Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective

Lawrence BROWN, Noah GANS, Avishai MANDELBAUM, Anat SAKOV, Haipeng SHEN, Sergey ZELTYN, and Linda ZHAO

A call center is a service network in which agents provide telephone-based services. Customers who seek these services are delayed in tele-queues. This article summarizes an analysis of a unique record of call center operations. The data comprise a complete operational history of a small banking call center, call by call, over a full year. Taking the perspective of queueing theory, we decompose the service process into three fundamental components: arrivals, customer patience, and service durations. Each component involves different basic mathematical structures and requires a different style of statistical analysis. Some of the key empirical results are sketched, along with descriptions of the varied techniques required. Several statistical techniques are developed for analysis of the basic components. One of these techniques is a test that a point process is a Poisson process. Another involves estimation of the mean function in a nonparametric regression with lognormal errors. A new graphical technique is introduced for nonparametric hazard rate estimation with censored data. Models are developed and implemented for forecasting of Poisson arrival rates. Finally, the article surveys how the characteristics deduced from the statistical analyses form the building blocks for theoretically interesting and practically useful mathematical models for call center operations.

KEY WORDS: Abandonment; Arrivals; Call center; Censored data; Erlang-A; Erlang-C; Human patience; Inhomogeneous Poisson process; Khintchine–Pollaczek formula; Lognormal distribution; Multiserver queue; Prediction of Poisson rates; Queueing science; Queueing theory; Service time.

1. INTRODUCTION

Telephone call centers are technology-intensive operations. Nevertheless, often 70% or more of their operating costs are devoted to human resources. Well-run call centers adhere to a sharply defined balance between agent efficiency and service quality; to do so, they use queueing-theoretic models. Inputs to these mathematical models are statistics concerning system primitives, such as the number of agents working, the rate at which calls arrive, the time required for a customer to be served, and the length of time customers are willing to wait on hold before they hang up the phone and abandon the queue. Outputs are performance measures, such as the distribution of time that customers wait "on hold" and the fraction of customers that abandon the queue before being served. In practice, the number of agents working becomes a control parameter, which can be increased or decreased to attain the desired efficiency–quality trade-off.

Estimates of these primitives are needed to calibrate queueing models, and in many cases the models make distributional assumptions concerning the primitives. In theory, the data required to validate and properly tune these models should be readily available, because computers track and control the minutest details of every call's progress through the system. It is thus surprising that operational data, collected at an appropriate level of detail, have been scarce. The data that are typically collected and used in the call center industry are simple *averages* calculated for the calls that arrive within fixed intervals of time, often 30 minutes. There is a lack of documented, comprehensive, empirical research on call center performance that uses more detailed data.

The immediate goal of our study is to fill this gap. In this article, we summarize a comprehensive analysis of operational data from a bank call center. The data span all 12 months of 1999 and are collected at the level of individual calls. Our data source consists of more than 1,200,000 calls that arrived at the center over the year. Of these, about 750,000 calls terminated in an interactive voice response unit (IVR or VRU), a type of answering machine that allows customers to serve themselves. The remaining 450,000 callers asked to be served by an agent; we have a record of the event-history of each of these calls.

This article is an important part of a larger effort to use both theoretical and empirical tools to better characterize call center operations and performance. It is an abridged version of the work of Brown et al. (2002a), which provided a more complete treatment of the results reported here. Mandelbaum, Sakov, and Zeltyn (2000) presented a comprehensive description of our call-by-call database. Gans, Koole, and Mandelbaum (2003) reviewed queueing and related models of call centers, and Mandelbaum (2001) provided an extensive bibliography.

1.1 Queueing Models of Call Centers

The simplest and most widely used queueing model in call centers is the so-called M/M/N system, sometimes referred to as Erlang-C (Erlang 1911, 1917). The M/M/N model is quite restrictive. It assumes, among other things, a steady-state environment in which arrivals conform to a Poisson process, service durations are exponentially distributed, and customers and servers are statistically identical and act independently of each other. It does not acknowledge, among other things, customer

Lawrence Brown is Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *lbrown@wharton.upenn.edu*). Noah Gans is Associate Professor, Department of Operations and Information Management, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *gans@wharton.upenn.edu*). Avishai Mandelbaum is Professor, Faculty of Industrial Engineering and Management, Technion, Haifa, Israel (E-mail: *avim@tx.technion.ac.il*). Anat Sakov is Postdoctoral Fellow, Tel-Aviv University, Tel-Aviv, Israel (E-mail: *sakov@post.tau.ac.il*). Haipeng Shen is Assistant Professor, Department of Statistics, University of North Carolina, Durham, NC 27599 (E-mail: *haipeng@email.unc.edu*). Sergey Zeltyn is Ph.D. Candidate, Faculty of Industrial Engineering and Management, Technion, Haifa, Israel (E-mail: *zeltyn@ie.technion.ac.il*). Linda Zhao is Associate Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *lzhao@wharton.upenn.edu*). This work was supported by National Science Foundation DMS-99-71751 and DMS-99-71848, the Sloane Foundation, Israeli Science Foundation grants 388/99 and 126/02, the Wharton Financial Institutions Center, and Technion funds for the promotion of research and sponsored research.

Journal of the American Statistical Association
March 2005, Vol. 100, No. 469, Applications and Case Studies
DOI 10.1198/016214504000001808

impatience and abandonment behavior, time-dependent parameters, customers' heterogeneity, or servers' skill levels. An essential task of contemporary queueing theorists is to develop models that account for these effects.

Queueing science seeks to determine which of these effects is most important for modeling real-life situations. For example, Garnett, Mandelbaum, and Reiman (2002) developed both exact and approximate expressions for $M/M/N+M$ (also called Erlang-A) systems, which explicitly model customer patience (time to abandonment) as being exponentially distributed. Empirical analysis can help us judge how well the Erlang-C and Erlang-A models predict customer delays, whether or not their underlying assumptions are met.

1.2 Structure of the Article

The article is structured as follows. Section 2 describes the call center under study and its database. Each of Sections 3–5 is dedicated to the statistical analysis of one of the stochastic primitives of the queueing system: Section 3 addresses call arrivals; Section 4, service durations; and Section 5, tele-queueing and customer patience. Section 5 also analyzes customer waiting times, a performance measure deeply intertwined with the abandonment primitive.

A synthesis of the primitive building blocks is typically needed for operational understanding. Toward this end, Section 6 discusses prediction of the arriving "workload," which is essential in practice for setting suitable service staffing levels.

Once each of the primitives has been analyzed, one can also attempt to use existing queueing theory, or modifications thereof, to describe certain features of the holistic behavior of the system. Section 7 concludes with analyses of this type. We validate some classical theoretical results from queueing theory and refute others.

Finally, we note that many statistical tests are considered throughout the article, which raises the problem of multiplicity (Benjamini and Hochberg 1995). When data from call centers are analyzed in support of operational decisions, the multiplicity problem must be addressed.

2. THE CALL CENTER OF BANK ANONYMOUS

The source of our data (Call Center Data 2002) is a small call center for one of Israel's banks. This center provides several types of basic services, as well as others, including stock trading and technical support, for users of the bank's Internet site. On weekdays (Sunday–Thursday in Israel) the center is open from 7 AM to midnight. During working hours, at most 13 regular agents, 5 Internet agents, and 1 shift supervisor may be working.

A simplified description of the path that each call follows through the center is as follows. A customer calls one of several telephone numbers associated with the call center, with the number depending on the type of service sought. Except for rare busy signals, the customer is then connected to a VRU and identifies herself. While using the VRU, the customer receives recorded information, both general and customized (e.g., an account balance). It is also possible for the customer to perform some self-service transactions here, and 65% of the bank's customers actually complete their service via the VRU. The other 35% indicate the need to speak with an agent. If an agent is free who is capable of performing the desired service, then the customer and the agent are matched to start service immediately. Otherwise, the customer joins the tele-queue.

Customers in the tele-queue are nominally served on a first-come, first-served (FCFS) basis, and customers' positions in queue are distinguished by the times when they arrive. In practice, the call center operates a system with two priorities—high and low—and moves high-priority customers up in queue by subtracting 1.5 minutes from their actual arrival times. Mandelbaum et al. (2000) compared the behavior of the two priority groups of customers.

While waiting, each customer periodically receives information on his or her progress in the queue. More specifically, he or she is told the amount of time that the first person in queue has been waiting, as well as his or her approximate location in the queue. The announcement is replayed every 60 seconds or so, with music, news, or commercials intertwined.

In each of the 12 months of 1999, roughly 100,000–120,000 calls arrived to the system, with 65,000–85,000 of these terminating in the VRU. The remaining 30,000–40,000 calls per month involved callers who exited the VRU indicating a desire to speak to an agent. These calls are the focus of our study. About 80% of those requesting service were in fact served, and about 20% were abandoned before being served.

Each call that proceeds past the VRU can be thought of as passing through up to three stages, each of which generates distinct data. The first of these is the *arrival* stage, which is triggered by the call's exit from the VRU and generates a record of an *arrival time*. If no appropriate server is available, then the call enters the *queueing* stage. Three pieces of data are recorded for each call that queues: the time it entered the queue, the time it exited the queue, and the manner in which it exited the queue, by being served or abandoning. In the last stage, *service*, the data recorded are the starting and ending times of the service. Note that calls that are served immediately skip the queueing stage, and calls that are abandoned never enter the service stage.

In addition to these time stamps, each call record in our database includes a categorical description of the type of service requested. The main call types are regular (PS in the database), stock transaction (NE), new/potential customer (NW), and Internet assistance (IN). Mandelbaum et al. (2000) described the process of collecting and cleaning the data and provided additional descriptive analysis of the data.

Over the year, two important operational changes occurred. First, in January–July, all calls were served by the same group of agents, but beginning in August, Internet (IN) customers were served by a separate pool of agents. Thus, in August–December, the center can be considered to be two separate service systems, one for IN customers and another for all other types. Second, as we discuss in Section 5, one aspect of the service time data changed at the end of October. In several instances, this article's analyses are based on only the November and December data. In other instances we have used data from August–December. Given the changes noted earlier, this ensures consistency throughout the manuscript. November and December were also convenient, because they contained no Israeli holidays. In these analyses, we also restrict the data to include only regular weekdays—Sunday–Thursday, 7 AM–midnight—because these are the hours of full operation of the center. We have performed similar analyses for other parts of the data, and in most respects the November–December results do not differ noticeably from those based on data from other months of the year.

3. THE ARRIVAL PROCESS

Figure 1 shows, as a function of time of day, the average rate per hour at which calls come out of the VRU. These are composite plots for weekday calls in November and December. The plots show calls according to the major call types. The volume of regular (PS) calls is much greater than that of the other three types; hence those calls are shown on a separate plot. [These plots were fit using the root–unroot method described by Brown, Zhang, and Zhao (2001), along with the adaptive free knot spline methodology of Mao and Zhao (2003). For a more precise study of these arrival rates, including confidence and prediction intervals, see our Sec. 6 and also Brown et al. 2001, 2002a,b.]

Note the bimodal pattern of PS call-arrival times in Figure 1. It is especially interesting that IN calls do not show a similar bimodal pattern and in fact have a peak volume after 10 PM. (This peak can be partially explained by the fact that Internet customers are sensitive to telephone rates, which significantly decrease in Israel after 10 PM, and that they also tend to be people who stay late.)

3.1 Arrivals Are Inhomogeneous Poisson

Common call center models and practice assume that the arrival process is Poisson with a rate that remains constant for blocks of time (e.g., half-hours), with a separate queueing model fitted for each block of time. A more natural model for capturing changes in the arrival rate is a time-inhomogeneous Poisson process. Following common practice, we assume that the arrival rate function can be well approximated as being piecewise constant.

We now construct a test of the null hypothesis that arrivals of given types of calls form an inhomogeneous Poisson process with piecewise constant rates. The first step in constructing our test involves breaking up the duration of a day into relatively short blocks of time, short enough so that the arrival rate does not change significantly within a block. For convenience, we used blocks of equal time length, L, although this equality assumption could be relaxed. One can then consider the arrivals within a subset of blocks—for example, blocks at the same time on various days or successive blocks on a given day. The former case would, for example, test whether the process is homogeneous within blocks for calls arriving within the given time span.

Let T_{ij} denote the jth ordered arrival time in the ith block, $i = 1, \ldots, I$. Thus $T_{i1} \le \cdots \le T_{iJ(i)}$, where $J(i)$ denotes the total number of arrivals in the ith block. Then define $T_{i0} = 0$ and

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

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

Let U_{ij} denote the jth (unordered) arrival time in the ith block. Then the assumed constant Poisson arrival rate within this block implies that, conditionally on $J(i)$, the unordered arrival times are independent and uniformly distributed, that is, $U_{ij} \overset{\text{iid}}{\sim} \mathrm{U}(0, L)$. Note that $T_{ij} = U_{i(j)}$. It follows that $\frac{L - T_{ij}}{L - T_{i,j-1}}$ are independent beta$(J(i) + 1 - j, 1)$ variables [see, e.g., problem 6.14.33(iii) in Lehmann 1986]. A standard change of variables then yields the conditional exponentiality of the R_{ij} given the value of $J(i)$. [One may alternatively base the test on the variables $R^*_{ij} = j(-\log \frac{T_{ij}}{T_{i,j+1}})$, where $j = 1, \ldots, J(i)$ and $T_{i,J(i)+1} = L$. Under the null hypothesis, these will also be independent standard exponential variables.]

The null hypothesis does not involve an assumption that the arrival rates of different intervals are equal or have any other prespecified relationship. Any customary test for the exponential distribution can be applied to test the null hypothesis. For convenience, we use the familiar Kolmogorov–Smirnov test, even though this may not have the greatest possible power against the alternatives of most interest. In addition, exponential Q–Q plots can be very useful in ascertaining goodness of fit to the exponential distribution.

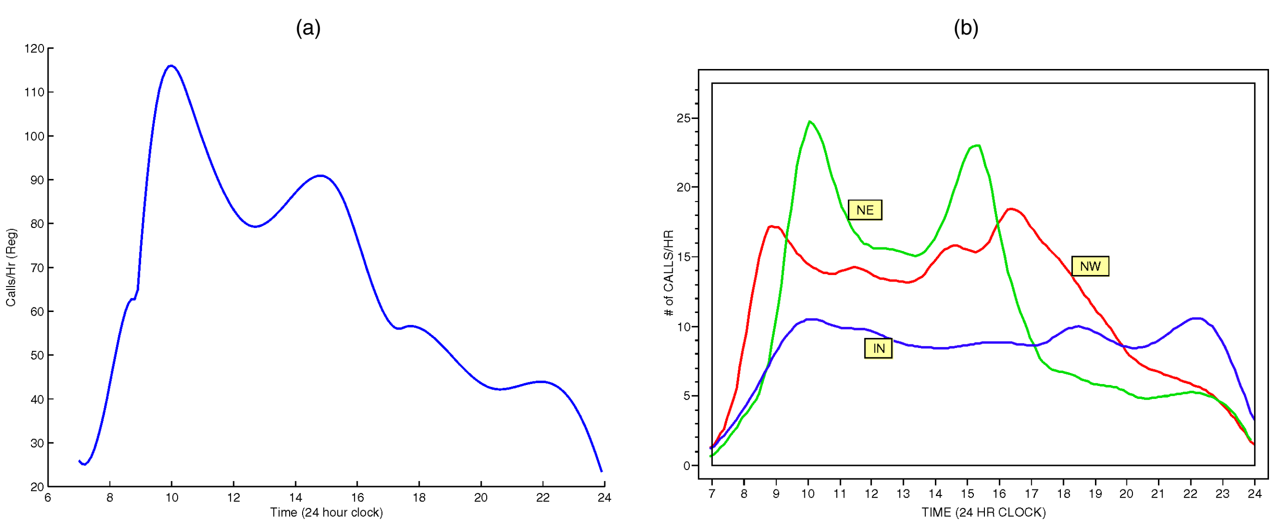

Figure 1. Arrivals in Calls/Hour by Time of Day, Weekdays in November–December. (a) PS calls; (b) IN, NW, and NE calls.

Brown et al. (2002a) presented quantile plots for a few applications of this test. For the PS data, we found it convenient to use $L = 6$ minutes. For the other types, we use $L = 60$ minutes, because these calls involved much lower arrival rates.

We omit the plots here to save space and because they demonstrate only minor deviations from the ideal straight-line pattern. One example involves arrival times of the PS calls arriving between 11:12 AM and 11:18 AM on all weekdays in November and December. A second example involves arrival of IN calls on Monday, November 23, from 7 AM to midnight. This was a typical midweek day in our dataset.

For both of the examples, the null hypothesis is not rejected, and we conclude that their data are consistent with the assumption of an inhomogeneous Poisson process for the arrival of calls. The respective Kolmogorov–Smirnov statistics have values $K = .0316$ (p value $\approx .8$ with $n = 420$) and $K = .0423$ (p value $\approx .9$ with $n = 172$). These results are typical of those that we obtained from various selections of blocks of the various types of calls involving comparable sample sizes. Thus overall, from tests of this nature, there is no evidence in this dataset to reject a null hypothesis that the arrival of calls from the VRU is an inhomogeneous Poisson process.

As an attempt to further validate the inhomogeneous Poisson character, we applied this method to the 48,193 PS calls in November and December in 6-minute blocks. With this large amount of data, one could expect to detect more than statistically negligible departures from the null hypothesis because of rounding of times in the data (to the nearest second) and because arrival rates are not exactly constant within 6-minute time spans. To compensate for the rounding, we "unrounded" the data before applying the test by adding independent uniform $(0, 1)$ noise to each observation. (This unrounding did noticeably improve the fit to the ideal pattern.) After the unrounding, the resulting Kolmogorov–Smirnov statistic was $K = .009$. This is a very small deviation from the ideal; nevertheless, the p value for this statistic with such a large $n = 48{,}963$ is $p \approx .00007$. [To provide an additional benchmark for evaluating the (lack of) importance of this value, we note that this same statistic with $n \approx 22{,}000$ would have had p value $\approx .05$, which is just acceptable.]

4. SERVICE TIME

The goal of a visit to the call center is the service itself. Table 1 summarizes the mean, standard deviation (SD), and median service times for the four types of service of main interest. The very few calls with service times >1 hour were not considered (i.e., we treat them as outliers). IN calls has little effect on the numbers. IN calls have the longest service times, with stock trading (NE) service calls next. Potential customers (NW) have the shortest service time (which is consistent with the nature of these calls). An important implication is that the workload that Internet consultation imposes on the system is greater than its share in terms of percent of calls. In earlier work (Brown et al. 2002a) we also verified that the full cumulative distributions of the service times are stochastically ordered in the same fashion as the means in Table 1.

4.1 Very Short Service Times

Figure 2 shows histograms of the combined service times for all types of service for January–October and for November–December. These plots resemble those for PS calls alone, because the clear majority of calls are for PS. We see that in the first 10 months of the year, the percentage of calls with service <10 seconds was larger than the percentage at the end of the year (7% vs. 2%).

Service times <10 seconds are questionable. And indeed, the manager of the call center discovered that short service times were primarily caused by agents who simply hung up on customers to obtain extra rest time. (The phenomenon of agents "abandoning" customers is not uncommon; it is often due to distorted incentive schemes, especially those that overemphasize short average talk-time or, equivalently, the total number of calls handled by an agent.) The problem was identified, and steps were taken to correct it in October 1999. For this reason, in the later analysis of service times, we focus on data from November and December. Suitable analyses can be constructed for the entire year by using a mixture model or, in a somewhat less sophisticated manner, by deleting from the service time analysis all calls with service times <10 seconds.

4.2 On Service Times and Queueing Theory

Most applications of queueing theory to call centers assume exponentially distributed service times as their default. The main reason for this is the lack of empirical evidence to the contrary, which leads one to favor convenience. Indeed, models with exponential service times are amenable to analysis, especially when combined with the assumption that arrival processes are homogeneous Poisson processes. This is the reason that M/M/N is the prevalent model used in call center practice.

In more general queueing formulas, the service time often affects performance measures through its squared coefficient of variation, $C_s^2 = \sigma_s^2/\mathrm{E}^2(S)$, $\mathrm{E}(S)$ is the average service time and σ_s is its standard deviation. For example, a common useful approximation for the average waiting time in an M/G/N model (Markovian arrivals, generally distributed service times, n servers), is given by

$$\text{E[Wait for M/G/N]} \approx \text{E[Wait for M/M/N]} \times \frac{(1 + C_s^2)}{2} \qquad (1)$$

(see Sze 1984; Whitt 1993). Note that for large call centers, this formula must be used with care, as discussed by Mandelbaum

Table 1. Service Time by Type of Service, Truncated at 1 Hour, November–December

	Overall	Regular service (PS)	Potential customers (NW)	Internet consulting (IN)	Stock trading (NE)
Mean	201	179	115	401	270
SD	248	189	146	473	303
Median	124	121	73	221	175

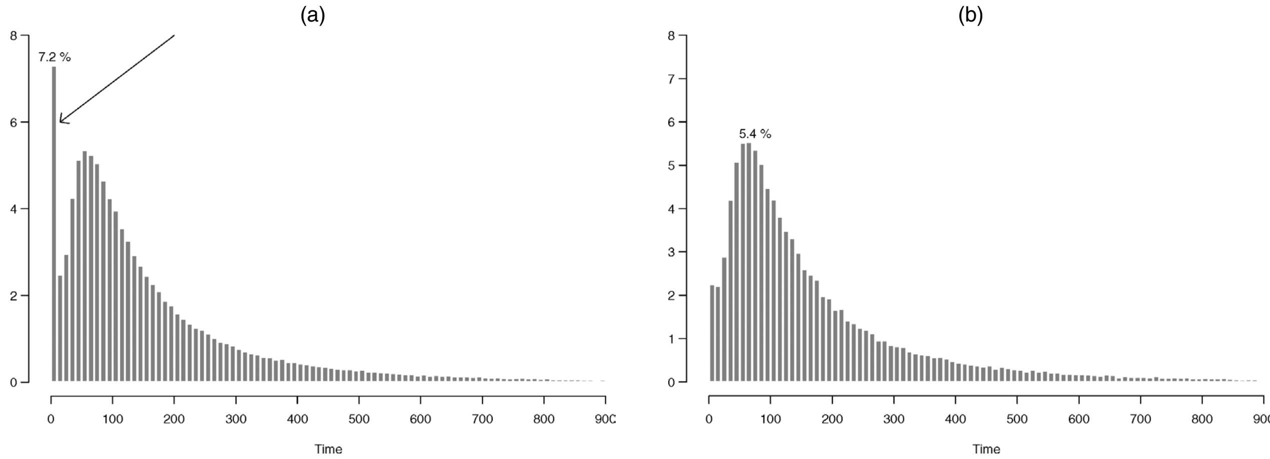

Figure 2. Distribution of Service Time. (a) January–October (mean, 185; SD, 238); (b) November–December (mean, 200; SD, 249).

and Schwartz (2002). Thus average wait with general service times is multiplied by a factor of $(1+C_s^2)/2$ relative to the wait under exponential service times. For example, if service times are in fact exponential, then the factor is 1. Deterministic service times halve the average wait of exponential. In our data, the observed factor is $(1+C_s^2)/2 = 1.26$.

4.3 Service Times Are Lognormal

Looking at Figure 2, we see that the distribution of service times is clearly not exponential, as is assumed by standard queueing theory. In fact, after separating the calls with very short service times, our analysis reveals a remarkable fit to the *lognormal* distribution.

Figure 3(a) shows the histogram of log(service time) for November and December, in which the short service phenomenon was absent or minimal. Superimposed is the best fitted normal density as provided by Brown and Hwang (1993). Figure 3(b) shows the lognormal Q–Q plot of service time. This does an amazingly good imitation of a straight line. Nevertheless, the Kolmogorov–Smirnov test decisively rejects the null hypothesis of exact lognormality. (The Kolmogorov–Smirnov statistic here is $K = .020$. This is quite small, but still much larger than the value of $K = .009$ that was attained for a similarly large sample size in the inhomogeneous Poisson test of Sec. 4.) We only provide the graphs to qualitatively support our claim of lognormality. Thus the true distribution is very close to lognormal, but is not exactly lognormal. (The most evident deviation is in the left tail of the histogram, where both a small excess of observations is evident and the effect of rounding to the nearest second further interferes with a perfect fit.) This is a situation where a very large sample size yields a statistically significant result, even though there is no "practical significance."

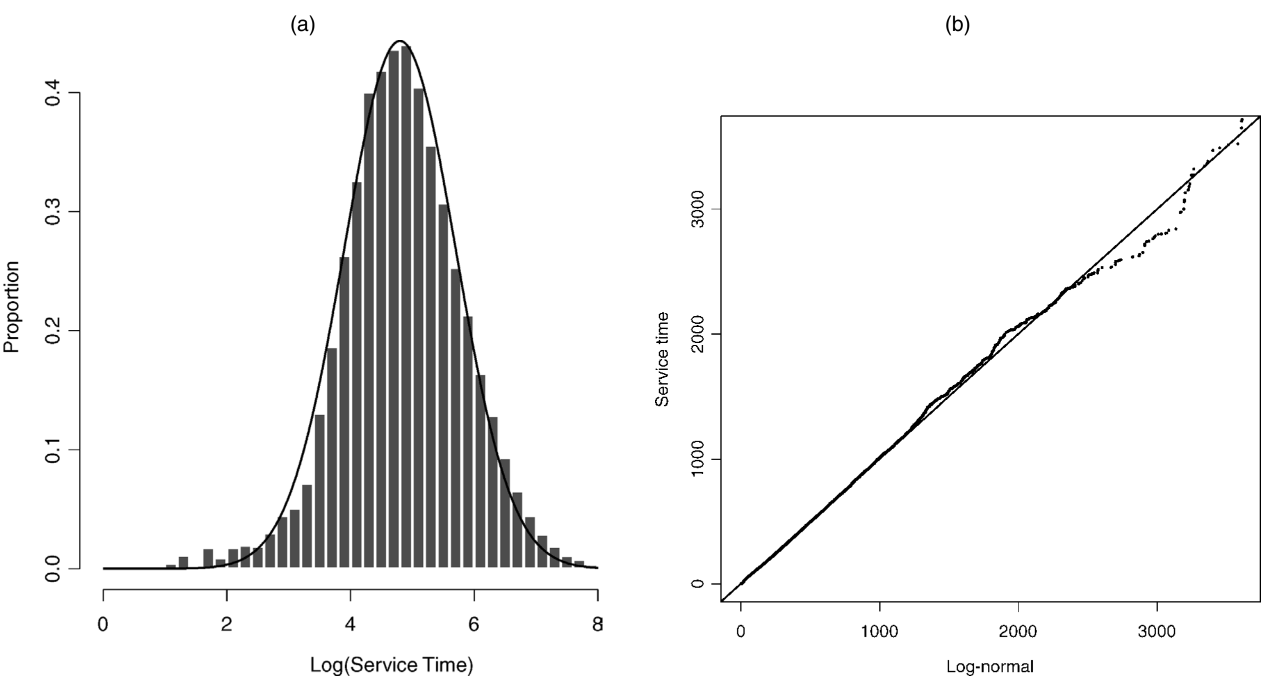

Figure 3. Histogram (a) and Q–Q Plot (b) of log(service time), November–December.

After excluding short service times, the strong resemblance to a lognormal distribution also holds for all other months. It also holds for various types of callers, even though the parameters depend on the type of call. This means that in this case, a mixture of lognormals is empirically lognormal, even though mathematically this cannot exactly hold. (See Mandelbaum et al. 2000, where the phenomenon is discussed in the context of the exponential distribution.) Brown and Shen (2002) gave a more detailed analysis of service times.

Lognormality of processing times has been occasionally recognized by researchers in telecommunications and psychology. Bolotin (1994) gave empirical results suggesting that the distribution of the logarithm of call duration is normal for individual telephone customers and a mixture of normals for "subscriber-line" groups. Ulrich and Miller (1993) and Breukelen (1995) provided theoretical arguments for the lognormality of reaction times using models from mathematical psychology. Mandelbaum and Schwartz (2000) used simulations to study the effect of lognormally distributed service times on queueing delays.

4.4 Regression of log(service times) on Time of Day

The important implication of the excellent fit to a lognormal distribution is that we can apply standard techniques to regress log(service time) on various covariates, such as time of day. For example, to model the mean service time across time of day, we can first model the mean and variance of the log(service time) across time of day, then transform the result back to the service time scale. [Shen (2002) gave a detailed analysis of service times against other covariates, such as the identities of individual agents (servers), as well as references to other literature involving lognormal variates.]

Let S be a lognormally distributed random variable with mean ν and variance τ^2. Then $Y = \log(S)$ will be a normal random variable with some mean μ and variance σ^2. It is well known that $\nu = e^{\mu+\sigma^2/2}$. This parameter (rather than μ or $\mu + \sigma^2/2$) is the primitive quantity that appears in calculations of offered load, as in Section 7. To provide a confidence interval for ν, we need to derive confidence intervals for μ and σ^2 or, more precisely, for $\mu + \sigma^2/2$.

For our call center data, let S be the service time of a call and let T be the corresponding time of day at which the call begins service. Let $\{S_i, T_i\}_{i=1}^n$ be a random sample of size n from the joint distribution of $\{S, T\}$ and sorted according to T_i. Then $Y_i = \log(S_i)$ will be the log(service time) of the calls, which are (approximately) normally distributed, conditional on T_i. We can fit a regression model of Y_i on T_i as $Y_i = \mu(T_i) + \sigma(T_i)\epsilon_i$, where $\epsilon_i|T_i$ are iid N(0, 1).

4.4.1 Estimation of $\mu(\cdot)$ and $\sigma^2(\cdot)$. If we assume that $\mu(\cdot)$ has a continuous third derivative, then we can use local quadratic regression to derive an estimate for $\mu(\cdot)$ (see Loader 1999). Suppose that $\hat{\mu}(t_0)$ is a local quadratic estimate for $\mu(t_0)$. Then an approximate $100(1-\alpha)\%$ confidence interval for $\mu(t_0)$ is $\hat{\mu}(t_0) \pm z_{\alpha/2}\mathrm{se}_\mu(t_0)$, where $\mathrm{se}_\mu(t_0)$ is the standard error of the estimate of the mean at t_0 from the local quadratic fit.

Our estimation of the variance function $\sigma^2(\cdot)$ is a two-step procedure. In the first step, we regroup the observations $\{T_i, Y_i\}_{i=1}^n$ into consecutive nonoverlapping pairs $\{T_{2i-1}, Y_{2i-1}; T_{2i}, Y_{2i}\}_{i=1}^{\lfloor n/2 \rfloor}$. The variance at T_{2i}, $\sigma^2(T_{2i})$, is estimated by a squared pseudoresidual, D_{2i}, of the form $(Y_{2i-1} - Y_{2i})^2/2$, a so-called "difference-based" estimate. The difference-based estimator that we use here is a simple one that suffices for our purposes. In particular, our method yields suitable confidence intervals for estimation of σ^2. More efficient estimators might improve our results slightly. There are many other difference-based estimators in the literature (see Müller and Stadtmüller 1987; Hall, Kay, and Titterington 1990; Dette, Munk, and Wagner 1998; Levins 2002).

In the second step, we treat $\{T_{2i}, D_{2i}\}_{i=1}^{\lfloor n/2 \rfloor}$ as our observed data points and apply local quadratic regression to obtain $\hat{\sigma}^2(t_0)$. Part of our justification is that under our model, the $\{D_{2i}\}$'s are (conditionally) independent given the $\{T_{2i}\}$'s. A $100(1-\alpha)\%$ confidence interval for $\sigma^2(t_0)$ is approximately $\hat{\sigma}^2(t_0) \pm z_{\alpha/2}\mathrm{se}_{\sigma^2}(t_0)$.

Note that we use $z_{\alpha/2}$, rather than a quantile from a chi-squared distribution, as the cutoff value when deriving the foregoing confidence interval. Given our large dataset, the degrees of freedom are large, and a chi-squared distribution can be well approximated by a normal distribution.

4.4.2 Estimation of $\nu(\cdot)$. We now use $\hat{\mu}(t_0)$ and $\hat{\sigma}^2(t_0)$ to estimate $\nu(t_0)$, as $e^{\hat{\mu}(t_0)+\hat{\sigma}^2(t_0)/2}$. Given that the estimation methods used for $\mu(t_0)$ and $\sigma^2(t_0)$, $\hat{\mu}(t_0)$ and $\hat{\sigma}^2(t_0)$, are asymptotically independent, we have

$$\mathrm{se}\big(\hat{\mu}(t_0) + \hat{\sigma}^2(t_0)/2\big) \approx \sqrt{\mathrm{se}_\mu(t_0)^2 + \mathrm{se}_{\sigma^2}(t_0)^2/4}.$$

When the sample size is large, we can assume that $\hat{\mu}(\cdot) + \hat{\sigma}^2(\cdot)/2$ has an approximately normal distribution. Then the corresponding $100(1-\alpha)\%$ confidence interval for $\nu(t_0)$ is

$$\exp\Big(\big(\hat{\mu}(t_0) + \hat{\sigma}^2(t_0)/2\big) \pm z_{\alpha/2}\sqrt{\mathrm{se}_\mu(t_0)^2 + \mathrm{se}_{\sigma^2}(t_0)^2/4}\Big).$$

4.4.3 Application and Model Diagnostics. In the analysis that follows, we apply the foregoing procedure to the weekday calls in November and December. The results for two interesting service types are shown in Figure 4. There are 42,613 PS calls and 5,066 IN calls. To produce the figures, we use the tricube function as the kernel and nearest-neighbor type bandwidths. The bandwidths are automatically chosen via cross-validation.

Figure 4(a) shows the mean service time for PS calls as a function of time of day, with 95% confidence bands. Note the prominent bimodal pattern of mean service time across the day for PS calls. The accompanying confidence band shows that this bimodal pattern is highly significant. The pattern resembles that for arrival rates of PS calls (see Fig. 1). This issue was discussed further by Brown et al. (2002a).

Figure 4(b) plots an analogous confidence band for IN calls. One interesting observation is that IN calls do not show a similar bimodal pattern. We do see some fluctuations during the day, but these are only mildly significant, given the wide confidence band. Also notice that the entire confidence band for IN calls lies above that of PS calls. This reflects the stochastic dominance referred to in the discussion of Table 1.

Standard diagnostics on the residuals reveal a qualitatively very satisfactory fit to lognormality, comparable with that in Figure 3.

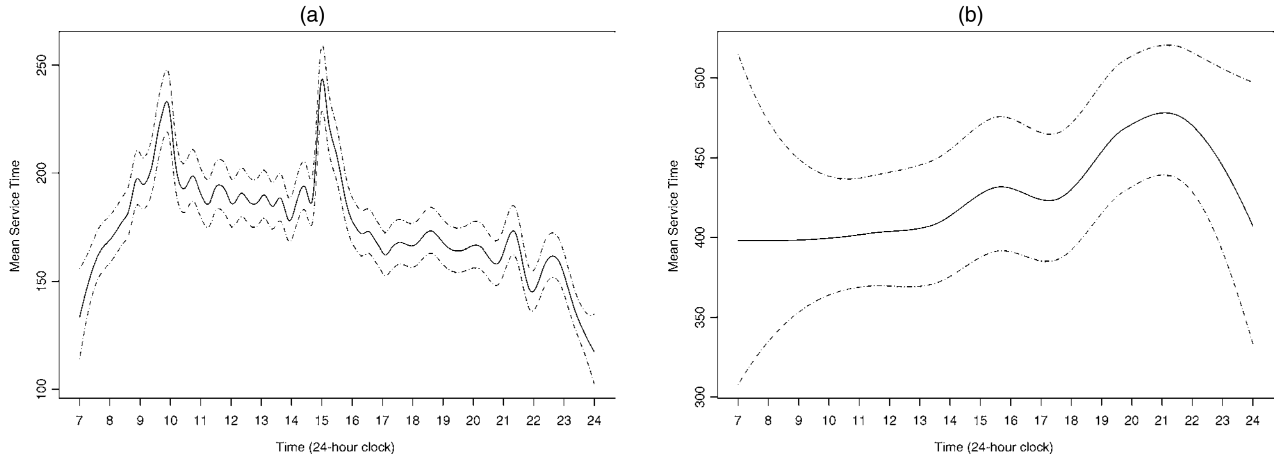

Figure 4. Mean Service Time (PS) (a) versus Time of Day [95% confidence interval (CI)], (b) Mean Service Time (IN) versus Time of Day (95% CI).

5. WAITING FOR SERVICE OR ABANDONING

In Sections 3 and 4 we characterized two primitives of queueing models, the arrival process and service times. In each case we were able to directly observe and analyze the primitive under investigation. We next address the last system primitive, customer patience and abandonment behavior, and the related output of waiting time. Abandonment behavior and waiting times are deeply intertwined.

There is a distinction between the time that a customer *needs* to wait before reaching an agent and the time that a customer is *willing* to wait before abandoning the system. The former is referred to as *virtual waiting time*, because it amounts to the time that a (virtual) customer, equipped with infinite patience, would have waited until being served. We refer to the latter as *patience*. Both measures are obviously of great importance, but neither is directly observable, and hence both must be estimated.

A well-known queueing-theoretic result is that in heavily loaded systems (in which essentially all customers wait and no one abandons), waiting time should be exponentially distributed. (See Kingman 1962 for an early result and Whitt 2002 for a recent text.) Although our system is *not* very heavily loaded, and in our system customers *do* abandon, we find that the observed distribution of time spent in the queue conforms very well to this theoretical prediction (see Brown et al. 2002a for further details).

5.1 Survival Curves for Virtual Waiting Time and Patience

Both times to abandonment and times to service are censored data. Let R denote the "patience" or "time willing to wait" and let V denote the "virtual waiting time," and equip both with steady-state distributions. One actually samples $W = \min\{R, V\}$, as well as the indicator $\mathbb{1}_{\{R<V\}}$, for observing R or V. One considers all calls that reached an agent as censored observations for estimating the distribution of R, and vice versa for estimating the distribution of V. We make the assumption that (as random variables) R and V are independent given the covariates relevant to the individual customer. Under this assumption, the distributions of R and V (given the covariates) can be estimated using the standard Kaplan–Meier product-limit estimator.

One may plot the Kaplan–Meier estimates of the survival functions of R (time willing to wait), V (virtual waiting time), and $W = \min\{V, R\}$ (see Brown et al. 2002a). There is a clear stochastic ordering between V and R in which customers are willing to wait (R) more than they need to wait (V). This suggests that our customer population consists of patient customers. Here we have implicitly, and only intuitively, defined the notion of a *patient customer*. (To the best of our knowledge, systematic research on this subject is lacking.)

We also consider the survival functions of R for different types of service. Again, a clear stochastic ordering emerges. For example, customers performing stock trading (type NE) are willing to wait more than customers calling for regular services (type PS). A possible empirical explanation for this ordering is that type NE needs the service more urgently. This suggests a practical distinction between tolerance for waiting and loyalty/persistency.

5.2 Hazard Rates

Palm (1953) was the first to describe impatience in terms of a hazard rate. He postulated that the hazard rate of the time willing to wait is proportional to a customer's irritation due to waiting. Aalen and Gjessing (2001) advocated dynamic interpretation of the hazard rate, but warned against the possibility that the population hazard rate may not represent individual hazard rates.

We have found it useful to construct nonparametric estimates of the hazard rate. It is feasible to do so because of the large sample size of our data (about 48,000). Figure 5 shows such plots for R and V.

The nonparametric procedure that we use to calculate and plot the figures is as follows. For each interval of length δ, the estimate of the hazard rate is calculated as

$$\frac{[\text{number of events during } (t, t+\delta]]}{[\text{number at risk at } t] \times \delta}.$$

For smaller time values, t, the numbers at risk and event rates are large, and we let $\delta = 1$ second. For larger times, when fewer

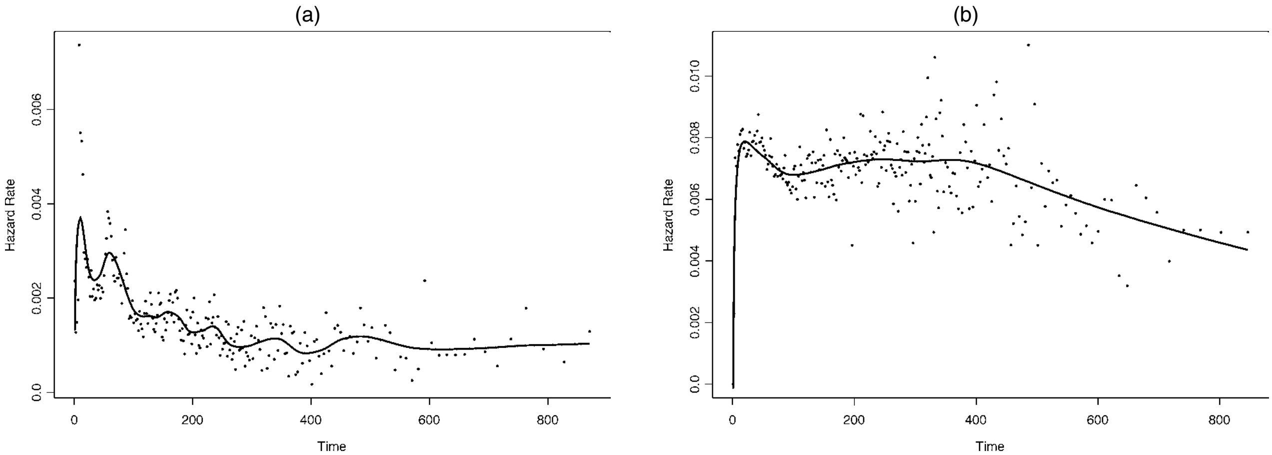

Figure 5. Hazard Rates for (a) the Time Willing to Wait for PS Calls and (b) Virtual Waiting Time, November–December.

are at risk, we use larger δ's. Specifically, the larger intervals are constructed to have an estimated expected number of events per interval of at least four. Finally, the hazard rate for each interval is plotted at the interval's midpoint.

The curves superimposed on the plotted points are fitted using nonparametric regression. In practice we used LOCFIT (Loader 1999), although other techniques, such as kernel procedures or smoothing splines, would yield similar fits. We choose the smoothing bandwidth by generalized cross-validation. (We also smoothly transformed the x-axis, so that the observations would be more nearly uniformly placed along that axis, before producing a fitted curve. We then inversely transformed the x-axis to its original form.) We experimented with fitting techniques that varied the bandwidth to take into account the increased variance and decreased density of the estimates with increasing time. However, with our data, these techniques had little effect, and thus we do not use here.

Figure 5(a) plots the hazard rates of the time willing to wait for PS calls. Note that it shows two main peaks. The first peak occurs after only a few seconds. When customers enter the queue, a "please wait" message, as described in Section 2, is played for the first time. At this point, some customers who do not wish to wait probably realize they are in a queue and hang up. The second peak occurs at about $t = 60$, about the time that the system plays the message again. Apparently, the message increases customers' likelihood of hanging up for a brief time thereafter, an effect that may be contrary to the message's intended purpose (or maybe not).

In Figure 5(b), the hazard rate for the virtual waiting times is estimated for all calls. (The picture for PS alone is very similar.) The overall plot reveals rather constant behavior and indicates a moderate fit to an exponential distribution. (The gradual general decrease in this hazard rate, from about .008 to .005, suggests an issue that may merit further investigation.)

5.3 Patience Index

Customer patience on the telephone is important, yet it has not been extensively studied. In the search for a better understanding of patience, we have found a *relative* definition to be of use. Let the means of V and R be m_V and m_R. One can define the *patience index* as the ratio m_R/m_V, the ratio of the mean time a customer is willing to wait to the mean time he or she needs to wait. The justification for calling this a "patience index" is that for experienced customers, the time that one needs to wait is in fact that time that one *expects* to wait. Although this patience index makes sense intuitively, its calculation requires the application of survival analysis techniques to call-by-call data. Such data may not be available in certain circumstances. Therefore, we wish to find an *empirical index* that will work as an auxiliary measure for the patience index.

For the sake of discussion, we assume that V and R are independent and exponentially distributed. As a consequence of these assumptions, we can demonstrate that

$$\text{Patience index} \triangleq \frac{m_R}{m_V} = \frac{\mathrm{P}(V < R)}{\mathrm{P}(R < V)}.$$

Furthermore, $\mathrm{P}(V < R)/\mathrm{P}(R < V)$ can be estimated by (number served)/(number abandoned), and we define

$$\text{Empirical index} \triangleq \frac{\text{number served}}{\text{number abandoned}}.$$

The numbers of both served and abandoned calls are very easy to obtain from either call-by-call data or more aggregated call center management reports. We have thus derived an easy-to-calculate empirical measure from a probabilistic perspective. The same measure can also be derived using the maximum likelihood estimators for the mean of the (right-censored) exponential distribution, applied separately to R and to V.

We can use our data to validate the empirical index as an estimate of the theoretical patience index. Recall, however, that the Kaplan–Meier estimate of the mean is biased when the last observation is censored or when heavy censoring is present. Nevertheless, a well-known property of exponential distributions is that their quantiles are just the mean multiplied by certain constants, and we use quantiles when calculating the patience index. In fact, because of heavy censoring, we sometimes do not obtain an estimate for the median or higher quantiles. Therefore, we used first quartiles when calculating the theoretical patience index.

The empirical index turns out to be a very good estimate of the theoretical patience index. For each of 68 quarter hours between 7 AM and midnight, we calculated the first quartiles of V and R from the survival curve estimates. We then compared the ratio of the first quartiles to that of (number of served) to (number of abandoned). The resulting 68 sample pairs had an R^2 of .94 (see Brown et al. 2002a for a plot). This result suggests that we can use the empirical measure as an index for human patience.

With this in mind, we obtain the following empirical indices for regular weekdays in November and December: PS $= 5.34$, NE $= 8.71$, NW $= 1.61$, and IN $= 3.74$. We thus find that the NE customers are the most patient, perhaps because their business is the most important to them. On the other hand, by this measure the IN customers are less patient than the PS customers. In this context, we emphasize that the patience index measures time willing to wait normalized by time needed to wait. In our case (as previously noted), the IN customers are in a separate queue from that of the PS customers. The IN customers on average are willing to wait slightly longer than the PS customers (see Brown et al. 2002a). However, they also need to wait longer, and overall their patience index is less than that of the PS customers.

Recall that the linear relationship between the two indices is established under the assumption that R and V are exponentially distributed and independent. As Figure 5(a) shows, however, the distribution for R is clearly not exponential. Similarly, Figure 5(b) shows that V also displays some deviation from exponentiality. Furthermore, sequential samples of V are not independent of each other. Thus we find that the linear relation is surprisingly strong.

Finally, we note another peculiar observation: The line does not have an intercept at 0 or a slope of 1, as suggested by the foregoing theory. Rather, the estimated intercept and slope are -1.82 and 1.35, which are statistically different from 0 and 1. We are working on providing a theoretical explanation that accounts for these peculiar facts, as well as an explanation for the fact that the linear relationship holds so well, even though the assumption of exponentiality does not hold for our data. (The assumption of independence of R and V may also be questionable.)

6. PREDICTION OF THE LOAD

This section reflects the view of the operations manager of a call center who plans and controls daily and hourly staffing levels. Prediction of the system "load" is a key ingredient in this planning. Statistically, this prediction is based on a combination of the observed arrival times to the system (as analyzed in Sec. 3) and service times during previous, comparable periods (as analyzed in Sec. 4).

In the discussion that follows we describe a convenient model and a corresponding method of analysis that can be used to generate prediction confidence bounds for the load of the system. More specifically, in Section 6.4 we present a model for predicting the arrival rate, and in Section 6.6 we present a model for predicting mean service time. In Section 6.7 we combine the two predictions to obtain a prediction (with confidence bounds) for the load according to the method discussed in Section 6.3.

6.1 Definition of Load

In Section 3 we showed that arrivals follow an inhomogeneous Poisson process. We let $\Lambda_j(t)$ denote the true arrival rate of this process at time t on a day indexed by the subscript j. Figure 1 presents a summary estimate of $\bar{\Lambda}.(t)$, the average of $\Lambda_j(t)$ over weekdays in November and December.

For simplicity of presentation, here we treat together all calls except the IN calls, because these were served in a separate system in August–December. The arrival patterns for the other types of calls appear to be reasonably stable in August–December. Therefore, in this section we use the August–December data to fit the arrival parameters. To avoid having to adjust for the short service time phenomenon noted in Section 4.1, we use only November and December data to fit parameters for service times. Also, here we consider only regular weekdays (Sunday–Thursday) that were not full or partial holidays.

Together, an arbitrary arrival rate $\Lambda(t)$ and mean service time $\nu(t)$ at t define the "load" at that time, $L(t) = \Lambda(t)\nu(t)$. This is the expected time units of work arriving per unit of time, a primitive quantity in building classical queueing models, such as those discussed in Section 7.

Briefly, suppose that one adopts the simplest M/M/N queueing model. Then, if the load is a constant, L, over a sufficiently long period, the call center must be staffed, according to the model, with at least L agents; otherwise, the model predicts that the backlog of calls waiting to be served will explode in an ever-increasing queue. Typically, a manager will need to staff the center at a staffing level that is some function of L—for example, $L + c\sqrt{L}$ for some constant c—to maintain satisfactory performance (see Borst, Mandelbaum, and Reiman 2004; Garnett et al. 2002).

6.2 Independence of $\Lambda(t)$ and $\nu(t)$

In Section 5.4.4 we noted a qualitative similarity in the bimodal pattern of arrival rates and mean service times. To try to explain this similarity, we tested several potential explanations, including a causal dependence between arrival rate and service times. We were led to the conclusion that such a causal dependence is not a statistically plausible explanation. Rather, we concluded that the periods of heavier volume involve a different mix of customers, a mix that includes a higher population of customers who require lengthier service. The statistical evidence for this conclusion is indirect and was reported by Brown et al. (2002a). Thus we proceed under the assumption that arrival rates and mean service times are conditionally independent given the time of day.

6.3 Coefficient of Variation for the Prediction of $L(t)$

Here we discuss the derivation of approximate confidence intervals for $\Lambda(t)$ and $\nu(t)$ based on observations of quarter-hour groupings of the data. The load, $L(t)$, is a product of these two quantities. Hence exact confidence bounds are not readily available from individual bounds for each of $\Lambda(t)$ and $\nu(t)$. As an additional complication, the distributions of the individual estimates of these quantities are not normally distributed. Nevertheless, one can derive reasonable approximate confidence bounds from the coefficient of variation (CV) for the estimate of L.

For any nonnegative random variable W with finite positive mean and variance, define the CV (as usual) by $CV(W) =$

$SD(W)/E(W)$. If U and V are two independent variables and $W = UV$, then an elementary calculation yields

$$CV(W) = \sqrt{CV^2(U) + CV^2(V) + CV^2(U) \cdot CV^2(V)}.$$

In our case, U and V correspond to Λ and ν. Predictions for Λ and ν are discussed in Sections 7.4 and 7.6. As noted earlier, these predictions can be assumed to be statistically independent. Also, their CVs are quite small (<.1). Note that $\hat{L}(t) = \hat{\Lambda}(t)\hat{\nu}(t)$, and, using standard asymptotic normal theory, we can approximate $CV(\hat{L})(t)$ as $CV(\hat{L})(t) \approx \sqrt{CV^2(\hat{\Lambda})(t) + CV^2(\hat{\nu})(t)}$.

This leads to approximate 95% CIs of the form $\hat{L}(t) \pm 2\hat{L}(t)CV(\hat{L})(t)$. The constant 2 is based on a standard asymptotic normal approximation of roughly 1.96.

6.4 Prediction of $\Lambda(t)$

Brown and Zhao (2001) investigated the possibility of modeling the parameter Λ as a deterministic function of time of day, day of week, and type of customer, and rejected such a model. Here we construct a random-effects model that can be used to predict Λ and to construct confidence bands for that prediction. The model that we construct includes an autoregressive feature that incorporates the previous day's volume into the prediction of today's rate.

In the model, which we elaborate on later, we predict the arrival on a future day using arrival data for all days up to that day. Such predictions should be valid for future weekdays on which the arrival behavior follows the same pattern as those for that period of data.

Our method of accounting for dependence on time and day is more conveniently implemented with balanced data, although it can also be used with unbalanced data. For convenience, we have thus used arrival data from only regular (nonholiday) weekdays in August–December on which there were no quarter-hour periods missing and no obvious gross outliers in observed quarter-hourly arrival rates. This leaves 101 days. For each day (indexed by $j = 1, \ldots, 101$), the number of arrivals in each quarter hour from 7 AM–midnight was recorded as N_{jk}, $k = 1, \ldots, 68$. As noted in Section 3, these are assumed to be Poisson with parameter $\Lambda = \Lambda_{jk}$.

One could build a fundamental model for the values of Λ according to a model of the form

$$N_{jk} = \text{Poisson}(\Lambda_{jk}), \qquad \Lambda_{jk} = R_j \tau_k + \varepsilon'_{jk}, \tag{2}$$

where the τ_k's are fixed deterministic quarter-hourly effects, the R_j's are random daily effects with a suitable stochastic character, and the ε'_{jk}'s are random errors. Note that this multiplicative structure is natural, in that the τ_k's play the role of the expected proportion of the day's calls that fall in the kth interval. This is assumed to not depend on the R_j's, the expected overall number of calls per day. (We accordingly impose the side condition that $\sum \tau_k = 1$.)

We instead proceed in a slightly different fashion that is nearly equivalent to (2), but is computationally more convenient and leads to a conceptually more familiar structure. The basis for our method is a version of the usual variance-stabilizing transformation. If X is a Poisson(λ) variable, then $V = \sqrt{X + \frac{1}{4}}$ has approximately mean $\theta = \sqrt{\lambda}$ and variance $\sigma^2 = \frac{1}{4}$. This is nearly precise even for rather small values of λ. [One instead could use the simpler form $\sqrt{X}$ or the version of Anscombe (1948) that has $\sqrt{X + \frac{3}{8}}$ in place of $\sqrt{X + \frac{1}{4}}$; only numerically small changes would result. Our choice is based on considerations of Brown et al. (2001).] Additionally, V is asymptotically normal (as $\lambda \to \infty$), and it makes sense to treat it as such in the models that follow. We thus let $V_{jk} = \sqrt{N_{jk} + \frac{1}{4}}$, and assume the model

$$\begin{aligned} V_{jk} &= \theta_{jk} + \varepsilon^*_{jk} \quad \text{with } \varepsilon^*_{jk} \overset{\text{iid}}{\sim} \mathrm{N}\big(0, \tfrac{1}{4}\big), \\ \theta_{jk} &= \alpha_j \beta_k + \varepsilon_{jk}, \\ \alpha_j &= \mu + \gamma V_{j-1,+} + A_j, \end{aligned} \tag{3}$$

where $A_j \sim \mathrm{N}(0, \sigma_A^2)$, $\varepsilon_{jk} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$, $V_{j,+} = \sum_k V_{jk}$, and A_j and ε_{jk} are independent of each other and of values of $V_{j',k}$ for $j' < j$. Note that α_j is a random effect in this model. Furthermore, the model supposes a type of first-order autoregressive structure on the random daily effects. The correspondence between (2) and (3) implies that this structure is consistent with an approximate assumption that

$$R_j = \left(\gamma \sum_k \sqrt{N_{j-1,k} + \frac{1}{4}} + A_j\right)^2.$$

The model is thus not quite a natural one in terms of R_j, but it appears more natural in terms of the V_{jk} in (3) and is computationally convenient.

The parameters γ and β_k need to be estimated, as do μ, σ_A^2, and σ_ε^2. We impose the side condition $\sum \beta_k^2 = 1$, which corresponds to the condition $\sum \tau_k = 1$. The goal is then to derive confidence bounds for $\theta_{jk} = \sqrt{\Lambda_{jk}}$ in (3), and squaring the bounds yields corresponding bounds for Λ_{jk}.

The parameters in the model (3) can easily be estimated by a combination of least squares and method of moments. Begin by treating the $\{\alpha_j\}$'s as if they were fixed effects and using least squares to fit the model

$$V_{jk} = \alpha_j \beta_k + (\varepsilon_{jk} + \varepsilon^*_{jk}).$$

This is an easily solved nonlinear least squares problem. It yields estimates $\hat{\alpha}_j$, $\hat{\beta}_k$, and $\hat{\sigma}^2$, where the latter estimate is the mean squared error from this fit. Then σ_ε^2 can be estimated by method of moments as

$$\hat{\sigma}_\varepsilon^2 = \hat{\sigma}^2 - \frac{1}{4}.$$

Then use the estimates $\{\hat{\alpha}_j\}$ to construct the least squares estimates of these parameters that would be appropriate for a linear model of the form

$$\hat{\alpha}_j = \mu + \gamma V_{j-1,+} + A_j. \tag{4}$$

This yields least squares estimates, $\hat{\mu}$ and $\hat{\gamma}$, and the standard mean squared error estimator, $\hat{\sigma}_A^2$, for the variance of A_j.

The estimates calculated from our data for the quantities related to the random effects are

$$\begin{aligned} &\hat{\mu} = 97.88, \qquad \hat{\gamma} = .6784 \\ &\qquad\qquad\qquad (\text{with corresponding } R^2 = .501), \\ &\hat{\sigma}_A^2 = 408.3, \qquad \hat{\sigma}_\varepsilon^2 = .1078 \quad (\text{because } \hat{\sigma}^2 = .3578). \end{aligned} \tag{5}$$

The value of R^2 reported here is derived from the estimation of γ in (3), and it measures the reduction in sum of squared error due to fitting the $\{\hat{\alpha}_j\}$ by this model, which captures the previous day's call volumes, $V_{j-1,+}$. The large value of R^2 makes it clear that the introduction of the autoregressive model noticeably reduces the prediction error (by about 50%) relative to that obtainable from a model with no such component, that is, one in which a model of the form (3) holds with $\gamma = 0$.

For a prediction, $\breve{\Lambda}_k$, of tomorrow's value of Λ_k at a particular quarter hour (indexed by k), one would use the foregoing estimates along with today's value of V_+. From (3), it follows that tomorrow's prediction is

$$\breve{\theta}_k = \hat{\beta}_k(\hat{\gamma} V_+ + \hat{\mu}) \tag{6}$$

as an estimate of

$$\theta_k = \beta_k(\gamma V_+ + \mu + A) + \varepsilon, \tag{7}$$

where $A \sim \mathrm{N}(0, \sigma_A^2)$ and $\varepsilon \sim \mathrm{N}(0, \sigma_\varepsilon^2)$ are independent. The variance of the term in parentheses in (7) is the prediction variance of the regression in (6). Denote this by $\operatorname{pred var}(V_+)$. The coefficient of variation of $\hat{\beta}_k$ turns out to be numerically negligible compared with other coefficients of variation involved in (6) and (7). Hence,

$$\operatorname{var}(\breve{\theta}_k) \approx \hat{\beta}_k^2 \times \operatorname{pred var}(V_+) + \hat{\sigma}_\varepsilon^2. \tag{8}$$

These variances can be used to yield confidence intervals for the predictions of θ_k. The bounds of these confidence intervals can then be squared to yield confidence bounds for the prediction of Λ_k. Alternatively, one may use the convenient formula $CV(\breve{\theta}_k^2) \approx 2 \times CV(\breve{\theta}_k)$, and produce the corresponding confidence intervals (see Brown et al. 2002a for such a plot).

We note that the values of $CV(\breve{\theta}_k^2)$ here are in the range of .25 (for early morning and late evening) down to .16 (for midday). Note also that both parts of (8) are important in determining variability; the values of $\operatorname{var}(\breve{\theta}_k)$ range from .14 (for early morning and late evening) up to .27 (for midday). The fixed part of this is $\hat{\sigma}_\varepsilon^2 = .11$, and the remainder results from the first part of (8), which reflects the variability in the estimate of the daily volume figure, A, in (3).

Correspondingly, better estimates of daily volume (perhaps based on covariates outside our dataset) would considerably decrease the CVs during midday but would not have much effect on those for early morning and late evening. [Incidentally, we tried including day of the week an additional covariate in the model (3), but with the present data this did not noticeably improve the resulting CVs.]

A natural suggestion would be to use a nonparametric model for the curve $\Lambda(t)$ in place of the binned model in (2) and (3). This suggestion is appealing, and we plan to investigate it. However, we have not so far succeeded in producing a nonparametric regression analysis that incorporates all of the features of the foregoing model and also provides theoretically unbiased prediction intervals.

The preceding model includes several assumptions of normality. These can be empirically checked in the usual way by examining residual plots and Q–Q plots of residuals. All of the relevant diagnostic checks showed good fit to the model. For example, the Q–Q plots related to A and ε support the normality assumptions in the model. According to the model, the residuals corresponding to ε_{jk} also should be normally distributed. The Q–Q plot for these residuals has slightly heavier-than-normal tails, but only 5 (out of 6,868) values seem to be heavily extreme. These heavy extremes correspond to quarter-hour periods on different days that are noticeably extreme in terms of their total number of arrivals.

6.5 Prediction of $\nu(t)$

In this section we also model the service time according to quarter-hour intervals. This allows us to combine (in Sec. 6.6) the estimates of $\nu(t)$ derived here with the estimates of $\Lambda(t)$ derived in Section 6.4, and to obtain rigorously justifiable, bias-free prediction confidence intervals. In other respects, the model developed in this section resembles the nonparametric model of Section 4.4.

We use weekday data from only November and December. The lognormality discussed in Section 4.3 allows us to model log(service times), rather than service times. Let Y_{jkl} denote the log(service time) of the lth call served by an agent on day j, $j = 1, \dots, 44$, in quarter-hour intervals k, $k = 1, \dots, 68$. In total, there are $n = 57{,}152$ such calls. (We deleted the few call records showing service times of 0 or $> 3{,}600$ seconds.) For purposes of prediction, we will ultimately adopt a model similar to that of Section 4.4, namely

$$Y_{jkl} = \mu + \kappa_k + \varepsilon_{jkl}, \qquad \varepsilon_{jkl} \sim \mathrm{N}(0, \sigma_k^2) \quad \text{(indep.)}. \tag{9}$$

Before adopting such a model, we investigated whether there are day-to-day inhomogeneities that might improve the prediction model. We did this by adding a random-day effect to the model in (9). The larger model had a partial $R^2 = .005$. This is statistically significant (p value $< .0001$) due to the large sample size, but it has very little numerical importance. We also investigated a model that used the day as an additional factor, but found no useful information in doing so. Hence in what follows, we use model (9).

The goal is to produce a set of confidence intervals (or corresponding CVs) for the parameter

$$\nu_k = \exp\left(\mu + \kappa_k + \frac{\sigma_k^2}{2}\right). \tag{10}$$

The basis for this is contained in Section 4.4, except that here we use estimates from within each quarter-hour time period, rather than kernel-smoothed estimates. This enables us to obtain rigorously justifiable, bias-free prediction confidence intervals. The most noticeable difference is that the standard error of σ_k^2 is now estimated by

$$\mathrm{se}_{\sigma_k^2} \approx \sqrt{\frac{2}{n_k - 1}} S_k^2, \tag{11}$$

where n_k denotes the number of observations within the quarter hour, indexed by k, and S_k^2 denotes the corresponding sample variance from the data within this quarter hour. This estimate is motivated by the fact that if $X \sim \mathrm{N}(\mu, \sigma^2)$, then $\operatorname{var}((X - \mu)^2) = 2\sigma^4$ (see Brown et al. 2002a for a plot of these prediction intervals).

CVs for these estimates can be calculated from the approximate (Taylor series) formula $CV^*(\hat{\nu}_k) \approx CV(\hat{\mu} + \hat{\kappa}_k + \frac{\hat{\sigma}_k^2}{2})$. (The

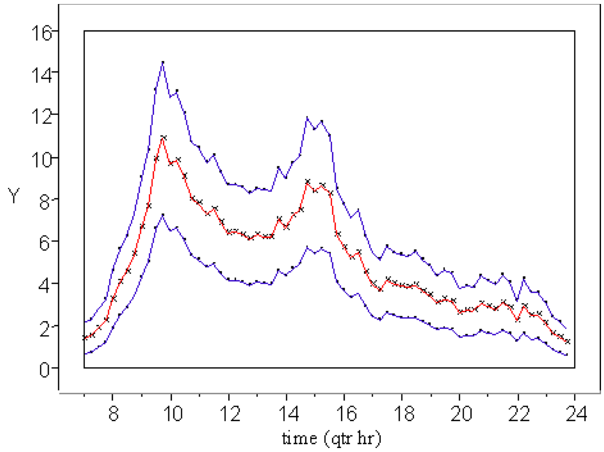

Figure 6. 95% Prediction Intervals for the Load, L, Following a Day With $V_+ = 340$.

intervals $\hat{\nu}_k \pm 1.96 \times \hat{\nu}_k \times CV^*$ agree with the foregoing to within 1 part in 200 or better.) The values of CV here range from .03 to .08. These are much smaller than the corresponding values of CVs for estimating $\Lambda(t)$. Consequently, in producing confidence intervals for the load, $L(t)$, the dominant uncertainty is that involving estimation of $\Lambda(t)$.

6.6 Confidence Intervals for $L(t)$

The confidence intervals can be combined as described in Section 6.3 to obtain confidence intervals for L in each quarter-hour period. Care must be taken to first convert the estimates of Λ and ν to suitable, matching units. Figure 6 shows the resulting plot of predicted load on a day following one in which the arrival volume had $V_+ = 340$.

The intervals in Figure 6 are still quite wide. This reflects the difficulty in predicting the load at a relatively small center such as ours. We might expect predictions from a large call center to have much smaller CVs, and we are currently examining data from such a large center to see whether this is in fact the case. Of course, inclusion (in the data and corresponding analysis) of additional informative covariates for the arrivals might improve the CVs in a plot such as Figure 6.

7. SOME APPLICATIONS OF QUEUEING SCIENCE

Queueing theory concerns the development of formal mathematical models of congestion in stochastic systems, such as telephone and computer networks. It is a highly developed discipline that has roots in the work of A. K. Erlang (Erlang 1911, 1917) at the beginning of the twentieth century. *Queueing science*, as we view it, is the theory's empirical complement; it seeks to validate and calibrate queueing-theoretic models via data-based scientific analysis. In contrast to queueing theory, however, queueing science is only starting to be developed. Although there exist scattered applications in which the assumptions of underlying queueing models have been checked, we are not aware of previous systematic effort to validate queueing-theoretic results.

One area in which extensive work has been done—and has motivated the development of new theory—involves the arrival processes of Internet messages (or message packets) (see, e.g., Willinger, Taqqu, Leland, and Wilson 1995; Cappe, Moulines, Pesquet, Petropulu, and Yang 2002; and the references therein). These arrivals have been found to involve heavy-tailed distributions and/or long-range dependencies (and thus differ qualitatively from the results reported in our Sec. 3).

In this section we use our call center data to produce two examples of queueing science. In Section 7.1 we validate (and refute) some classical theoretical results. In Section 7.2 we demonstrate the robustness (and usefulness) of a relatively simple theoretical model, namely the M/M/N+M (Erlang-A) model, for performance analysis of a complicated reality, namely our call center.

7.1 Validating Classical Queueing Theory

We analyze two congestion laws: first, the relationship between patience and waiting, which is a byproduct of Little's law (Zohar, Mandelbaum, and Shimkin 2002; Mandelbaum and Zeltyn 2003), and then the interdependence between service quality and efficiency, as it is manifested through the classical Khintchine–Pollaczek formula (see, e.g., eq. 5.68 in Hall 1991).

7.1.1 On Patience and Waiting. Here we consider the relationship between average waiting time and the fraction of customers that abandon the queue. To do so, we compute the two performance measures for each of the 3,867 hourly intervals that constitute the year. Regression then shows that a strong linear relationship exists between the two, with a value of $R^2 = .875$.

Indeed, if W is the waiting time and R is the time a customer is willing to wait (referred to as *patience*), then the law

$$\%\ \text{Abandonment} = \frac{\mathrm{E}(W)}{\mathrm{E}(R)} \tag{12}$$

is provable for models with *exponential* patience, like those of Baccelli and Hebuterne (1981) and Zohar et al. (2002). However, exponentiality is not the case here (see Fig. 5).

Thus the need arises for a theoretical explanation of why this linear relationship holds in models with *generally distributed* patience. Similarly, the identification and analysis of situations in which nonlinear relationships arise remains an important research question. [Motivated by the present study, Mandelbaum and Zeltyn (2003) pursued both directions.]

Under the hypothesis of exponentiality, we use (12) to estimate the average time that a customer is willing to wait in a queue, an absolute measure of customer patience. (Compare this with the relative index defined in Sec. 5.3.) From the inverse of the regression-line slope, we find that the average patience is 446 seconds in our case.

7.1.2 On Efficiency and Service Levels. As fewer agents cope with a given workload, operational efficiency increases. The latter is typically measured by the system (or agents') "occupancy," the average utilization of agents over time. Formally, this is defined as

$$\rho = \frac{\lambda_{\text{eff}}}{N\mu}, \tag{13}$$

where λ_{eff} is the effective arrival rate (namely, the arrival rate of customers who get served), μ is the service rate [$\mathrm{E}(S) = 1/\mu$ is the average service time], and N is number of active agents either serving customers or available to do so. Thus the staffing

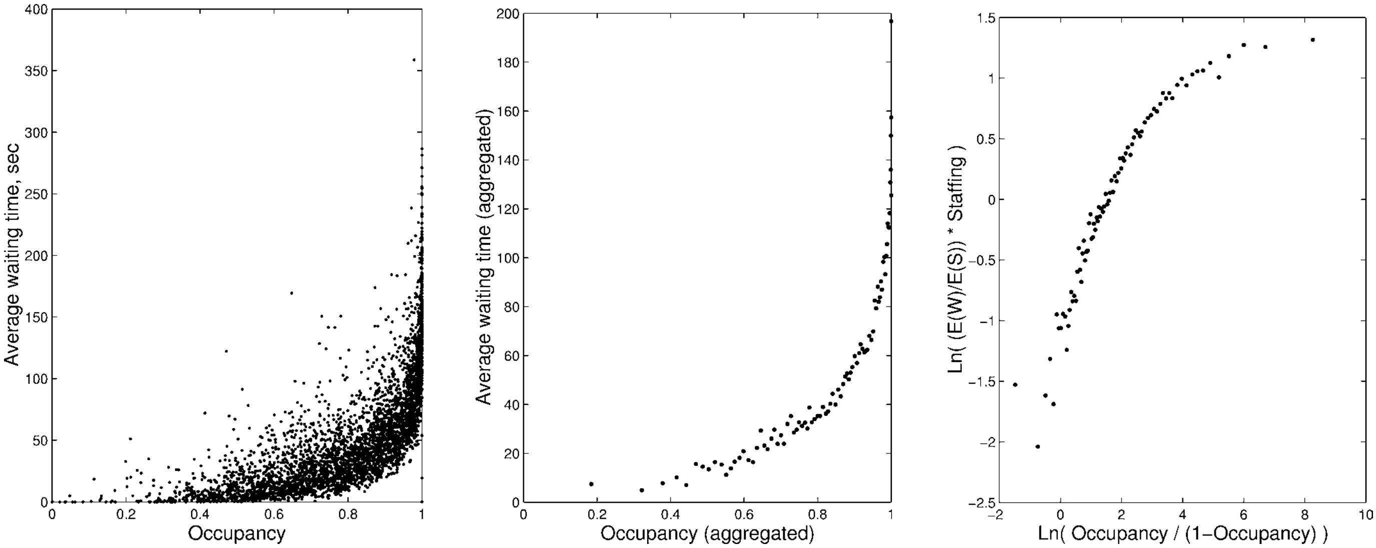

Figure 7. Agents' Occupancy versus Average Waiting Time.

level N is required to calculate agents' occupancy. Neither occupancies nor staffing levels are explicit in our database, however, so we derive indirect measures of these from the available data (see Brown et al. 2002a for details).

The three plots of Figure 7 depict the relationship between average waiting time and agents' occupancy. The first plot shows the result for each of the 3,867 hourly intervals over the year. The second and third plots emphasize the patterns by aggregating the data. (The hourly intervals were ordered according to their occupancy, and adjacent groups of 45 were then averaged together.)

The classical Khintchine–Pollaczek formula suggests the approximation

$$\mathrm{E}(W) \approx \frac{1}{N}\frac{\rho}{1-\rho}\frac{1+C_s^2}{2}\mathrm{E}(S), \tag{14}$$

which is a further approximation of (1) (see, e.g., Whitt 1993). Here C_s denotes the coefficient of variation of the service time, and ρ denotes the agents' occupancy.

The third plot of Figure 7 tests the applicability of the Khintchine–Pollaczek formula in our setting by plotting $N \cdot \mathrm{E}(W)/\mathrm{E}(S)$ versus $\rho/(1-\rho)$. To check whether the two plots exhibit the linear pattern implied by (14), we display an aggregated version of the data as a scatterplot on a logarithmic scale. This graph pattern is not linear. This can be explained by the fact that classical versions of Khintchine–Pollaczek formula are not appropriate for queueing systems with abandonment.

Note that queueing systems with abandonment usually give rise to dependence between successive interarrival times of *served* customers, as well as between interarrival times of served customers and service times. For example, long service times could engender massive abandonment and, therefore, long interarrival times of served customers. A version of the Khintchine–Pollaczek formula that can potentially accommodate such dependence was derived by Fendick, Saksena, and Whitt (1989). Theoretical research is needed to support the fit of these latter results to our setting with abandonment, however.

7.2 Fitting the M/M/N + M Model (Erlang-A)

The M/M/N model (Erlang-C), by far the most common theoretical tool used in the practice of call centers, does not allow for customer abandonment. The M/M/N + M model (Palm 1943) is the simplest abandonment-sensitive refinement of the M/M/N system. Exponentially distributed, or Markovian, customer patience (time to abandonment) is added to the model, hence the "+M" notation. This requires an estimate of the average duration of customer patience, $1/\theta$, or, equivalently, an individual abandonment rate, θ. Because it captures abandonment behavior, we call M/M/N + M the "Erlang-A" model (see Garnett et al. 2002 for further details). The 4CallCenters software (4CallCenters 2002) provides a valuable tool for implementing Erlang-A calculations.

The analysis in Sections 4 and 5 shows that in our call center, both service times and patience are not exponentially distributed. Nevertheless, simple models have often been found to be reasonably robust in describing complex systems. We therefore check whether the M/M/N + M model provides a useful description of our data.

7.3 Using the Erlang-A Model

We now validate the Erlang-A model against the overall hourly data used in Section 7.1. We consider three performance measures: probability of abandonment, average waiting time, and probability of waiting (at all). We calculate their values for our 3,867 hourly intervals using exact Erlang-A formulas, then aggregate the results using the same method used in Figure 7. The resulting 86 points are compared against the line $y = x$.

As before, the parameters λ and μ are easily computed for every hourly interval. For the overall assessment, we calculate each hour's average number of agents, N. Because the resulting N's need not be integral, we apply a continuous extrapolation of the Erlang-A formulas, obtained from relationships developed by Palm (1943).

For θ, we use formula (12), valid for exponential patience, to compute 17 hourly estimates of $1/\theta = \mathrm{E}(R)$, one for each of the

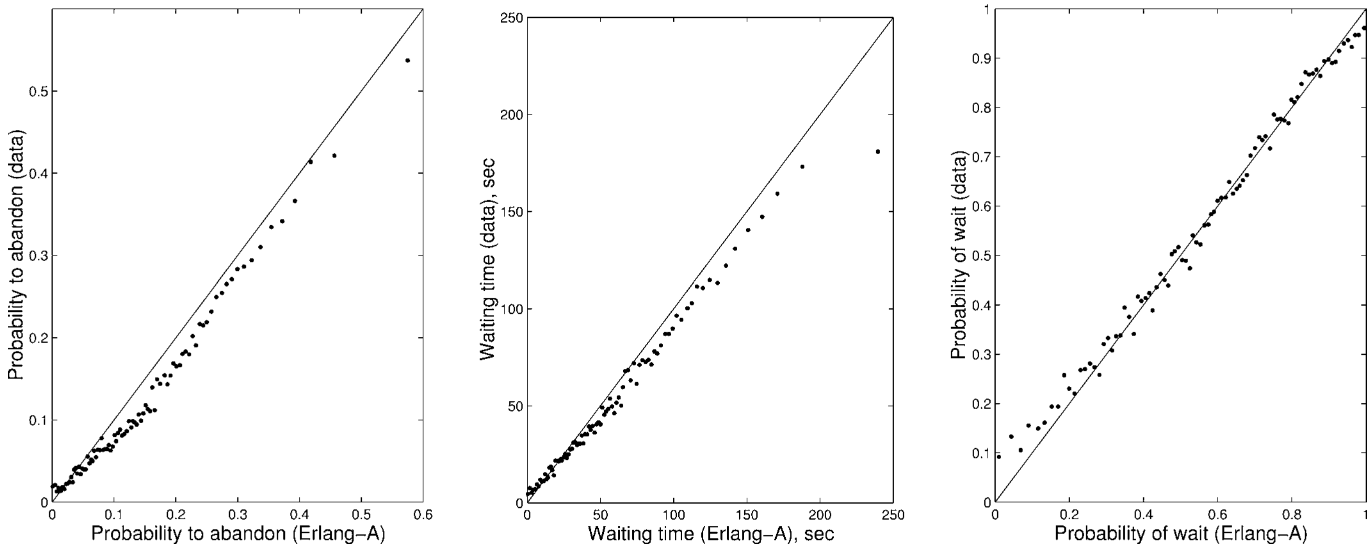

Figure 8. Erlang-A Formulas versus Data Averages.

17 1-hour intervals 7–8 AM, 8–9 AM, ..., 11–12 PM. The values for E(R) ranged from 5.1 minutes (8–9 AM) to 8.6 minutes (11–12 PM). We judged this to be better than estimating θ individually for each of the 3,867 hours (which would be very unreliable) or, at the other extreme, using a single value for all intervals (which would ignore possible variations in customers' patience over the time of day; see Zohar et al. 2002).

The results are displayed in Figure 8. The first two graphs show a relatively small yet consistent overestimation with respect to empirical values, for moderately and highly loaded hours. (We plan to explore the reasons for this overestimation in future research.) The rightmost graph shows a very good fit everywhere except for very lightly and very heavily loaded hours. The underestimation for small values of P{Wait} can probably be attributed to violations of work conservation (idle agents do not always answer a call immediately). Summarizing, it seems that these Erlang-A estimates can be used as useful upper bounds for the main performance characteristics of our call center.

7.4 Approximations

Garnett et al. (2002) developed approximations of various performance measures for the Erlang-A ($\mathrm{M/M/N+M}$) model. Such approximations require significantly less computational effort than exact Erlang-A formulas. The theoretical validity of the approximation was established by Garnett et al. (2002) for large Erlang-A systems. Although this is not exactly our case, the plots that we have created nevertheless demonstrate a good fit between the data averages and the approximations.

In fact, the fits for the probability of abandonment and average waiting time are somewhat superior to those in Figure 8 (i.e., the approximations provide somewhat larger values than the exact formulas). This phenomenon suggests two interrelated research questions of interest: how to explain the overestimation in Figure 8, and how to better understand the relationship between Erlang-A formulas and their approximations.

The empirical fits of the simple Erlang-A model and its approximation turn out to be very (perhaps surprisingly) accurate. Thus for our call center—and those like it—using Erlang-A for capacity-planning purposes could and should improve operational performance. Indeed, the model is already beyond typical current practice (which is Erlang-C dominated), and one aim of this article is to help change this state of affairs.

[Received November 2002. Revised February 2004.]

REFERENCES

Aalen, O. O., and Gjessing, H. (2001), "Understanding the Shape of the Hazard Rate: A Process Point of View," *Statistical Science*, 16, 1–22.

Anscombe, F. (1948), "The Transformation of Poisson, Binomial and Negative-Binomial Data," *Biometrika*, 35, 246–254.

Baccelli, F., and Hebuterne, G. (1981), "On Queues With Impatient Customers," in *International Symposium on Computer Performance*, ed. E. Gelenbe, Amsterdam: North Holland, pp. 159–179.

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Pratical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Ser. B, 57, 289–300.

Bolotin, V. (1994), "Telephone Circuit Holding Time Distributions," in *14th International Tele-Traffic Conference (ITC-14)*, Amsterdam: Elsevier, pp. 125–134.

Borst, S., Mandelbaum, A., and Reiman, M. (2004), "Dimensioning Large Call Centers," *Operations Research*, 52, 17–34; downloadable at *http://iew3.technion.ac.il/serveng/References/references.html*.

Breukelen, G. (1995), "Theorectial Note: Parallel Information Processing Models Compatible With Lognormally Distributed Response Times," *Journal of Mathematical Psychology*, 39, 396–399.

Brown, L. D., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2002a), "Statistical Analysis of a Telephone Call Center: A Queueing Science Perspective," technical report, University of Pennsylvania, downloadable at *http://iew3.technion.ac.il/serveng/References/references.html*.

Brown, L. D., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2002b), "Multifactor Poisson and Gamma–Poisson Models for Call Center Arrival Times," technical report, University of Pennsylvania.

Brown, L. D., and Hwang, J. (1993), "How to Approximate a Histogram by a Normal Density," *The American Statistician*, 47, 251–255.

Brown, L., and Shen, H. (2002), "Analysis of Service Times for a Bank Call Center Data," technical report, University of Pennsylvania.

Brown, L., Zhang, R., and Zhao, L. (2001), "Root Un-Root Methodology for Nonparametric Density Estimation," technical report, University of Pennsylvania.

Brown, L., and Zhao, L. (2002), "A New Test for the Poisson Distribution," *Sankhyā*, 64, 611–625.

Call Center Data (2002), Technion, Israel Institute of Technology, downloadable at *http://iew3.technion.ac.il/serveng/callcenterdata/index.html*.

Cappe, O., Moulines, E., Pesquet, J. C., Petropulu, A. P., and Yang, X. (2002), "Long-Range Dependence and Heavy-Tail Modeling for Teletraffic Data," *IEEE Signal Processing Magazine*, 19, 14–27.

Dette, H., Munk, A., and Wagner, T. (1998), "Estimating the Variance in Nonparametric Regression: What Is a Reasonable Choice?" *Journal of Royal Statistical Society*, Ser. B, 60, 751–764.

Erlang, A. (1911), "The Theory of Probability and Telephone Conversations," *Nyt Tidsskrift Mat. B*, 20, 33–39.

——— (1917), "Solutions of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges," *Electroteknikeren*, 13, 5–13 [in Danish].

Fendick, K., Saksena, V., and Whitt, W. (1989), "Dependence in Packet Queues," *IEEE Transactions on Communications*, 37, 1173–1183.

4CallCenters Software (2002), downloadable at *http://iew3.technion.ac.il/serveng/4CallCenters/Downloads.html*.

Gans, N., Koole, G., and Mandelbaum, A. (2003), "Telephone Call Centers: Tutorial, Review, and Research Prospects," *Manufacturing and Service Operations Management*, 5, 79–141.

Garnett, O., Mandelbaum, A., and Reiman, M. (2002), "Designing a Call-Center With Impatient Customers," *Manufacturing and Service Operations Management*, 4, 208–227.

Hall, P., Kay, J., and Titterington, D. (1990), "Asymptotically Optimal Difference-Based Estimation of Variance in Nonparametric Regression," *Biometrika*, 77, 521–528.

Hall, R. (1991), *Queueing Methods for Services and Manufacturing*, Englewood Cliffs, NJ: Prentice-Hall.

Iglehart, D., and Whitt, W. (1970), "Multiple Channel Queues in Heavy Traffic, I and II," *Advances in Applied Probability*, 2, 150–177, 355–364.

Kingman, J. F. C. (1962), "On Queues in Heavy Traffic," *Journal of the Royal Statistical Society*, Ser. B, 24, 383–392.

Lehmann, E. L. (1986), *Testing Statistical Hypotheses* (2nd ed.), New York: Chapman & Hall.

Levins, M. (2002), "On the New Local Variance Estimator," unpublished doctoral thesis, University of Pennsylvania.

Loader, C. (1999), *Local Regression and Likelihood*, New York: Springer-Verlag.

Mandelbaum, A. (2001), "Call Centers: Research Bibliography With Abstracts," technical report, Technion, Israel Institute of Technology, downloadable at *http://iew3.technion.ac.il/serveng/References/references.html*.

Mandelbaum, A., Sakov, A., and Zeltyn, S. (2000), "Empirical Analysis of a Call Center," technical report, Technion, Israel Institute of Technology, downloadable at *http://iew3.technion.ac.il/serveng/References/references.html*.

Mandelbaum, A., and Schwartz, R. (2002), "Simulation Experiments With $M/G/100$ Queues in the Halfin–Whitt (Q.E.D) Regime," technical report, Technion, Israel Institute of Technology, downloadable at *http://iew3.technion.ac.il/serveng/References/references.html*.

Mandelbaum, A., and Zeltyn, S. (2003), "The Impact of Customers Patience on Delay and Abandonment: Some Empirically-Driven Experiments With the $M/M/N+G$ Queue," submitted to *OR Spectrum*, Special Issue on Call Centers, downloadable at *http://iew3.technion.ac.il/serveng/References/references.html*.

Mao, W., and Zhao, L. H. (2003), "Free Knot Polynomial Spline Confidence Intervals," *Journal of the Royal Statistical Society*, Ser. B, 65, 901–919.

Müller, H., and Stadtmüller, U. (1987), "Estimation of Heteroscedasticity in Regression Analysis," *The Annals of Statistics*, 15, 610–625.

Palm, C. (1943), "Intensitatsschwankungen im Fernsprechverkehr," *Ericsson Technics*, 44, 1–189 [in German].

——— (1953), "Methods of Judging the Annoyance Caused by Congestion," *Tele*, 4, 189–208.

Serfozo, R. (1999), *Introduction to Stochastic Networks*, New York: Springer-Verlag.

Shen, H. (2002), "Estimation, Confidence Intervals and Nonparametric Regression for Problems Involving Lognormal Distribution," unpublished doctoral thesis, University of Pennsylvania.

Sze, D. (1984), "A Queueing Model for Telephone Operator Staffing," *Operations Research*, 32, 229–249.

Ulrich, R., and Miller, J. (1993), "Information Processing Models Generating Lognormally Distributed Reaction Times," *Journal of Mathematical Psychology*, 37, 513–525.

Whitt, W. (1993), "Approximations for the $GI/G/m$ Queue," *Production and Operations Management*, 2, 114–161.

——— (2002), *Stochastic-Process Limits*, New York: Springer-Verlag.

Willinger, W., Taqqu, M. S., Leland, W. E., and Wilson, D. V. (1995), "Self-Similarity in High-Speed Packet Traffic: Analysis and Modeling of Ethernet Traffic Measurements," *Statistical Science*, 10, 67–85.

Zohar, E., Mandelbaum, A., and Shimkin, N. (2002), "Adaptive Behavior of Impatient Customers in Tele-Queues: Theory and Empirical Support," *Management Science*, 48, 556–583.