Data-Stories about (Im)Patient Customers in Tele-Queues

Avishai Mandelbaum
Faculty of Industrial Engineering & Management,
Technion, Haifa 32000, Israel
avim@ie.technion.ac.il

Sergey Zeltyn
IBM Research Lab, Haifa 31905, Israel
sergeyz@il.ibm.com

April 2, 2013

Abstract

Credible queueing models of human services acknowledge human characteristics. A prevalent one is the ability of humans to abandon their wait, for example while waiting to be answered by a telephone agent, waiting for a physician's checkup at an emergency department, or waiting for the completion of an Internet transaction. Abandonments can be very costly, to either the service provider (a forgone profit) or the customer (deteriorating health after leaving without being seen by a doctor), and often to both. *Practically*, models that ignore abandonment can lead to either over- or under-staffing; and in well-balanced systems (e.g. well-managed telephone call centers), the "fittest (needy) who survive" and reach service are rewarded with surprisingly short delays. *Theoretically*, the phenomenon of abandonment is interesting and challenging, in the context of Queueing Theory and Science as well as beyond (e.g. Psychology). Last, but not least, queueing models with abandonment are more robust and numerically stable, when compared against their abandonment-ignorant analogues.

For our relatively narrow purpose here, abandonment of customers, while queueing for service, is the operational manifestation of customer patience, perhaps impatience, or (im)patience for short. This (im)patience is the focus of the present paper. It is characterized via the distribution of the time that a customer is willing to wait, and its dynamics are characterized by the hazard-rate of that distribution.

We start with a framework for comprehending impatience, distinguishing the times that a customer expects to wait, is required to wait (offered wait), is willing to wait (patience time), actually waits and felt waiting. We describe statistical methods that are used to infer the (im)patience time and offered wait distributions. Then some useful queueing models, as well as their asymptotic approximations, are discussed.

In the main part of the paper, we discuss several "data-based pictures" of impatience. Each "picture" is associated with an important phenomenon. Some theoretical and practical problems that arise from these phenomena, and existing models and methodologies that address these problems, are outlined.

The problems discussed cover statistical estimation of impatience, behavior of overloaded systems, dependence between patience and service time, and validation of queueing models. We also illustrate how impatience changes across customers (e.g. VIP vs. regular customers), during waiting (e.g. in response to announcements) and through phases of service (e.g. after experiencing the answering machine over the phone). Our empirical analysis draws data from repositories at the Technion SEELab, and it utilizes SEEStat - its online EDA (Exploratory Data Analysis) environment. SEEStat and most of our data are internet-accessible, which enables reproducibility of our research.

Contents

1 Introduction

Customer impatience plays a central role in the analysis of many service systems. For example, it occurs on the phone while waiting to be answered by an agent, in the emergency department of a hospital while waiting for a nurse or a physician, or on the Internet while going through an online purchase process. Impatience is manifested via customers' *abandonment* which, in our opinion, is often one of the most important measures of operational performance. Indeed, abandonment is an example of a *subjective* operational measure, through which customers inform the system of their perception of the service that they are being offered or received. In essence, the time until abandonment provides an answer to the question "how long is the service worth waiting for?"

In this paper, we focus on impatience and abandonment from telephone queues. (We do not consider reneging due to impatience after service starts; this phenomenon is interpreted as unsuccessful service, not as abandonment per se.) Telephone queues constitute a prevalent example of *invisible* queues, where customers do not have information on the queue state unless the service provider is able and willing to provide it. This invisibility has important psychological ramifications. For example, in the case of prolonged waiting, a face-to-face customer will become irritated when first observing a long queue. However, once an informed decision to wait is made, the customer would often feel increasingly more positive while advancing in the queue. In contrast, an uninformed telephone customer will not experience frustration at the beginning, but negative emotions (irritation) will, most likely, accumulate during waiting.

There exists significant theoretical research on models that incorporate abandonment and can potentially be applied to telephone queues. Two surveys, Gans et al. [22] and Aksin et al. [4] provide numerous references. However, there is a gap between the abundance of theoretical models and the scarce research on patience patterns in tele-queues. Interestingly, the landscape of research on telephone abandonment differs from that on abandonment from emergency departments: in the latter, ample statistical work has been performed on customers who Left Without Being Seen (LWBS; e.g. Fernandes et al. [21], Hobbs et al. [27]), but the analytical modeling side are yet undeveloped.

We shall discuss several research directions that could be pursued to fill this gap between theory and practice in tele-services. These research directions will be presented via a sequence of data-based case studies. Most data for our cases was extracted from the databases of the Service Enterprise Engineering Laboratory (SEELab), in the Faculty of Industrial Engineering & Management at the Technion [53]. Our empirical research was enabled by SEEStat, which is SEESlab's online environment for EDA (Tukey's Exploratory Data Analysis [58]). SEEStat and some of SEELab's data-bases are publicly accessible -

this is a modest attempt to address the proprietary and reproducibility challenges in present scientific research [18], following the principle advocated in [16] (see Appendix B below).

Outline of the paper. In Section 2, we present several key concepts that guide us in our research on impatience. First, *patience time* and *offered wait* are introduced as the building blocks of patience modeling from the operational viewpoint. *Censored data analysis* is the main statistical approach to the estimation of patience and offered wait distributions. The *hazard rate* is a basic patience characteristic from the perspective of customer behavior. Finally, we discuss several key queueing models and their operational regimes.

In concert with the title of the paper, Section 3 consists of data-based stories on different aspects of impatience in tele-queues. Each story is followed by a discussion and references to relevant literature. Specifically, Subsection 3.1 presents a unique example of a day with 100% abandonment and discusses fitting of patience distributions. Subsection 3.2 explores overloaded systems with significant (10's %) abandonment. The impact of information that a customer gathers through personal experience or automatic announcements is discussed in Subsection 3.3. Subsection 3.4 relates patience and Interactive Voice Response (IVR) experience. Subsections 3.5 and 3.6 consider the influence on patience of service-time durations and customer priorities, respectively. Subsection 3.7 discusses fitting of the M/M/n+M (Erlang-A) queueing model to call-center data.

Finally, Section 4 presents our personal perspective on some worthwhile research directions, in the area of (im)patience while waiting in tele-queues. The paper ends with two appendices: Appendix A on accessing SEELab data and Appendix B on the precise sources that support our data-stories.

2 Comprehending and Modeling Impatience in Tele-Queues

2.1 The anatomy of waiting for service

Our experience (as opposed to a deserving psychological research) suggests that the following five "waiting times" (or durations) could serve as a skeleton for a framework that captures the impatience phenomenon:

1. *Anticipated-Time*: Time that a customer *expects*, or *anticipates* to wait. This time is formed through individual accumulated experiences, either in a specific call center or through a general conception of what telephone services are or ought to be like.

2. *Patience or Impatience*: Time that a customer is *willing to wait*. If this patience time elapses and the customer is still waiting in the tele-queue, abandonment takes place. We denote this patience time by τ, conceptualizing it as a random variable with cumulative distribution function G.

   The patience τ is a cornerstone of queueing models that incorporate abandonment. Depending on a model and its application, $\tau \sim G$ is determined upon arrival to the tele-queue or, possibly, it is reevaluated due to exogenous events during waiting (e.g. automatic announcements, which will be discussed in Subsection 3.3).

3. *Offered-Wait*: Time that a customer is *required to wait*, which we denote by a random variable V. Alternatively, V is the time that a customer with infinite patience (sometimes referred to as a test-customer) would wait until reaching a server. This time depends on the call center environment and protocols, e.g. staffing levels and customer priority.

4. *Waiting-Time*: Time that a customer *actually waits*. Note that this time, denoted by W_q, is the minimum between the above preceding times: $W_q = \min\{\tau, V\}$.

5. *Perceived-Time*: Time that a customer *perceives waiting*. This time can be significantly different from actual wait: it is, in some sense, a subjective operational summary of the service experience (Munichor and Rafaeli [45], Katz et al. [30]).

The above general framework applies iteratively during customers' experience of telephone services: perceived waiting time affects the anticipated waiting time prior to the subsequent visit, and so on.

Our five waiting times can be divided into two groups. Anticipated and perceived waiting times are mainly related to the *psychological aspect* of wait. While acknowledging their importance as research targets, here we do not address them (though the concept of anticipated wait is relevant for Subsection 3.3). The triple (τ, V, W_q) embodies the *operational aspect* of waiting, and it constitutes the focus of the present paper.

Note that the full framework, with five waiting times, can be reduced to the operational framework (τ, V, W_q) through the following two reasonable assumptions: customers are familiar with the service system and its waiting, hence (roughly) "anticipated = offered = V"; and they are "unemotional" about their waiting, hence "perceived = actual = W_q".

2.2 Inferring patience and offered wait

The waiting time W_q is the only directly observable element in the triple (τ, V, W_q). To elaborate, if the offered wait exceeds impatience then the customer abandons and one directly

observes $W_q = \tau$. In the language of Statistical Survival Analysis, the offered wait is then *right-censored*: the available information is the value of W_q and the event $\{W_q = \tau < V\}$. Otherwise, the offered wait is directly observed and the patience-time observation is right-censored: in other words, for customers who were served, one has only lower bounds V on their patience τ.

The Kaplan-Meier estimator for survival functions represents the classical non-parametric approach for analyzing right-censored data (see, for example, Cox and Oakes [17]). A more recent approach that we embrace is Aalen, Borgan and Gjessing [1], which is based on a process-view of event histories. Brown et al. [15] applied the Kaplan-Meier approach to estimate survival functions of patience times and offered waits in a small Israeli call center. They also developed non-parametric techniques for inferring the corresponding hazard-rates.

We advocate the use of both the survival function and the hazard rate for patience inference. The practical significance of the former (probability that patience exceeds t units of time) is clear. In particular, we shall later use survival functions to (stochastically) compare the patience of customers, for example VIP vs. Regular customers. However, hazard rates are superior to survival functions for deciphering the experience of customers during their wait. This pertains to both isolated events (e.g. reaction to announcements while waiting) or trends (e.g. increasing/decreasing hazard rate indicates growing/diminishing impatience over time). Following Palm [47], we thus use the hazard rate as a means for *dynamically* describing impatience, or in Palm's words *customer irritation*, as a function of their waiting time. Readers will discover and hopefully agree that hazard rate profiles, over time-periods or across customer-types, lead to important insights which, moreover, are often rather surprising. (One should beware though that pitfalls can arise from the analysis of, what Aalen and Gjessing [2] refers to as, "the rather elusive concept of hazard rate", especially when modeling groups vs. individuals. We leave this important subject for future research, and refer the reader to Aksin et al. [5], who relate these issues to real call-center data.)

The mean and moments of customer patience also provide important information. However, mean and variance estimates, derived via the Kaplan-Meier approach, are often unreliable (see the Appendix in [66]). For example, the mean is usually computed as the area (integral) under the survival function. Therefore, the value depends on how the distribution's tail is estimated, which is challenging in view of censoring.

Recall that the hazard rate of an absolutely-continuous distribution is defined via $h(t) = f(t)/S(t)$, $t \geq 0$; here f is the patience density, $S = 1 - F$ is the survival function, and F is the cumulative distribution function. Figures 1 and 2 display the estimators of the hazard-rates of τ and V, for a small Israeli call center and a large U.S. call center, respectively (the left-hand plot in Figure 1 is similar to the one in [15]). The points in the plots correspond

to hazard rate estimates for each second; then these seconds-level estimates are smoothed to infer the hazard rate curve.

Several interesting phenomena emerge from the graphs. In the Israeli call center, two distinctive abandonment peaks are observed in the left-hand graph. It turns out that these peaks, reflecting an increased tendency to abandon, occur after customers listen to automatic announcements; the influence of announcements on abandonment patterns will be further studied in Section 3.3. The patience hazard rate also seems to decrease for large values of wait. The offered wait in Figure 1 has a relatively stable hazard rate (except for the first few seconds); recall that constant hazard rate is a manifestation of the memoryless property, which characterizes the exponential distribution if it prevails over $[0, \infty)$.

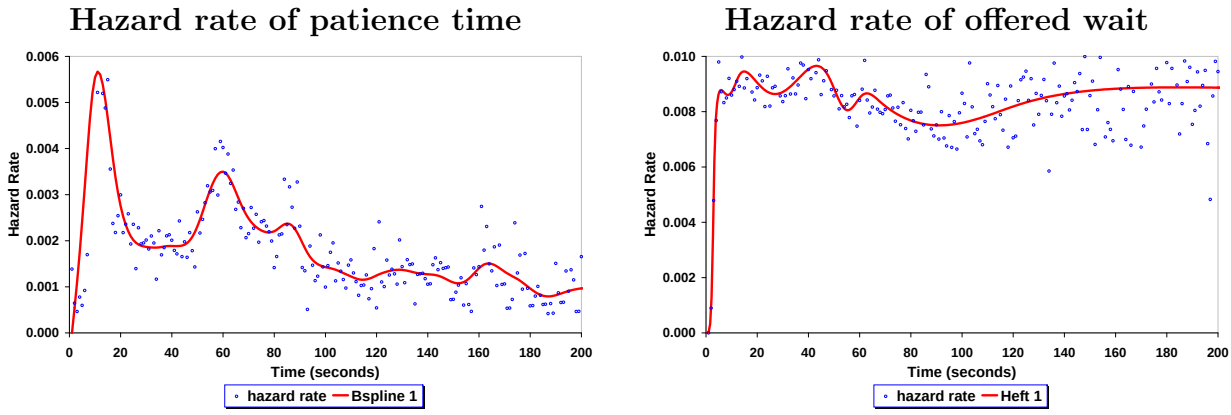


Figure 1: **Patience and offered wait in an Israeli call center**

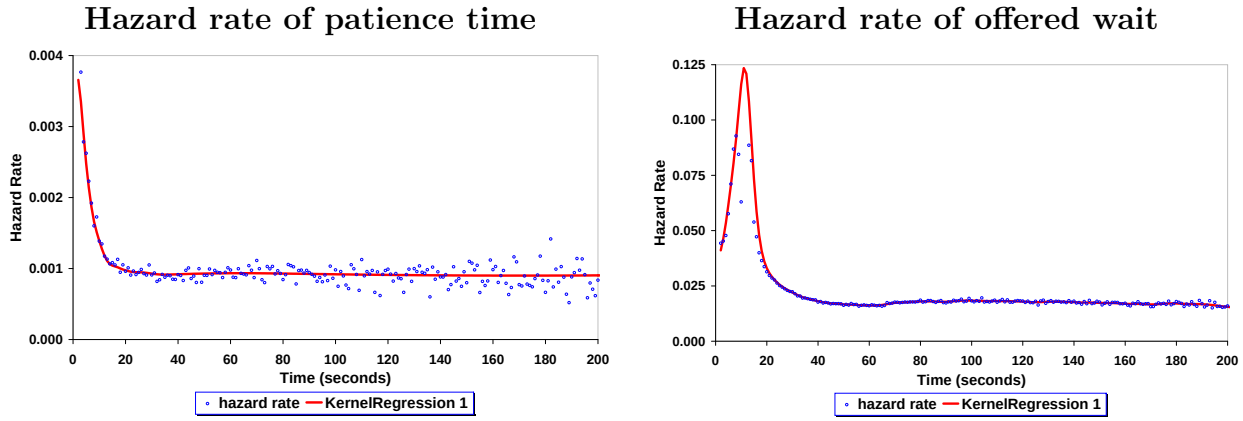


Figure 2: **Patience and Offered Wait in a U.S. Call Center**

In contrast to the Israeli call center, the left-hand plot of the U.S. data shows significant abandonment during the first several seconds of wait, which is then followed by a stable hazard rate (exponential-like distribution). The offered wait plot in Figure 2 displays a clear peak in serviced customers after approximately ten seconds of wait. In reality, the call center under consideration consists of four smaller call centers in as many cities, which are interconnected through the following dynamics: An inbound call first seeks service in its local call center; if delayed in its local queue for about 10 seconds, the call is placed in an *inter-queue* (multi-location queue) which can then be served by all call centers. This example reveals a high service level (there is a relatively small fraction of delays beyond 11 seconds), and it underscores the importance of service protocols (load-balancing among call centers, in our case).

2.3 Basic Queueing models with impatience.

As mentioned above, the patience time τ and the offered wait V are the main "heroes" of our data-based stories. If one resorts to queueing theory for accommodating and modeling the phenomenon of abandonment, the patience time distribution typically serves as an input characteristic of the chosen model, and the offered wait is an output of model analysis. (Exceptions are models where customers are optimizing their wait, such as Mandelbaum and Shimkin [38], to which we return in the sequel.) It is usually assumed that τ and V are independent, which is natural for tele-queues without announcements. The pair (τ, V) is then the basis for calculating various operational measures, in steady state. Examples include the mean waiting time $E[min\{\tau, V\}]$, and the fraction abandoning $P\{\tau < V\}$; or $P\{5 \text{ seconds } < \tau < V\}$, which counts as abandonments only those who waited more than, say, 5 seconds, as well as their mean waiting time $E[\tau | 5 \text{ seconds } < \tau < V]$.

The simplest queueing model that acknowledges impatience is the Erlang-A model, often denoted M/M/n+M (Figure 3). It builds on the classical Erlang-C (M/M/n) queue, which assumes Poisson arrivals at rate λ, iid exponential service times with rate μ, n iid servers, and independence between arrivals and services. Erlang-A then adds exponential patience times with mean $1/\theta$, independent of everything else.

Erlang-A was introduced by Palm [48] ([41] provides a recent survey). The exponential-patience assumption was relaxed by Baccelli and Hebuterne [12]: they allowed generally distributed iid patience times, denoting their model M/M/n+G. In both models, formulae for steady-state performance measures are derived, including queue-length and offered wait. Erlang-A calculations are mostly based on the incomplete Gamma function, while M/M/n+G formulae require numerical integration [43]. Generalizing M/M/n+G further, especially allowing service times to be generally distributed (e.g. M/G/n+G), gives rise to intractable models, which naturally raises the need for approximations. Approximations

are also useful for the tractable Erlang-A and M/M/n+G as they extract model essentials. This, in turn, yields theoretical and practical insights, which are unavailable through exact yet cumbersome mathematical expressions.

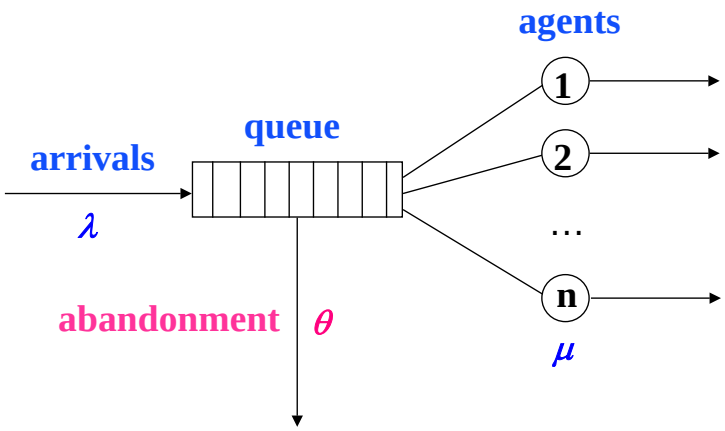


Figure 3: **Schematic Representation of the Erlang-A Model**

2.4 Many-server asymptotics

We restrict the discussion to approximations in steady-state (as opposed to time-varying, such as Mandelbaum et al. [37]). Most approximations to queues with abandonments have been asymptotic. In those relevant to tele-queues, the asymptotic analysis is carried out as the number of servers n (e.g. agents in a call center) increases indefinitely - *many-server asymptotics*. The resulting approximation is then determined by the limiting offered-load-per-server $\rho = \lim_{n\to\infty} \frac{\lambda_n}{n\mu}$, assuming μ remains fixed. Specifically, $\rho < 1$ leads to a Quality-Driven (QD) regime, where operational service level is high at the cost of low servers' efficiency (utilization) and system performance is similar to one without abandonment. In contrast, $\rho > 1$ corresponds to an Efficiency-Driven (ED) regime while $\rho = 1$, if appropriately staffed, gives rise to a Quality- and Efficiency-Driven (QED) regime that attains high levels of *both* service-level and server-efficiency. (The QED regime differs from conventional heavy-traffic; in the latter, $\rho = 1$ via λ and μ increasing indefinitely while n remains fixed, as in Ward and Glynn [60].)

The QED and ED regimes have been most popular research-wise.

- The QED regime is often referred to as the Halfin-Whitt regime, after the authors of [26] where it was formalized in the context of the M/M/n (Erlang-C) queue. The most practical characterization of this regime is in term of its so called square-root staffing rule $n = R + \beta\sqrt{R}$, where $R = \lambda/\mu$ is the offered load. (Note that the computation

of R could demand more elaborate approaches, especially for time-dependent models). The QED work of Halfin and Whitt was extended to M/M/n+M by Garnett et al. [23]. It was further refined in Zhang et al. [65] and generalized to M/M/n+G by Zeltyn and Mandelbaum [64].

- The ED regime corresponds to many-server systems with a significant fraction of abandonments. Being rooted in fluid-approximations (as opposed to diffusion approximations in the QED case), the ED regime is much more amenable to extensions of the basic M/M/n+M model, for example multiple customer and server types or an uncertain arrival rate; e.g. see Bassamboo et al. [13] and Bassamboo and Zeevi [14]. We shall further discuss ED system management in Section 3.2.

But the QED and ED regimes cover only a small part of the asymptotic landscape of Erlang-A, which has been greatly expanded in recent years. To this end, Ward [62] and Gurvich et al. [25] develop universal approximations, with each accommodating several regimes simultaneously.

2.5 Factors that affect impatience

Empirical studies of abandonment give rise to many interesting and often surprising issues. A central theme in these studies has been the search for factors that affect impatience. The factors of which we are aware can be usefully and roughly classified into the following categories:

- *Customer category.* Are VIP customers more or less patient than regular ones? What are the differences, if any, in patience patterns between people of various cultural backgrounds, needs, etc.?

- *Call category.* Will customers that need a longer service be more patient? What are the differences, if any, in patience patterns between service types for the same customer? How does the time-of-day (e.g. morning vs. evening) affect patience?

- *Overall experience.* Anticipated waiting time is an obvious factor that can be integrated into queueing models. It is clear though that more qualitative factors that pertain to accumulated experience (e.g. past quality of service) could also be of importance.

- *Local experience.* Here one encounters many interesting questions such as the impact of automatic announcements, IVR experience and history of recent retrials.

It is impossible, within a single paper, to provide empirical examples for the influence of all these miscellaneous factors. Nevertheless, we have tried to touch on most of them through

our data-stories, as told in the following section. (One important omission is the impatience of customers, as it evolves through successive redials that follow either a service or an abandonment; see Khudyakov, Gorfine and Feigin [32] for a recent analysis.)

3 Data Stories of Customer Impatience

3.1 Fitting Patience Distribution

Can one observe an uncensored patience distribution? The left-hand plot in Figure 4 summarizes the tele-waiting experience, of close to 12,000 customers of an Israeli Bank during November 24, 2008. These customers were seeking phone service from their bank's call center, being unaware of the fact that, due to equipment malfunction, agents were unable to accept their calls. As far as we know, the customers in this call center did not hear any announcements on estimated wait, neither on regular days nor specifically on November 24, 2008. The histogram is that of their waiting time until they gave up and abandoned (hung up). This is a rare data set, from which one can learn that phone customers can in fact be very patient. The mean is about five minutes, with some customers willing to wait close to half an hour. On the other hand, a significant fraction of customers abandon during their first several seconds of waiting. The right-hand plot in Figure 4 shows the patience distribution for waiting times that exceed 10 seconds.

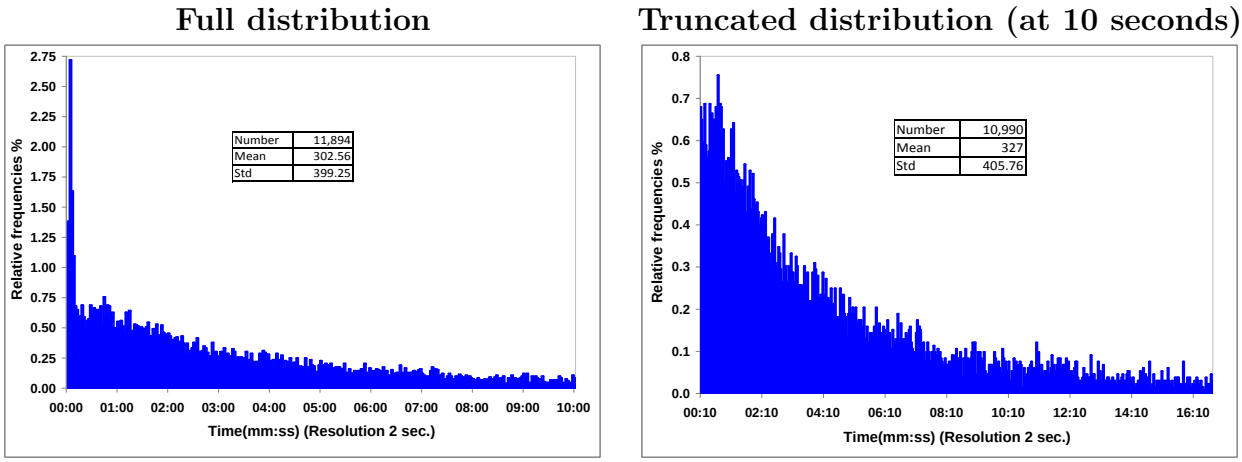
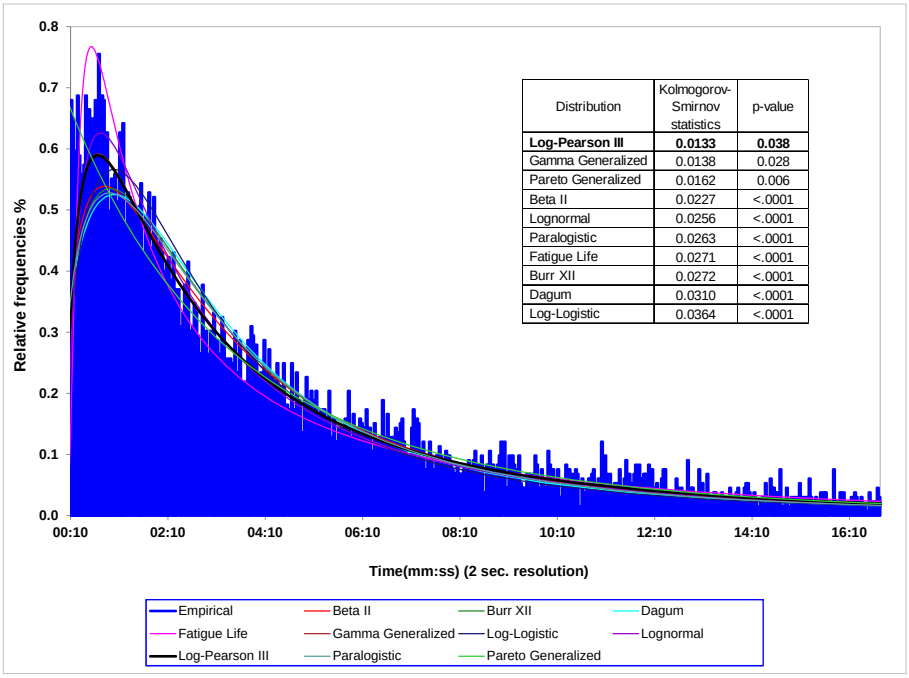


Figure 4: **Financial Call Center. Patience at a day with no service**

We now fit several theoretical distributions to our uncensored data. There are clear practical and theoretical reasons why this is an important and interesting exercise: informed choice of an appropriate queueing model, better understanding of customer impatience,

etc. (An additional reason will be provided below, when discussing Figure 7.) The SEEStat software [59] of the Technion SEELab provides us with statistical tools to perform our fitting.

The left-hand histogram in Figure 4 demonstrates a complex pattern near the origin. For clarity, we thus start with the simpler, truncated right-hand plot[1]. Figure 5 displays the ten distributions with the best fit, including a table with the output of the Kolmogorov-Smirnov test. (Two other tests imply very similar rankings.) The Log-Pearson III distribution provides the highest p-values, and the Generalized-Gamma distribution is second best. While these are useful options, neither reveal an excellent fit (thus the corresponding hypotheses are rejected at the 5% confidence level).



Distribution	Kolmogorov-Smirnov statistics	p-value
Log-Pearson III	**0.0133**	**0.038**
Gamma Generalized	0.0138	0.028
Pareto Generalized	0.0162	0.006
Beta II	0.0227	<.0001
Lognormal	0.0256	<.0001
Paralogistic	0.0263	<.0001
Fatigue Life	0.0271	<.0001
Burr XII	0.0272	<.0001
Dagum	0.0310	<.0001
Log-Logistic	0.0364	<.0001

Figure 5: **Patience at a day with no service. Fitting various distributions**

Our uncensored data set provides an opportunity to fit a model for the tail of the patience distribution. (In a reasonably managed call center, the patience tail is typically heavily censored, e.g. 1-5% of the customers abandon hence over 95% of the observations are censored, which makes the tail hard to estimate.) Figure 6 presents fitting results for patience times that exceed 10 minutes. The Generalized Pareto distribution offers an almost perfect fit. (In contrast to all other distributions, its p-values of SEEStat tests far exceed 0.05.) Since Generalized Pareto is a power-tailed distribution, we have now learned that customers, who

[1]SEEStat uses a number of statistical tests for inferring goodness-of-fit. The main ones are Kolmogorov-Smirnov, Anderson-Darling and Cramer von-Mises tests. The fitting algorithm covers approximately 50 theoretical distributions, and it automatically produces ranked goodness-of-fit measures.

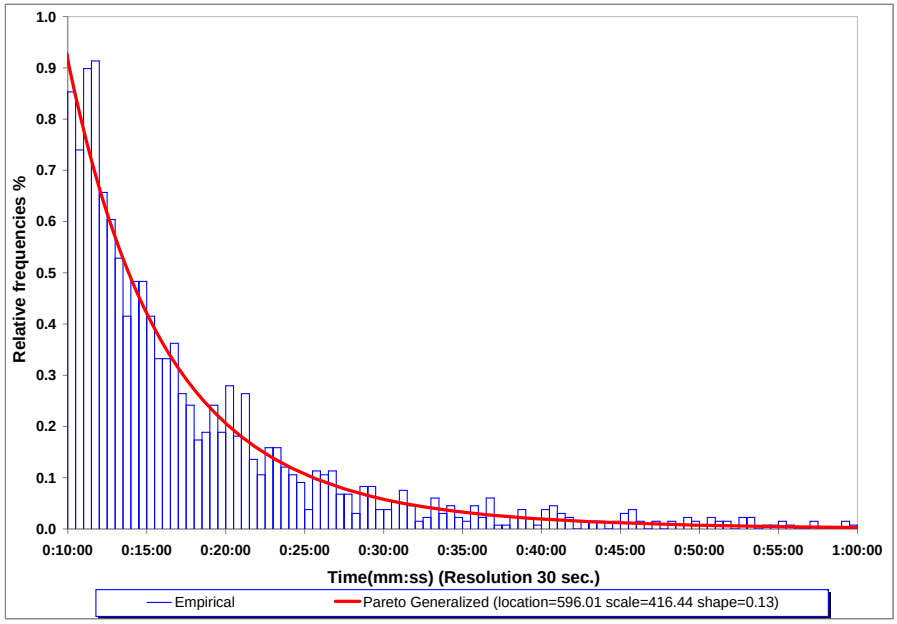


Figure 6: **Patience at a day with no service. Fitting distribution tail**

have already waited for a significant time, tend to remain increasingly patient. We call this phenomenon "accepting one's fate": given that a customer waits t unit of time, the expected residual patience time increases with t; this phenomenon is also characterized by a (tail) decreasing hazard rate of patience.

3.1.1 Balking customers: an atom at the origin of the patience distribution

In a final exercise of distribution fitting, we use the mixture-fitting tool that SEEStat offers. This is a very useful feature when the customer population is heterogeneous, namely constituting a mixture of groups with differing characteristics. For example, in our case, a relatively large fraction of highly impatient customers leave after a few seconds of wait. Figure 7 and Table 1 show the fitting results for the mixture of five lognormal distributions. Table 1 demonstrates that the two last distributions have very small mixture weights and can most likely be ignored. When performing goodness-of-fit tests (SEEStat does it automatically), it turns out that the mixture goodness-of-fit for the overall distribution (Figure 7) is better than a single-distribution goodness-of-fit for the truncated distribution. (Recall the table within Figure 5.) Specifically, three statistical tests based on differences between empirical and theoretical distributions were performed for mixtures: Anderson-Darling, Cramer-von Mises and Kolmogorov-Smirnov. Two tests out of three (except Kolmogorov-Smirnov) imply non-significant p-values (0.11 for Anderson-Darling and 0.73 for Cramer-von Mises).

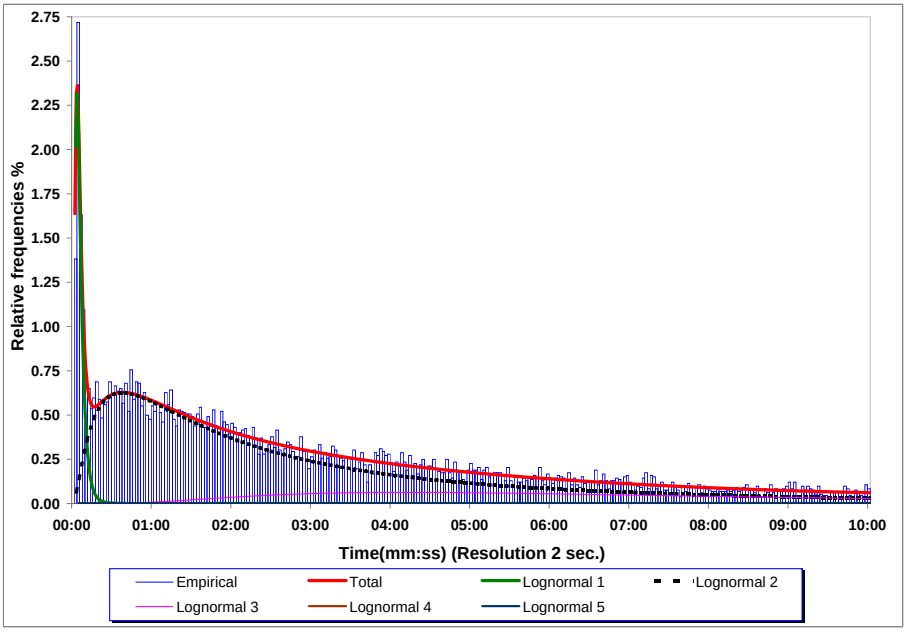


Figure 7: **Patience at a day with no service. Fitting mixtures**

Note the first distribution in Table 1, with its mean equal to 6 seconds. It corresponds to the above-mentioned peak of short waiting times. This suggests a practical mixture-based test for estimating the proportion of *balking customers*: they are defined as those who are unwilling to wait at all, hence they disconnect once they realize that they can not get service immediately upon arrival to the queue. (Practically in our data, it takes such customers several seconds until they disconnect.) In terms of the patience distribution, balking customers are manifested via an atom of the patience distribution at the origin. (Such patience models often require a special theoretical approach; e.g. see Zeltyn and Mandelbaum [64].) If we reasonably assume that the balking customers correspond to the first line of Table 1, their proportion is approximately 8% of the overall population.

Table 1: **Fitting patience distribution with lognormal mixtures**

Mixing proportions	Mean (sec)	Standard deviation (sec)
8.08%	6.11	3.53
70.18%	251.58	399.25
19.27%	514.87	399.25
1.25%	787.62	135.34
1.22%	1512.66	399.25

3.1.2 Some practical observations

In general, the shape of the patience distribution is idiosyncratic. Together with its moments, the distribution strongly depends on a specific call center, customer type and various additional factors. (See Section 2.5 for a discussion of such factors.) Scarce existing research on actual patience sheds some light on its behavior. The seminal paper of Palm [47] analyzed a case of "uncensored" patience, similar to the one considered above, and suggested the Weibull distribution for impatience modeling. Kort [34] arrived at a similar conclusion based on laboratory testing of patience behavior. Baccelli and Hebuterne [12] analyzed data from Roberts [52] and fitted the Erlang distribution with three phases. Finally, the patience distribution in Brown at al. [15] appears to not be covered by any conventional parametric model.

According to our experience in data analysis of modern call centers, the following observation generally prevails: Tele-customers are *very patient*; average patience is commonly in the range of 5–20 minutes. Another useful natural way of expressing average patience is in terms of its relation to the average service time. (Indeed, in many queueing models, Erlang-A in particular, it is the ratio of the two, as opposed to each separately, that appears in formulas.) Here our experience suggests values for this ratio that exceed 2, namely average patience is at least twice as long as average service time.

As expected, there are exceptions to the above rule, which raises the need for an ad hoc empirical analysis to better accompany specific queueing applications. This fact, and hopefully our examples, provide sufficient incentives to study empirical patience distributions. As mentioned, these patience distributions are typically idiosyncratic. It is hence unreasonable to expect that a single classical parametric family of distributions, or several such families for that matter, will encapsulate them. Nevertheless, even simple parametric models of patience have been found very useful, as described in Section 3.7. Note also that, as a rule, patience data are highly censored and their fitting raises nonstandard statistical challenges. The survival analysis tools (methods for fitting censored data) in SEEStat [59] can take one a long way toward addressing these challenges.

3.2 The Efficiency-Driven (ED) regime: call centers with ample abandonment

Figure 8 summarizes daily performance at a call center of an Israeli Telecom service provider, where the solid line shows the hourly arrival rate and the dashed line is the number of served customers per hour. One observes that only 24% of the calls were answered while the rest abandoned. One can further note the following:

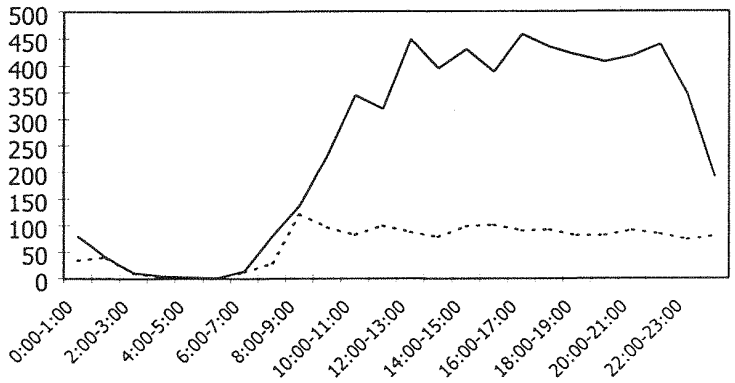


Figure 8: **Demand explosion in a call center of a Telecom company**

- Service capacity is about 100 calls per hour while service demand, during most of the day, exceeds 450 calls per hour; the call center has been clearly malfunctioning.
- Another confirmation (in numerical data that is not displayed here) that phone-customers could be highly patient: During the peak hour, from 13:00–14:00, the average waiting time (across all customers) was close to 15 minutes, while there were customers who waited over 20 minutes.

The cause of this "overload phenomenon" is the following: The Marketing branch of the Telecom company in question invested a lot of effort in promotions that led to a very significant increase in the arrival volume. However, this effort was uncoordinated with the Operations branch, in charge of the call center. Consequently, adequate measures to increase call center staffing were not taken which led to poor performance over approximately a month, as displayed in Figure 8.

The example in Figure 8 is extreme. However, many call centers are often managed in the overloaded regime: Almost all calls have to wait and the waiting times in the tele-queue are of the same order (or even larger) than the service times. An overloaded service system has the connotation of being inadequately managed, in that capacity is not well-matched with demand. But this need not be the case. Oftentimes, due to policy and union constraints, staffing levels are rigid (as opposed to flexible) over time-periods where overloading occurs - which inevitably leads to overloading. Moreover, economies of scale enable overloaded moderate-to-large call centers to still deliver acceptable service levels. Their operational regime is then Efficiency-Driven (ED) [23, 61] or its relative ED+QED, introduced in Mandelbaum and Zeltyn [42].

Figure 9 displays arrival rates of customers to Telesales service in a large U.S. bank during eight days in October 2001. One observes that the arrival volume during the second week is

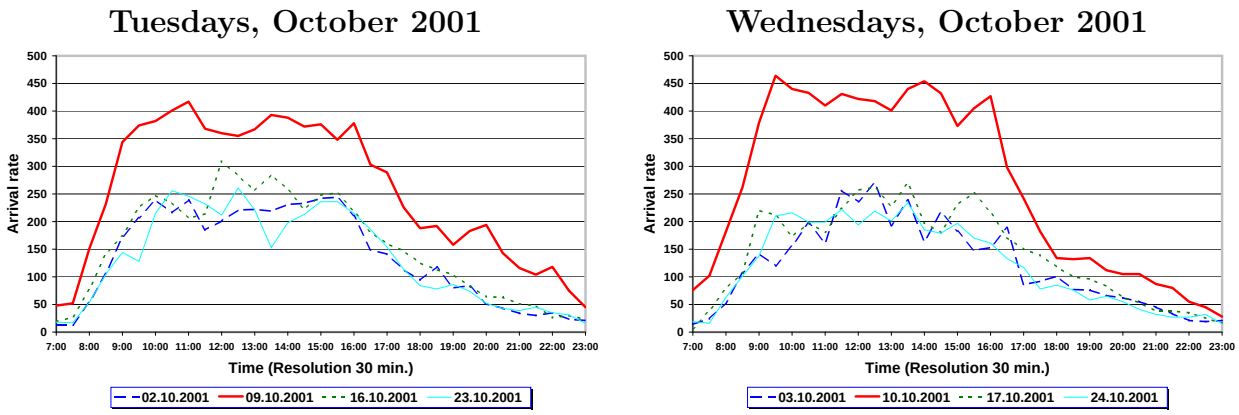


Figure 9: **US Bank. Arrival rate of Telesales customers**

much higher than during the other weeks. In fact, Monday, October 8, was "Columbus" day and, as is frequently the case, a surge in arrivals is observed after holidays. In this case, the surge was more significant than on average. Very likely, a marketing promotion took place during these days as well.

Do we observe a corresponding increase in the number of agents? Figure 10 shows that agents were added on October 9–10, but to a lesser extent than the arrival rate; especially on October 10, where the arrival rate during the most overloaded hours was larger than on the previous day yet the number of agents was smaller.

Finally, Figure 11 displays two key performance measures on these two days. The overall probability to abandon is higher on October 10 than on October 9: 51% vs. 43%. Average waiting times are almost the same: 543 seconds on October 9 vs. 539 seconds on October 10. It looks as though the customers become less patient towards the midday of October 10, possibly because they are already aware of the low service level that is offered.

This example underscores the importance of the ED regime: it often arises in practice due to short-term (up to several days) surges in the arrival rate, as in Figure 9. (An operationally prevalent solution to such surges is outsourcing on demand.) This is to be contrasted with a long-term surge, which seems to be the case in Figure 8, and against which an adequate number of agents must be hired.

3.2.1 Some ED observations, on theory and practice

Staffing in the overloaded or ED regime differs from that in balanced systems (where a significant fraction of customers get served without delay). On one hand, its supporting theory is easier in that *fluid models* suffice to capture overloaded performance; these models are, in general, more tractable than QED or exact models. On the other hand, two potentially complicated phenomena become more pronounced in overloaded systems.

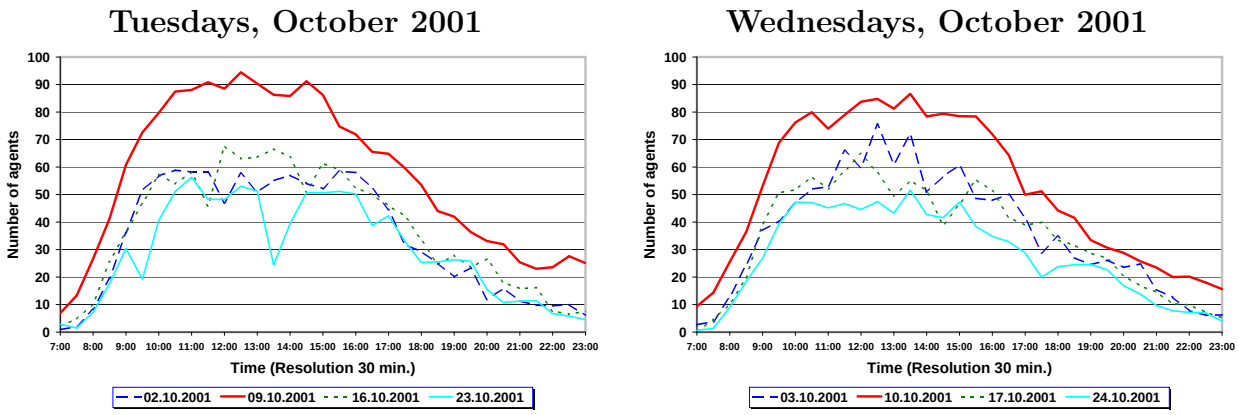


Figure 10: **US Bank. Telesales customers. Number of agents**

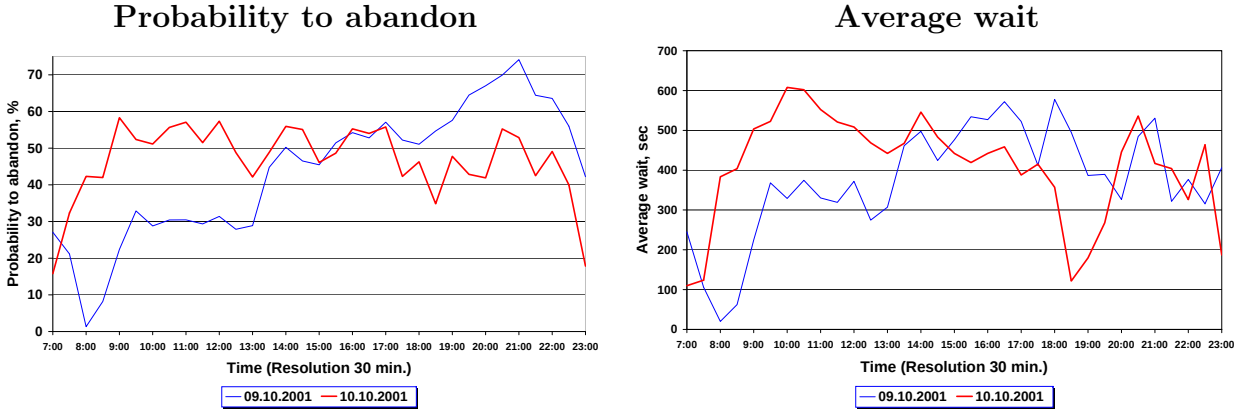


Figure 11: **US Bank. Performance measures on October 9–10, 2001**

First, one must not resort to the conventional stationary independent period-by-period (SIPP) approach to staffing. (SIPP uses steady-state models to support staffing during disjoint basic intervals, independently across interval, e.g. 48 half-hour intervals over the day or similarly 15-minute intervals; see Green et al. [24].) However, in an overloaded call center, a significant fraction of customers could be served during a basic-interval that in fact occurs later than the one they arrived at, which creates dependence across intervals. (In some examples that we tested, SIPP staffing under such undesirable circumstances could lead to overstaffing that exceeds 10% of the workforce.) This naturally calls for time-dependent models which, unfortunately, are notoriously intractable unless one resorts to their approximations. As a convincing example, time-dependent approximations yield staffing rules that in fact stabilize performance: for example, Feldman et al. [20] stabilize the delay probability, and Liu and Whitt [35] attend to the probability of abandonment, the latter being more

relevant for the ED regime.

Potential dependence between service and patience times (e.g. willingness to wait more for longer services) presents an additional challenge. Indeed, when a significant fraction of customers abandon, it is not immediately clear that the handle time of the served customers is representative of the abandoning population. We shall return to this topic in Section 3.5.

3.3 Impatience, announcements and customer expectations

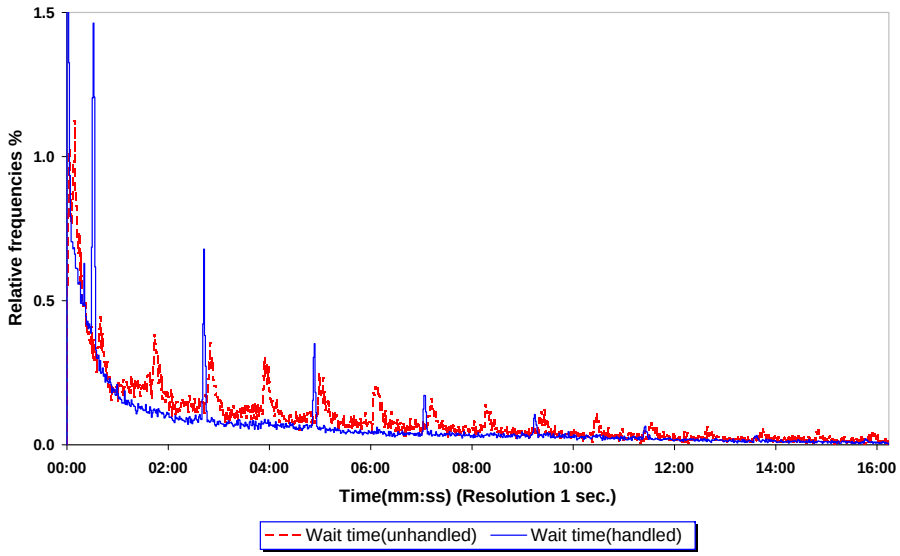


Figure 12: **Patience and offered wait in a US bank**

In the Introduction, we discussed the difference between visible and invisible queues. It stems from the fact that customers in invisible queues do not see their waiting peers. There are, however, two factors that sometimes mitigate the difference. First, experienced customers arrive to queue with certain wait expectations. Second, many call centers provide announcements regarding the customer's state in the queue.

Figure 12 presents histograms of (censored) waiting times for handled (served) and unhandled (abandoning) calls, which correspond to a specific service type in a US bank. We observe peaks in both histograms: the abandonment peaks occur about every minute while service peaks have a two-minute period. Using the survival analysis techniques for "uncensoring," described in Section 2, we created hazard-rate estimates of patience time and offered wait. Figure 13 displays these estimates at 1-second resolution, after smoothing. The remarkable peaks for both abandoning and served customers are observed again.

Offered-wait peaks: Similar peaks in hazard rates of offered-wait have been observed in other data sets. These peaks have been typically associated with times at which customers enjoy

an upgrade in their priority status, e.g. periodically every 1-minute. Such a priority upgrade increases the likelihood of being served, which is manifested by a local peak in the hazard rate of the offered-wait. In our present case, the cause is somewhat different. The service in question was related to a stock exchange operation and thus provided by high-skilled agents. These agents served customers according to a protocol that connected them once customers waited specific amounts of time (which can be derived from the peaks of offered wait). Therefore, agents can be idle for some time before attending to waiting customers. Note that this violates conventional call center protocols, under which service must start if there are waiting customers and there is at least one idle agent (work-conserving strategies, in queueing parlance).

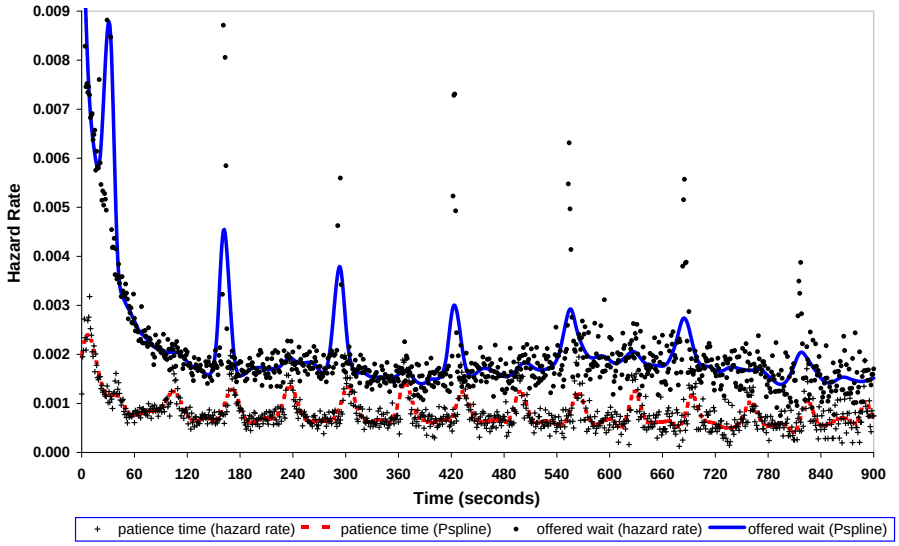


Figure 13: **Hazard rates of patience and offered wait in a US bank**

Patience peaks: Peaks in patience are also prevalent in call center applications. Recall Figure 1 in Section 2, where such a peak arises after an automatic announcement at 60 seconds. Similarly, such peaks are expected to arise periodically if announcements occur periodically. These announcements vary in nature, sometimes providing information on expected waiting time, system load, number-in-queue or combinations of these; the announcement can also be informative at various levels, from attempting at being precise (e.g. "Your waiting time is estimated to be T seconds") to confusing (e.g. "You are number X in queue and the first in queue has been waiting T minutes") or uninformative (e.g. "Your call is important to us"). Regardless, announcements during waiting have been often found to encourage or trigger abandonments, even when the goal of the announcement is precisely the opposite, namely encourage the customer to remain waiting for service.

A more refined example of announcement effects is Figure 14, taken from [19]. It shows (Kaplan-Meier) survival-function estimates of the patience of customers from a small Israeli bank. The customers were exposed to an announcement at the outset of their wait, and they are here grouped according to type of the announcement; the majority (around 90%), who did not get an announcement at all, constitutes a separate group. The announcements included six different estimates of waiting time, referred to as SmartQ in the figure legend (less than 1 min, 1 minute, ..., 4 minutes, more than 5 minutes). These wait estimates correlated positively with the actual wait. Note, however, that the actual waiting times of many customers deviated significantly from their estimated waits. Announcements of long waits included also a recommendation to return to the IVR service but only a small fraction of customers did so. Note that announcements take place *before* service and not *during* it. The two settings can, in general, imply different psychology of abandonment.

Several interesting conclusions can be (cautiously) drawn from Figure 14 (x-axis units are seconds). First, customers who did not receive any announcement appear to be the least patient. Second, customers who are promised a short wait, of less than 1 minute, become impatient at some point (after about 2.5 minutes). This phenomenon is psychologically plausible. Finally, customers with the longest estimated wait (> 5 min) seem to be relatively impatient; however, this can be attributed to the small number of observations in this group.

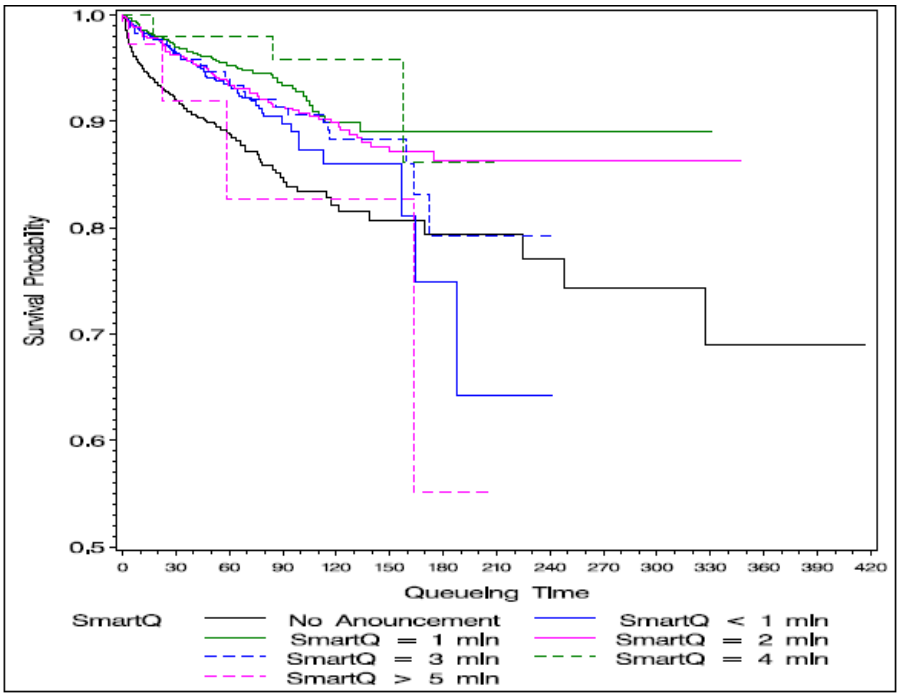


Figure 14: **Israeli bank. Survival functions for different announcements**

3.3.1 Patience control

The phenomena observed above raises a number of interesting questions. If the goal of announcements is to help retain customers in their tele-queue, do they typically achieve this goal? What is a good or best announcement strategy, if the goal is to help customers reach an "informed choice" on waiting vs. abandonment? What would it be if the goal is to maximize revenues for the service provider? And is there an announcement strategy that simultaneously attains the goals of both customers and their service provider?

Recently, a considerable number of papers on the relation between announcements and abandonment have appeared. The conventional assumption is that an announcement takes place when a customer joins the queue. This announcement can be based on real-time wait estimates, steady-state models or even be biased in order to maximize profits. The corresponding models can rely on game theory considerations; see, for example, Allon et al. [3] and references therein. Descriptive behavior models comprise an alternative to a rational decision paradigm (e.g. Armony et al. [8]). For both types of models, practical validation is of major importance and seems to be lacking. Moreover, we are unaware of any research on announcement systems with *multiple*, say periodic, updating announcements, and these are abundant in practice.

A final observation is the emergence of announcement systems that offer customers the option of "abandoning," which is then followed by a later return of their call. Such a system creates a service *inventory* of calls-to-be-returned. It frees customers from waiting and it alleviates provider's waiting costs (e.g. 1-800 costs). The call-return strategy can be either customer- or system-driven: in the former case, the customer specifies a convenient return-time; in the latter, a fair option is to retain (more or less) the customer position in queue in the sense that their call will be returned around the time that they would have reached service, had they not abandoned.

A related research direction studies the behavior of customers who adapt their patience time to service expectations. Examples of information available to experienced customers include the mean offered wait (Zohar et al. [66]) and distribution of the offered wait (Mandelbaum and Shimkin [38]). Armony and Maglaras [7] study a model where customers are provided, upon arrival, with information on anticipated delay and then choose between the options of waiting, abandoning, or receiving a callback.

3.4 Impatience and the IVR experience

The vast majority of customers of large call centers start their service process at the IVR (Interactive Voice Response unit), which is sometimes referred to as VRU (Voice Response

Unit). A very significant fraction of these customers (for example, around 70-80% in the banking sector) also complete their *self-service* process at the IVR, which circumvents any contact with a human agent. An interesting "abandonment" phenomenon arises when a customer originally seeks self-service at the IVR, but then opts out to agent-service due to, say, poor design or insufficient capabilities of the IVR system. (A similar problem is also prevalent in Internet services, both stand-alone or those that are inter-connected to call centers.) Empirically, the *opt-out* phenomenon is difficult to infer directly - it has to do with customer intentions - which could be the reason behind the scarcity of its research.

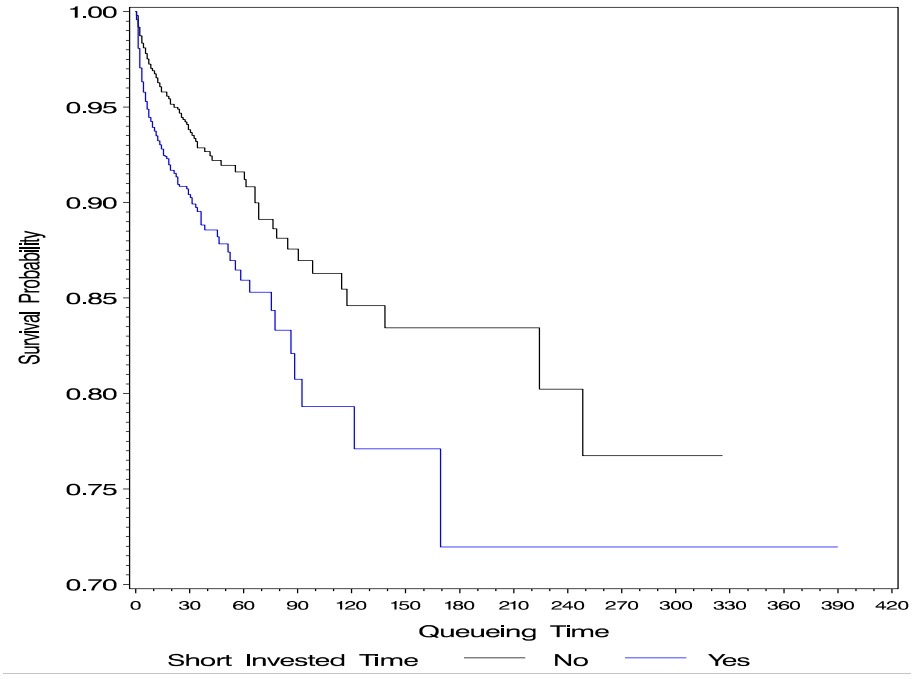


Figure 15: **Patience survival functions of customers according to invested times of customers**

Customers who seek service from an agent, after their IVR experience, must often wait in tele-queues. A natural question arises regarding their patience in these *post-IVR* phases of service, in particular the effect of the IVR experience on abandonment. A starting point for research on the subject could be the few queueing models that incorporate IVR, for example Srinivasan et al. [55], Khudyakov et al. [31], Tezcan and Behzad [57]. These models, as well as preliminary queueing-related empirical analysis of IVR services, are surveyed in Yuviler [63].

Figure 15 from Feigin [19] is based on data from a small Israeli call center. Essentially all customers there had to go through an IVR phase. Customers who also sought agent service

listened to a post-IVR message and were transferred to a tele-queue. Define the *invested time* of a customer before entering the tele-queue as the IVR time plus duration of post-IVR message.

How does the invested time affect patience during tele-queue waiting? The upper estimated survival curve in Figure 15 corresponds to customers who invested more than 100 seconds and the lower curve corresponds to the rest: those that already invested more time appear to be more patient. Of course, one should be cautious with such conclusions: There are likely to be unknown factors (e.g. more or less important call type) that influence both pre-agent invested time and patience at the tele-queue. Clearly, the issue of post-IVR factors that affect patience calls for further statistical research.

3.5 Impatience and service time

Independence between stochastic building-blocks of a service model is a standard assumption in queueing theory. Specifically, in queues with abandonment, service and patience times are typically assumed to be independent. Is this a realistic assumption? One can raise at least three plausible reasons against it. First, a longer service could, on average, be more important for a customer and, in turn, service importance naturally affects patience. Second, longer waits can potentially lead to shorter service times for customers whose allowable sojourn time over the phone is restricted. Third, customers' irritation from prolonged waits can influence their behavior during calls (e.g. the customer-agent exchange starts with complaints and apologies) and can ultimately prolong call duration.

This dependence between patience and service was empirically studied in Reich [50], who proposed a framework to address it and analyzed some of its operational (queueing-related) consequences. First, [50] offers a statistical test that checks for independence between service and patience times. Then, if dependence is not rejected, the following non-linear regression model attempts to capture the relation between service and waiting times:

$$S_i \;=\; g(W_i) + \epsilon_i,$$

where S_i and W_i are the service and the waiting time of customer i, respectively; $g(w) = E[S|\tau > W = w]$ is the mean service duration of those who waited exactly w units of time and were served. Significantly, this $g(\cdot)$ is *observable*. Under independence between (τ, S) and V, which is plausible for invisible queues (V is unobservable to customers during their waiting), g in fact simplifies to $g(w) = E[S|\tau > w]$ [50]. Figure 16 shows an estimate of $g(w)$ for the Retail customers in a large US bank. (First, $g(w)$ was inferred for each second, which is given by the individual points; then these points were smoothed via a cubic spline with 5 knots.)

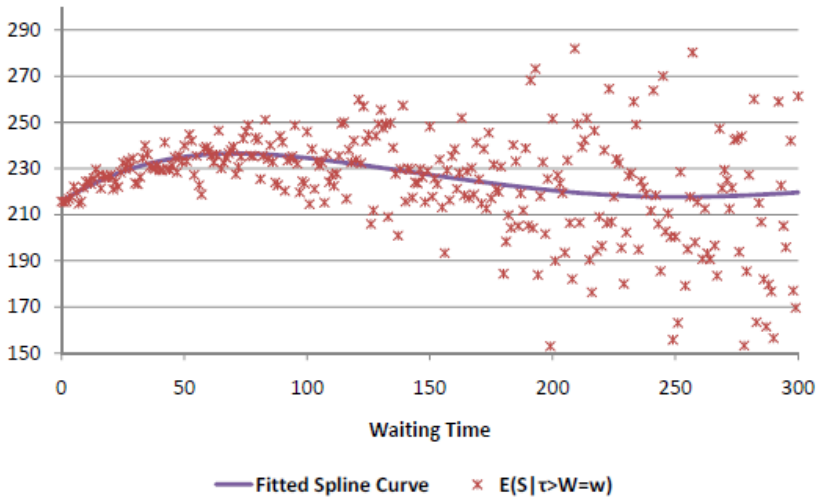


Figure 16: **Mean service time as a function of waiting time**

We observe an apparently mild non-monotone dependence between $g(\cdot)$ and waiting time (values range between about 210 seconds and 235 seconds). This, however, could (and in fact does) reflect a much wider range for $E[S|\tau = w]$, which is the function one is ultimately seeking: the mean service duration of those whose patience is w units of time.

One could, in principle, infer the whole distribution of S conditioned on τ. Restricting the analysis to the mean, for simplicity, this can be done via the following formula [50]:

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

where $f_\tau(\cdot)$ is the patience-time density and $h_\tau(\cdot)$ is its hazard rate. It turns out that an accurate estimation of $g'(w)$ (a derivative) and $f_\tau(w)$ (based on censored data) is non-trivial. This, in turn, complicates the reliable inference of $E[S|\tau = w]$, which is still work-in-progress [51]. As already indicated, preliminary results show that $E[S|\tau = w]$ strongly depends on w, and exhibits significantly larger variation than the $g(w)$ in Figure 16; see Figures 7.14 and 7.18 in [50].

An additional important topic of Reich [50] relates to calculation of the offered-load under dependence between service and patience times. It also relates to abandonment as this offered-load must account for the (potential) service times of those who abandoned (and hence were not served). Offered-load calculations are of great practical importance. For example, they are required to correctly apply the square-root staffing rule in the QED regime.

3.6 Impatience and customer priority

Many commercial call centers differentiate customers according to their business value. Consequently, they design higher service levels for their VIP customers, both in terms of wait (shorter) and agent skills (VIP agents). The comparison of patience between VIP and regular customers yields interesting empirical findings. These point to an uncharted research direction which could shed light on the practices and psychology of call center customers.

On the surface, various factors can affect the patience of VIP customers in different ways: A high-priority customer can be more irritated when encountering wait and, hence, be *less* patient; VIP customers plausibly need more important services under more demanding circumstances, when compared against regular customers, which would render them *more* patient; and VIP customers potentially trust the system more, which would also lead to more patience, etc. We reiterate that practical research on this and similar topics has been scarce. To the best of our knowledge, Mandelbaum et al. [40] remains the only relevant source, and the following "story" is adapted from there.

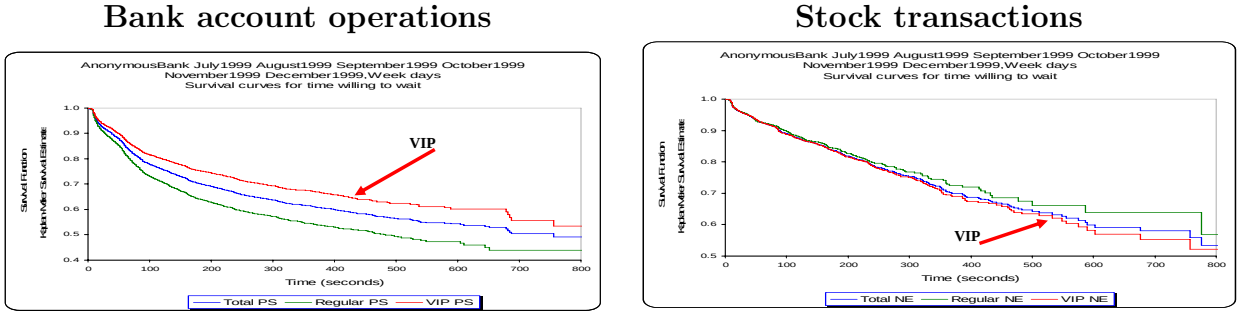


Figure 17: **Israeli bank. Impatience of VIP and regular customers**

Figure 17 compares the estimated survival functions of patience, for VIP and Regular customers in an Israeli bank. The middle curve in each plot presents the survival function for all customers. The actual service policy was as follows: High-priority customers were moved up in queue by subtracting 1.5 minutes from their actual arrival times; then a first-come-first-served discipline was enforced across all customers. Two service types were considered. The left-hand graph displays survival curves for standard banking operations. We observe that VIP customers are more patient in this case. In contrast, VIP customers are less patient that Regulars while waiting for a stock-exchange service. One can speculate on the reasons for these differences but, we believe, that "correct" answers can be only derived through a combination of statistical and psychological research. As a final comment, our data reveal a stochastic order between the patience distributions of VIP vs. Regular customers; but, as seen in [40], there is in fact an ordering between the hazard-rates for the bank account

operations – Regular being above VIP - which is stronger than the order between survival functions.

3.7 M/M/n+M model validation

In our previous data stories, we presented a rich spectrum of impatience patterns but without any detailed discussion on their integration into queueing model, specifically if the ultimate goal is applications of these models. Our last "story" now discusses the applicability of Erlang-A, or M/M/n+M, which is the most basic model that accommodates impatience and hence abandonment. Our experience suggests that it is used by some advanced workforce management software, as a basic block of their staffing engine. (Surprisingly, Erlang-C (M/M/n) is still used by some, perhaps many, who then must become very "creative" in order to generate information about abandonment from their patience-ignorant model.)

Consider a call center, or one of its departments, which consists of a group of (more or less) homogeneous agents, who are all dedicated to serve a homogeneous population of customers. The following two questions then naturally arise:

1. Does M/M/n+M describe the reality of such a call center with practical precision?
2. Is it reasonable to use M/M/n+M for staffing design? What is an appropriate methodology in case of an affirmative answer?

The two questions seem almost identical, but they are not. We shall mainly discuss the first question while only briefly touching on the second one. To this end, we use data from an Israeli Telecom company, maintained at the Technion SEELab [53].

3.7.1 The First Question

The following approach was pursued to address the first question. A basic interval length was first chosen (half-hour in our case), and steady-state formulae were assumed applicable at this resolution. Then, for every interval i, historical data provided estimates of the four M/M/n+M parameters: arrival rate λ_i, service rate μ_i, impatience rate θ_i and staffing level n_i. Specifically, the actual values for a specific interval can be used to estimate its λ_i, μ_i and n_i. For the call center under study, data on the arrival and service rates was easily accessible, and we also had reliable estimates for the number of agents. (The latter information is typically harder to obtain than information on arrival rates and service times.)

In contrast to arrivals and services, estimation of the impatience rate θ_i cannot be performed separately for an individual interval; the number of observations is simply too small (e.g. there can be no abandonment in a specific interval). We thus used a well-known linear

relation between the probability to abandon and average wait, which prevails for queueing systems with exponential patience: $\mathrm{P}\{\text{Abandon}\} = \theta \cdot \mathrm{E}[W]$ (e.g. [41], where the robustness of this estimation procedure, beyond exponential patience, is explained). This relation can be applied to any data subset for which a single patience distribution is found to hold. In our case, $\theta_i \equiv \theta$ was found sufficient for our purposes (equal patience at all times).

Having estimated the 4 parameters for each historical interval, we then measured performance measures such as the delay probability, probability to abandon and average wait. Now we substituted the input parameters into M/M/n+M formulae and compared steady-state performance predictions against the actual observations.

Due to stochastic variability, it is overoptimistic to expect a good fit at the resolution of a single interval (1/2 hour). Then several types of performance aggregation can be performed: by time period (day, week), by time-of-day (say, all 9–9:30 Thursday intervals during several weeks) etc. Should we expect to get an unbiased relation and a good fit using the described approach? Some interesting preliminary results on this issue were presented in Brown et al. [15] but, in that research, independent data on staffing levels was unavailable and an indirect method was used to estimate them. (As mentioned above, the analysis in this section uses data from a different call center, which did provide reliable information on the number of agents.)

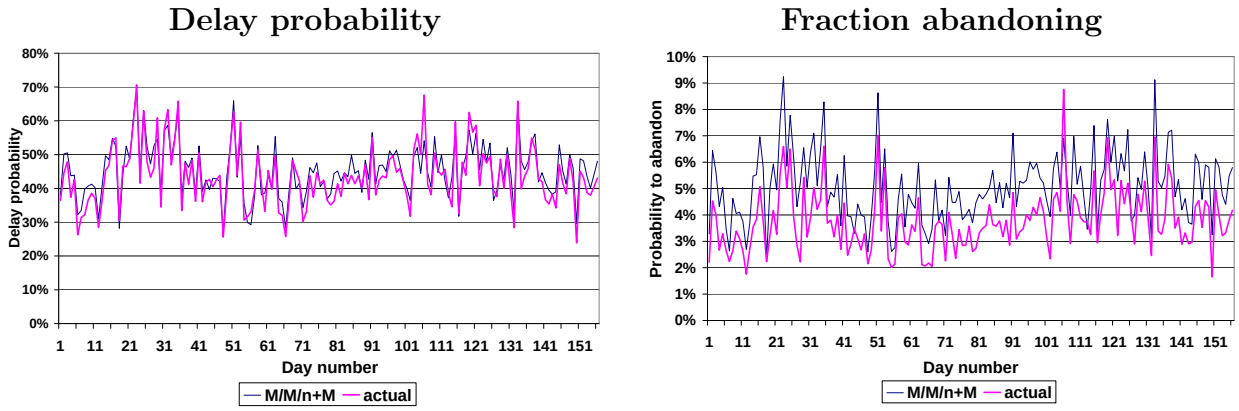


Figure 18: **Comparisons between M/M/n+M predictions and actual performance**

Figure 18 displays two comparisons between actual values and M/M/n+M predictions for one of the considered service types. A daily data aggregation was performed. We observe a very good fit for the delay probability, but a biased "pessimistic" M/M/n+M predictions for the fraction abandoning.

To contemplate the reasons for the above two findings, consider a single interval, some performance measure P, and assume that the M/M/n+M model prevails. Then denote

the steady-state expected performance by $P_{ss}(\lambda_i, \mu_i, n_i, \theta)$ (assuming, for simplicity, that the patience parameter is known and identical across all time intervals). The method described above uses $P_{ss}(A_i, 1/\bar{S}_i, n_i, \theta)$, where A_i is the actual arrival rate and $\bar{S}_i$ is the actual mean service time. In general, $\mathrm{E}[P_{ss}(A_i, 1/\bar{S}_i, n_i, \theta)] \neq P_{ss}(\lambda_i, \mu_i, n_i, \theta)$, hence the presented method is likely to be biased.

On the other hand, it is plausible that a better fit would be observed on the left-hand graph for the delay probability. The reason is that the delay probability is approximately linear in the arrival rate, over the relevant performance regime (in the QED regime, the delay probability ranges between 30% and 70 %; see Figure 7 in [41]). In contrast, the probability to abandon is convex in the arrival rate over the small values of fraction-abandoning in the QED regime. This convexity implies upward-biased estimates due to Jensen's inequality. The ongoing research of Mandelbaum et al. [39] addresses the problem and studies some preliminary methods to correct this bias. Jouini et al. [29] provide a relevant recent reference on the fit of Erlang-A extensions to reality. They fit Erlang-A with balking and M/M/n+G with hyperexponential patience to actual call center data, and report a very good correspondence with reality.

3.7.2 The Second Question

Concerning the second question posed above, a standard approach to staffing applications of the queueing models is as follows: All model parameters for a certain basic interval, except the staffing level, are predicted. Then, given the staffing level, system performance can be estimated. Inversely, staffing levels that correspond to desired performance (goals) can be computed, assuming implicitly that the predicted values of all other parameters are accurate over the whole planning horizon. (This horizon spans several weeks into the future in most call centers.) But in fact, this accuracy assumption is typically false, at least in the case of arrival rates (e.g. Steckley et al. [56]). Hence, even if the system is really M/M/n+M, it can still be non-trivial to staff it according to M/M/n+M formulae. In the case of "uncertain forecast", models with random arrival rates are relevant. As mentioned in Section 2 above, overloaded systems are typically much more tractable for the corresponding approximations. See Maman [36] for an alternative approach, which generalizes the QED regime in order to capture over-dispersion in the arrival rate.

4 Some Challenges for (Im)Patience Research

Several open research problems and directions have already been raised throughout the case studies of Section 3. We now conclude with some additional themes, the research of which

could advance the understanding of customer (im)patience and their abandonment from tele-queue.

4.1 Understanding empirical patience

The distribution of customer (im)patience is a central building block of queueing models that acknowledge abandonment. However, we are still lacking a clear comprehensive understanding of its operational scope and the best statistical way to capture it - in particular, what types of distributions arise in practice, and are hazard rates their best representations? (Recall the discussion in Section 2.2 on "the elusive concept of hazard rate".) What complicates the task is that, frequently, one is actually observing a mixture of populations, with differing characteristics, yet without refined information about the individual constituents of the mixture. The Frailty approach (Khudyakov et al. [32]) has been proved successful to capture such heterogeneity. An alternative is phase-type modeling, as in Ishay [28] which is consistent with the approach advocated in [1].

As far as queueing applications are concerned, one can cope with customer heterogeneity via asymptotic models - these reveal the essentials that could render the models robust with respect to distributional assumptions on (im)patience. A convincing example in this regard is that, in the QED and QD regimes, merely the value of the patience-density at the origin could capture queueing performance. One must be careful though to account for additional characteristics that affect the asymptotic models, for example balking (as in Section 3.1; see Zeltyn and Mandelbaum [64] for further details.)

4.2 Control of queues with (im)patient customers

How to account for (im)patience when deciding on the priority of customers who are waiting for service? Except for special circumstances, the answers become more complex than those which assume away abandonments. For example, Atar et al. [10] proved, for the multi-type Erlang-A in the ED regime, that the $c\mu$-rule minimizes *abandonment* costs (c denotes cost per abandonment; note that, in [10], the focus is *waiting* costs, in which case the index is $c\mu/\theta$); but [33] presents a clear example of the challenges that arise when exponentiality assumptions are relaxed. The latter paper also offers an updated survey of asymptotic control for queues with abandonments. Non-asymptotic work can be found in [6, 44] and the references therein.

In the above research, the (im)patience distribution is taken as given, sometimes with restrictive assumptions (e.g. exponential patience, increasing hazard rates). The present paper can support or refute these assumptions for a specific call center or, in turn, suggest new ones. Moreover, emerging data resources on (im)patience and abandonment would also

enable service policies that are tailored to the experience of the individual customer: for example, assign high priority to a customer who abandoned frequently in recent past; or aim to serve a specific customer at the time that is tailored to this specific customer's value.

4.3 Approximations and their empirical validation

In Section 2.3, we briefly discussed asymptotic approximations of queueing systems with impatient customers. Our focus was on many-server queues, which offer a natural modeling-environment for tele-service operations. Here the ED and QED regimes have been the most popular, and they have enjoyed extensions and refinements. Examples include ED+QED [42], hazard-rate scaling [49] and non-degenerate slowdown [11]. Given the existing asymptotic landscape, some guidance through this proliferation of asymptotic regimes is called for. We have in mind validation efforts in the spirit of those in Section 3.7 and are now in search of a regime that fits well a real system. And as in Section 3.7, empirical analysis of (im)patience will be essential, but with some additional asymptotic guidance (e.g. for QED queues, as indicated above, it suffices to estimate the value of the (im)patience density at the origin rather than inferring the whole (im)patience distribution.) A preliminary validation of QED approximations was carried out by Brown at al. [15], but systematic studies are yet to be performed.

4.4 On the assumptions of survival analysis

While discussing inference of (im)patience τ and offered-wait V, we explained that one is censoring the other. In fact, it is possible to obtain all observations of V since, at day's end, one can reconstruct their values for abandoning customers (specifically: how long it would have taken, for abandoning customers, to reach service had they stayed in queue). On the other hand, the (im)patience of served customers must be *uncensored*, and here one resorts to survival analysis.

However, the classical assumptions of survival analysis could be easily violated in service settings. For example, different observations of V are typically dependent (a long wait of a customer suggests a long wait of adjacent customers); observations of τ are dependent when they arise from the same customer (e.g. (im)patience that evolves during redials [32]); and the independence of V and τ is no longer natural when waiting customers are exposed to information about their queues, in particular, in face-to-face services. Further research is thus required for developing inference tools for (im)patience that accommodate the idiosyncracies of services. Readers are referred to the Appendix of [66] for some discussion on inference of τ and V, including methods of making these inferences more robust.

4.5 Scope of (im)patience, in tele-queues in beyond

In this survey, we have focused on (im)patience while waiting for a telephone service. Similarly, one could consider (im)patience during a chat/email service (human service) or IVR/Internet service (automatic self-service), with consequent abandonments. And one could expand beyond tele-services, for example to healthcare. In modeling mass-casualty events [44] or long queues for surgeries [9, 54], a central performance measure is death-rate, which naturally corresponds to abandonment-rate; and, as already mentioned, abandonments in Emergency Departments correspond to patients who Left Without Being Seen (LWBS, e.g. [27]). Interestingly, in healthcare applications, the natural time-resolution for measuring (im)patience ranges from hours to months; this is in contrast to tele-services where, as apparent from our data stories, the "right" resolution is seconds.

References

[1] Aalen O., Borgan O. and Gjessing H. (2008) *Survival and Event History Analysis*, Springer. 2.2, 4.1

[2] Aalen O. and Gjessing H. (2001) Understanding the shape of the hazard rate: a process point of view. *Statistical Science*, 16(1), 1-22. 2.2

[3] Allon G., Bassamboo A. and Gurvich I. (2011) We Will be Right with You: Managing Customers with Vague Promises. To appear in *Operations Research*. 3.3.1

[4] Aksin Z., Armony M. and Mehrotra V. (2007) The Modern Call-Center: A Multi-Disciplinary Perspective on Operations Management Research, *Production and Operations Management*. Special Issue on Service Operations in Honor of John Buzacott (G. Shanthikumar and D. Yao, eds.), 16(6), 655–688. 1

[5] Aksin Z., Ata B., Emadi S. and Su C. (2011) Structural Estimation of Callers' Delay Sensitivity. Working paper. 2.2

[6] Argon N.T., Ziya S., and Righter R. (2008) Scheduling Impatient Jobs in a Clearing System with Insights on Patient Triage in Mass-Casualty Incidents, *Probability in the Engineering and Informational Sciences*, 22(3), 301–332. 4.2

[7] Armony M. and Maglaras C. (2004) Contact Centers with a Call-Back Option and Real-Time Delay Information, *Operations Research*, 52(4) 527–545. 3.3.1

[8] Armony M., Shimkin N. and Whitt W. (2009) The Impact of Delay Announcements in Many-Server Queues with Abandonment, *Operations Research*, 57(1), 66–81. 3.3.1

[9] Armstrong P. (2000) Unrepresentative, invalid and misleading: are waiting times for elective admission wrongly calculated? *Journal of Epidemiology and Biostatistics*, 5(2), 117–123. 4.5

[10] Atar R., Giat C. and Shimkin N. (2010) The $c\mu/\theta$ rule for many-server queues with abandonment, *Operations Research*, 58(5), 1427–1439. 4.2

[11] Atar R. (2012) A diffusion regime with nondegenerate slowdown. *Operations Research*, 60(2), 490-500. 4.3

[12] Baccelli F. and Hebuterne G. (1981) On Queues with Impatient Customers. In: Kylstra F.J. (Ed.), *Performance '81*. North-Holland Publishing Company, 159–179. 2.3, 3.1.2

[13] Bassamboo A., Randhawa R. and Zeevi A. (2010) Capacity Sizing under Parameter Uncertainty: Safety Staffing Principles Revisited, *Management Science*, 56(10), 1668–1686. 2.4

[14] Bassamboo A. and Zeevi A. (2009) On a Data-Driven Method for Staffing Large Call Centers, *Operations Research*, 57(3), 714–726. 2.4

[15] Brown L.D., Gans N., Mandelbaum A., Sakov A., Shen H., Zeltyn S. and Zhao L. (2005) Statistical Analysis of a Telephone Call Center: A Queueing Science Perspective, *Journal of the American Statistical Association (JASA)*, 100(469), 36–50. 2.2, 3.1.2, 3.7.1, 4.3

[16] Buckheit J. and Donoho D. L. (1995) Wavelab and reproducible research, *Wavelets and Statistics*, Editor A. Antoniadis, Springer NY, 55–81. 1, B

[17] Cox D.R. and Oakes D. (1984) *Analysis of Survival Data*, Chapman and Hall. 2.2

[18] Donoho, D. L. (2010) An invitation to reproducible computational research, *Oxford Journals*, 11(3), 385–388. 1

[19] Feigin P.D. (2006) Analysis of Customer Patience in a Bank Call Center. Working paper. 3.3, 3.4

[20] Feldman Z., Mandelbaum A., Massey W. and Whitt W. (2008) Staffing of Time-Varying Queues to Achieve Time-Stable Performance, *Management Science*, 54(2), 324–338. 3.2.1

[21] Fernandes C.M.B., Price A. and Christenson J.M. (1997) Does Reduced Length of Stay Decrease the Number of Emergency Department Patients Who Leave Without Seeing a Physician? *The Journal of Emergency Medicine*, 15(3), 397–399. 1

[22] Gans N., Koole G. and Mandelbaum A. (2003) Telephone Call Centers: A Tutorial and Literature Review. Invited review paper, *Manufacturing and Service Operations Management*, 5(2), 79–141. 1

[23] Garnett O., Mandelbaum A. and Reiman M. (2002) Designing a Telephone Call-Center with Impatient Customers, *Manufacturing and Service Operations Management*, 4, 208–227. 2.4, 3.2

[24] Green L.V., Kolesar P.J. and Whitt W. (2007) Coping with Time-Varying Demand when Setting Staffing Requirements for a Service System, *Production and Operations Management*, 16(1), 13–39. 3.2.1

[25] Gurvich I., Huang J. and Mandelbaum A. (2012) Excursion-based universal approximations for the Erlang-A queue in steady-state. Submitted for publication. Downloadable at `http://www.kellogg.northwestern.edu/faculty/gurvich/personal/ErlangExcursions.pdf`. 2.4

[26] Halfin S. and Whitt W. (1981) Heavy-Traffic Limits for Queues with Many Exponential Servers, *Operations Research*, 29, 567–588. 2.4

[27] Hobbs D., Kunzman S.C., Tandberg D. and Sklar D. (2000) Hospital Factors Associated With Emergency Center Patients Leaving Without Being Seen, *The American Journal of Emergency Medicine*, 18(7), 767–772. 1, 4.5

[28] Ishay E. (2002) Fitting Phase-Type Distributions to Data from a Telephone Call Center. M.Sc. Thesis, Technion. Downloadable at `http://ie.technion.ac.il/serveng/References/Thesis.pdf`. 4.1

[29] Jouini O., Koole G. and Roubos A. (2011) Performance Indicators for Call Centers with Impatience. Submitted for publication. 3.7.1

[30] Katz K.L., Larson B.M., and Larson R.C. (1991) Prescription for the waiting-in-line blues: entertain, enlighten, and engage. *Sloan Management Review*, 32(2), 44–53. 5

[31] Khudyakov P., Feigin P. and Mandelbaum A. (2010) Designing a Call Center with an IVR (Interactive Voice Response), *Queueing Systems*, 66(3), 215–237. 3.4

[32] Khudyakov P., Gorfine M. and Feigin P.D. (2010) Test for Equality of Baseline Hazard Functions for Correlated Survival Data using Frailty Models. Submitted for publication. 2.5, 4.1, 4.4

[33] Kim J. and Ward A.R. (2012) Dynamic Scheduling of a GI/GI/1+GI Queue with Two Customer Classes. Submitted for publication. 4.2

[34] Kort B.W. (1983) Models and methods for evaluating customer acceptance oftelephone connections. *Proc. IEEE GLOBECOM 83*, San Diego, CA. IEEE, New York, 706-714. 3.1.2

[35] Liu Y. and Whitt W. (2011) Stabilizing Customer Abandonment in Many-Server Queues with Time-Varying Arrivals. Submitted to *Operations Research*. 3.2.1

[36] Maman S. (2009) Uncertainty in the Demand for Service: The Case of Call Centers and Emergency Departments. M.Sc. Thesis, Technion. Downloadable at `http://iew3.technion.ac.il/serveng/References/Thesis_Shimrit.pdf`. 3.7.2

[37] Mandelbaum A., Massey W.A. and Reiman M. (1998) Strong Approximations for Markovian Service Networks. *Queueing Systems: Theory and Applications (QUESTA)*, 30, 149–201. 2.4

[38] Mandelbaum A. and Shimkin N. (2000) A Model for Rational Abandonments from Invisible Queues, *Queueing Systems: Theory and Applications (QUESTA)*, 36, 141–173. 2.3, 3.3.1

[39] Mandelbaum A., Plonsky O. and Zeltyn Z. (2012) Validation of Erlang-A Model on Call Center Data. Working paper. 3.7.1

[40] Mandelbaum A., Sakov A. and Zeltyn S. (2000) Empirical Analysis of a Call Center. Technical report, Technion. Downloadable at `http://ie.technion.ac.il/serveng/References/ccdata.pdf`. 3.6, 3.6

[41] Mandelbaum A. and Zeltyn S. (2007) Service Engineering in Action: The Palm/Erlang-A Queue, with Applications to Call Centers. In: Spath D., Fähnrich, K.-P. (Eds.), *Advances in Services Innovations*, Springer-Verlag, 17–48. 2.3, 3.7.1, 3.7.1

[42] Mandelbaum A. and Zeltyn S. (2009) Staffing many-server queues with impatient customers: constraint satisfaction in call centers. *Operations Research*, 57(5), 1189–1205. 3.2, 4.3

[43] Mandelbaum A. and Zeltyn S. (2009) The M/M/n+G Queue: Summary of Performance Measures. Downloadable at `http://ie.technion.ac.il/serveng/References/MMNG_formulae.pdf`. 2.3

[44] Mills A. F., Argon N. and Ziya S. (2011) Resource-Based Patient Prioritization in Mass-Casualty Incidents. Submitted to *Manufacturing & Service Operations Management*. 4.2, 4.5

[45] Munichor N. and Rafaeli A. (2007) Numbers or Apologies? Customer Reactions to Telephone Waiting Time Fillers, *Journal of Applied Psychology*, 92(2), 511–518. 5

[46] Nadjharov E., Trofimov V., Gavako I and Mandelbaum A. (2013) Bank Call Center EDA via SEEStat 3.0 to Reproduce "Data-Stories about (Im)Patient Customers in Tele-Queues". Technical report. Downloadable at `iew3.technion.ac.il/serveng/References/reproducing_impatience_data_stories.docx`. B

[47] Palm C. (1953) Methods of Judging the Annoyance Caused by Congestion, *Tele*, 4, 189–208. 2.2, 3.1.2

[48] Palm C. (1957) Research on Telephone Traffic Carried by Full Availability Groups, *Tele*, 1(107) pp. (English translation of results first published in 1946 in Swedish in the same journal, which was then entitled *Tekniska Meddelanden fran Kungl. Telegrafstyrelsen.*) 2.3

[49] Reed J.E. and Tezcan T. (2012) Hazard Rate Scaling of the Abandonment Distribution for the GI/M/n+GI Queue in Heavy Traffic. *Operations Research*, 60(4), 981-995. 4.3

[50] Reich M. (2011) The Workload Process: Modelling, Inference and Applications. M.Sc. Thesis, Technion. Downloadable at `http://iew3.technion.ac.il/serveng/References/Michael_Reich_Thesis_withlinks.pdf`. 3.5

[51] Reich M., Mandelbaum A., Ritov, Y. (2012) On the Relation Between (Im)Patience and Service Durations in Call Centers. In Progress. 3.5

[52] Roberts J.W. (1980). Recent observations of subscriber behavior. *Annals of Telecommunications*, 35 (3-4), 113–124. 3.1.2

[53] SEE: Website of the Technion's "Service Enterprise Engineering" Research Center, `http://ie.technion.ac.il/Labs/Serveng/`. 1, 3.7, A, B

[54] Sobolev B. and Kuramoto L. (2007) Analysis of Waiting-Time Data in Health Services Research. Springer. 4.5

[55] Srinivasan R., Talim J. and Wang J. (2004) Performance analysis of a call center with interacting voice response units. *TOP*, 12, 91–110. 3.4

[56] Steckley S.G., Henderson S.G. and Mehrotra V. (2009) Forecast Errors in Service Systems, *Probability in the Engineering and Informational Sciences*, 23, 305–332. 3.7.2

[57] Tezcan T. and Behzad B. (2012) Robust design and control of call centers with flexible IVR systems. *Manufacturing & Service Operations Management*, to appear. 3.4

[58] Tukey, J. W. (1977) Exploratory Data Analysis, Addison Wesley. 1

[59] Trofimov V., P.D. Feigin, Mandelbaum A., Ishay E., and Nadjharov E. (2006) DATA MOdel for Call Center Analysis: Model Description and Introduction to User Interface. Technion, Israeli Institute of Technology, Technical Report. Downloadable at `http://ie.technion.ac.il/Labs/Serveng`. 3.1, 3.1.2

[60] Ward A.R. and Glynn P.W. (2005) A diffusion approximation for a GI/GI/1 queue with balking or reneging. *Queueing System: theory and Applications (QUESTA)*, 50(4), 371-400. 2.4

[61] Whitt, W. (2004) Efficiency-Driven Heavy-Traffic Approximations for Many-Server Queues with Abandonments. *Management Science*, 50(10), 1449–1461. 3.2

[62] Ward A.R. (2011) Asymptotic Analysis of Queueing Systems with Reneging: A Survey of Results for FIFO, Single Class Models. *Surveys in Operations Research and Management Science*, 16(1), 1–14. 2.4

[63] Yuviler N. (2011) Modeling and Designing an IVR via Phase Type Distributions. M.Sc. Research Proposal, Technion. Downloadable at `http://iew3.technion.ac.il/serveng/References`. 3.4

[64] Zeltyn S. and Mandelbaum A. (2005) Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue. *Queueing Systems: Theory and Applications (QUESTA)*, 51, 361–402. 2.4, 3.1.1, 4.1

[65] Zhang B., van Leeuwaarden L.S.H. and B. Zwart. (2012) Staffing call centers with impatient customers: Refinements to many-server asymptotics. *Operations Research*, 60(2), 461-474. 2.4

[66] Zohar E., Mandelbaum A. and Shimkin N. (2002) Adaptive behavior of impatient customers in tele-queues: theory and empirical support. *Management Science*, 48(4), 566–583. Appendix available through the Extended version in `http://webee.technion.ac.il/people/shimkin/publications.html`. 2.2, 3.3.1, 4.4

A Data Repositories and EDA Tools at the SEELab

SEELab is a research laboratory that opened in 2007 at the Technion [53]. (SEE stands for "Service Enterprise Engineering" [53].) At its beginning, SEELab focused on cleaning, archiving and analyzing transaction-level call center data (event logs). Later, the data framework was extended to additional types of service systems, such as health care, internet and face-to-face services.

Currently, SEELab databases include call-by-call data from four large call centers: a U.S. bank, two Israeli banks and an Israeli mobile-phone company. Three of the four databases cover periods of 2-3 years; the fourth one, which is presently the most active, has about 1.5 years data-worth.

The EDA environment of the SEELab is SEEStat - a software suit that enables real-time statistical analysis of service data at second-to-month resolutions. SEEStat implements many statistical algorithms: parametric distribution fitting and selection, fitting of distribution mixtures, survival analysis (mainly for estimating customers' impatience), and more - all these algorithms interact seamlessly with all the databases. SEEStat interacts with SEEGraph, a pilot-environment for creating and displaying date-based (queueing) networks.

Two call center data-bases are publicly accessible, via the internet (at the SEELab server): from a small Israeli bank and a relatively large U.S. bank. The latter covers the operational history of close to 220 million calls; out of these, about 40 million were served by (up to 1000) agents and the rest by a VRU (answering machine).

SEEStat Online: The connection protocol to SEELab data, for any research or teaching purpose, is simply as follows: go to the SEELab webpage
`http://ie.technion.ac.il/Labs/Serveng`;
then, either via the link **SEEStat Online**, or directly through
`http://seeserver.iem.technion.ac.il/see-terminal`, and complete the registration procedure. Within a day or so, you will receive a confirmation of your registration, plus a password that allows you access to SEEStat, SEELab's EDA environment, and via SEEStat to the above mentioned databases. Note that your confirmation email includes two attachments: a trouble-shooting document and a self-taught tutorial. We propose that you print out the tutorial, connect to SEEStat and then let the tutorial guide you, hands-on, through

SEEStat basics - this should take no more than 1 hour.

B Sources Description for Some Figures in the Paper

Some of our data sets are in the public domain and more will become available in the future. It is thus feasible, and of utmost significance, to provide references that enable other researchers to reproduce and validate the results of the present paper. Indeed, it is our strong belief that empirically-based research, specifically in Operations Research and Operations Management, should overcome the challenge of proprietary data, and strive for reproducibility, the latter being a well-established principle in traditional sciences. Accordingly, our data is publicly available [53], and Table 2 provides the sources of the figures that were created via SEEStat. (The names of the data sets coincide with their names in SEEStat. Detailed instructions for reproducing our figures are provided in Nadjharov et al. [46].) Quoting the principle guiding [16]: "When we publish articles containing figures which were generated by computer, we also publish the complete software environment which generates the figures."

Table 2: **Source data for Figures produced via SEEStat**

Figure	Data set	Time period	Customer type	Performance measures
Figure 1	AnonymousBank	Nov-Dec 1999 Weekdays	PS (Retail)	Patience and offered wait Estimate of hazard rate
Figure 2	USBank	Jan-Oct 2003 Weekdays	Retail	Patience and offered wait Estimate of hazard rate
Figure 4	ILBank	Nov 24, 2008	Private	Waiting time Histograms for full and truncated (10 sec) distributions
Figure 5	ILBank	Nov 24, 2008	Private	Waiting time Distributions fit; data truncated at 10 sec
Figure 6	ILBank	Nov 24, 2008	Private	Waiting time Tail distribution fit for generalized Pareto
Figure 7	ILBank	Nov 24, 2008	Private	Waiting time Fitting mixtures by proportions grid algorithm
Figure 9	USBank	Tue and Wed on Oct 2001	Telesales	Arrivals to queue
Figure 10	USBank	Tue and Wed on Oct 2001	Telesales	Agents on line
Figure 11	USBank	Oct 9-10, 2001	Telesales	Unhandled proportion, Average wait time
Figure 12	USBank	Dec 2002, Weekdays	Quick&Reilly	Wait time (handled, unhandled)
Figure 13	USBank	Dec 2002, Weekdays	Quick&Reilly	Wait time (handled, unhandled) survival curve estimate; polynomial smoothing spline