau = w)$ be continuous:*

1. *If $E(S|\tau = w)$ is an increasing (decreasing) monotone function, then $g(w)$ is also an increasing (decreasing) monotone function. The opposite is not necessarily true.*
2. *$g(w)$ is an increasing monotone function if and only if $g(w) \geq E(S|\tau = w)$.*
   $g(w)$ is an decreasing monotone function if and only if $g(w) \leq E(S|\tau = w)$.
3. *$g(w)$ converges to a limit L if and only if $E(S|\tau = w)$ converges to L*

$$L = \lim_{w \to \infty} g(w) = \lim_{w \to \infty} E(S|\tau = w)$$

Proof:

1. Assume that $E(S|\tau = w)$ is an increasing monotone function, and let $0 \leq w_1 < w_2$. We define $F^c_\tau(w) = \int_{u=w}^{\infty} f_\tau(u)\, du$, the survival function of the patience at w. Then, From equation 4.4,

$$\begin{aligned} g(w_2) - g(w_1) &= \frac{\int_{u=w_2}^{\infty} f_\tau(u)\, E(S|\tau = u)\, du}{F^c_\tau(w_2)} - \frac{\int_{u=w_1}^{\infty} f_\tau(u)\, E(S|\tau = u)\, du}{F^c_\tau(w_1)} \\ &= \left(\frac{1}{F^c_\tau(w_2)} - \frac{1}{F^c_\tau(w_1)}\right) \int_{u=w_2}^{\infty} f_\tau(u) \cdot E(S|\tau = u)\, du \\ &\quad - \frac{\int_{u=w_1}^{w_2} f_\tau(u)\, E(S|\tau = u)\, du}{F^c_\tau(w_1)}\,. \end{aligned}$$

From the assumptions that $0 \leq w_1 < w_2$ and $E(S|\tau = w_1) < E(S|\tau = w_2)$ follow:

(a) $\dfrac{1}{F^c_\tau(w_2)} - \dfrac{1}{F^c_\tau(w_1)} > 0$

(b) $$\begin{aligned} \int_{u=w_2}^{\infty} f_\tau(u)\, E(S|\tau = u)\, du &> \int_{u=w_2}^{\infty} f_\tau(u)\, E(S|\tau = w_2)\, du \\ &= E(S|\tau = w_2)\, F^c_\tau(w_2) \end{aligned}$$

(c) $$\begin{aligned} \int_{u=w_1}^{w_2} f_\tau(u)\, E(S|\tau = u)\, du &< \int_{u=w_1}^{w_2} f_\tau(u)\, E(S|\tau = w_2)\, du \\ &= E(S|\tau = w_2)\, [F^c_\tau(w_1) - F^c_\tau(w_2)] \end{aligned}$$

Then,

$$\begin{aligned} g(w_2) - g(w_1) &> \left(\frac{1}{F^c_\tau(w_2)} - \frac{1}{F^c_\tau(w_1)}\right) E(S|\tau = w_2)\, F^c_\tau(w_2) \\ &\quad - \frac{E(S|\tau = w_2)\, [F^c_\tau(w_1) - F^c_\tau(w_2)]}{F^c_\tau(w_1)} = \\ &= E(S|\tau = w_2) \left[1 - \frac{F^c_\tau(w_2)}{F^c_\tau(w_1)} - \left(1 - \frac{F^c_\tau(w_2)}{F^c_\tau(w_1)}\right)\right] = 0 \end{aligned}$$

Hence, we deduce that $g(w_1) < g(w_2)$ for any $0 \leq w_1 < w_2$. The same proof can be applied for $E(S|\tau = w)$ decreasing monotone.

We show now a counterexample for the opposite direction:
Consider the following increasing monotone function, which is a variation on the logistic function $g(w) = a + \frac{b}{1 + c\, e^{-d(w-s)}}$, $w \geq 0$, $a, b, c, d, s \in \mathbb{R}_+$, and assume that τ is exponentially distributed, with mean $1/\theta$.
By Equation (4.5) the appropriate $E(S|\tau = w)$ is $a + \frac{b}{1 + c\, e^{-d(w-s)}} - \frac{bcde^{-d(w-s)}}{\theta\,(1 + c\, e^{-d(w-s)})^2}$.
This model is valid if a, b, c, d and s are such that $E(S|\tau = w) \geq 0$ for any $w \geq 0$. By taking $g(w) = 0.06 + \frac{0.01}{1 + 1 \cdot e^{-50(w-0.1)}}$, we show that $E(S|\tau = w)$ is not monotone, as shown in Figure 4.2.

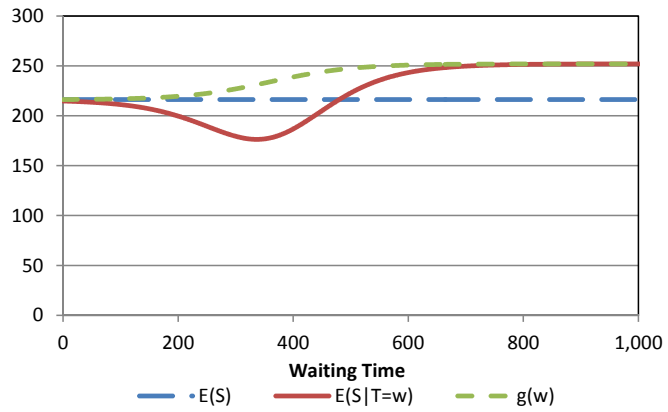


Figure 4.2: $g(w)$ is an increasing monotone function, but $E(S|\tau = w)$ is not. The figure shows the model $g(w) = 0.06 + \frac{0.01}{1 + 1 \cdot e^{-50(w-0.1)}}$, where $\tau \sim exp(8)$.

2. This property follows directly from the second bullet of Proposition 2.

3. This property follows from the third claim of Proposition 2.

This concludes our proof. ■

Remark 4.1 *All the results, presented in Section 4.1, can be applied for any transformation T on S which satisfies that $E(T(S))$ exists (i.e. $E(|T(S)|) < \infty$) and that $E(T(S)|\tau = w)$ is continuous with respect to w. Then Equation (4.5) takes the form:*

$$E(T(S)|\tau = w) = E(T(S)|\tau > w) - \frac{\frac{\partial}{\partial w} E(T(S)|\tau > w)}{h_\tau(w)}\,, w > 0\,. \tag{4.7}$$

In particular, this is true for $T(S) = S^k, k \in \mathbb{N}$, *as long as the* k^{th} *moment of* S *exists, and for* $T(S) = \mathbb{1}_{\{S \leq s\}}$, *providing us with explicit expressions for the conditional moments and the conditional cdf of* S *as follows:*

1. $E(S^k|\tau = w) = E(S^k|\tau > w) - \frac{\frac{\partial}{\partial w}E(S^k|\tau > w)}{h_\tau(w)}, w > 0\,.$

2. $P(S \leq s|\tau = w) = P(S \leq s|\tau > w) - \frac{\frac{\partial}{\partial w}P(S \leq s|\tau > w)}{h_\tau(w)}, w > 0\,.$

From the last insight, one observes that by knowing the distribution of S given $\tau > w$, for all $w > 0$, the distribution of S given $\tau = w$ can be obtained. However, if the distribution of S given $\tau > w$ is to be estimated, then one should be careful with applying this model. While dealing with empirical issues, we identified two significant sources of failure:

1. **Noise:** The proposed model can be applied for a variety of service time distributions, including mixtures of distributions. However, in case of a mixture, an important assumption is that the relationship between the service time and the patience are identical in all the mixture's components. Ignoring this fact may lead to extreme and unreliable behavior of the model.

2. **Estimation od distributions:** The estimation of the conditional distribution of the service time, given that the patience is greater than the waiting time, requires an estimation of a distribution at any waiting time value. Moreover, long tails of the distributions are often expected. If possible, a choice of parametric distributions is recommended. The estimation problem is then reduced to the estimation of the parameters of these distributions and smoothing them over the waiting time.

We discuss these issues extensively in Chapter 7.

4.2 Testing the Relationship

At this stage, we present methods to test the null hypothesis that $E(S|\tau)$ is a constant (i.e. if $E(S|\tau = w) \equiv E(S), \forall w \geq 0$). Since $E(S|\tau = w)$ is not observable, the test will be applied to $g(w) = E(S|\tau > W, W = w)$ which, under the null hypothesis, is equal to $E(S|\tau = w)$. Moreover, from Proposition 2 one can easily observe that $E(S|\tau = w)$ is

constant if and only if $g(w)$ is constant. We would like to apply the model described in Section 4.1 only if the null hypothesis is rejected. Otherwise, we assume that the service time and the patience are independent.

We propose here two test methods:

First Method: Based on a one-way ANOVA. In this method, we divide the customers who got service into several groups by their waiting times, in a manner that the number of observations in each group is similar. Then, we test the hypothesis that the average service time in all the groups is equal, against the alternative hypothesis that at least one group has a different average service time.

Second Method: The second proposed method relies on permutation tests. We want to test the significance of the relationship between the service time and the patience. We are interested in testing whether the observed relationship between these two variables, through our operational data sample, is stronger than expected due to chance.
Assuming that there is no relation between S and τ, implies that the variance of the conditional expectation $E(S|\tau)$ equals zero. Let $G(W)$ be a random variable which takes the value $g(w) = E(S|\tau > W, W = w)$, according to the density function $f_{W|\tau>W}(w)$. Then, it suffices to test if the variance of $G(W)$ can be assumed to be zero. Notice that, by taking only the observations in which the service time is greater than zero, we regard only the probability space that is conditional on the event $\tau > W$. Therefore, $f_{W|\tau>W}(w)$ can be also explicitly estimated as the empirical density of the waiting times in the population of the served customers.

From the equality $Var(g(W)) = E(g^2(W)) - E^2(g(W))$, we can choose the statistic for our test to be

$$T = \int_{u=0}^{\infty} g^2(u)\, f_{W|\tau>W}(u)\, du - \left[\int_{u=0}^{\infty} g(u)\, f_{W|\tau>W}(u)\, du \right]^2 .$$

The permutation distribution of the statistic T is constructed by calculating its values $t_1, t_2, \ldots, t_K$, from a large number K of resamples from the original data. A resample, which is consistent with the null hypothesis, is drawn by permuting the observed service times and waiting times at random. In this manner, we generate a random pairing between the service times and the waiting times. For each sample, we calculate the value of the above statistic. The p-value is then approximated by the proportion of samples with the

value of this statistic larger than the original sample's statistic, denoted by t^*. Explicitly:

$$p - value \approx \frac{\sum_{i=1}^{K} \mathbb{I}_{\{t_i > t^*\}}}{K} \,. \tag{4.8}$$

Notice that, for given data, the estimation of $f_{W|\tau>W}$ is not affected by the random permutation. As a result, $\int_{u=0}^{\infty} g(u)\, f_{W|\tau>W}(u)\, du$ is equal for any permutation, since it is simply the average over all the observations. This leads us to a simplified version of the test statistic, which is the empirical second moment:

$$T = \int_{u=0}^{\infty} g^2(u)\, f_{W|\tau>W}(u)\, du \,. \tag{4.9}$$

In practice, we consider a discretization of the values of w, by separating the range of w into several groups. For each group, we estimate both the probability of being in the group and the average service time of the observations in the group. The statistic's value for a certain permutation is then calculated as the sum over all the w's groups, of the probability of being in the group multiplied by the square of the average service time within it.

Chapter 5

Estimation and Prediction of the Offered-Load Process/Function

5.1 Estimation of The Offered-Load

In Section 3.1, we gave a formal description of the $M_t/GI/N_t+GI$ queue , and its analysis. We now use these foundations and theoretical results in estimation and prediction of the offered-load process. The estimation of the offered-load process is used for both off-line analysis of system performance, and for prediction of the future offered-load. As discussed in Chapter 3, offered-load prediction is essential for determining appropriate staffing levels, and achieving a predetermined service performance.

Based on the results of Chapter 3, the estimation of $L(t)$ and $R(t)$ presents challenges. Moreover, one should be careful with estimating $R(t)$, by taking into account the *potential* service times of abandoning customers.

Consider the case where we have all the operational data, for every call that reached our service system, and we wish to estimate $L(t)$. A naive solution is to estimate it from available historical records, with the imputation of the service times of those who abandoned, based on the observed service times of the served customers. This could be good enough, unless the *potential service times of customers who abandoned* (and therefore their service times are not available) *have different distributions from the one of those who waited and reached service*. Furthermore, in a scenario where the patience time and the service time are dependent, the naive solution is likely to face *biased sampling* issues, as

discussed in the introduction to Chapter 4. This might be a very realistic possibility, that is not supported by the assumptions of the $M_t/GI/N_t+GI$ queue, where the service time and patience time are assumed to be independent.

Hence, we would like first to test the relationship between patience and service time. If the service time and the patience time can be assumed to be independent, then we will assume that the service times of those who abandoned are distributed similarly to those who were served, regardless of their patience time. In this case, we will impute their service times according to the observable empirical service time distribution of the served customers. Otherwise, if the service time and the patience time *can not be assumed to be independent*, we would like to have a generalization of the $M_t/GI/N_t+GI$ queue model, in order to take into account that relationship. Then we shall propose a method to impute the service times of those who abandoned, for whom we observe their patience from the operational data. This will be discussed in Section 5.1.2.

5.1.1 The Case of No Abandonment

When analyzing the offered-load of a queueing system which does not encounter abandonment, we consider two different models, that were described in Section 3.1:

1. $\mathbf{M_t/GI/\infty}$ queue: This queue is characterized by an infinite number of servers. In this model, all customers enter service immediately. Here the offered-load process, L, can be observed as the number of customers in service at any time t, and the offered-load function, R, can be simply estimated by the average number of customer in service over several periods that are assumed to be equal in distribution.

2. $\mathbf{M_t/GI/n_t}$ queue: The second model is the $M_t/GI/n_t$ queue, where all customers have infinite patience. In this case, the offered load *definitely should not* be regarded as the average number of customers in service. The reason is that customers do not necessarily start service immediately upon their arrivals. For customers who have positive waiting time before starting service, the servers encounter their service period in a delay of their waiting time. We assume that customers characteristics are not influenced by the system state. Consequently, their service times do not depend on their waiting times and, therefore, the operational data that is kept by the system, includes all the required information on service times. The estimation of the offered-load function can be carried in any of the following two methods:

First method: The simplest way to estimate the offered-load process is to calculate the number of customers in a corresponding $M_t/GI/\infty$ queue. This is done by eliminating the customers' waiting times and shifting their service period to start right upon their arrivals to the queue. Then, in order to estimate $L(t)$, it is left to calculate the number of customers in service in the adjusted data. This is actually an implementation of Formula (3.1). An estimate of the offered-load function can be achieved by averaging the offered-load process over all days that are assumed to be identically distributed.

Second method: Another method to estimate the offered-load function is based on the following representation, shown in Theorem 3.1:

$$R(t) = \int_{-\infty}^{t} [1 - G(t-u)]\lambda(u)du\,.$$

Here, we need first to estimate the survival function of the service time, $1 - G(t)$, $t \geq 0$, and the arrival rate $\lambda(t)\,, t \geq 0$. Finally, we shall approximate numerically the integral in this expression at any time point t.

We suggest the following method to approximate the expression above of the offered-load: Consider the following sequences:

- $0 = t_0 < t_1 < t_2 < \ldots$ - Time sequence for which we estimate the arrival rate in any time interval $[t_{i-1}, t_i)$, $i = 1, 2, \ldots$. We assume for simplicity that for any $i \geq 1$, $t_i - t_{i-1} = d$ (i.e. all the time intervals are of the same length) and that the arrival rate is nearly constant over any time interval, but may differ across time intervals. Denote λ_i as the arrival rate per d time units in the interval $[t_{i-1}, t_i)$.
- $0 = s_0 < s_1 < s_2 < \ldots$ - Service time sequence for which we estimate the measure $G^c(s_k) = 1 - G(s_k)$, $k = 1, 2, \ldots$, where G is the cdf of the service time. Again, we assume that for any $k \geq 1$, $s_k - s_{k-1} = r$. We also set that $d = r \cdot n$, $n \in \mathbb{N}$. Namely, we require that d is an integer multiple of r. Finally, G^c is assumed to be approximately linear within any of the time intervals, as described in Figure 5.1(b).

Now, assume that the system is empty at time 0. For simplicity, we seek to estimate the offered load at time $t = t_i > 0$, for some $i = 1, 2, \ldots$. The proportion of arrivals from time $t - s_k$, that are still in service at time t, is $G^c(t - s_k)$. The arrival rate in $[t - s_k, t - s_{k-1})$ is $\lambda_k(t) = \lambda_{i-\lceil \frac{k}{n} \rceil}$ per d time units, or $\lambda_k(t)\,\frac{r}{d}$ per r time units (which is

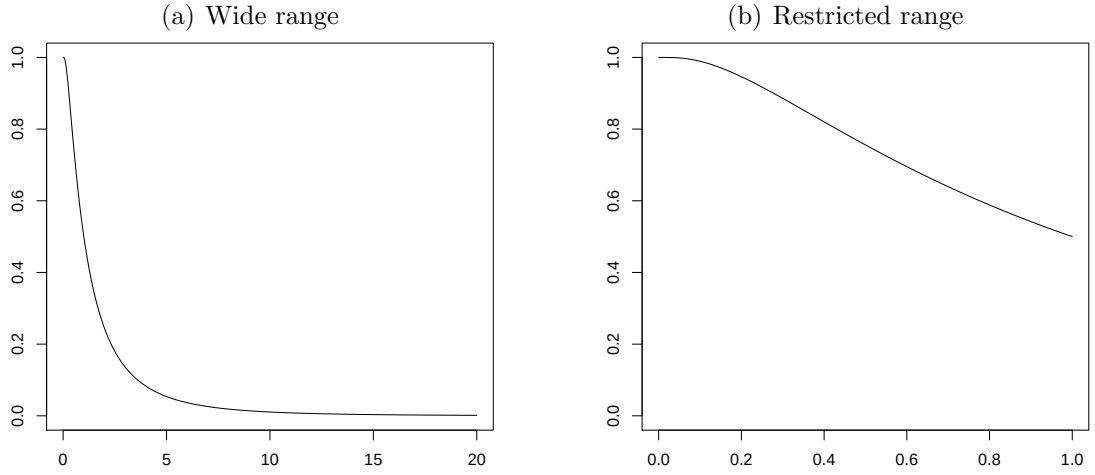


Figure 5.1: The service time distribution, G, is Log-Normal with $\mu = 0$ and $\sigma = 1$. In (a) is shown that $1 - G()$ decreases very fast to zero. Moreover, numerical calculations show that $\int_{50}^{\infty} [1 - G(u)]du < 0.001$. Figure (b) shows that, for time intervals of length 0.1, $1 - G$ is approximately linear.

the time interval where we assume that G is linear). Thus, from the assumptions above, the rate of arrivals from the time interval $[t - s_k, t - s_{k-1})$, that remain in service after time t, is approximately

$$\tilde{\lambda}_k(t) = \lambda_k(t) \frac{r}{d} \frac{G^c(t - s_k) + G^c(t - s_{k-1})}{2}. \tag{5.1}$$

We deduce that $R(t) \approx \sum_{\{k=1,\ldots,j:s_j=t\}} \tilde{\lambda}_k(t)$.

At this stage we wish to remove the assumption of an empty system at time $t = 0$. For this purpose, we look for the shortest time interval K, such that estimation of the offered-load at the end of the interval can be done by using only the observations that arrived to the system during this interval. Now, one shall decide on a value ϵ, representing the desired upper bound of the error of the estimator.

Assume that the arrival rate λ is bounded by the value M (and that $E(S) < \infty$). Then, choose the value K such that

$$K = \text{argmin} \left\{ v \in \mathbb{R} : \; M \cdot \int_{v}^{\infty} [1 - G(u)]du < \epsilon \right\}.$$

Then, the empty system assumption at time 0 can be omitted in the estimation of $R(t)$,

for $t > K$. The above expression is then reduced to $R(t) \approx \sum_{\{k=1,\ldots,j:s_j=K\}} \tilde{\lambda}_k(t)$, for any $t > K$. In many service systems, such as call centers or service over the counter, K can be taken to be a few hours, whereas in other types of service systems, such as hospitals, it might be a few days or even weeks.

Notice that in case that the service time distribution varies over the day (but assumed to be constant at the same time across different days), the expression for the service time survival function $G^c(s)$, in Formula (5.1), can be replaced with a time dependent $G^c_t(s)$, where t stands for the arrival time of a customer.

Emergency Department Example

The following example is based on data from a Medium-size Emergency Department (ED) of an Israeli Hospital, with an average monthly arrival rate of 6,400 patients (referred to as 'Hospital 3' in Marmor [33], Chapter 3). The time period of our analysis is January-July 2003. First, we calculate the arrival rate to the emergency department as the mean number of arrivals over same days in different weeks. The calculation is taken in partition into half hour time intervals. Figure 5.2 reveals that the arrival function has a similar pattern over weekdays (with an increased rate on Sundays), but a different pattern on Fridays and Saturdays.

Here, we refer to the beds of the ED as the servers, and the service time is the Length-of-Stay (LOS) of patients. Hence, the offered-load is calculated as the average number of patients in the ED (i.e. the number of occupied beds). An important fact is that the service time is dependent on the arrival rate, due to the changes in the availability of the department's staff, and on similar discharge times, but patients are not blocked from being admitted to the ED.

The first estimation of the offered-load is done by averaging the load over different weekdays, partitioned into half-hour intervals (solid black line in Figures 5.3 and 5.4). The second graph in Figure 5.3 shows the estimated average load, using the second method in Section 5.1.1, where the service time is assumed homogenous over all time intervals. Since we claim that the service time is dependent on the occupation level and the discharge times, this assumption is not valid. Indeed, one observes in Figure 5.3 that, in the early morning hours, the average-load is consistently greater then its calculation using the second method. For that reason, we estimated the distribution of the service time for each time interval separately (over all days). The result is given in Figure 5.4, where

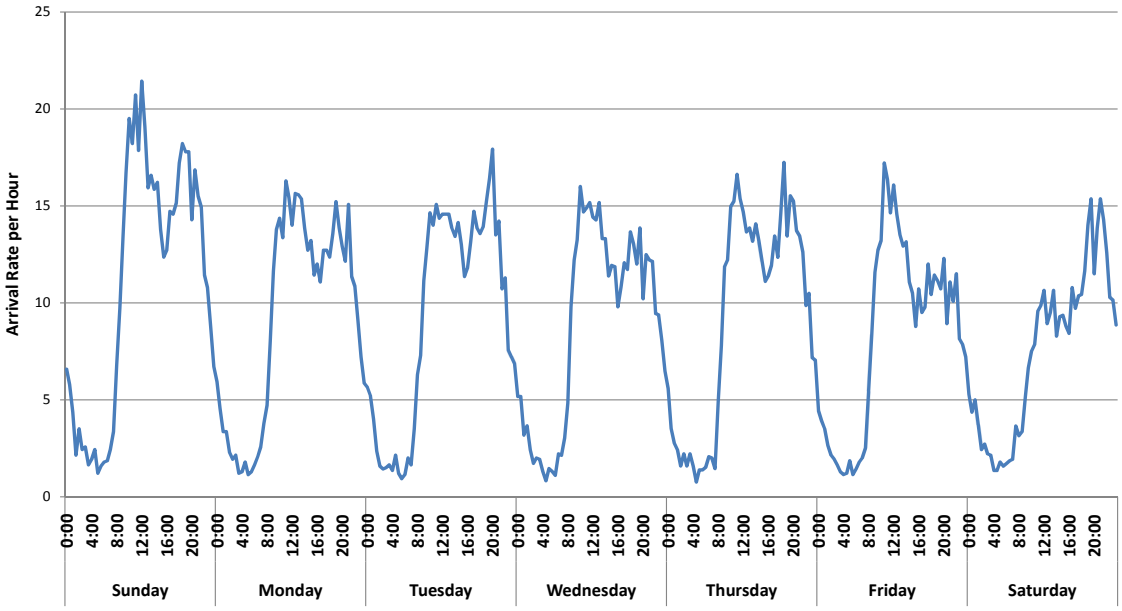


Figure 5.2: The arrival Rate to an emergency department of an Israeli Hospital. The arrival rate is calculated as the mean number of arrivals over same days in different weeks in partition into half hour time intervals.

the two methods yield similar estimators, apart from some small mismatches which arise mainly on weekends.

5.1.2 The Case in the Presence of Abandonment

As mentioned before, the main lack of information in the analysis of the offered load, when dealing with abandons, is that not all customers' service times are available through the operational data. The next section suggests several methods to overcome this problem. In both methods, we either impute the service times of those who abandoned or estimate the unconditional distribution of the service time. Our goal is to take into account a possible relationship between patience and service time in the estimation of the offered-load.

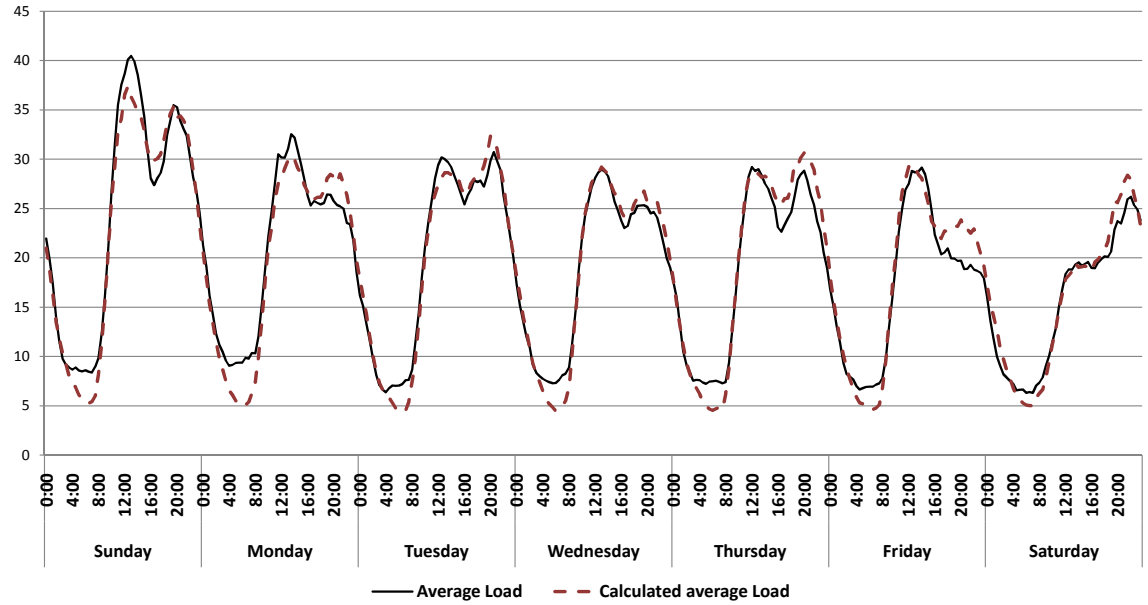


Figure 5.3: An estimation of the average-load in an emergency department of an Israeli Hospital. Once it is calculated as the mean of the load process and once it is calculated using the second method in Section 5.1.1, where the service time is assumed homogenous for all the time intervals.

5.1.2.1 Imputing Service Times

The following method for estimating the offered-load is based on the first method, described in section 5.1.1, for the $M_t/GI/n_t$ queue with no abandons. We assume that we observe the arrival times of all customers. For those served we observe their service times and their patience time censored by their waiting time, and for those abandoning we observe their patience. At this stage, we need to impute the service times of the abandoning customers. Then, we have an arrival time and a service time for each customer, and the rest of the estimation of the offered-load process and the offered-load function is done according to the first method in Section 5.1.1.

In order to impute the service time of a customer who abandoned, one can first estimate this customer's service time distribution. Then we randomly sample the value of that customer's service time from that distribution. Here we apply the testing method described in Section 4.2, in order to test if patience and service time are dependent. If the null hypothesis, that there is no relationship between the service time and the patience, is not rejected, we assume that all the customers' service times are equally distributed. Then, the service time of the abandoning customers is generated from the service time distri-

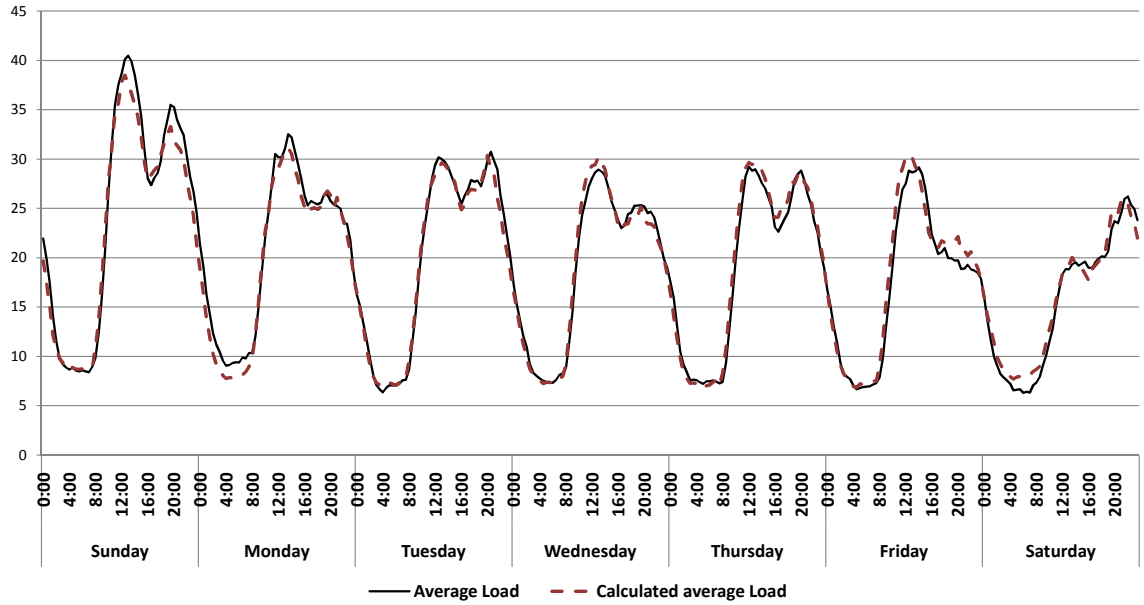


Figure 5.4: An estimation of the average-load in an emergency department of an Israeli Hospital. Once it is calculated as the mean of the load process and once it is calculated using the second method in Section 5.1.1, where the service time is assumed heterogenous between time intervals (but homogenous over weekdays).

bution that is observed from the served customers. Otherwise, if it is rejected, then we shall use the conditional service time distribution given the known patience time. Since a good non-parametric estimator for the conditional service time distribution may not be available, due to lack of observations (see our empirical and simulations analysis in Chapter 7), one can assume a parametric distribution which seems to be similar to the observed distribution of the service time of served customers. Then, the parameters of the conditional service time distribution, as a function of the patience, can be estimated using the empirical moments, following Remark 4.1.

5.1.2.2 Using the Service Times of Non-Waiting Customers

When the estimation of the offered-load process is required, then call by call data is needed, whereas in the estimation of the offered-load function only averages are necessary. We discover in Chapter 7 that, under the procedure applied in that chapter, the estimation of $E(S|\tau = w)$ is unreliable. However, as mentioned in Proposition 1, the unconditional distribution of the service time is expressed through the service time of

non-waiting customers. Practically, it can be estimated from calls that enjoy very short waiting times (e.g. less then 5 seconds). Following the second method in Section 5.1.1, one can estimate the offered-load function in terms of the arrival process and the unconditional distribution of the service time, without taking into account the relationship between patience and service time. Moreover, if the arrival process is a homogeneous Poisson process, then the mean service time itself is sufficient.

5.2 Prediction of the Offered-Load Process

An important aspect in managing service systems is determining appropriate staffing level. In other words, a manager wishes to know how many servers to assign to work at any time, in order to satisfy a predetermined service level. This amounts to predicting the amount of work that is expected to be in the system. However, the amount of work in the system is determined by the arrivals to the system, the service times, the patience of the customers, the routing of customers in the system and by the agents who treat them. Under the assumptions of the $M_t/GI/N_t + GI$ queue, the square-root staffing rule, presented in section 3.1.4, stabilizes the performance of the system over the day. Under this rule, the staffing level should be calculated with respect to the offered-load [36]. This result suggests that a prediction of the offered-load process is essential for determining staffing levels.

Usually staffing levels are determined in advance, several days ahead, up to one day in advance. However, it can be beneficial to update the staffing level dynamically, exploiting up-to-date system information. This can be done by dynamically revising of the offered-load prediction during the day. We refer to this dynamic prediction as *intra-day prediction*.

As discussed by Shen [10], a manager can use the revised prediction of the offered-load, by adjusting the staffing level within the day. If the suggested staffing level is higher than the actual staffing level, then the manager may offer overtime work to agents on duty, direct calls to agents who work from home or in a back-office, or restrict agent activities to only handling incoming calls. In the opposite case, when the staffing level is higher than required, the manager may dismiss agents early or move them to handle other activities.

Most research on prediction of the offered-load process in service systems focuses on the prediction of the arrival process. Then, approximations of the offered-load is made, based

on the distribution of the service time, as described in Section 3.1.4. In service systems that face high variability in demand, these approximations sometimes tend to be poor. In this context, there are several papers which investigated the dynamic prediction of the arrival process, and which are presented in the literature review in Section 2.2. These methods produce a prediction of the number of arrivals during short time intervals of the day, up to one day in advance. Then, as new data is gathered at the beginning of the day, the arrival process up to a certain time point is taken as input, which helps updating the prediction of the arrival process up to the end of the day.

Based on a work of Goldberg et al. [32], *we propose to apply the prediction of the offered-load process directly.* Previously in this chapter, an estimation of the offered-load process was presented. We use the estimates of the offered-load processes in previous days, as the input for the prediction models. Weinberg [12] and Shen [11] exploit in their forecast the properties of an assumed Poisson arrival process. Whereas the arrivals to the system can be often assumed to follow a Poisson process, the offered-load process is definitely not Poisson. However, in [32] the predicted process is not restricted to being a Poisson process.

In their work, they consider the problem of forecasting the continuation of a curve, using functional data techniques. Here, the arrival process or the offered-load process are assumed to be sampled from some underlying smooth curve. Such curves are collected previously to the forecasting period. Then, they estimate the continuation of a new curve, given its beginning, using the behavior of the previously collected curves. We refer the reader to details on this method in their paper: "The Best Linear Unbiased Estimators for Continuation of a Function" (2010) [32].

Chapter 6

Staffing the $M_t/GI^*/n_t + GI$ Queue

After the analysis of the relationship between the patience (waiting) of customers and their service times, it is interesting to obtain insight into the influence of this relationship on models for determining required staffing levels.

First, we consider the case described by Whitt in [29], where staffing is set by aiming to immediately answer all calls. Here, the waiting times are expected to be eliminated, causing all customers to be served, independently of their patience. In this case, the relationship between patience and service time does not play a role in the model. Thus, all the conclusions of [29] are valid, and the only issue that is left to be applied is to estimate the unconditional service time of the customers (Since we assume that the input data for the model is drawn from a service system where there are both waiting times and abandons). See Section 2.2 for a brief overview of [29].

Now, consider a service system where considerable waiting is required. The staffing rule and other conclusions of the Erlang-A model, take into account the arrival rate, the service times and the fact that some customers shall abandon, depending on the quality of service. An essential assumption for this model is that the patience and the service times are independent, and that both of them are not dependant on the system's state. *However, in systems where patience and service times are dependent, the service times that the system faces vary as the required waiting time is changing, which is influenced by the staffing level*, as demonstrated in Figure 6.1.

Denote the service time of a served customer, in a steady state system with N servers, by $S^*(N)$. Then, for a given staffing level N, the mean service time of served customers

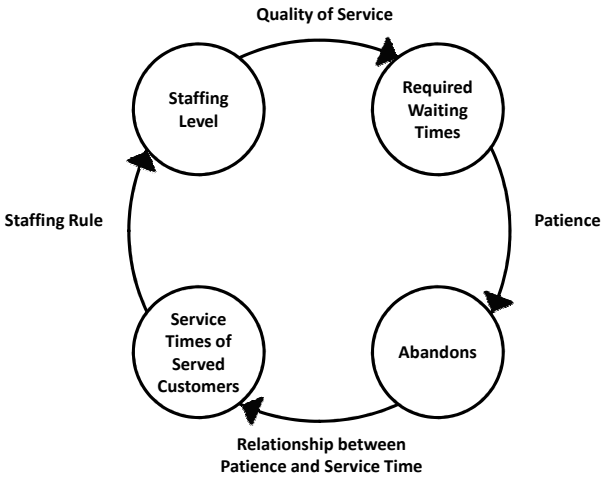


Figure 6.1: Diagram of the dependencies between the staffing level and the staffing rule parameters. The mean service time is an input for the (square-root) staffing-rule; but this mean is also determined by the staffing level, assuming the presence of a relationship between patience and service times.

is given by:

$$E(S^*(N)) = \int_{w=0}^{\infty} g(w) \cdot f_W(w)\, dw\,, \tag{6.1}$$

where $g(w) = E(S|\tau > W = w)$ and f_W is the pdf of the waiting time.

Consider a service system, in which the queue discipline is FCFS, and assume that patience and the required waiting time are independent, and that staffing level is fixed. Then the quality of service is determined by the service times of the served customers only (along with the arrivals and patience times of all customers). Hence, we make the hypothesis that staffing level and quality of service are related through the value of $E(S^*(N))$, and not through $E(S)$ as given by the Garnett function (2.4): specifically the Garnett function works in this case, with the mean service time $1/\mu$ taken to be $E(S^*(N))$. This hypothesis is analyzed in the next subsection, using simulations for two types of relationships. Further analysis, including possibly theory, is left for future research.

A main limitation of the above, for applications, is that the distribution of the waiting time is unknown and depends on the staffing level. Nevertheless, for a specific system, where all the required parameters are known, $E(S^*(N))$ can be estimated, for several values of N, using simulations.

6.1 Analysis of Service Parameters via Simulations

This section is taken from [24].

6.1.1 Description of the ISA Algorithm

The **Iterative-Staffing-Algorithm (ISA)**, developed by [35], determines time-dependent staffing levels aiming to achieve a given constant-over-time delay probability, α. *In our implementation, we add the feature of generating a random service time from a conditional service distribution, which is determined by the patience of a customer.* For simplicity of the results, we consider in our simulations only the case where the arrival rate is constant, although the following formalization applies to a time-varying arrival rate.

For implementation of the algorithm, we start with an $M_t/GI^*/n_t + G$ queue, with homogeneous customers. We assume that service times are dependent on patience times, but both patience and service times are independent of the arrival rate or system's state. The arrivals are according to a nonhomogeneous Poisson arrival process, with the arrival function $\lambda = \{\lambda(t),\ 0 \leq t \leq T\}$.

To start, we fix an arrival rate function $\lambda(t)$, a patience distribution , a service time distribution and the relationship between patience and service time, and finally a time horizon $[0, T]$. Although the algorithm is formulated in continuous time, staffing decisions are made at discrete times. This is achieved by dividing the time-horizon into small intervals of length Δ. The number of servers is constant within each such interval.

The service discipline is FCFS, and servers follow an exhaustive service discipline: a server that finishes a shift in the middle of a service will complete the service and sign out only when finished.

Denote:

- $n_t^{(i)}$ - The staffing level at time t in iteration i (equal for all replications of a certain iteration).
- $L_t^{(i)}$ - The random total number of customers in the system at time t under the staffing level n_t^i.

We estimate the distribution of $L_t^{(i)}$ for each i and t by performing multiple (5000) independent replications. We start with infinitely many servers, at the first iteration. Since this is a simulation, we choose a large finite number of servers, ensuring that the probability of delay (i.e., of having all servers busy upon arrival) is negligible for all t.

The algorithm iteratively performs the following steps, until convergence is obtained. Here convergence means that the staffing level does not change much after an iteration. (Practically, they are allowed to change by some threshold τ, which we take to be 1.)

1. Given the i^{th} staffing function $\{n_t^{(i)} : 0 \leq t \leq T\}$, evaluate the distribution of $L_t^{(i)}$, for all t, using simulation.

2. For each t, $0 \leq t \leq T$, let $n_t^{(i+1)}$ be the least number of servers such that the delay probability constraint is met at time t, i.e.

$$n_t^{(i+1)} = \text{argmin}\left\{n \in \mathbb{N} : E\left[\mathbb{I}_{\{L_t^{(i)} > n\}}\right] \ \leq \ \alpha\right\}.$$

3. If there is a negligible change in the staffing from iteration i to iteration $i+1$, then stop; i.e., if

$$\| \, n^{i+1} - n^i \, \| \equiv \max\left\{|n_t^{(i+1)} - n_t^{(i)}| : 0 \leq t \leq T\right\} \leq \tau,$$

   then stop and let $n_t^{(i+1)}$ be the proposed staffing function. Otherwise, advance to the next iteration, i.e., replace i by $i+1$ and go back to step 1.

Let n_t^{ISA} denotes the the final staffing level at time t and L_t^{ISA} denotes the (random) number in system at time t with that staffing function n_t^{ISA}.

The implementation of the algorithm is written in C++ and is based on code written by Z. Feldman (see [34], cf. [35]). For more features of the program, see the beginning of Chapter 7.2.

6.1.2 Short-Term versus Long-Term Performance Measures

Assume a service system with an arrival rate $\lambda(t)$, that varies from interval to interval. Moreover, assume that the system at time interval t is staffed by n_t servers, according

to some staffing rule. Analyzing such service system involves calculation of expected performance measures in two different contexts. The first relates to performances on a specific replication (assuming that a single replication stands for a specific day), and the second to performances in the long-run (say, one month period). Following Henderson [23], we call these two cases *short-term performance measures*, and *long-term performance measures*, respectively.
The short-term and long-term performance measures differ in terms of how one should weight the performance. Since more customers experience the performance associated with a large number of arrivals and vice versa, considering a period of days (long-term) requires weighing by the daily number of customers.

Let Υ_{t,n_t} be a system performance measure (e.g the delay probability or the queue length) under staffing of n_t. Assume a set of d days, and denote:

- $Arrived_t^j$ - Number of arrivals at time interval t on day j, $j = 1, 2, \ldots, d$.
- v_{t,n_t}^j - The value of the performance measure Υ_{t,n_t}, at time interval t on day j .

The short-term value of Υ_{n_t}, based on the d days data, is simply the arithmetic average

$$\bar{\Upsilon}_{t,n_t} \triangleq \frac{\sum_{j=1}^{d} v_{t,n_t}^j}{d}. \tag{6.2}$$

The calculation of the long-term value requires differentiation between two classes of performance measures. If a performance measure relates to the system state, for example, the queue length and the offered-load, one would average its values over the period of d days, as in (6.2). But the appropriate long-term measures that relate to customers experience (e.g. the delay probability and average waiting time) weight the performance by the number of arrivals on each day:

$$\widetilde{\Upsilon}_{t,n_t} \triangleq \frac{\sum_{j=1}^{d} Arrived_t^j \cdot v_{t,n_t}^j}{\sum_{j=1}^{d} Arrived_t^j}. \tag{6.3}$$

6.1.3 Calculation of Performance Measures

Here we present an extensive procedure for estimating performance measures in ISA. Subscript t indicates the t^{th} partition interval of size Δ; superscript j indicates the j^{th} replication.

Denote:

- $Reps$ - The total number of replications.
- $Arrived_t^j$ - Number of arrivals during time interval t in replication j.
- $Delayed_t^j$ - Number of customers who arrived at time interval t in replication j and did not start their service immediately.
- $Abandoned_t^j$ - Number of customers who arrived at time interval t in replication j and eventually abandoned.
- $WaitingTime_t^j$ - The total waiting time of customers who arrived at time t in the replication j.
- L_t^j - The total number of customers in the system at the end of interval t in replication j.
- Q_t^j - The queue length in interval t in replication j.
- $Busy_t^j$ - The total time the servers were working during the interval t in the j^{th} replication.
- n_t - The number of servers during the interval t (fixed over the interval, for all replications).
- ρ_t^j - The servers utilization during interval t in replication j. $\rho_t^j = \frac{Busy_t^j}{\Delta \cdot n_t}$.

Tables 6.1 and 6.2 summarize the formulae for calculations of the performance measures in the simulation software. Table 6.1 contains the performance measures which are related to the customers, and presents the calculations for both short-term (arithmetic average) and long-term (weighted average). Table 6.2 contains the performance measures which are related to the system, and calculated by arithmetic averages.

Performance Measure	Short-Term	Long-Term
Delay Probability, $\mathbf{P_t\{W_q > 0\}}$	$\frac{1}{Reps} \cdot \sum_{j=1}^{Reps} \frac{Delayed_t^j}{Arrived_t^j}$	$\frac{\sum_{j=1}^{Reps} Delayed_t^j}{\sum_{j=1}^{Reps} Arrived_t^j}$
Abandonment Probability, $\mathbf{P_t\{Ab\}}$	$\frac{1}{Reps} \cdot \sum_{j=1}^{Reps} \frac{Abandoned_t^j}{Arrived_t^j}$	$\frac{\sum_{j=1}^{Reps} Abandoned_t^j}{\sum_{j=1}^{Reps} Arrived_t^j}$
Average Waiting Time, $\mathbf{E_t[W]}$	$\frac{1}{Reps} \cdot \sum_{j=1}^{Reps} \frac{WaitingTime_t^j}{Arrived_t^j}$	$\frac{\sum_{j=1}^{Reps} WaitingTime_t^j}{\sum_{j=1}^{Reps} Arrived_t^j}$
Average Waiting Time Given Waiting, $\mathbf{E_t[W \mid W > 0]}$	$\frac{1}{Reps} \cdot \sum_{j=1}^{Reps} \frac{WaitingTime_t^j}{Delayed_t^j}$	$\frac{\sum_{j=1}^{Reps} WaitingTime_t^j}{\sum_{j=1}^{Reps} Delayed_t^j}$

Table 6.1: Calculating Performance Measures - Customers

Performance Measure	Simulation Calculation
Offered-Load, $\mathbf{R_t}$ (calculated at iteration 1)	$\frac{\sum_{j=1}^{Reps} L_t^j}{Reps}$
Server Utilization, $\rho_\mathbf{t}$	$\frac{\sum_{j=1}^{Reps} \rho_t^j}{Reps}$
Queue Length, $\mathbf{Q_t}$	$\frac{\sum_{j=1}^{Reps} Q_t^j}{Reps}$

Table 6.2: Calculating Performance Measures - System

6.2 Examples

ISA was applied to three types of $M/M^*/n_t + M$ queueing systems, all with similar parameters, but they differ by the relationship between patience and service time: in one model, the mean service time is an increasing monotone function of the patience of a customer; in another model, the mean service time is a decreasing monotone function of the patience of a customer; in the last model, the patience and the service time are independent - this model serves as a control for the simulation results and as a baseline for comparing the two other models.

Description of Models:

- The arrivals to the system are according to a homogeneous Poisson process with arrival rate of 100 customers per time unit.
- The (unconditional) mean service time is equal to 1 time unit. The service time conditional on the patience of a customer is exponentially distributed.
- Customer patience is distributed exponentially with mean equal to 1 time unit.
- The queue discipline is FCFS.
- The running horizon is 24 time units and performance statistics are collected after the 4th time unit to make sure that the system reaches a steady state.
- All the values are calculated as an average of 5000 replications.

The parametrization of the relationship between the patience and the service time follows Example 1. The formulae for the conditional mean service time as a function of the patience ($E(S|\tau = w)$) and as a function of the waiting time ($g(w) = E(S|\tau > W = w)$), for all three models, are given in Table 6.3. These functions are plotted in Figure 6.2 (Increasing Monotone Function) and in Figure 6.3 (Decreasing Monotone Function).

Relationship Type	E(S)	$E(S|\tau = w)$	g(w)
No Relation	1	1	1
Increasing Monotone Function	1	$1.2 - e^{-4 \cdot w}$	$1.2 - 0.2 \cdot e^{-4 \cdot w}$
Decreasing Monotone Function	1	$0.8 + e^{-4 \cdot w}$	$0.8 + 0.2 \cdot e^{-4 \cdot w}$

Table 6.3: Examples of service time conditional on patience and on waiting - three models.

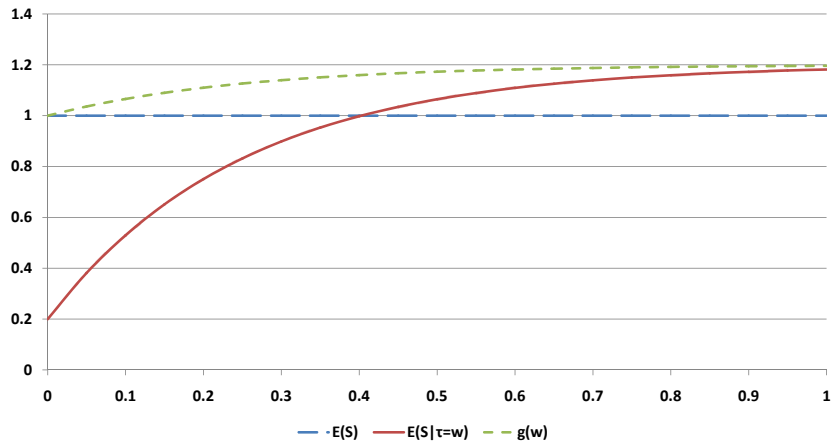


Figure 6.2: Monotone increasing example - service time conditional on patience and on waiting.

Next, we introduce some interesting properties of performance measures that are related to the dependency between patience and service times of customers, as they arise via simulation results of the above described models.

6.2.1 Analysis of Service Times of Served Customers

As discussed in Chapter 4 and shown in Figures 6.2 and 6.3, when there is a relationship between patience and service time, the mean service time of customers varies as a function of their waiting times. For example, in the model of the increasing monotone function, which is described in Table 6.3, the patience and the service time are positively correlated. If we consider all the customers who waited exactly w time units and were served, we

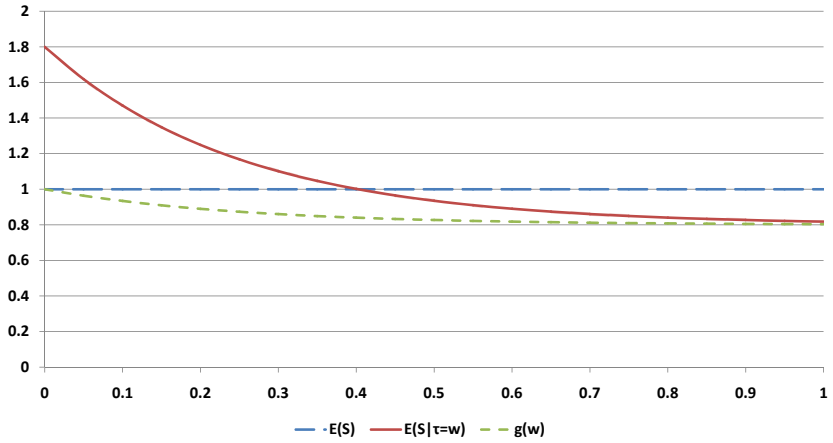


Figure 6.3: Monotone decreasing example - service time conditional on patience and on waiting.

do not account for customers with patience less than w. As a result, we tend to omit the service times of customers with shorter patience times. Thus, in this example, as the waiting time grows, we consider only groups of customers with higher expected service times.

When comparing systems with the same parameters, apart from their staffing level, a higher staffing level yields better quality of service (i.e. shorter waiting times and less abandons) and more served customers, whose expected service times are hence shorter. Consequently, with higher levels of staffing, the servers observes more shorter service times. Figures 6.4 and 6.5 demonstrate, for each model, the change in the mean service times of served customers as a function of the staffing level and of the probability of waiting, for both the increasing and decreasing monotone function models.

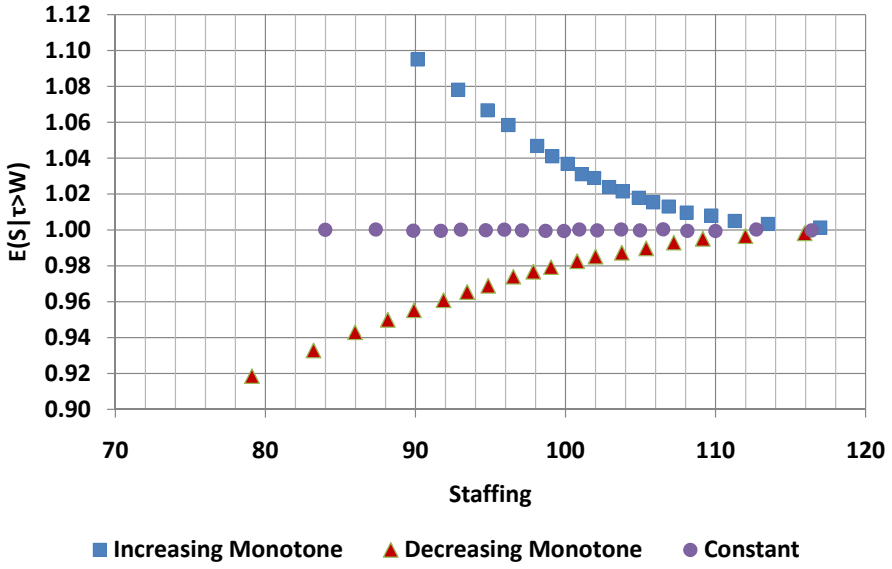


Figure 6.4: Mean service times of served customers, as a function of the staffing level. In all simulations the offered-load is $R = 100$.

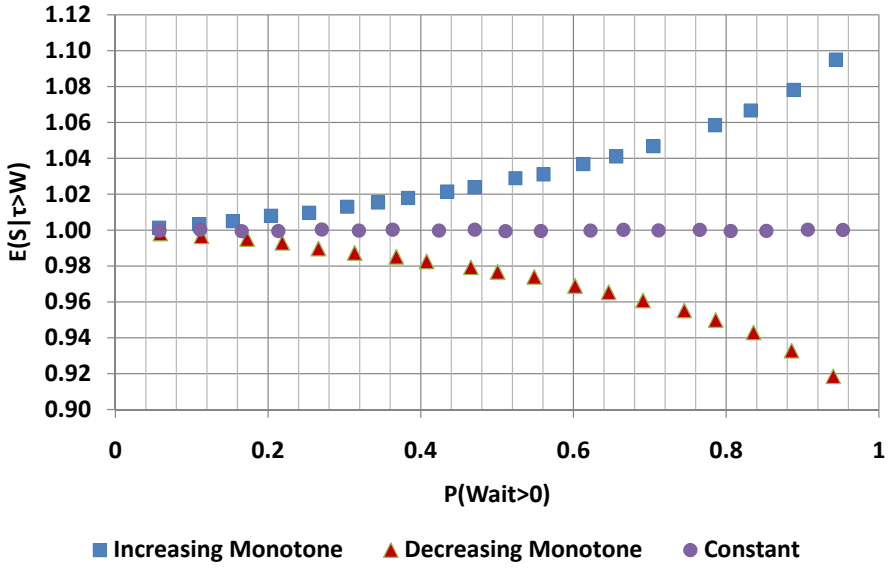


Figure 6.5: Mean service times of served customers, as a function of the probability of waiting. In all simulations the offered-load is $R = 100$.

6.2.2 Analysis of the Square-Root Rule

At the beginning of this chapter, we claimed that using the (unconditional) mean service time may not be sufficient, when applying a staffing rule for a service system where the patience and the service times are dependent. Furthermore, we suggested that an appropriate candidate to replace the measure of the mean service time is the mean service time of only the served customers (i.e. the service times that are observed by the system). As stated before, this measure depends on the required waiting times (determined by the staffing level).

We now look for an analogue staffing rule to the rule presented in Theorem 2.2 (which is appropriate for an Erlang-A queue where patience and service time are independent).

Define $n = R + \beta\sqrt{R}$, where $R = \lambda \cdot E(S)$ is the offered-load function and β is the Quality-of-Service (QoS) parameter.

Then we define the implied QoS parameter as

$$\beta^{ISA} \equiv \frac{n^{ISA} - R}{\sqrt{R}}. \tag{6.4}$$

Figure 6.6 plots the implied QoS parameter for each of the three models (Increasing monotone, Decreasing monotone and constant functions, as described at the beginning of Section 6.2). One observes from the figure that in the constant mean service time case, the QoS grade falls exactly on the Garnett Function (2.4), in the increasing monotone function case, the QoS grade of Garnett Function is lower then the implied QoS grade (i.e. a higher staffing level is required in order to satisfy the same probability of waiting); finally, in the decreasing monotone function case, the QoS grade of Garnett Function is higher then the implied QoS grade.

We relate this phenomenon to the fact that the actual mean service time that the system faces (due to served customers) is different from the unconditional mean service time, as described at the beginning of this chapter and in Subsection 6.2.1. Notice that the offered-load function is not influenced by the relationship since it assumes a queueing system with infinitely many servers, where patience is not expressed.

Now, we define a modified offered-load expression, given by:

$$R^* = \lambda \cdot E(S^*(N)), \tag{6.5}$$

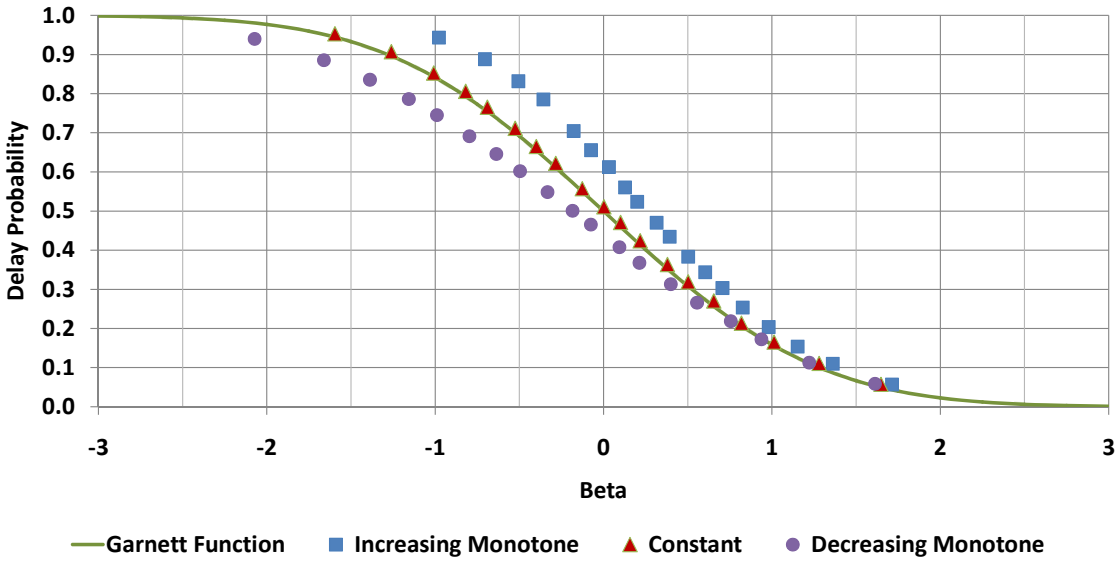


Figure 6.6: Delay probability vs β for increasing monotone, decreasing monotone and constant $E(S|\tau = w)$.

where $E(S^*(N))$ is the mean service time of served customers only, as defined in Formula (6.1). Figures 6.7 and 6.8 show that, using the modified offered-load, in the expression of the implied quality of service grade, instead of the offered-load, places it back on the Garnett-Function. This validates our hypothesis that the staffing level and the quality of service are related through the value of $E(S^*(N))$ and not through $E(S)$.

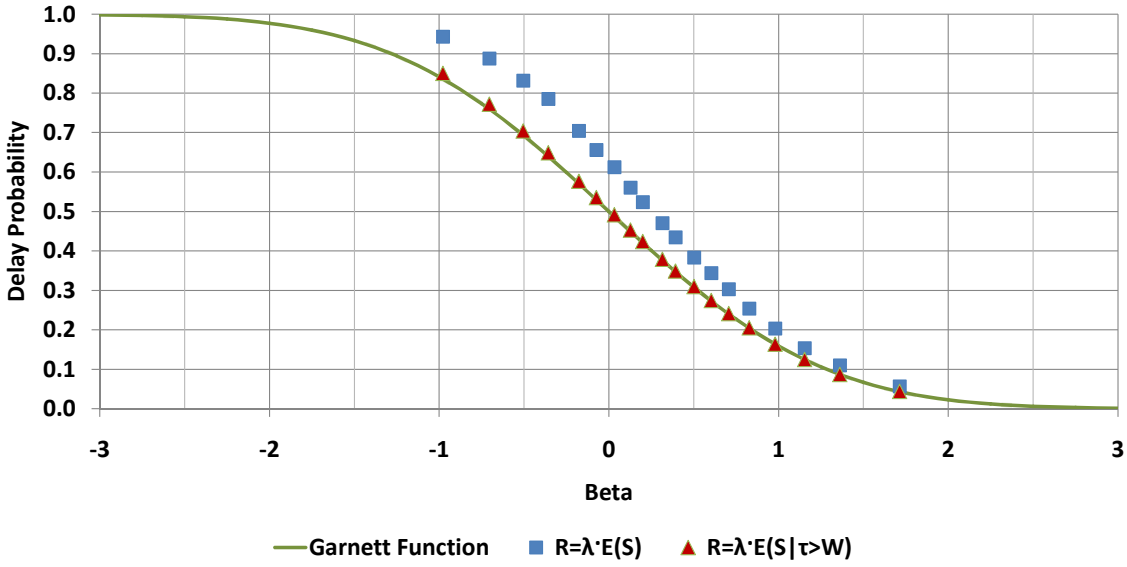


Figure 6.7: Delay probability vs β and modified *beta* for increasing monotone $E(S|\tau = w)$.

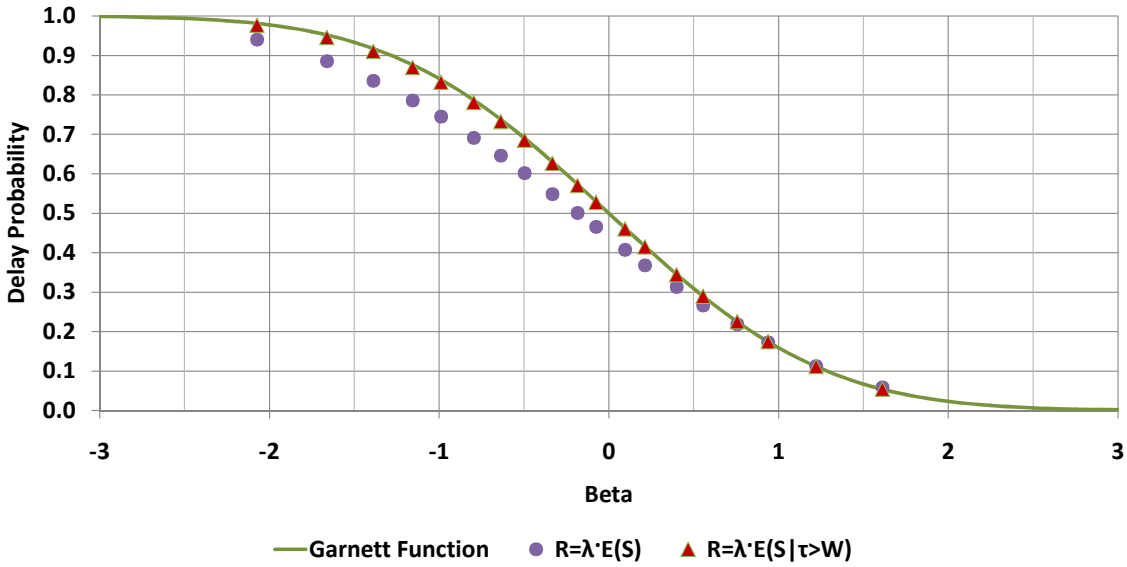


Figure 6.8: Delay probability vs β and modified *beta* for decreasing monotone $E(S|\tau = w)$.

6.2.3 Other Performance Measures

At this stage, we summarize several additional performance measures and analyze the relationship between them. For the Erlang-A queueing model, these relationships are asymptotically explored (see Section 2.1.3). However, in the $M_t|GI^*|n_t + G$ queue, the relationships may differ from those of the Erlang-A model, as demonstrated in Figure 6.9. This is explained by the fact that the service time distribution of the served customers is different from the prior distribution of the service time of any random customer. Notice that from Figures 6.9(c) and 6.9(f), apparently the relationships between the expected waiting time and the probability to abandon, and the relationship between the expected waiting time and the average queue length, are consistent with these relationships in the Erlang-A model.

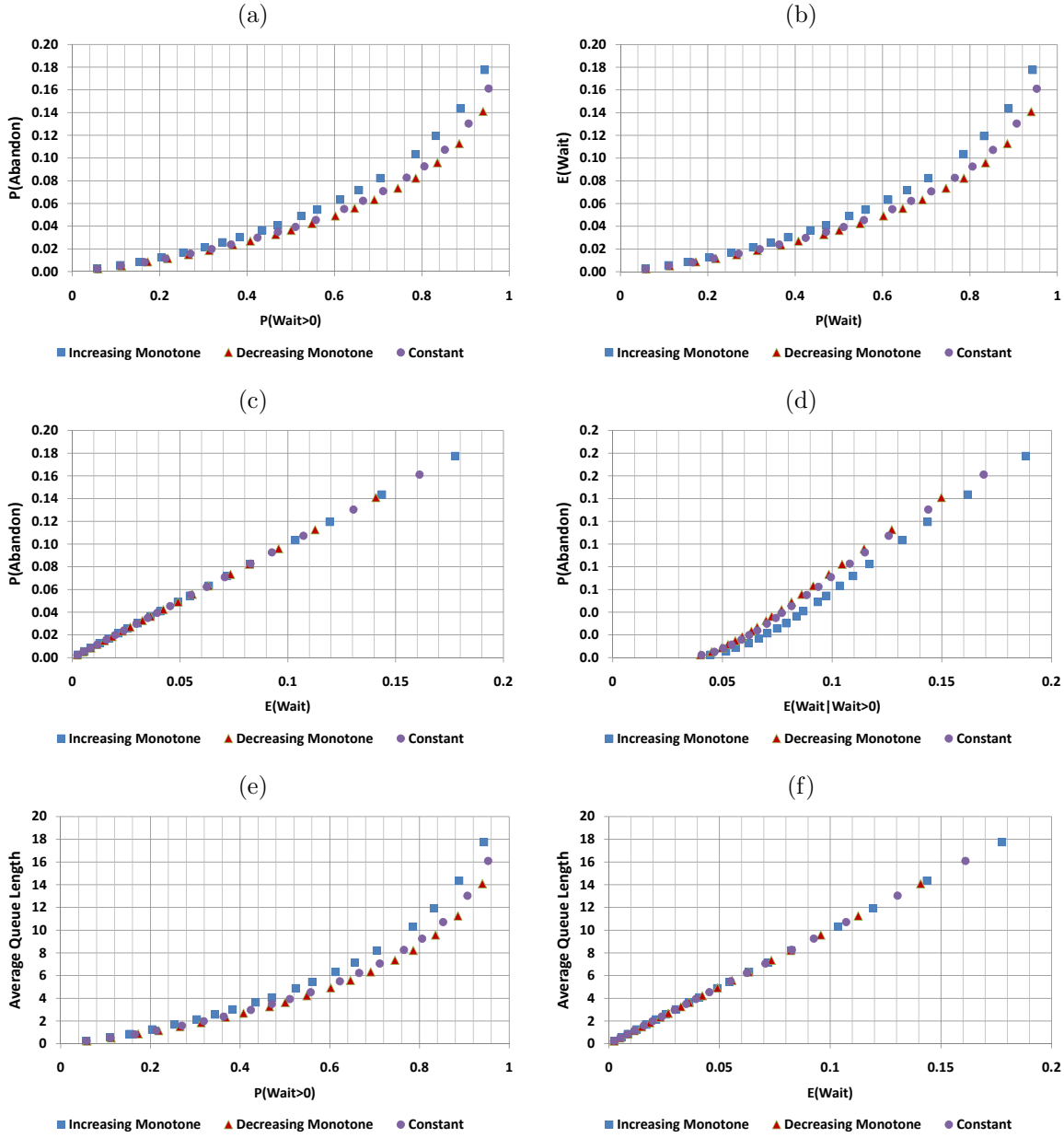


Figure 6.9: Various plots of the relationships between performance measures, according to simulation results of the three models.

Chapter 7

Data and Simulation Based Analysis

In this section, we test our theory against empirical and simulated results. The first stage was to identify a data set, where a relationship between the patience of customers and their service times exists. We scanned this relation in databases of call centers of a U.S. Bank and an Israeli Telecom company, available at the Technion SEE Laboratory. For a variety of service groups, a period of half a year was considered. Then, all served calls where categorized into groups by their waiting times. We plotted the average service time as a function of the waiting time (in seconds). In this manner, we could identify some service groups in which there is a strong evidence of the sought-after relation, as demonstrated in Figure 7.1.

In Subsection 7.1, we focus on analyzing the data set of the Retail service group in the U.S. Bank. As shown below, there are many challenges that arise in the analysis of empirical data (i.e. noise, multi-type customers and heterogeneity among customers, variance and dimensionality issues). These topics are covered in this section. To understand better the reasons and the effects of these problems, we bring in Subsection 7.2 simulation results, where all service parameters are controlled and the above issues can be then isolated.

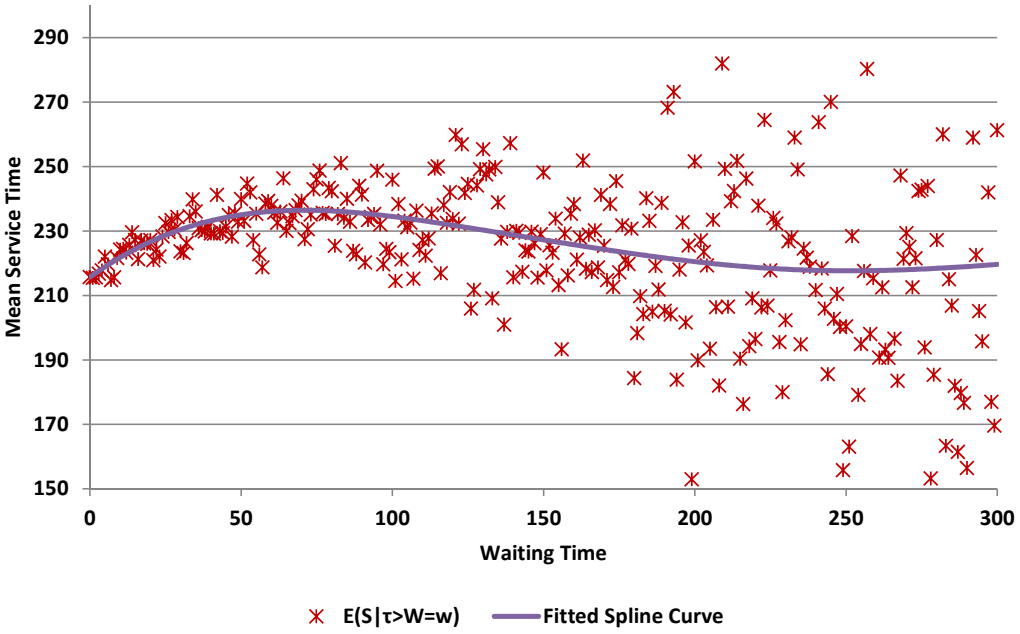


Figure 7.1: A plot of the mean service time, of served customers, as a function of their waiting time, in the U.S. Bank Retail line. The fitted spline is a cubic spline with 5 knots.

7.1 Empirical Results

7.1.1 The Data

Our empirical analysis is based on a data set, originating at a call center of a large North American commercial bank (U.S. Bank). The bank data originated from the ongoing Data MOdels for Call Center Analysis (DataMOCCA) research project, conducted at the SEE[1] Laboratory of the Technion. (For more information on SEE and the DataMOCCA project see [9].)

The call center of U.S. Bank provides service for various types of operations, such as Retail, Business, Telesales, Online Banking and Consumer Loans. We restrict our analysis to the Retail Banking service type, which covers 68% of the total incoming calls. The call center operates seven days a week, 24 hours a day. On regular weekdays, the majority of the daily incoming calls arrive between 07:00 and 21:00, where the arrival rate increases sharply between 07:00 and 10:00. Then, from 10:00 to 16:00 the number of arrivals slowly

[1]SEE = Service Enterprise Engineering

decreases. Finally, the number of arrivals constantly decreases until the late night hours that are characterized by small arrival volumes.

The time period of the analysis is taken over all weekdays (Monday through Friday) between January 1st, 2006 and June 30th, 2006. In order to keep the calls as homogeneous as possible, we consider only calls that entered the system between the hours 10:00 and 16:00. The total number of observations is then 2,722,129, out of which 2,683,418 calls where served. Each call has an arrival time, a waiting time and a service time or an abandon indicator. The waiting and service times are stored in the database in seconds resolution.

7.1.2 Testing the relationship between patience and service time

In the first stage we wish to explore the relationship between patience and service time. Figure 7.1 describes the average service time of served customers as a function of their waiting time. The waiting time in the plot is truncated by the value of 300 seconds, since the number of observations with waiting time greater then 300 is relatively low (less then 100 observations for any waiting time value, in seconds), which cause a high variation of the means. Looking at Figure 7.1 gives rise to the suspicion that the service time of customers indeed depends on their patience. Moreover, service time and patience seem to be positively correlated.

Here, we test if the dependency is significant, using the permutation test described in Section 4.2. We omit all non-waiting observations, more explicitly, with waiting time less than or equal to 1 second, which do not carry any information on the patience and consequently do not reflect on the relationship between patience and service time. The remaining observations are then divided into 9 groups, by the rank of their waiting times, such that in each group the number of observations is similar. The summary of the groups boundaries and frequencies is given in Table 7.1. Then, the test statistic is calculated once on the original data and 4000 additional times on random pairing permutations, in order to generate a null distribution (where no relationship between the patience and the waiting time exists) of the test statistic. The original permutation statistic (with value of 46,140.62) is greater than any of the other 4000 random pairing permutations (p-value=0). Thus, we reject the assumption that the service time and patience are independent. The null distribution of the test statistic, generated by the random pairing permutations, is shown in the histogram of Figure 7.2.

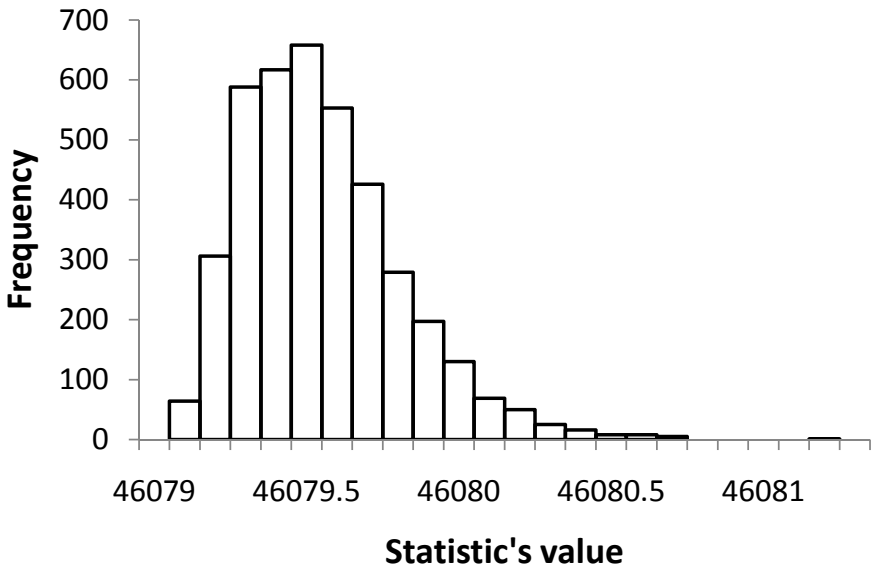


Figure 7.2: A histogram for the distribution of the test statistic under the null hypothesis of no relation between patience and service time. The statistic's distribution is computed by 4000 permutations of random pairing between waiting times and service times of served customers.

Group	Minimal Value	Maximal Value	Frequency	Probability
1	2	3	96243	0.11
2	4	5	94091	0.11
3	6	7	103581	0.12
4	8	9	92460	0.11
5	10	11	145366	0.17
6	12	12	92301	0.11
7	13	20	85967	0.10
8	21	62	85691	0.10
9	63	5702	61114	0.07

Table 7.1: Ranks of served customers' waiting times in the Retail line of a U.S. Bank. The table includes, for each rank, the minimum and maximum values of the waiting times, the number of observations and their proportion out of total.

7.1.3 Analysis of the relationship

The test results from the previous subsection leads us to the analysis of the relationship between patience and service time. We start with inference on the distribution of service time, focusing on non-waiting calls. From Figure 7.3, it seems that the service time

distribution involves possibly three Log-normal distributions. A possible explanation for the mixture may be three different types of calls that reach the business line. For example: the very short times are system noise (i.e. calls that hang up right upon answer); the second type of up to 20 seconds are of customers who reached a wrong service line; the third main calls are of the customers who get the service that they were aiming at.

Since we know that the service time distribution depends on patience, in Figure 7.4 we also examine the changes in the distribution of the service time as the waiting time grows. In order to validate if a Log-normal distribution can be assumed, we also estimate a mixture of three Log-Normal distributions, using an EM algorithm which is described in [13]. A plausible assumption that we raise is that, in this case, only the major component of the mixture (i.e. the higher service time distribution) is dependent on the waiting time and the other two shorter service time distributions are independent of the waiting time. We do not analyze this conjecture further in this work, but if it is valid, then one should account for this assumption, and apply the relationship inference only to the third component. Ignoring this phenomenon may cause poor inference on the relationship between patience and service time, especially for short waiting (patience) times.

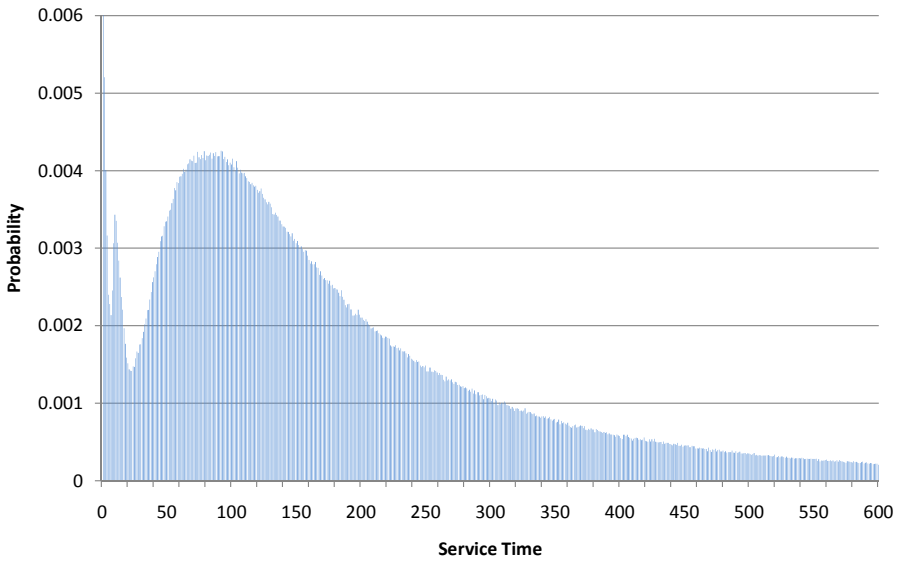


Figure 7.3: The distribution of the service time of all served customers.

Now, we estimate the mean service time of customers who waited w seconds and were served. In order to estimate $E(S|\tau = w)$, using Formula (4.5), we also estimate a cubic smooth spline for the mean service time as a function of the waiting time. The spline was run in the statistical program 'R', using the smooth.spline() function, with 5 knots. The

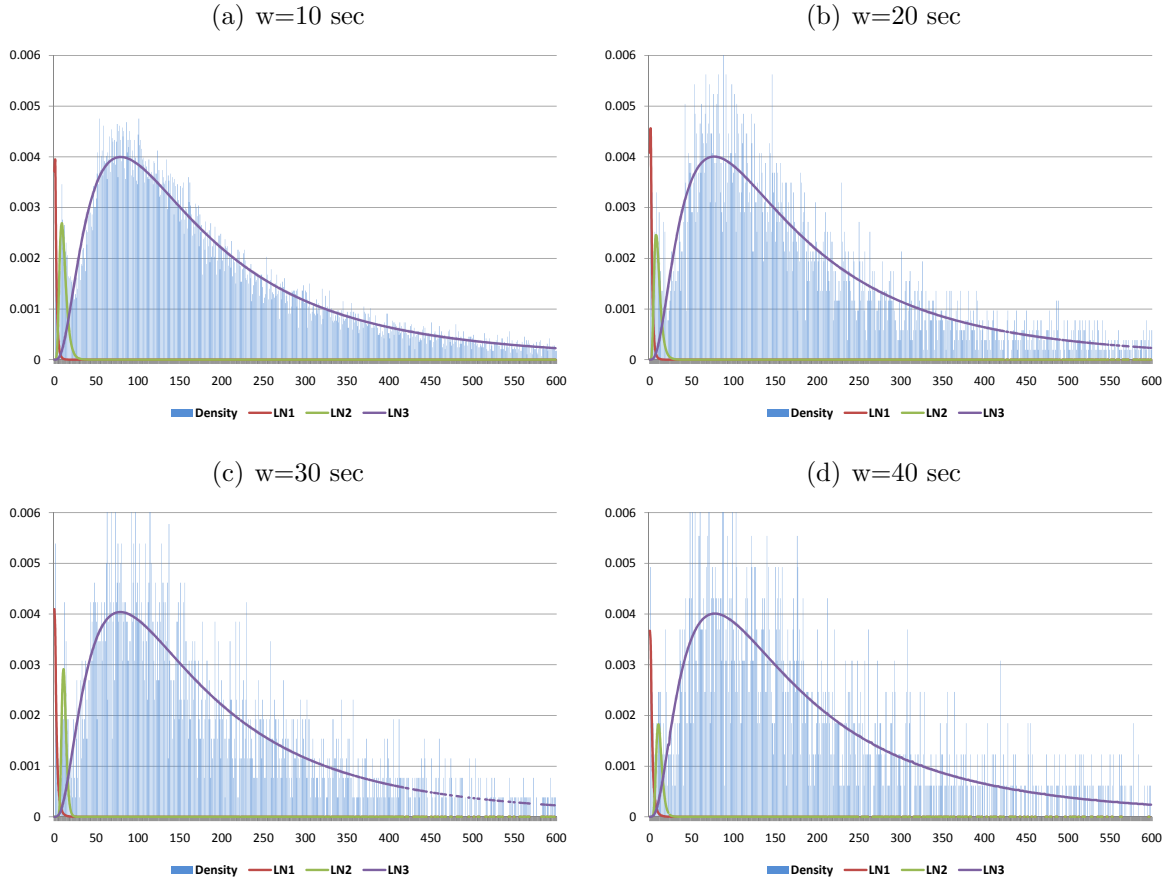


Figure 7.4: Distributions of the service times of served customers, given the waiting times of the customers. Each plot stands for a different value of waiting time w. For each chart, an estimate for the 3 Log-Normal distribution components of the mixture is plotted.

spline function estimator is chosen here since it is designed to handle smooth functions with no specific structure. Moreover, it enables one to simply extract the derivatives (first and second derivatives in a cubic spline) of the functions, which is required according to Formula (4.5). We chose a spline with only 5 knots in order to keep the first derivative of the spline (which is required in Formula (4.5)) as smooth as possible.

The spline fit for $g(w) = E(S|\tau > W = w)$ is plotted in Figure 7.1, at the beginning of this chapter. From the figure, one observes that in the first 70 seconds of the waiting time, the mean service time is constantly growing. Then, it mildly decreases until reaching waiting times of 250 seconds. From then on, there is a fairly small number of observations, but we assume that the means of service time of served customers with patience greater than 250 seconds is constant, as a function of patience. The spline's derivative with respect to the waiting times is given in Figure 7.5.

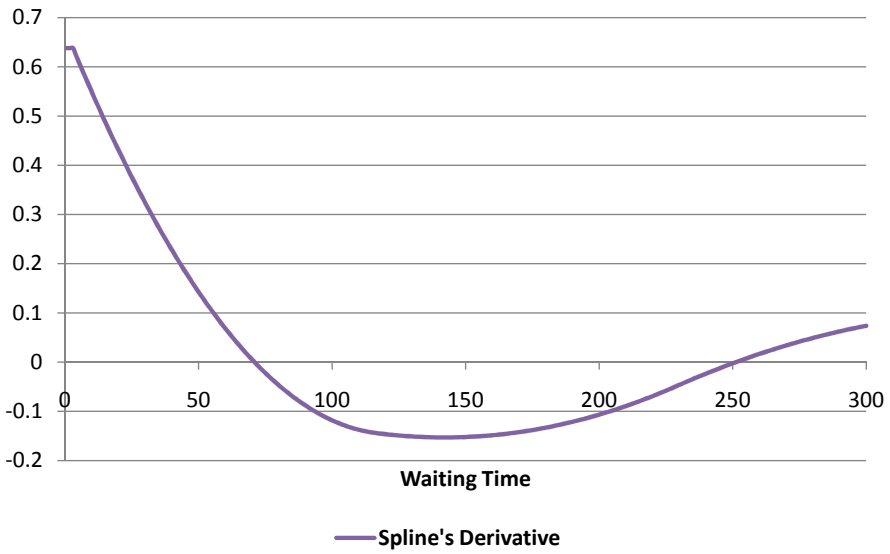


Figure 7.5: A plot of the derivative of the fitted spline for the mean service time, of served customers, as a function of their waiting times.

The last component that is left to be estimated according to Formula (4.5) is the hazard rate of the patience, denoted by $h_\tau(w)$. We extract the Kaplan-Meier (KM) product limit estimator for the hazard rate. We also create a smooth estimator for the hazard rate function using the Hazard Estimation with Flexible Tails (HEFT) algorithm of Kooperberg, Stone and Truong [5], which is also implemented in 'R'. The KM and HEFT estimators are shown in Figure 7.6.

At this stage, with all the components of (4.5) estimated, we construct the estimate for $E(S|\tau = w)$ (We using the HEFT estimator for the hazard rate function). As shown in Figure 7.7, the estimator is not stable and, moreover, achieves unrealistic measure for the service time as it goes below zero for low patience times. We do not have a solution for this problem, up to this time (see proposed future research in Chapter 8), but we assume that this phenomenon is related to an error of the estimator for the spline's derivative, together with the fact that the hazard rate is relatively low, causing a multiplication of this error by approximately 1,000, in the rightmost component in Formula (4.5).

From this point to the end of the chapter, we use simulation results in order to analyze the performance of our estimator.

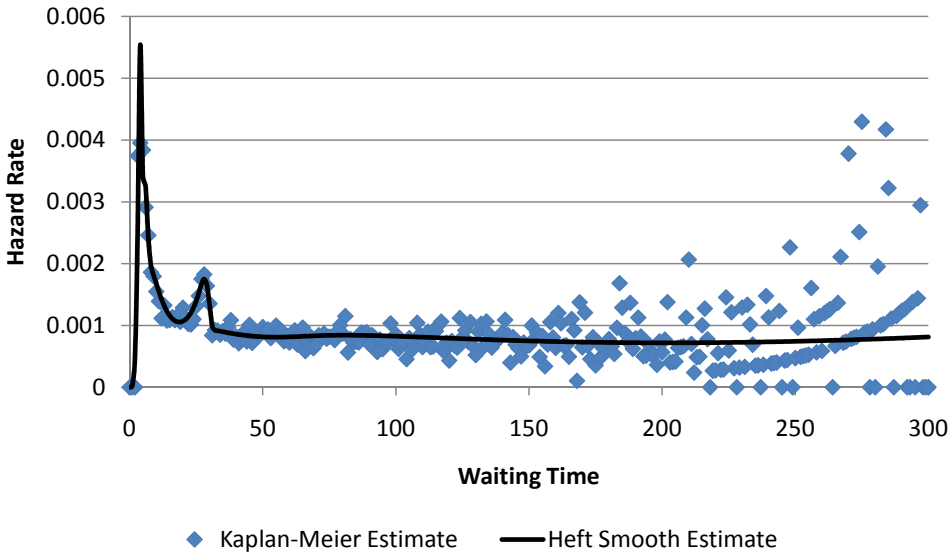


Figure 7.6: Estimators for the hazard rate function of patience: Kaplan-Meier estimator and a smooth HEFT estimator.

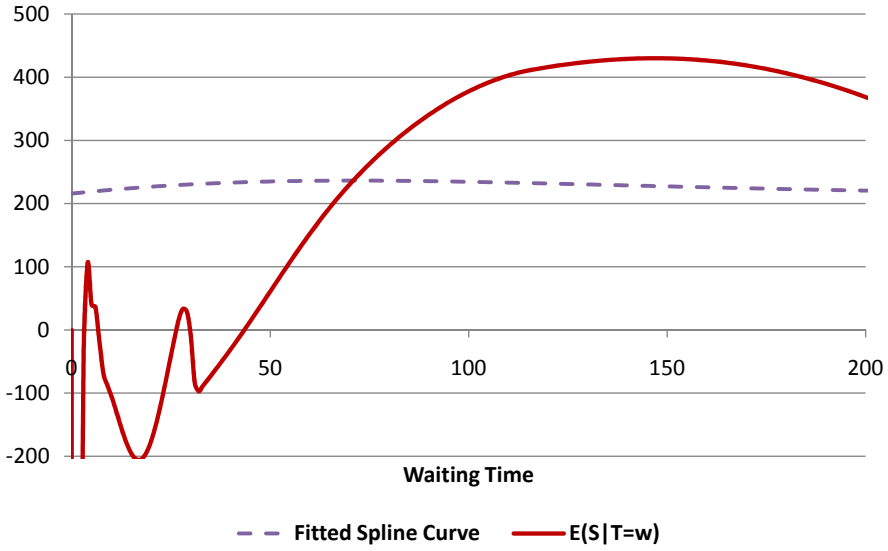


Figure 7.7: The estimator for the mean service time as a function of the patience of a customer, in comparison with the mean service time, of served customers, as a function of the waiting time.

7.2 Simulation Results

In this section, we investigate, through simulations, some of the properties that arise from the relationship between patience and service time. The simulations where built using the C++ programming language. The program consists of customer types, arrival processes, a queue engine, a service engine and a simulation body that operates the simulation. Before running a simulation, one can define the simulation length, number of replications of the simulation, staffing and customers types and their arrival processes. Any customer type is represented by a patience distribution, a service distribution (which may depend on the generated value of the patience of a customer) and priority. An arrival process (assumed to be a Poisson Process) corresponds to a customer type (only one per process) and an arrival rate which can be time dependent.

The engines are designed to handle customers who arrive to the system from all the defined arrival processes, and support different priorities of customers. They also record the operational information of any call transaction. Finally, all the above components are combined into the simulation body, that runs a pre-defined number of replications of the simulation. Then, it collects statistics from the simulation and records them. The results are kept for each customer type separately, and also for the total arrivals to the system. Another feature of the simulation is an iterative staffing algorithm, which is based on the ISA of [36] (see Subsection 3.1.4).

7.2.1 The Model

We wish to simulate a system similar to the system in our original data. Therefore, the simulation parameters are chosen so that the performances of the two systems will be similar. The chosen model is according to Example 1 in Section 4.1, where customers' patience (τ) is exponentially distributed, with mean $\frac{3600}{8} = 450$ seconds. The conditional distribution of the service time given the patience is Log-Normal. The conditional expectation is given by the following formula (in seconds):

$$E(S|\tau = w) = 230 \cdot (1.2 - e^{-w\cdot(\frac{29}{3600}-\frac{8}{3600})}) .$$

Here, according to Example 1, the mean service time of the observed (served) customers who waited w seconds is: $g(w) = E(S|\tau > W = w) = 230 \cdot (1.2 - \frac{8}{29} \cdot e^{-w\cdot(\frac{29}{3600}-\frac{8}{3600})})$ and $E(S) = 230 \cdot (1.2 - \frac{8}{29}) \approx 212.55$ seconds. The values of the Log-Normal parameters are chosen in several combinations as described below. In order to reconstruct the mean service time conditional on the patience of a customer, we use Formula (4.5):

$$E(S|\tau = w) = g(w) - \frac{g'(w)}{h_\tau(w)},$$

The derivative of $g(w)$ with respect to w is simply:

$$\frac{\partial}{\partial w} g(w) = 230 \cdot \frac{8}{29} \cdot e^{-w\cdot(\frac{29}{3600}-\frac{8}{3600})} \cdot \left(\frac{29}{3600} - \frac{8}{3600}\right)$$

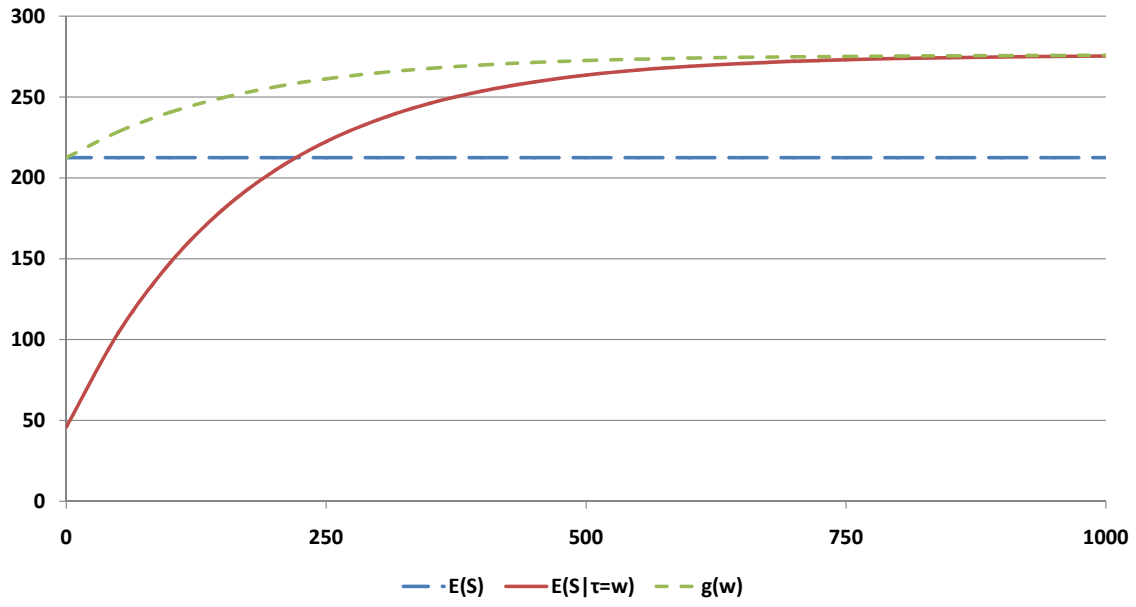


Figure 7.8: A comparison between $E(S|\tau = w)$ and $E(S|\tau > W = w)$, where $E(S|\tau = w) = 230 \cdot (1.2 - e^{-w\cdot(\frac{29}{3600}-\frac{8}{3600})})$, and the patience is exponentially distributed with mean $\frac{3600}{8} = 450$.

Assume that the service time of a customer with patience $\tau = t$ is a Log-Normal random variable, denoted by $S|\tau = t \sim LogNorm(\mu(t), \sigma^2)$, with a *pdf*

$$f_{S|\tau}(s|t) = \frac{1}{s\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln s - \mu(t))^2}{2\sigma^2}}, \; s, t > 0.$$

Then,

- $E(S|\tau = t) = e^{\mu(t)+\frac{\sigma^2}{2}}$.
- $E(S^2|\tau = t) = e^{\sigma^2} e^{\mu(t)+\frac{\sigma^2}{2}} = e^{\sigma^2} E^2(S|\tau = t)$.
- $Var(S|\tau = t) = (e^{\sigma^2} - 1)e^{\mu(t)+\frac{\sigma^2}{2}} = (e^{\sigma^2} - 1)E^2(S|\tau = t)$.

Let $E(S|\tau = t)$ be of the form:

$$E(S|\tau = t) = a \cdot (b - e^{-t\cdot(\beta-\theta)}),\ \beta > \theta,\ a > 0,\ b > 1.$$

This can be achieved by choosing $\mu(t) = \ln(a \cdot (b - e^{-t\cdot(\beta-\theta)})) - \dfrac{\sigma^2}{2}$. From Example 1, $E(S|\tau > W = t) = a \cdot (b - \dfrac{\theta}{\beta} \cdot e^{-t\cdot(\beta-\theta)})$. With an extension of Formula (4.4), supported in Remark 4.1,

$$E(S^2|T > W = w) = \frac{\int_{u=w}^{\infty} f_\tau(u) \cdot E(S^2|\tau = u)\, du}{\int_{u=w}^{\infty} f_\tau(u)\, du}. \tag{7.1}$$

Calculations show that, in this case,

$$\begin{aligned} E(S^2|T > W = w) &= e^{(\sigma^2)} \cdot a^2 \cdot \left[b^2 - 2 \cdot b\frac{\theta}{\beta} \cdot e^{-t\cdot(\beta-\theta)} - \frac{\theta}{\theta - 2\cdot\beta} \cdot e^{-2\cdot t\cdot(\beta-\theta)}\right] \\ &= e^{\sigma^2} \cdot E^2(S|T > W = w) - a^2 \cdot e^{(\sigma^2)} \cdot e^{-2\cdot t\cdot(\beta-\theta)} \cdot \left[\frac{\theta^2}{\beta^2} + \frac{\theta}{2\beta - \theta}\right] \end{aligned}$$

Then, we extract the following expression for the variance of the conditional service time

$$\begin{aligned} Var(S|\tau > W = w) &= E(S^2|T > W = w) - E^2(S|T > W = w) \\ &= (e^{\sigma^2} - 1) \cdot E^2(S|T > W = w) - a^2 \cdot e^{(\sigma^2)} \cdot e^{-2\cdot t\cdot(\beta-\theta)} \cdot \left[\frac{\theta^2}{\beta^2} + \frac{\theta}{2\beta - \theta}\right] \end{aligned}$$

Here, we refer to the values of the parameters of the Log-Normal Distribution, that are described at the beginning of this subsection. Figure 7.9 describes, for several values of σ, the differences between the two variance expressions, $Var(S|\tau = w)$ and $Var(S|\tau > W = w)$ as a function of the waiting time. One observes that under the assumptions of the described model, for customers who face short-to-medium waiting times, the variance of the observed service time, $Var(S|\tau > W = w)$, is much higher than the true variance of the service time of a customer with patience w, $Var(S|\tau = w)$. Moreover, the gap between

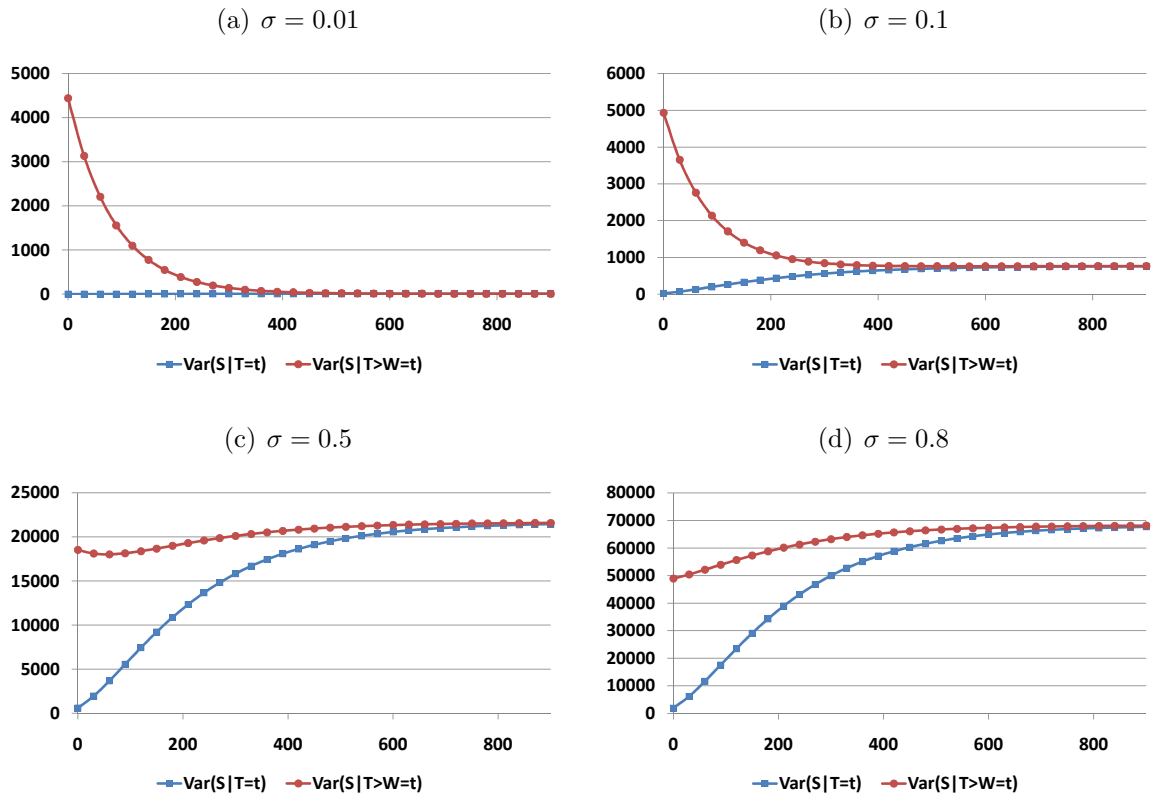


Figure 7.9: A comparison between $Var(S|\tau = w)$ and $Var(S|\tau > W = w)$ with different values of σ. In all figures, the conditional service time of a customer with patience w is Log-Normal distributed, $S|\tau = w \sim LogNorm(\mu(w), \sigma^2)$, with mean $E(S|\tau = w) = 230 \cdot (1.2 - e^{-w \cdot (\frac{29}{3600} - \frac{8}{3600})})$, and the patience is exponential, $\tau \sim exp(\frac{8}{3600})$.

these variaces expands as the value of σ grows. Finally, these expressions converge to the same value where the expression of the conditional service time, $E(S|T = w)$, converges to its limit. There, the distribution of the service time is equal for all waiting times. Recall that the means of the conditional service time are not dependent on the values of σ^2 and are shown in Figure 7.8.

7.2.2 Simulation Analysis of the Model

At this stage, we analyze some of the properties of the model that is described in Subsection 7.2.1 through simulation. The arrival process is a homogeneous Poisson process with rate 3,500 arrivals per hour. The simulation's duration is 300 hours. Important prerequisites for the analysis of the relationship between patience and service time is that

there is a considerable number of observations of customers who receive service, and that their waiting times span the desired range of the patience times. Staffing according to the QD Regime (see Operational Regimes in Subsection 2.1.3) causes very short waiting times, whereas in the ED Regime we lack short waiting times of served customers. Hence, we run the simulation in the QED Regime. We ran the iterative staffing algorithm with a target of probability of waiting that equals 0.8. The final staffing is then set to 205 agents.

In a single run of the simulation, there are approximately 1,050,000 observations, from which 5 percent abandon. The first hour of the simulation is omitted, assuming that the system is stabilized from then on. We present the summary of one realization of the simulation. There are 1,050,092 arrivals, where 1,046,508 arrived after the first hour. During the 299 hours, there where 54,081 customers who abandoned and 216,042 non-waiting customers. The waiting time histogram is shown in Figure 7.10. Note that the number of observations for each waiting time above 110 seconds is considerably low (less than 100 observations for any waiting time).

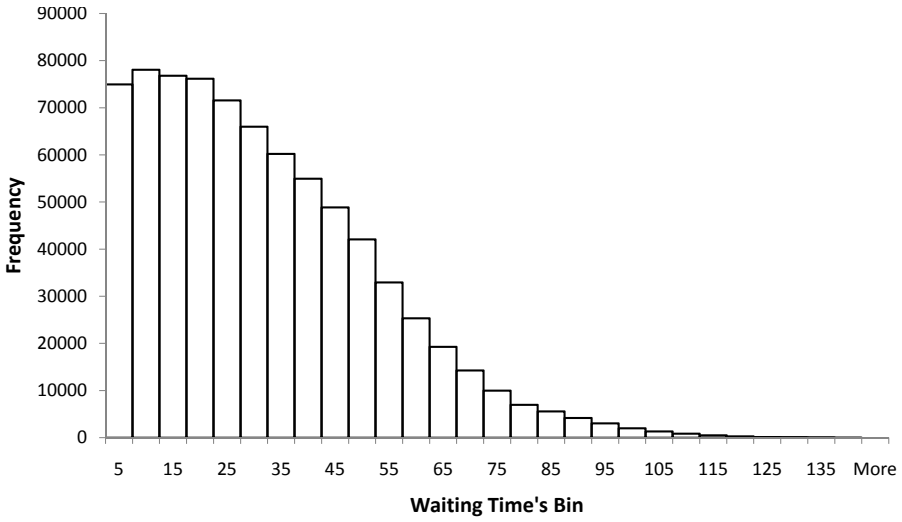


Figure 7.10: A histogram of the waiting times in a single run of the simulation.

Now, we estimate the conditional mean service time of served customers, as a function of their waiting times. Here again, as in Section 7.1, we use the smooth.spline() function in 'R' in order to build a cubic smooth spline, with 5 knots, for $E(S|\tau > W = w)$.

In Figure 7.11, we observe that the spline estimator fits well the theoretic function of $g(w) = E(S|\tau > W = w)$ in the range of waiting times between 0 and 100 seconds. To be more explicit, for any waiting time in this range, the error of the fit is less then 1

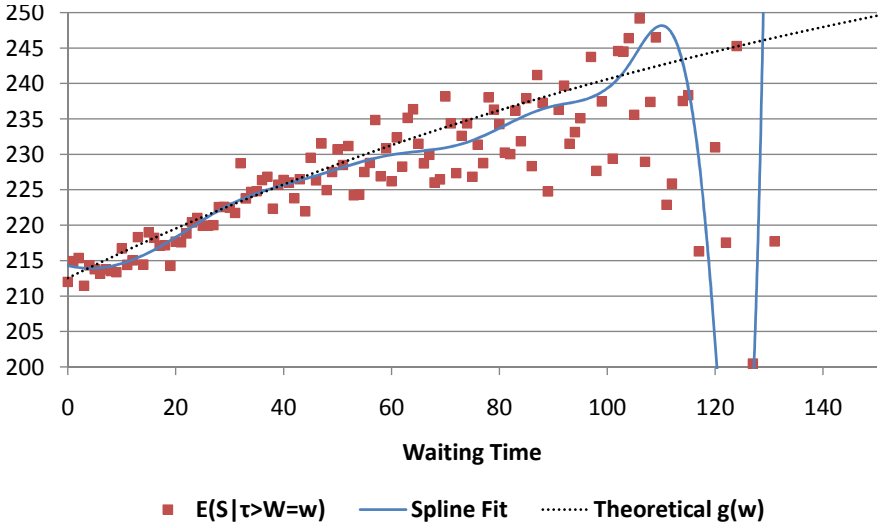


Figure 7.11: A comparison between the mean service time of served customer as a function of their waiting times, the fitted cubic spline for them and the theoretical function. For visualization purpose (scaling), the frame of the chart does not include extreme values. Notice that the number of observations is decreasing as the waiting time grows.

percent of the theoretical value. As discussed before in this section, there is a low number of observations with waiting time above 100 seconds. Together with the high variance of the observations, shown in Subsection 7.2.1 (the standard deviation of $g(w)$ is of the same order as its value), the estimation in this range is unreliable.

On the other hand, the derivative of the estimator for $g(w)$ seems to be more noisy, as demonstrated in Figure 7.12. The theoretical values of the derivative for values of w between 20 and 100 seconds range from 0.2 to 0.4, whereas the absolute error of the estimator, as calculated from the simulation, is often between 0.1 and 0.2.

According to Formula (4.5), the derivative of the estimator for $g(w)$ should be divided by the estimator of the hazard rate function of the patience at w, $h_\tau(w)$. We exploit the fact that, in the model, the patience is exponentially distributed with mean $\frac{1}{\theta}$. The hazard rate is then simply the parameter θ itself:

$$h_\tau(w) = \frac{f_\tau(w)}{1 - F_\tau(w)} = \frac{\theta \cdot e^{-\theta w}}{e^{-\theta w}} = \theta \,.$$

From the definition of the model, the value of θ is $\frac{8}{3600}$. Consequently, the value of the derivative of the estimator for $g(w)$ is multiplied, according to the right expression of (4.5), by the estimator for $1/\theta$, which is approximately $\frac{3600}{8} = 450$ (the exact estimator

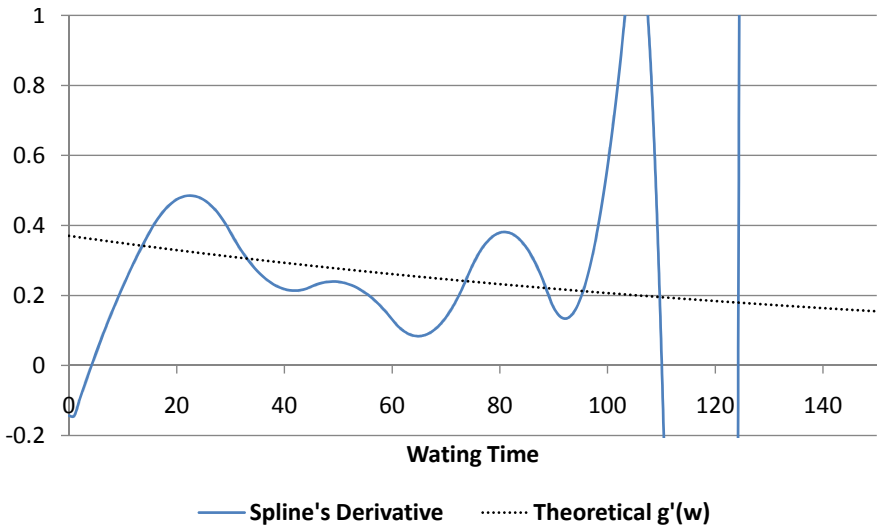


Figure 7.12: The fitted spline's derivative against the derivative of the theoretical function of $g(w)$.

for θ will be given immediately). It follows, that the value of the error of the estimator for $\dfrac{g'(w)}{h_\tau(w)}$, for these waiting times, is often in the range 45-90 seconds, which is close to the order of the theoretical value of $E(S|\tau = w)$.

At this stage, estimation of the hazard rate function of the patience is required. A maximum likelihood estimator (MLE) for θ can be simply derived as follows:

Assume that there are m customers who abandoned and n served customers. Define t_i, $i = 1 \dots m$, the waiting time of customer i who abandoned, and s_j, $j = 1 \dots n$, the waiting time of served customer j. Then we observe that customer i had a patience of exactly t_i seconds. On the other hand, customer j has patience more than s_j. The Likelihood function is then given by

$$\begin{aligned} L(\theta) &= \left[\prod_{i=1}^{m} f_\tau(t_i)\right] \cdot \left[\prod_{j=1}^{n} (1 - F_\tau(s_j))\right] \\ &= \left[\prod_{i=1}^{m} \theta \cdot e^{-\theta t_i}\right] \cdot \left[\prod_{j=1}^{n} e^{-\theta s_j}\right] \\ &= \theta^m \cdot e^{-\theta(\sum_{i=1}^{m} t_i + \sum_{j=1}^{n} s_j)} \end{aligned}$$

We perform a logarithmic transformation of the likelihood function:

$$\ln(L(\theta)) = m \cdot \ln(\theta) - \theta(\sum_{i=1}^{m} t_i + \sum_{j=1}^{n} s_j)\,.$$

Finally, setting the value of the derivative to zero yields the MLE for the parameter θ which, for the exponential distribution, is also the hazard rate function at any point:

$$\hat{\theta} = \frac{m}{\sum_{i=1}^{m} t_i + \sum_{j=1}^{n} s_j}\,.$$

Practically, the estimator for θ is the number of abandons divided by the total sum of all the waiting times of both served and abandoning customers. Figure 7.13 shows the Kaplan-Meier and the maximum likelihood estimators for the hazard rate of the patience, in comparison with the theoretical value of the patience according to the simulation.

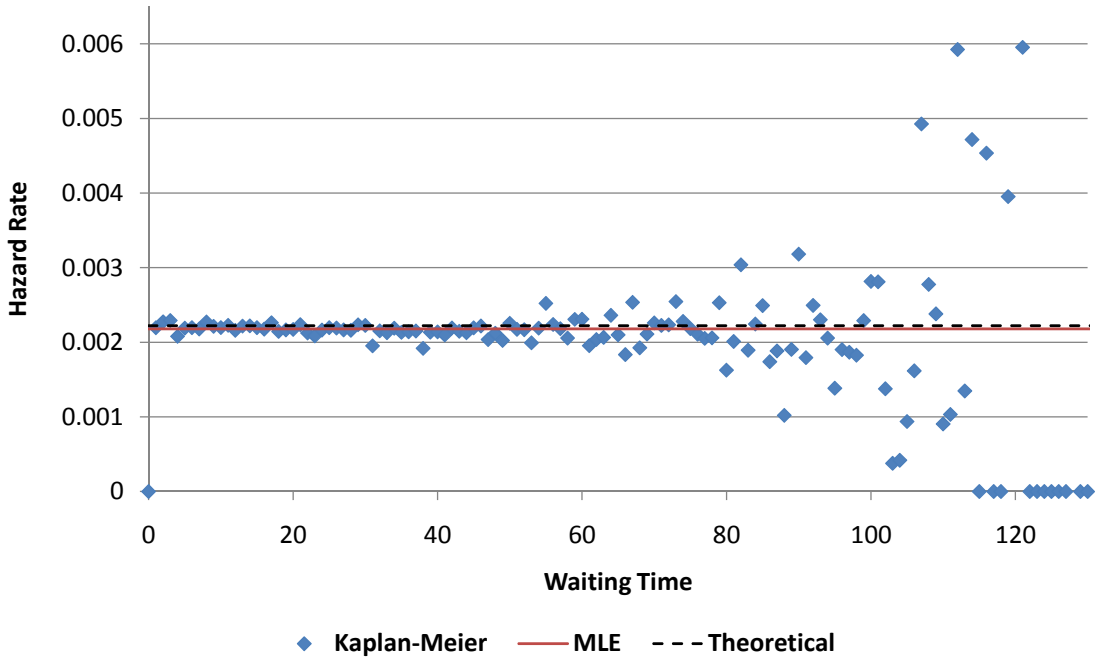


Figure 7.13: The Kaplan-Meier and the maximum likelihood estimators for the hazard rate of the patience vs the theoretical value of the hazard rate.

Since we are interested in the expression $\frac{g'(w)}{h_\tau(w)}$, we use the MLE for $\frac{1}{\theta}$ which, according to the simulation results, gets the value of 458.83. That means that the error of the estimator here is 8.83 seconds. Thus, if we used the real value of $g'(w)$, then the maximal error of the rightmost expression in Formula (4.5) was lower then 4 seconds.

Figure 7.14 shows the estimator against the theoretical function of the mean service time, as a function of the patience. It is clear that the estimator is very noisy and yields unrealistic values, and that we cannot rely on it in our analysis.

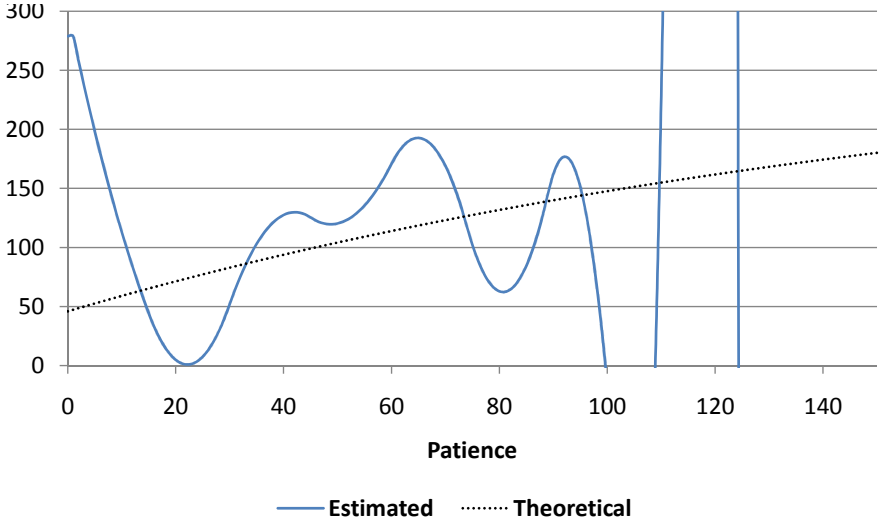


Figure 7.14: The estimator for $E(S|\tau = w)$. The plot shows an unreliable behavior of the estimator, with extreme variation and values close to (and below) zero.

From the above analysis, we deduce that the main source for the error in the process of estimating $E(S|\tau = w)$ is the estimation of the spline's derivative, which is multiplied by $\frac{1}{h_\tau(w)}$. A possible solution for this obstacle is to have more constrains on the estimated function for $g(w)$, such as assuming a monotonic function, limiting the variation of the first derivative or finding a moderate function that fits the empirical data (i.e. a linear combination of low order polynomial functions or of exponential functions). In the next subsection, we analyze some properties of the estimators that are described above, using empirical distributions.

7.2.2.1 Empirical Distributions

In order to analyze the properties of the estimators from the previous subsection, we repeat the simulation 400 times in order to generate a distribution of the relevant estimators. In each simulation, we repeat the estimation process that was described above. In this manner, we are able to create confidence intervals and to verify if the estimators are biased.

We start with the hazard rate estimator. Observing Figure 7.15 reveals that the MLE estimator for $\frac{1}{\theta}$ is biased, with all values from the 400 simulations greater than the theoretical value, 450. To be more explicit, the bias of the estimator is 8.71 and its standard deviation estimate is 2.03. The bias can be caused because of time-aggregation when converting the patience data from continuous time into grouped seconds units. Petersen discusses in [28] the time-aggregation bias in the hazard rate estimation.

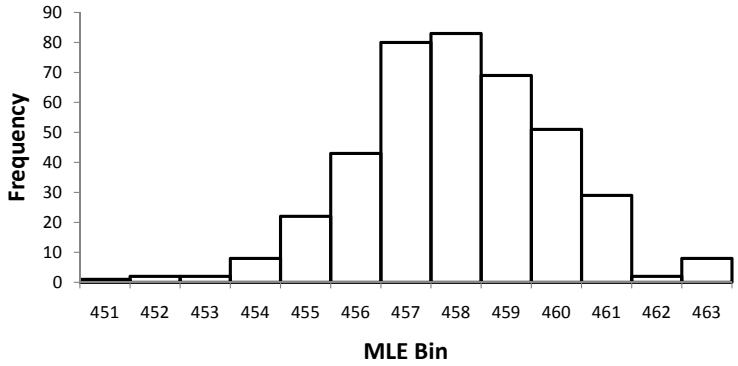


Figure 7.15: A histogram of the maximum likelihood estimators for the hazard rate of the patience.

In Figures 7.16-7.18, the 95% confidence intervals are drawn for the estimators that are relevant in the estimation of $E(S|\tau = w)$. All of the estimators (apart from the hazard rate estimator) seems to be unbiased, and perform better in the range of waiting times between 20 and 80 seconds. From Figure 7.16 one observes that the spline estimator for $g(w) = E(S|\tau > W = w)$ fits well the theoretical function, with a narrow confidence interval which hardly gets wider as σ (and consequently the variance of the service time) grows. Nevertheless, with the increase in σ, the spline estimators seems to be less smooth, as shown in Figure 7.17, where the confidence intervals for the derivative are drawn. Finally, 7.18 reveals that when all the above estimators are combined into Formula 4.5, the aggregated estimator for $E(S|\tau = w)$ is very sensitive to an increase in the service time variance and to the dimension (number of records per waiting time) of the data.

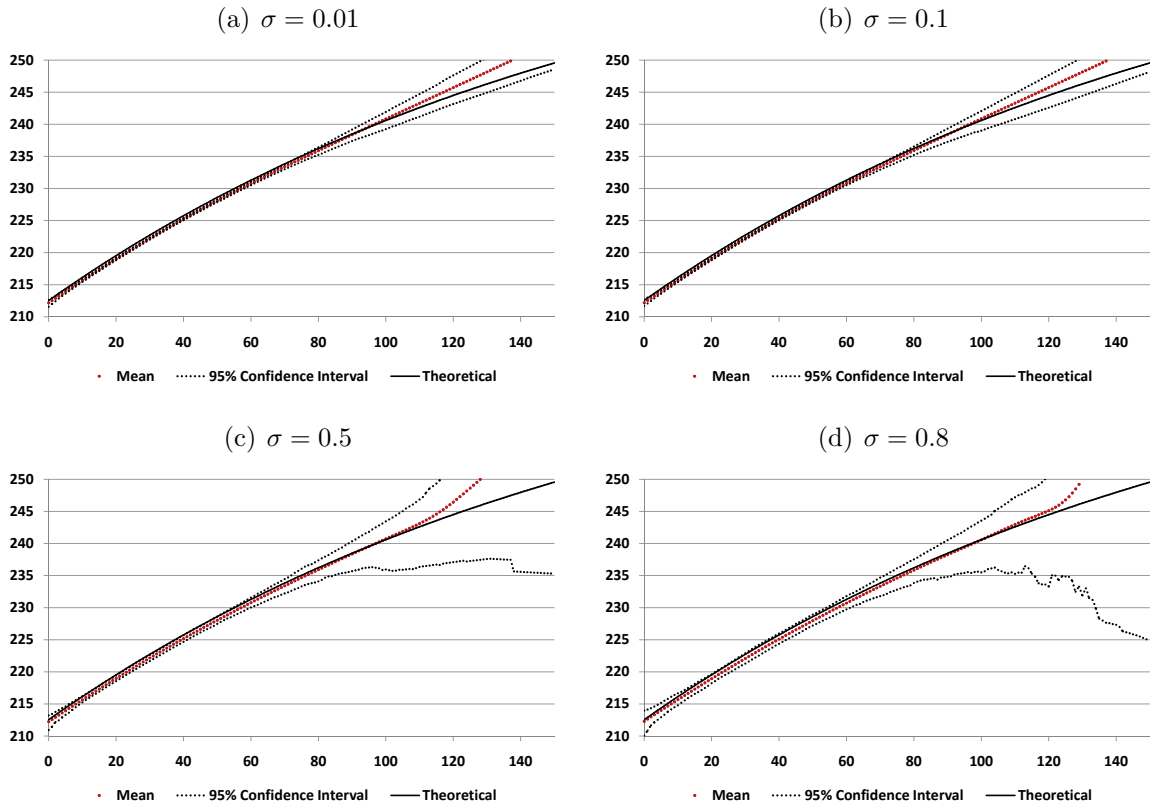


Figure 7.16: 95% percent confidence intervals of the spline estimator for $g(w) = E(S|\tau > W = w)$, for different values of σ.

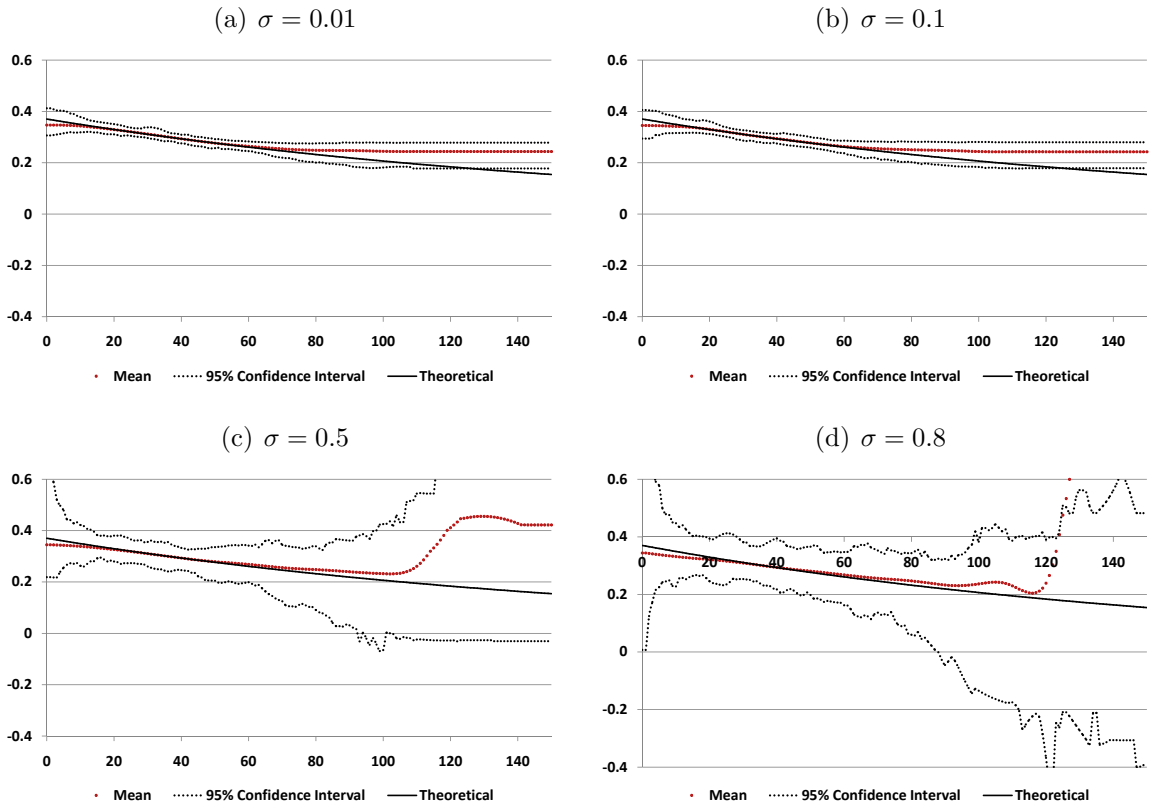


Figure 7.17: 95% percent confidence intervals of the derivative of the spline estimator for $g(w)$, for different values of σ. As σ grows the scale of the error is increasing.

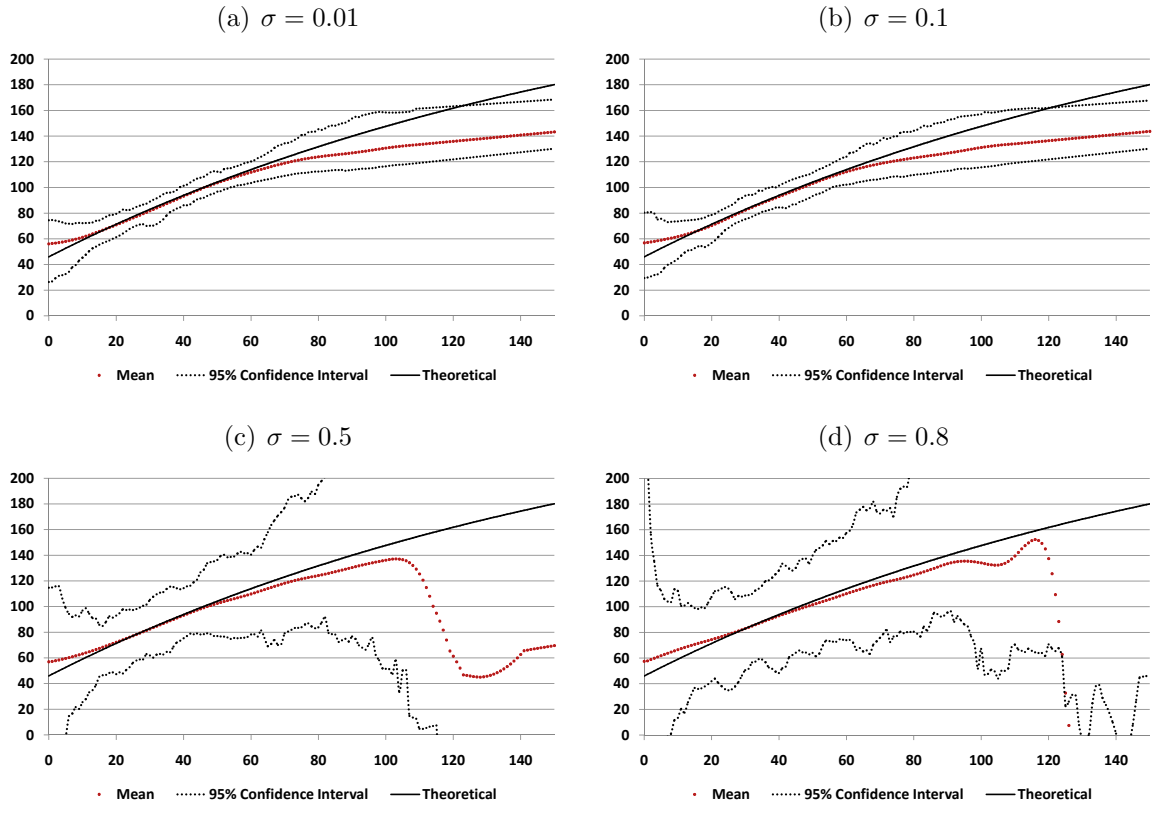


Figure 7.18: 95% percent confidence intervals of the estimator for $E(S|\tau = w)$, for different values of σ.

Chapter 8

Future Research

Interesting extensions are possible in the following directions:

- **Refinement of the estimation of $\mathbf{E(S|\tau = w)}$**
  The results that are presented in Chapter 7 show that in a typical service system, that faces a relationship between patience and service time, our described estimation procedure may be poor. An insight that emerged from our empirical study is that the main source of error is related to the estimation error of the first derivative of $g(w)$ (see Section 4.1), which is divided by the relatively small expression of the hazard rate, $h_\tau(w)$.
  Since the pattern of the relationship between the patience and the service time is expected to be smooth, with very slow change in the mean service time over patience, one should try to model it with more constraints than those of the cubic spline regression. For example, a linear combination of polynomial or exponential functions can be attempted. Naturally, the assumptions of the model (e.g. choosing smooth function for $E(S|\tau = w)$ and varifying that $E(S) < \infty$) should be maintained.

- **Construct a framework for the analysis of performance measures of the $\mathbf{M_t/GI^*/n_t + GI}$ queue**
  In Chapter 6 we showed that staffing rules and many performance measures are influenced by a possible relationship between the patience and the service time. We believe there is much more to be done in this direction.

- **Further simulation-based analysis**
  We use ISA to explore some of the properties that arise from the relationship between

patience and service time. It is interesting to verify whether these results are valid for other service time and patience distributions. Furthermore, one can check if stabilizing the probability of delay in a $M_t/GI^*/n_t + GI$ queue (by time varying staffing) yields other time-stable performance measures, similarly to the $M_t/GI/n_t+GI$ queue as shown in [36].

- **Validating the proposed procedures for estimating the offered-load**
  In Chapter 5.1 we propose methods to estimate the offered-load process and offered-load function, in several types of service systems. One could and should validate the described procedures using simulations.

- **Apply the presented model to other databases**
  There are many service systems, where the service time of customers depends on their patience. We are interested in applying the model to databases of other service systems in a variety of business fields.

Bibliography

[1] Mandelbaum A. and Zeltyn S., *Staffing many-server queues with impatient customers: constraint satisfaction in call centers*, (2007), Under revision for Operations Research. 2.1.3, 2.1.3

[2] Prekopa A., *On secondary processes generated by random point distributions of poisson type*, Annales Univ. Sci Budapest de Eotvos Nom. Sectio Math **1** (1958), 153–170. 3.1.1

[3] Erlang A.K., *The theory of probabilities and telephone conversations*, Nyt Tidsskrift Mat. (1909), D 20 33–39. 2.1, 2.1.1

[4] Fralix B.H. and Riano G., *Another look at transient versions of littles law, and $M/G/1$ preemptive last-come-first-served queues*, Eurandom Report No. 2008-042, Eindhoven: Eurandom, 15 pp. (2008). 5

[5] Kooperberg C., Stone C.J., and Truong Y.K., *Hazard regression*, Journal of the American Statistical Association **90** (1995), 7894. 7.1.3

[6] Palm C., *Research on telephone traffic carried by full availability groups*, Tele **1, 107** (1957). 2.1

[7] Bertsimas D. and Mourtzinou G., *Transient laws of non-stationary queueing systems and their applications*, Queueing Systems **25** (1997), 115–155. 5

[8] Litvak E., McManus M.L., and Cooper A., *Root cause analysis of emergency department crowding and ambulance diversion in Massachusetts*, Boston University Program for Management Variablity in Health Care Delivery (2002). 3.1.3

[9] SEE Center (Service Enterprise Engineering), `http://iew3.technion.ac.il/serveng/References`, and `http://ie.technion.ac.il/Labs/Serveng`. 7.1.1

[10] Shen H., *On modeling and forecasting time series of smooth curves*, (To appear). 5.2

[11] Shen H. and Huang J. Z., *Interday forecasting and intraday updating of call center arrivals*, MANUFACTURING SERVICE OPERATIONS MANAGEMENT **10** (2008), no. 3, 391–410. 2.2, 5.2

[12] Weinberg J., Brown L.D., and Stroud J.R., *Bayesian forecasting of an inhomogeneous poisson process with applications to call center data*, Journal of the American Statistical Association **Vol. 102** (2007). 2.2, 5.2

[13] Bilmes J.A., *A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models*, Technical Report of the International Computer Science Institute, Berkeley, CA (1998). 7.1.3

[14] Green L., *Capacity planning and management in hospitals*, Operations Research and Helth Care, (Brandwau et al editors) (2004), 14–41. 1

[15] Rozenshmidt L., *On priority queues with impatient customers: Stationary and time-varying analysis*, M.Sc. Thesis, Technion (2007). 3.1.3

[16] Channouf N., L'Ecuyer P., Avramidis A., and Ingolfsson A., *The application of forecasting techniques to modeling emergency medical system calls in calgary, alberta*, Health Care Manegement Science **10, No.1** (2007), 25–45. 1

[17] Gans N., Koole G., and Mandelbaum A., *Telephone call centers: Tutorial, review and research prospects*, Manufacturing and Service Operations Management **5**, 79–141. 2.1.3

[18] Garnett O., Mandelbaum A., and Reiman M., *Designing a call center with impatient customers*, Manufacturing and Service Operations Management **4(3)** (2002), 208–227. 2.1.2, 2.2

[19] Aldor-Noiman S., *Forecasting demands for a telephone call center: Analysis of desired versus attainable precision.*, M.Sc. Thesis, Technion (2006). 2.2

[20] Aldor-Noiman S., Feigin P.D., and Mandelbaum A., *Workload forecasting for a call center: Methodology and a case study*, The Annals of Applied Statistics **3** (2009), no. 4, 1403–1447. 2.2

[21] Eick S., Massey W. A., and Whitt. W., *The physics of the $M_t|G|\infty$ queue*, Operations Research **41, No.4** (1993), 731–742. 2.2

[22] Halfin S. and Whitt W., *Heavy-traffic limits for queues with many exponential servers*, Operations research 29 (1981), 567–587. 2.1

[23] Henderson S., Mehrotra V., and Steckley S., *Performance measures for service systems with a random arrival rate*, Proceedings of the 2005 Winter Simulation Conference (2005). 6.1.2

[24] Maman S., *Uncertainty in the demand for service: The case of call centers and emergency departments*, M.Sc. Thesis, Technion (2009). 6.1

[25] Steckley S., Henderson S., and Mehrotra V., *Service system planning in the presence of a random arrival rate*, Doawnloadble from: http://legacy.orie.cornell.edu/~shane/pubs/stehenmeh04.pdf (2004). 1

[26] Zeltyn S. and Mandelbaum A., *Call centers with impatient customers: many-server asymptotics of the $M|M|n+G$ queue*, QUESTA **51** (2005), 361–402. 2.1.2

[27] Eick S.G., Massey W.A., and Whitt W., *The physics of the $M_t/G/\infty$ queue*, Operations Research **41** (1993), no. 4, 731–742. 3.1.1, 6

[28] Petersen T., *Time-aggregation bias in continuous-time hazard-rate models*, Sociological Methodology **21** (1991), 263–290. 7.2.2.1

[29] Whitt W., *Dynamic staffing in a telephone call center aiming to immediately answer all calls*, Operations Research Letters **24** (1999), 307–314. 2.2, 6

[30] ———, *Staffing a call center with uncertain arrival rate and absenteeism*, Production and Operations Management **15, No.1** (2006), 88–102. 6

[31] Wolff R. W., *Poisson arrivals see time averages*, Operations Research **30, No.2** (1982), 223–231. 2.1.1

[32] Goldberg Y., *The best linear unbiased estimators for continuation of a function*, Submitted to the Annals of Applied Statistics (2010). 1.1, 5.2

[33] Marmor Y., *Emergency-departments simulation in support of service-engineering: Staffing, design, and real-time tracking*, Ph.D. Thesis, Technion - Israel Institute of Technology (2009). 5.1.1

[34] Feldman Z., *Staffing of time-varying queues to achieve time-stable performance.*, Downloadable from http://iew3.technion.ac.il/serveng/References. (2004). 6.1.1

[35] Feldman Z., Mandelbaum A., Massey W. A., and Whitt W., *Staffing of time-varying queues to achieve time-stable performance*, Management Science **54, No.2** (2007), 324–338. 2.2, 6.1.1

[36] Feldman Z., Mandelbaum A., Massey W.A., and Whitt W., *Staffing of time-varying queues to achieve time-stable performance*, Management Science (2007). 3.1, 3.1.4, 5.2, 7.2, 8