Probability in the Engineering and Informational Sciences, **23**, 2009, 305–332. Printed in the U.S.A.
doi:10.1017/S0269964809000187

FORECAST ERRORS IN SERVICE SYSTEMS

SAMUEL G. STECKLEY
The Mitre Corporation
McLean, VA 22102-7508
E-mail: sgsteckley@gmail.com

SHANE G. HENDERSON
School of Operations Research and Industrial Engineering
Cornell University
Ithaca, NY 14853
E-mail: sgh9@cornell.edu

VIJAY MEHROTRA
Department of Decision Sciences
San Francisco State University
San Francisco, CA 94132-4156
E-mail: vjm@sfsu.edu

We investigate the presence and impact of forecast errors in the arrival rate of customers to a service system. Analysis of a large dataset shows that forecast errors can be large relative to the fluctuations naturally expected in a Poisson process. We show that ignoring forecast errors typically leads to overestimates of performance and that forecast errors of the magnitude seen in our dataset can have a practically significant impact on predictions of long-run performance. We also define short-run performance as the random percentage of calls received in a particular period that are answered in a timely fashion. We prove a central limit theorem that yields a normal-mixture approximation for its distribution for Markovian queues and we sketch an argument that shows that a normal-mixture approximation should be valid in great generality. Our results provide motivation for studying staffing strategies that are more flexible than the fixed-level staffing rules traditionally studied in the literature.

© 2009 Cambridge University Press 0269-9648/09 $25.00

1. INTRODUCTION

Service systems, such as inbound call centers, involve the provision of a service to customers by service agents (staff). A key question in such systems is how to select the number of agents in order to provide quality service to customers while keeping staffing costs at an acceptable level. Staff scheduling processes and algorithms for such systems typically operate on a 1-week cycle and involve three main steps:

1. forecasting customer arrival rates
2. choosing agent levels that will ensure satisfactory service
3. constructing shifts that cover the levels selected in step 2 and assigning these shifts to staff

The forecasts obtained in step 1 are almost always point estimates that are then taken as exact parameter values in subsequent steps. However, as we demonstrate through the analysis of data from several different call centers, there can be substantial errors in the forecasts. The question of how to deal with such errors when determining the target agent levels is the subject of this article.

The issues associated with forecast errors and staffing levels have certainly been recognized previously. Gans, Koole, and Mandelbaum [10] discussed this issue as part of a survey of the area of call center design and management. Grassmann [12] modeled forecast errors using a random arrival rate. Thompson [24] and Jongbloed and Koole [14] gave methods for determining target staffing when the arrival rate is random. Whitt [27] suggested a particular form of a random arrival rate for capturing forecast uncertainty. Chen and Henderson [7] examined the potential impact of ignoring arrival rate variability on performance predictions. Ross [20, Chap. 4] developed extensions to the "square-root staffing rule" to account for a random arrival rate. Avramidis, Deslauriers, and L'Ecuyer [1] developed several different arrival process models and compared their fit to call center data. They found that performance measures depend fairly strongly on the form of the arrival rate process. Deslauriers, L'Ecuyer, Pichitlamken, Ingolfsson, and Avramidis [9] showed that it is appropriate in their setting to weight performance by the (random) arrival rate. Brown, Zhang, and Zhao [5] developed an autoregressive model for the arrival rate that can capture correlation across different time periods within the same planning cycle. Mehrotra, Ozluk, and Saltzman [16] modeled arrival rate variability as part of a framework for intraday forecast and schedule updating. Harrison and Zeevi [13] developed an economic model based on attaching costs to abandonment and agent levels. Mathematical support for their model is given in Bassamboo, Harrison, and Zeevi [2]. Whitt [29] gave an economic analysis for a special case of the Harrison–Zeevi model, offering two computational approaches for estimating performance. In addition to a random arrival rate, Whitt [29] dealt explicitly with absenteeism, which he modeled through a random number of servers being available. We do not adopt an economic model here, instead working directly with performance measures associated with the waiting time distribution of a "typical" customer. We do not consider a random number

of servers, although it is possible to capture that phenomenon in a straightforward manner within our framework. Robbins, Medeiros, and Dunn [19] gave simulation results for a model with a random arrival rate in the quality-driven, efficiency-driven and quality-and-efficiency-driven regimes, showcasing the potential impact of random arrival rates on key performance measures through empirical results.

We view the main contributions of this article, relative to these earlier contributions, as follows.

- We show that forecast errors are substantial in a large call center dataset, suggesting that it may be worthwhile explicitly modeling forecast errors. To the best of our knowledge, ours is the first study that has access to the call volume forecasts that were used for staffing decisions and makes use of these to quantify forecast errors.
- We introduce a model of forecast errors and describe how to fit it to data.
- We clarify what performance measures one should compute when taking forecast errors into account and show that, in general, ignoring forecast errors leads to optimistic estimates of something we term *long-run* performance. Long-run performance is essentially the fraction of calls that are answered in a reasonable amount of time, where the average is calculated over a large number of periods in which the forecasted arrival rate is constant.
- We introduce *short-run* performance measures and explain how to compute them. Roughly speaking, these measures tell us "what might happen tomorrow." More precisely, short-run performance is essentially the distribution of the random fraction of calls received in a given period that have reasonable waiting times. Why is this distribution important? For staffing purposes, attaining a long-run performance goal seems very reasonable. However, some call centers might be "risk averse" with regard to delivering high-quality service, in the sense that they would like to ensure that, with high probability, customer waiting times are not too large. This is the case for certain hotel reservation, financial planning, and emergency services call centers, for example. The long-run performance measure can be viewed as an average, and so masks such risk. The distribution of the fraction of satisfactory calls in a period reveals much more information about the potential customer experience.
- We demonstrate empirically that forecast errors can have a practically significant impact on performance, reinforcing the results of Robbins et al. [19].

Our motivation comes from the inbound call center setting and so we will refer to customers and calls interchangeably.

The article is organized as follows. In Section 2 we analyze a dataset and show that forecast errors can be substantial. We also provide a model of forecast errors and outline how to fit the model to data. Section 3 describes how to compute the long-run performance of the system and shows that, in general, one can expect performance

estimates to be optimistically estimated if forecast errors are ignored. It concludes by computing long-run performance for some reasonable parameter regimes that we believe are representative of the case in practice, showing that forecast errors can have substantial impact on long-run performance. Section 4 then looks into short-run performance measures, showing that even when long-run performance is reasonable, there can be a nonnegligible probability of a very poor customer experience. Finally, Section 5 discusses the implications of our results for service system planning.

Some of the results in this article were first reported in Steckley, Henderson, and Mehrotra [23].

2. THE PRESENCE OF FORECAST ERRORS

Our dataset contains 9 weeks of call records from four different call centers. Two of the call centers are customer service operations: one that receives inquiries about order shipment and another that takes calls about a line of medical products. The other two call centers are outsourcers that handle calls for different companies in a wide variety of industries on a contract basis. For each queue and each 15-min interval, we examine both the forecasted call volumes used to drive the agent staffing and scheduling process and the actual call volumes that were received during each time period.

The forecasts were created using commercial software that allows its users to compute weighted averages of any combination of previous weeks' data and also to manually scale and/or edit any individual call forecast value quite easily. The analysts responsible for creating these forecasts used both historical call volume data as well as additional information about external factors that is not reflected in the historical data. For the two customer service call centers, this additional information typically includes informal input from other parts of the organization, such as marketing and operations groups, about factors that might influence customers' propensity to make service calls. For the two outsourcer call centers, additional input into the forecasting process typically includes information from client companies about their own in-house staffing levels as well as information about known or anticipated customer issues. The call volume forecasts were typically made 2–4 weeks prior to the actual days on which the calls arrived.

In the following, we describe the process by which we analyze these data to examine forecasting errors. We also propose a model for call arrivals and explain how to use such a dataset to determine the model parameters.

In the context of the staffing algorithms mentioned at the outset of the article, the 1-week planning horizon is typically split into p equal-length periods, which are indexed by j, $j = 1, 2, \ldots, p$. These p periods span the hours of the week for which the call center is open and accepting inbound calls; for example, for a call center that operates 7 days per week for 24 h each day, p is between 168 (1-h periods) and 672 (15-min periods). Letting t denote the length of the periods, we assume without loss of generality, by selecting the time units, that $t = 1$.

In our dataset we have several weeks of data, so we have multiple instances of each period j. Let n denote the number of weeks of data, which are indexed by $k, k = 1, 2, \ldots, n$. Now, let $\lambda(j,k)$ denote the forecasted expected number of calls received in period j in week k ($j = 1, \ldots, p, k = 1, \ldots, n$). Notice that $\lambda(j,k)$ can be interpreted as the predicted arrival rate of calls in period j of week k. Let $N(j,k)$ denote the actual number of calls received in period j of week k.

It seems reasonable to assume that $N(j,k)$ is Poisson distributed with mean $\lambda(j,k)$. To understand why, recall that the Palm–Khintchine theorem (see Whitt [28] for a basic version of this result, and Cinlar [8] for a more general version) asserts that the superposition of a large number of independent point processes is approximately a Poisson process, possibly nonhomogeneous. Since the arrival process can be viewed as the superposition of the arrivals generated by each individual in a surrounding population, this result seems applicable to our situation. Therefore, in any given period of any given day, it seems reasonable to model the arrival process as Poisson, which implies that the number of arrivals that are seen in a period has a Poisson distribution. Indeed, this is the standard model that is used, almost without comment, in the queuing and staffing literatures, and a formal hypothesis test of essentially this assumption on call center data in Brown, Gans, Mandelbaum, Sakov, et al. [4] did not reject the hypothesis. We therefore view this assumption as very safe.

If our forecasts are correct, then these, together with the Palm–Khintchine theorem, imply that $(N(j,k) : j, k \geq 1)$ consists of independent Poisson $(\lambda(j,k))$ random variables. In exploring the impact of forecasting errors, we thus begin by testing whether our data support this hypothesized model.

There are many well-known tests to determine whether a set of data reflects an independent and identically distributed (i.i.d.) sample from a Poisson distribution with fixed mean, as discussed in Brown and Zhao [6]. However, in our setting the means vary. Brown et al. [5] used a variance-stabilizing transformation. We employ an approach that seems more appropriate for our particular setting. Define

$$Z(j,k) = \frac{N(j,k) - \lambda(j,k)}{\sqrt{\lambda(j,k)}}.$$

Under our hypothesis, the $Z(j,k)$s are independent, have mean 0 and variance 1, and for $\lambda(j,k)$ large, $Z(j,k)$ is approximately normally distributed.

We restrict attention to periods j where the forecasted arrival rate is on the order of 50 calls per period or more, to ensure that the normal approximation is appropriately accurate. This gave a total of 120 different time periods j across several different queues and weeks, giving over 2000 Z values. A histogram of the resulting Z values is given in Figure 1 and is clearly nonnormal. This visual conclusion can be verified by a hypothesis test but we did not bother because the conclusion is apparent. To probe further, we tested 120 different hypotheses: one for each of the different time periods. Each hypothesis was of the form $(Z(j,k) : 1 \leq k \leq n)$ consisting of i.i.d. normal(0, 1)

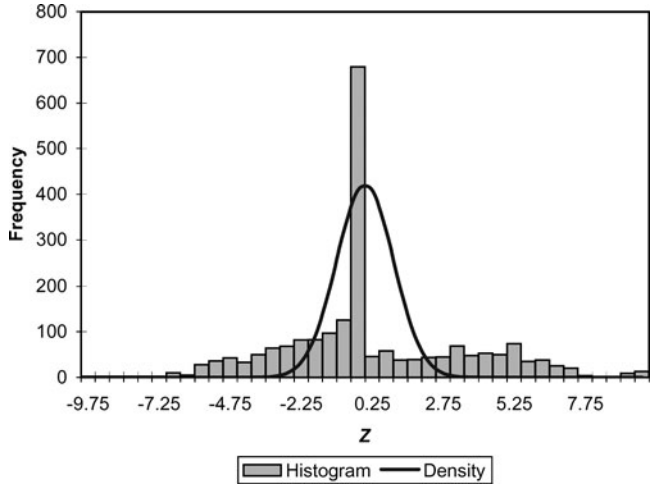

FIGURE 1. Histogram of Z values for 120 different periods and multiple weeks and an appropriately normalized density of a standard normal random variable.

random variables. We computed

$$S_j = \frac{\sum_{k=1}^{n} Z(j,k)}{\sqrt{n}}$$

and rejected the null hypothesis if $|S_j|$ was too large. When tested at the 99% level, we rejected 75 of the 120 hypotheses.

Based on these results, we see that the hypothesis that $N(j,k)$ is Poisson with parameter $\lambda(j,k)$ is, in many cases, simply incorrect. It is difficult to argue with the above Palm–Khintchine reasoning that $N(j,k)$ is Poisson distributed, so our test results suggest that the rates associated with these random variables are not as predictable as one might hope. We will take this a step further and model the arrival rates as being random.

To examine the impact of such arrival rate variability on system performance, we adopt a model of the arrival process that is essentially the one proposed in Whitt [27]. Suppose that $N(j,k)$ is Poisson distributed with mean $\Lambda(j,k)$, where $\Lambda(j,k) = B(j,k)\lambda(j,k)$. Here, $(B(j,k) : 1 \le j \le p, 1 \le k \le n)$ is a set of identically distributed random variables with mean 1 that are independent of everything else. The $B(j,k)$'s can be interpreted as "busyness" factors that indicate how busy a particular period is relative to the forecast.

One might assume that the $B(j,k)$ values arising on a particular day are identical, but from day to day are independent. Avramidis et al. [1] fitted a version of this model to their data. (They did not have access to call volume forecasts, so they also fitted nominal arrival rates $\lambda(j)$, which then play the role of our $\lambda(j,k)$'s. We believe that

the extra degree of freedom allowed us by using forecasts is important.) They found that the correlations between periods on a given day from the estimated model were stronger than those seen in the data.

Brown et al. [5] fitted a generalization of this model, where the busyness factors follow an autoregressive process from day to day, but again they did not have access to forecasts. The autoregressive component improved the fit to data considerably. More recently, Weinberg, Brown, and Stroud [25] fitted another generalization of this model to data, again without forecasts. Shen and Huang [22] describe a method for forecasting arrivals and updating those forecasts within a single day.

The key difference between our application of this model and previous applications is that the busyness factor represents an adjustment to *forecasts* rather than to *sample means*. The forecasts themselves almost certainly depend on data recorded until the forecasts are made, but, typically, they also reflect information not contained in the data, such as the timing of promotions, adjustments to the size of the customer base, and so forth.

We adopt the assumption that the $B(j,k)$'s are independent of one another, so that they are in fact i.i.d. The assumption is reasonable if the forecasts capture any effects that give rise to the sample correlations often seen in data; for example, if call volumes are trending upward over a series of weeks, then one would see a positive correlation in call volumes for adjacent periods. However, if the forecasts remove that trend, then the positive correlation would disappear.

Another reason why this assumption seems reasonable in our context is that we study the staffing problem one period at a time. In other words, we only need the marginal distribution of the call arrivals in a single period. From that perspective then, dependencies between periods do not play a role.

Although our model has limitations, it is easily understood and interpreted. It is therefore well suited to our goals, which are to get some sense of the magnitude of forecast errors and to understand the impact of forecast errors on true performance. One could certainly imagine developing more complex, and accurate, models of forecast error. We do not do so here for simplicity and for clarity of exposition. We expect that many of our conclusions in later sections still apply under more complex and accurate models.

Under our assumption that the $B(j,k)$'s are i.i.d. in both j and k, there is no need to maintain the distinction between weeks and periods. Accordingly, let us move to a single index k that indexes periods. For each period, we have a forecast $\lambda(k)$, an unobserved busyness index $B(k)$, and an actual volume of calls $N(k)$ that is assumed to be conditionally Poisson distributed with mean $\lambda(k)B(k)$, $k = 1, \ldots, n$.

To fit this model to historical data about actual and forecasted call volume, we adapt a parametric approach [14] that leads to simplifications in the analysis. Suppose that $(B(k) : 1 \leq k \leq n)$ are i.i.d. gamma random variables with shape parameter α and scale parameter $1/\alpha$, so that $B(k)$ has density $f(b;\alpha) = \alpha^{\alpha} b^{\alpha-1} e^{-\alpha b}/\,\Gamma(\alpha)$ and $\mathbb{E}\,B(1) = 1$ for any choice of $\alpha > 0$. We want to estimate α from the data $(N(k) : 1 \leq k \leq n)$ and forecasts $(\lambda(k) : 1 \leq k \leq n)$.

Let $L(\cdot)$ be the likelihood function for α. Then

$$\begin{aligned}
L(\alpha) &= \prod_{k=1}^{n} \int_0^\infty f(b(k);\alpha) \frac{e^{-b(k)\lambda(k)}(b(k)\lambda(k))^{N(k)}}{N(k)!}\, db(k) \\
&= \text{const} \prod_{k=1}^{n} \int_0^\infty (\alpha^\alpha b(k)^{\alpha-1} e^{-\alpha b(k)} / \Gamma(\alpha)) e^{-b(k)\lambda(k)} b(k)^{N(k)}\, db(k) \\
&= \text{const} \prod_{k=1}^{n} \left[\frac{\alpha^\alpha}{\Gamma(\alpha)} \int_0^\infty b(k)^{N(k)+\alpha-1} e^{-b(k)(\lambda(k)+\alpha)}\, db(k) \right] \\
&= \text{const} \prod_{k=1}^{n} \frac{\alpha^\alpha \Gamma(\alpha + N_k)}{\Gamma(\alpha)(\alpha+\lambda(k))^{\alpha+N(k)}},
\end{aligned}$$

where the last step follows from properties of the Gamma function and "const" contains terms that do not involve α or $b(k)$ for any k.

The likelihood function $L(\cdot)$ is easily numerically maximized, because it is one dimensional. If, however, one wishes to avoid this optimization, then a method-of-moments estimator of α can be derived as follows.

The marginal distribution of $N(k)$ is negative binomial with parameters α and $(1+\lambda(k)/\alpha)^{-1}$. It therefore has mean $\lambda(k)$ and variance $\lambda(k)(1+\lambda(k)/\alpha)$. Hence, $Z(k) = (N(k)-\lambda(k))/\sqrt{\lambda(k)}$ has mean 0 and variance $1+\lambda(k)/\alpha$.

Consider the sample variance s_n^2 of $(Z(k) : 1 \le k \le n)$. This is a sample variance of random variables that are *not* i.i.d. Letting $\bar{Z}_n$ and $\bar{\lambda}_n$ denote the sample means of $(Z(k) : 1 \le k \le n)$ and $(\lambda(k) : 1 \le k \le n)$, respectively, we get that

$$\begin{aligned}
s_n^2 &= \frac{1}{n-1} \sum_{k=1}^{n} (Z(k) - \bar{Z}_n)^2 \\
&\approx \frac{1}{n} \sum_{k=1}^{n} Z(k)^2 \qquad \textbf{(1)} \\
&\approx \frac{1}{n} \sum_{k=1}^{n} EZ(k)^2 \qquad \textbf{(2)} \\
&= \frac{1}{n} \sum_{k=1}^{n} (1+\lambda(k)/\alpha) \\
&= 1 + \bar{\lambda}_n/\alpha.
\end{aligned}$$

The approximation (1) assumes that the sample mean $\bar{Z}_n$ is essentially zero. The approximation (2) assumes that we have enough observations at each forecast level

that the strong law applies (at least approximately). Hence, we estimate α by

$$\hat{\alpha}_n = \frac{\bar{\lambda}_n}{s_n^2 - 1}.$$

The above derivation may appear to be somewhat heuristic in nature, but the estimator obtained is consistent under mild conditions. The following result is proved in Appendix A.

PROPOSITION 2.1: *Suppose that* $0 < \alpha \leq \infty$ *and* $0 < \lambda_* \leq \lambda(k) \leq \lambda^* < \infty$ *for deterministic bounds* λ_* *and* λ^*. *Then* $\hat{\alpha}_n \to \alpha$ *as* $n \to \infty$ *a.s.*

Remark 2.2: We explicitly allow $\alpha = \infty$, which corresponds to zero forecast error (i.e., $B \equiv 1$).

Remark 2.3: The condition that the forecasts be uniformly bounded away from zero and ∞ is easily stated and simplifies the proof slightly. The condition can be relaxed, as it is used only to bound certain infinite sums in the proof, and these sums could be bounded using weaker, but less transparent, conditions.

We used the method-of-moments approach above to estimate α for the 120 different time periods mentioned earlier. A histogram of the resulting estimated α values is given in Figure 2. The bin labeled "More" can be viewed as the number of time periods where the data implied an approximately deterministic arrival rate. Most of the values are less than 35. To get a sense for what this means from a practical standpoint, recall that $N(k)$ has mean $\lambda(k)$ and variance $\lambda(k)(1 + \lambda(k)/\alpha)$. So with a deterministic arrival rate, the variance is equal to the mean, and this is the level of variability that staffing levels are typically chosen to handle. However, if $\alpha = 25$, say, and the forecast $\lambda(k)$ is 50 or more (as it is for the data used earlier), then the variance can be three times the mean or more. It seems apparent that such a high degree of variability will have a serious impact on call center performance; that is the subject of the remainder of the article.

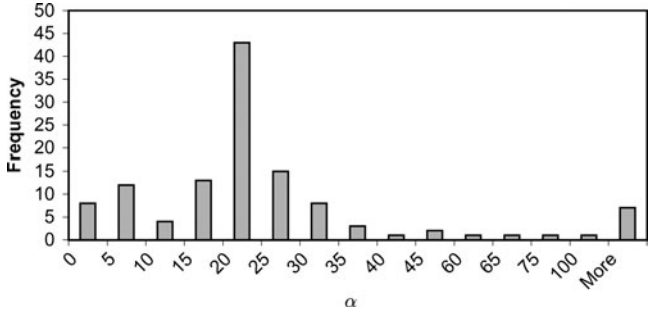

FIGURE 2. Histogram of estimated α values from 120 sets of data.

3. LONG-RUN PERFORMANCE MEASURES

In the previous section we developed a model of the customer arrival process that explicitly models forecast errors. We now adopt that model as truth and determine how to compute appropriate performance measures.

Ultimately, we want to compute the staffing level (number of agents) required in each period to ensure satisfactory service performance. To do so, we need to be able to compute performance for any given staffing level. For the remainder of this section, we implicitly assume a given staffing level and look at the question of how to compute performance.

Two standard performance measures are the fraction of customers who abandon and the fraction of customers who would experience a delay in queue of at most τ seconds were they willing to wait at least that long (and, hence, not abandon beforehand). Both measures can be handled in the framework below. We focus on the latter measure. Common choices for τ are 20 s (a moderate delay) and 0 s (no delay). Clearly, such performance measures depend on the set of customers over which the fraction is computed. In the present section we focus on a "long-run" interpretation, where we consider a large number of instances of periods like the one in question. The reason is that, in practice, staffing levels are usually selected to achieve a given long-run performance. (If the fraction is computed over customers that arrive in a single instance of the period, then the fraction is a random variable. We will study this "short-run" case in more detail in the next section.)

Consider a sequence of instances of the period with a common call volume forecast λ and a common staffing level. Let $S(k)$ denote the number of satisfactory calls (calls that are answered within the time limit τ) out of a total of $N(k)$ calls that are received in instance k of the period. Over n instances, the fraction of satisfactory calls is then

$$\frac{\sum_{k=1}^{n} S(k)}{\sum_{k=1}^{n} N(k)}. \tag{3}$$

Our assumptions that the forecasts and staffing levels are the same in all instances and that the busyness parameters are i.i.d. ensures that $((S(k), N(k)) : k \geq 1)$ consists of i.i.d. random elements. Now, $0 \leq \mathbb{E}\, S(1) \leq \mathbb{E}\, N(1) < \infty$. Dividing both the numerator and denominator of (3) by n and taking the limit as $n \to \infty$, the strong law then implies that the long-run fraction of satisfactory calls is

$$\frac{\mathbb{E}\, S(1)}{\mathbb{E}\, N(1)}. \tag{4}$$

However, how do we compute this ratio?

Let $B(k)$ be the busyness parameter associated with the kth instance of the period. Recall that we select time units so that the period is of length 1. Note that

$$\mathbb{E}\, N(1) = \mathbb{E}\, \mathbb{E}[N(1)|B(1)] = \mathbb{E}[\lambda B(1)] = \lambda, \tag{5}$$

since $\mathbb{E}\,B(1) = 1$. Computing $\mathbb{E}\,S(1)$ is more difficult. We again condition on $B(1)$ to obtain $\mathbb{E}\,S(1) = \mathbb{E}\,s(\lambda B(1))$, where $s(\gamma)$ is the expected number of satisfactory calls in the period when the arrival rate is γ. Our initial goal is an expression for $s(\gamma)$.

Fix the arrival rate to be deterministic and equal to γ (for now). Let $(X(u;\gamma) : u \geq 0)$ be a Markov process used to model the call center when there is a fixed arrival rate γ. In specialized cases, one can take X to be the process giving the number of customers in the system, but it might be more complicated. For technical reasons we take $X(\cdot;\gamma)$ to have sample paths that are left-continuous with right limits. Suppose that a customer arriving at time u will receive satisfactory service if and only if $X(u;\gamma) \in A$ for some set of states A.

Example 3.1: A common model of a call center is an $M/M/c + M$ queue (i.e., the Erlang-A model). There are c servers, service times are exponentially distributed, and the arrival process is Poisson. Customers are willing to wait an exponentially distributed amount of time (the "patience time") in the queue and abandon if they do not reach a server by that time. Here, we take $X(u;\gamma)$ to be the number of customers in the system at time u. Then $X(\cdot;\gamma)$ is a continuous-time Markov chain (CTMC). Suppose that a service is considered satisfactory if and only if the customer immediately reaches a server. Then we can take $A = \{0, 1, 2, \ldots, c-1\}$; that is, a service is satisfactory if and only if the number of customers in the system is $c - 1$ or less when the customer arrives.

Example 3.2: Consider the same model as Example 3.1 but now define a service to be satisfactory if and only if a customer would reach a server in at most $\tau > 0$ s, assuming she does not abandon. The state space of the CTMC defined in Example 3.1 is no longer rich enough to determine, upon a customer arrival, whether that customer will receive satisfactory service or not. We might turn to a different Markov process in such a case. Without loss of generality, suppose that as soon as a customer arrives, the patience and service times for that customer are sampled and therefore known. Since customers are served in FIFO order, we can determine, for every customer who has arrived by time u, whether that customer will abandon or not, and if not, which agent the customer will be served by. Let $V_m(u;\gamma)$ denote the "work in process" for agent m at time u, $m = 1, \ldots, c$. The quantity $V_m(u;\gamma)$ gives the time required for agent m to complete the service of all customers in the system at time u that are, or will be, served by agent i. Let $X(u;\gamma)$ be the vector $(V_m(u;\lambda) : 1 \leq m \leq c)$. The process $X(\cdot;\gamma)$ is a Markov process and we can take $A = \{v : \min_{m=1}^{c} v_m \leq \tau\}$, so that a service is satisfactory if and only if at least one server will be available to answer a call within τ s of a customer's arrival.

We denote the period we are studying as the interval $[0, 1]$. Let $\mathbb{P}_\varphi(\cdot)$ denote the probability measure when the Markov process has initial distribution φ. Let ν and π be respectively the distribution of the Markov process at time 0 (i.e., the state of the system at the start of the period) and the stationary distribution (assumed to

exist and be unique). Proposition 3.3 serves as a foundation for the use of steady-state approximations for performance measures in both the deterministic and random arrival rate contexts and is proved in Appendix A.

PROPOSITION 3.3: *Under the above conditions,*

$$s(\gamma) = \gamma \int_0^1 \mathbb{P}_\nu(X(u;\gamma) \in A)\, du.$$

If $\nu = \pi$, so that the Markov process is in steady state at time 0, then

$$s(\gamma) = \gamma f(\gamma),$$

where $f(\gamma) = \mathbb{P}_\pi(X(0;\gamma) \in A)$ is the steady-state probability that the system is in A. We can interpret $f(\gamma)$ as the long-run fraction of customers who receive satisfactory service if the arrival rate remains constant at γ.

Suppose that we adopt the steady-state approximation $s(\gamma) \approx \gamma f(\gamma)$. Here, γ is the expected number of customer arrivals in the period and $f(\gamma)$ is the long-run fraction of customers that receive satisfactory service. From (4) and (5), we see that

$$\frac{\mathbb{E}\,S(1)}{\mathbb{E}\,N(1)} = \frac{\mathbb{E}\,s(\lambda B(1))}{\lambda\,\mathbb{E}\,B(1)} \approx \frac{\mathbb{E}[\lambda B(1) f(\lambda B(1))]}{\lambda\,\mathbb{E}\,B(1)}. \tag{6}$$

Expression (6) simplifies slightly since $\mathbb{E}\,B(1) = 1$ and λ appears twice, but we have chosen to represent it in this way to clearly show the weighting. The fact that one should weight $f(\lambda B(1))$ by $\lambda B(1)$ is well known. It is implicit (and at times explicit) in the work of Harrison and Zeevi [13] and Whitt [29], for example. Chen and Henderson [7] did not perform this weighting in their analysis.

What are the consequences of ignoring a randomly varying arrival rate when predicting performance in a service system? In that case, we would first estimate an assumed-to-be-deterministic arrival rate. The most commonly used estimator is the sample mean of the number of arrivals, and this converges to λ almost surely as the data size increases (again assuming a large number of periods with forecasted arrival rate λ). We would then estimate performance as $f(\lambda)$.

Together with (6), Proposition 3.4 establishes that if f is decreasing and concave over the range of $\lambda B(1)$, then we will overestimate performance if a random arrival rate is ignored. The function f is, in great generality, decreasing in λ. For many models, it is also concave, at least in the region of interest; see Chen and Henderson [7].

PROPOSITION 3.4: *Suppose that f is decreasing and concave on the range of $\lambda B(1)$. Then*

$$\frac{\mathbb{E}[\lambda B(1) f(\lambda B(1))]}{\lambda\,\mathbb{E}\,B(1)} \le f(\lambda).$$

PROOF: We have that

$$\mathbb{E}[\lambda B(1) f(\lambda B(1))] \leq \lambda \, \mathbb{E}\, B(1) \, \mathbb{E} f(\lambda B(1)) \quad \textbf{(7)}$$

$$\leq \lambda \, \mathbb{E}\, B(1) f(\lambda E B(1)) \quad \textbf{(8)}$$

establishing the result. The inequality (7) follows since f is decreasing (see, e.g., Whitt [26]), and (8) uses Jensen's inequality. ■

For certain models and distributions of $B(1)$, we might be able to compute (6) exactly. In general though, this will not be possible. In such a case, we can use numerical integration. The problem is quite straightforward since f is typically easily computed and the integral $\mathbb{E}[\lambda B(1) f(\lambda B(1))]$ is one dimensional.

In our numerical examples, we model the call center as an $M/M/c + M$ queue (i.e., the Erlang-A model). As in Section 2, assume that $(B(k) : 1 \leq k \leq n)$ are i.i.d. gamma random variables with shape parameter α and scale parameter $1/\alpha$.

We compute both the long-run fraction of customers who would experience a delay in the queue of at most τ time units were they willing to wait at least that long and the fraction of customers who abandon. Note that when $\tau = 0$, a call is considered satisfactory if and only if it reaches a server immediately. These performance measures are computed using both the steady-state approximation described earlier and a simulation-based estimate. For the simulation estimate, we use an extensive warm-up period in which the parameter settings are identical to those used in the simulation of the actual period. Therefore, our data reflect steady-state performance.

We considered representative combinations of α and the forecasted call volume λ. Note that when $\alpha = \infty$, the arrival rate is deterministic and is given by λ. Let μ and θ be the service rate per hour and the abandonment rate per hour, respectively. The chosen values for μ and θ are typical for call centers. An abandonment rate of 0 corresponds to the case in which there is no abandonment, in which case the call center is modeled as an $M/M/c$ queue. For all scenarios, we let t, the length of the period, be 1 h.

We also need to choose the number of servers. This was chosen to be the minimum value such that the fraction of satisfactory ($\tau = 20$ s) calls is at least 80% in the case where the arrival rate is deterministic and equal to λ. This reflects the situation in practice where forecast error is ignored when setting staffing levels.

Both the simulation-based estimates and steady-state approximations for long-run performance (long-run fraction of satisfactory calls) are reported in Table 1. The simulation results are accurate to approximately two decimal places, and so are reported only to that accuracy. Due to space considerations, we present only selected scenarios. This selection illustrates the essential characteristics and trends seen in the results as a whole. The left-hand side of Table 1 identifies the parameter values in the scenarios. The right-hand side describes the performance. The first performance column gives the fraction of customers receiving service immediately, the second column is the fraction of customers who would receive service within 20 s were they willing to wait that long, and the third column is the fraction of customers who abandon.

TABLE 1. Approximations and Simulation-Based Estimates (in Parentheses) of Long-Run Performance

					Performance		
α	λ	μ	θ	c	$\tau = 0$	$\tau = 20$	Abandon
∞	500	12	0	48	0.75 (0.74)	0.83 (0.82)	— (—)
25	500	12	0	48	0.54 (0.55)	0.58 (0.60)	— (—)
∞	500	12	6	46	0.69 (0.68)	0.81 (0.80)	0.02 (0.02)
25	500	12	6	46	0.56 (0.57)	0.64 (0.65)	0.05 (0.05)
∞	500	12	12	45	0.68 (0.67)	0.81 (0.80)	0.03 (0.03)
25	500	12	12	45	0.57 (0.58)	0.67 (0.68)	0.07 (0.06)

The steady-state approximations and simulation-based estimates are, as they should be, very similar. Any differences are due to a combination of warm-up error in the simulation, the usual error in simulation estimates, and numerical error in computing the approximation. When $\alpha = \infty$, the arrival rate is deterministic and given by λ. In such cases, performance ($\tau = 20$ s) is very close to 0.8, as expected because the number of servers was chosen to ensure this. When there is variability in the arrival rate, as in the case when $\alpha = 25$, the fraction of calls answered within τ time units decreases. Additionally in some cases, the fraction declines significantly, underscoring the danger of ignoring a randomly varying arrival rate. As for the long-run fraction of abandoning calls, Table 1 shows that when there is variability in the arrival rate, the fraction of calls that abandon is significantly higher than when the arrival rate is deterministic.

The results also indicate that the degradation in the fraction of calls answered within τ time units is less when abandonment is modeled and is further reduced as the rate of abandonment θ increases. We believe this is because abandonment helps to reduce the queue length, so that patient customers reach service more quickly. Note that we see this "positive" impact from abandonment even though abandoning customers are counted as "unsatisfactory."

The results also show that, not surprisingly, as the rate of abandonment increases, the long-run fraction of abandoning calls increases. So a large abandonment rate tends to positively impact the long-run fraction of calls answered before τ but to negatively impact the long-run fraction of calls that abandon.

4. SHORT-RUN PERFORMANCE

The numerical results in the previous section show that long-run performance can be seriously impacted by forecast errors. This suggests that the observed (or "short-run")

performance on any given day could be even worse. By short-run performance we mean the distribution of $S(1)/N(1)$ (i.e., the distribution of the fraction of satisfactory calls in a single period $[0, t]$ on a single day).

There is a positive probability that $N(1) = 0$, but it is very small for even modest arrival rates. In any case, we can just define $0/0 = 1$ arbitrarily to ensure that $S(1)/N(1)$ is a proper random variable. The random variable $S(1)/N(1)$ is supported on the rationals, so its exact probability mass function is likely to be "spiky" and difficult to interpret. We give approximations for its distribution that are likely to be more informative.

Suppose that conditional on $B(1)$, the period is long enough that the fraction of satisfactory calls is close to its steady-state mean $f(\lambda B(1))$, where λ is the forecasted arrival rate. This transformation of the random variable $\lambda B(1)$ is our first approximation. It ignores the "process variability" that arises even for a fixed arrival rate. In numerical experiments reported in Steckley et al. [23] we found that this first approximation indicates general trends but that it is important to account for process variability to get a clearer picture.

Fortunately, we can refine this approximation to take into account process variability (i.e., the fact that the observed fraction of satisfactory calls will not be exactly equal to the steady-state mean). The key to the refinement is a central limit theorem (CLT) for $S(1)/N(1)$ under a fixed arrival rate, showing that the ratio is approximately normally distributed. Such a CLT then implies that the unconditional distribution of $S(1)/N(1)$ is approximately a mixture of normals. We prove such a CLT for a specific model, and we also give a nonrigorous argument suggesting that the CLT should hold in great generality. We begin with the rigorous result.

Let the arrival rate γ be fixed. Suppose that our goal is to answer calls immediately. Suppose further that the number-in-system process $X = (X(s) : s \geq 0)$ can be modeled as an irreducible CTMC on the finite state space $\{0, 1, \ldots, d\}$, where $d > c$.

Remark 4.1: It is not essential that the state space be finite, but it simplifies the proof of the CLT below. In particular, an appropriately integrable solution to Poisson's equation always exists in the finite-state space case, but with a countable state space, one must impose additional conditions. Rather than be diverted by a necessarily lengthy discussion of such conditions, we prefer to make the finite-state space assumption, which has the added benefit of making our model more closely match reality. (One never has an infinite number of trunk lines!)

Let $M(s)$ be the number of state transitions in the CTMC X by time s and let $Y = (Y_n : n \geq 0)$ be the embedded discrete-time Markov chain that arises by observing the CTMC X just after state transitions. Then we can write

$$\frac{S(1)}{N(1)} \approx \frac{U_{M(t)}}{V_{M(t)}}, \tag{9}$$

where

$$U_n = \frac{1}{n}\sum_{i=1}^{n} I(Y_i = Y_{i-1} + 1, Y_{i-1} \leq c - 1)$$

and

$$V_n = \frac{1}{n}\sum_{i=1}^{n} I(Y_i = Y_{i-1} + 1).$$

Here, $I(\cdot)$ is the indicator function that is 1 if its argument is true and 0 otherwise, so that U_n gives the fraction of the first n transitions that correspond to an arriving customer finding a server available. Similarly, V_n gives the fraction of the first n transitions that correspond to an arrival joining the system. Notice that V_n does not count blocked customers. That is why the relation in (9) is not an equality. When d is large enough that few customers are turned away, the approximation should be very good.

Let $\Rightarrow$ denote convergence in distribution and $\mathcal{N}(a, b)$ denote a normally distributed random variable with mean a and variance b.

THEOREM 4.2: *Under the assumptions given above,*

$$\sqrt{\lambda s}\left(\frac{U_{M(s)}}{V_{M(s)}} - \frac{u}{v}\right) \Rightarrow \mathcal{N}(0, \sigma^2(\gamma))$$

as $s \to \infty$, where u, v, and $\sigma^2(\gamma)$ are specified in the following proof.

PROOF: The proof has three steps. The key step is to establish the joint CLT

$$\sqrt{n}\left(\begin{pmatrix} U_n \\ V_n \end{pmatrix} - \begin{pmatrix} u \\ v \end{pmatrix}\right) \Rightarrow \mathcal{N}(0, \Sigma) \tag{10}$$

as $n \to \infty$, where $\mathcal{N}(0, \Sigma)$ denotes a Gaussian random vector with mean 0 and covariance matrix Σ, and u, v, and Σ are specified below. The final two steps consist of applying a random time change and then the delta method.

To establish (10), we apply a Markov chain CLT (see, e.g., Meyn and Tweedie [17, Thm. 17.4.4]). Consider the (irreducible, finite-state-space) Markov chain $\tilde{Y} = (\tilde{Y}_i : i \geq 0)$, where $\tilde{Y}_i = (Y_i, Y_{i+1})$. We can write

$$U_n - u = \frac{1}{n}\sum_{i=0}^{n-1} \tilde{h}_1(\tilde{Y}_i)$$

and

$$V_n - v = \frac{1}{n}\sum_{i=0}^{n-1} \tilde{h}_2(\tilde{Y}_i),$$

where

$$\tilde{h}_1(x,y) = I(y = x+1, x \le c-1) - u$$

and

$$\tilde{h}_2(x,y) = I(y = x+1) - v.$$

Let $\tilde{\pi}$ be the stationary distribution of $\tilde{Y}$. We choose u and v to be steady-state means, so that $\tilde{\pi}\tilde{h}_i = \sum_{(x,y)} \tilde{\pi}(x,y)\tilde{h}_i(x,y) = 0$ for $i = 1, 2$. Let $\tilde{P}$ be the transition matrix of $\tilde{Y}$, and let $\tilde{g}_1$ and $\tilde{g}_2$ solve Poisson's equation

$$\tilde{P}\tilde{g}_i(x,y) = \tilde{g}_i(x,y) - \tilde{h}_i(x,y)$$

for $i = 1, 2$ and all (x, y).

We now wish to apply the Markov chain CLT [17, Thm. 17.4.4] to obtain (10). That result applies only to univariate processes, but the result extends to our multivariate case through an application of the Cramér–Wold device (e.g., Billingsley [3, Thm. 7.7]). The Cramér–Wold device asserts that a sequence of $\mathbb{R}^m$-valued random vectors $\{\xi_n\}$ converges in distribution to the $\mathbb{R}^m$-valued random vector ξ if and only if each linear combination of the components of the vector ξ_n converges in distribution to the corresponding linear combination of the components of ξ. Let a and b be arbitrary constants and note that

$$\begin{aligned} a(U_n - u) + b(V_n - v) &= \frac{1}{n}\sum_{i=0}^{n-1}[a\tilde{h}_1(\tilde{Y}_i) + b\tilde{h}_2(\tilde{Y}_i)] \\ &= \frac{1}{n}\sum_{i=0}^{n-1}\tilde{h}(\tilde{Y}_i), \end{aligned}$$

where $\tilde{h} = a\tilde{h}_1 + b\tilde{h}_2$. Furthermore, since $\tilde{P}$ is a linear operator, $\tilde{g} = a\tilde{g}_1 + b\tilde{g}_2$ solves Poisson's equation $\tilde{P}\tilde{g} = \tilde{g} - \tilde{h}$. Theorem 17.4.4 of Meyn and Tweedie [17] then allows us to conclude that

$$\sqrt{n}(a(U_n - u) + b(V_n - v)) \Rightarrow aW_1 + bW_2$$

as $n \to \infty$, where W is a two-dimensional normal random vector with mean 0 and covariance matrix Σ defined by

$$\begin{aligned} \Sigma_{ij} &= \mathbb{E}_{\tilde{\pi}}[(\tilde{g}_i(\tilde{Y}_1) - \tilde{P}\tilde{g}_i(\tilde{Y}_0))(\tilde{g}_j(\tilde{Y}_1) - \tilde{P}\tilde{g}_j(\tilde{Y}_0))] \\ &= \mathbb{E}_{\tilde{\pi}}[\tilde{g}_i(\tilde{Y}_0)\tilde{h}_j(\tilde{Y}_0) + \tilde{h}_i(\tilde{Y}_0)\tilde{g}_j(\tilde{Y}_0) - \tilde{h}_i(\tilde{Y}_0)\tilde{h}_j(\tilde{Y}_0)]. \end{aligned} \tag{11}$$

(Equality (11) follows as in Meyn and Tweedie [17, Eq. 17.47].) The Cramér–Wold device then immediately implies the CLT (10).

In fact, we obtain a stronger result, namely a functional CLT. Moreover, $M(s)/s \to \beta$ as $s \to \infty$ a.s., where $\beta > 0$ is the long-run rate of transitions in the

CTMC X. We can then directly apply the random-time-change result [3, Thm. 17.1] to obtain

$$\sqrt{M(s)}\left(\begin{pmatrix} U_{M(s)} \\ V_{M(s)} \end{pmatrix} - \begin{pmatrix} u \\ v \end{pmatrix}\right) \Rightarrow \mathcal{N}(0, \Sigma)$$

as $s \to \infty$. The converging-together lemma [3, Problem 1, p. 28] then implies that

$$\sqrt{\beta s}\left(\begin{pmatrix} U_{M(s)} \\ V_{M(s)} \end{pmatrix} - \begin{pmatrix} u \\ v \end{pmatrix}\right) \Rightarrow \mathcal{N}(0, \Sigma)$$

as $s \to \infty$.

The final step applies the delta method (e.g., Serfling [21, p. 122]), using the function $\phi(x, y) = x/y$, to conclude that

$$\sqrt{\beta s}\left(\frac{U_{M(s)}}{V_{M(s)}} - \frac{u}{v}\right) \Rightarrow \mathcal{N}(0, \eta^2),$$

where

$$\begin{aligned} \eta^2 &= \nabla\phi(u, v)^T \Sigma \nabla\phi(u, v) \\ &= \frac{\Sigma_{11} - 2(u/v)\Sigma_{12} + (u/v)^2\Sigma_{22}}{v^2}. \end{aligned}$$

(The second equality uses the fact that $\nabla\phi(u, v) = v^{-1}(1, -u/v)$.) Setting $\sigma^2(\gamma) = \gamma\eta^2/\beta$ yields the result. ■

Equation (9) and Theorem 4.2 establish that conditional on $B(1)$, the fraction $S(1)/N(1)$ is approximately normally distributed with mean $f(\lambda B(1))$ and variance $\sigma^2(\lambda B(1))/(\lambda B(1)t)$. So we can approximate the distribution of $S(1)/N(1)$ by the normal mixture $\mathcal{N}(f(\lambda B(1)), \sigma^2(\lambda B(1))/(\lambda B(1)t))$.

Remark 4.3: The variance of this normal mixture is

$$\text{var}\, f(\lambda B(1)) + \mathbb{E}\,\frac{\sigma^2(\lambda B(1))}{\lambda B(1)t},$$

which can be viewed as a decomposition of the variance into contributions from arrival rate variability and process variability respectively.

To compute the distribution of this normal mixture we need to be able to compute the constant $\sigma^2(\gamma)$, which, in turn, depends on β and η^2 (which also depend on γ). In Appendix B we sketch how to compute σ^2. The derivation exploits the strong relationships between the two-step Markov chain $\tilde{Y}$ and the single-step Markov chain Y and between the CTMC X and its embedded chain Y.

Remark 4.4: The above derivation is for the fraction of calls answered immediately. One can perform a very similar derivation that yields a mixture-of-normals approximation for the abandonment rate. The key ideas are as follows.

In addition to the above definitions, let Z_n be 0, 1, or 2, depending on whether the event that caused the nth state change in the CTMC X (from state Y_{n-1} to Y_n) was an arrival, a service completion, or an abandonment, respectively. Then if $A(1)$ is the number of abandoning calls in the period, we can write

$$\frac{A(1)}{N(1)} \approx \frac{U_{M(t)}}{V_{M(t)}},$$

where

$$U_n = \frac{1}{n}\sum_{i=1}^{n} I(Y_i = Y_{i-1} - 1, Z_i = 2)$$

and

$$V_n = \frac{1}{n}\sum_{i=1}^{n} I(Y_i = Y_{i-1} + 1).$$

One then proves a joint CLT for (U_n, V_n) using the same technique used for Theorem 4.2, via the Markov chain $\tilde{Y} = (\tilde{Y}_i : i \geq 0)$, where $\tilde{Y}_i = (Y_i, Y_{i+1}, Z_{i+1})$. The proof is again completed through a random-time change and an application of the delta method.

To get a sense of the quality of the mixture-of-normals approximations for the distribution of $S(1)/N(1)$, we return to the $M/M/c + M$ call center model discussed at the end of Section 3. All parameters were chosen in the same manner as earlier, and, again, we used a warm-up period in the simulations to ensure that our results reflect steady-state performance.

Due to space considerations, we present only selected scenarios, but the essential characteristics and trends seen in the selected scenarios hold for all of the scenarios we considered. Figure 3 plots the simulation-based estimate of the distribution (histogram) along with the approximation for a particular scenario where abandonment is not modeled. The first and last bar of the histogram correspond to the observed $S(1)/N(1)$ values that were exactly 0 and 1, respectively. The density of the approximation has been truncated at 0 and 1 and the probability of the truncated regions are plotted as histogram bars, one to the left of 0 and the other to the right of 1.

Both the approximation and the simulation-based histogram show that the distribution of $S(1)/N(1)$ spikes at 0 and 1 and has little density at the intermediate values. Here, the arrival rate, given by $\lambda B(1)$, rarely takes on values that the staffing level is designed to handle. More frequently, the arrival rate is too large or too small for the given staffing level. Therefore, performance is bimodal: either very poor or very good, with little chance of moderate performance. Note that the approximation tracks the simulation-based results quite well.

In Figure 4 we consider a scenario with abandonment. We no longer see the bimodal behavior. However, the distribution still indicates that there is significant probability that the fraction of calls answered immediately will be quite low. Just as

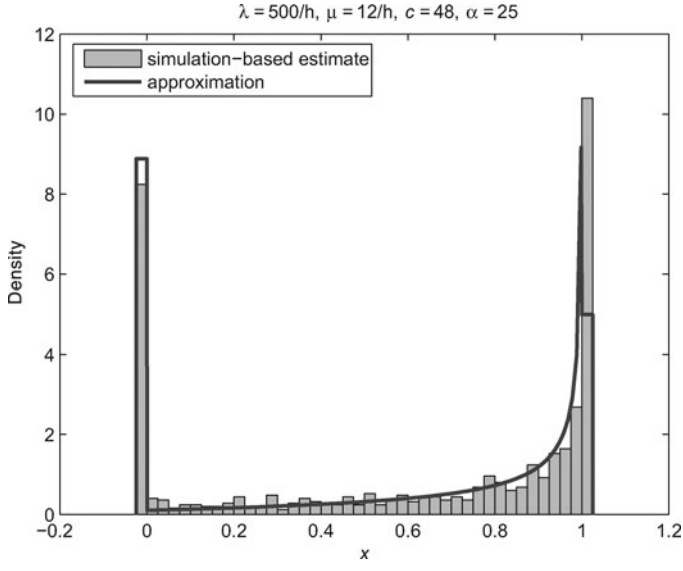

FIGURE 3. Plots of the distribution estimates for the fraction of calls answered immediately when $\lambda = 500$, $\mu = 12$, $c = 48$, $\theta = 0$, and $\alpha = 25$.

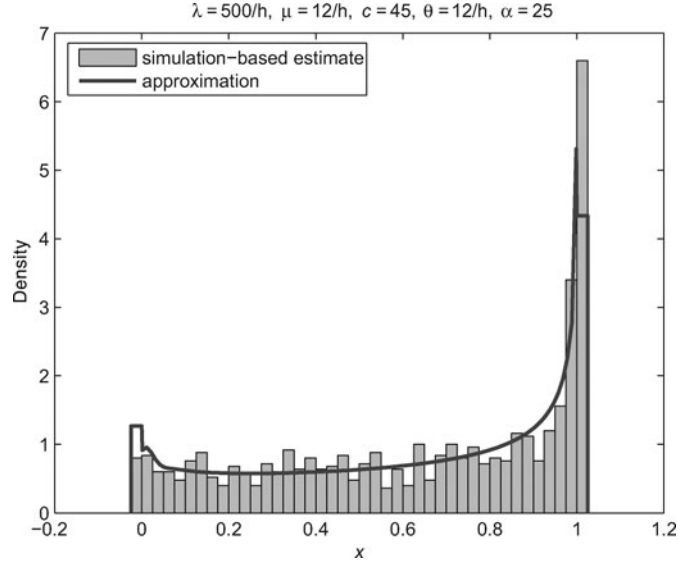

FIGURE 4. Plots of the distribution estimates for the fraction of calls answered within τ when $\lambda = 500$, $\mu = 12$, $c = 45$, $\theta = 12$, $\tau = 0$, and $\alpha = 25$.

we saw in the long-term calculations, abandonment improves the short-run fraction of satisfactory calls. The intuition here is again that abandonment keeps the queue short, so that customers who do get served reach a server quickly.

Now, suppose that $S(1)/N(1)$ gives the fraction of calls that abandon. As mentioned earlier, we can approximate the distribution of the fraction of calls that abandon by another mixture of normals. The approximation and the simulation-based estimate are plotted in Figure 5 for a particular scenario. Notice that the abandonment rate on a given day can be higher than 15% with nontrivial probability.

The above normal-mixture result is for the special case where the call center can be modeled as a discrete-state-space CTMC. However, the result can be expected to hold far more generally. We now sketch an argument that makes that assertion more clear. We once again condition on the arrival rate $\lambda B(1)$ in the period.

Let $(X(u;\gamma) : u \geq 0)$ denote the underlying Markov process with customer arrival rate γ. Let $T_i(\gamma)$ denote the time of the ith customer arrival when the arrival rate is γ. Define $Z_i(\gamma) = X(T_i(\gamma);\gamma)$ to be the state of the Markov process just before the ith customer arrival. The ith customer receives satisfactory service if and only if $Z_i(\gamma) \in A$, where A is the set defining satisfactory service (see the discussion immediately before Example 3.1). So $S(1)/N(1)$ has the same distribution as

$$\frac{1}{N(t;\gamma)} \sum_{i=1}^{N(t;\gamma)} I(Z_i(\gamma) \in A),$$

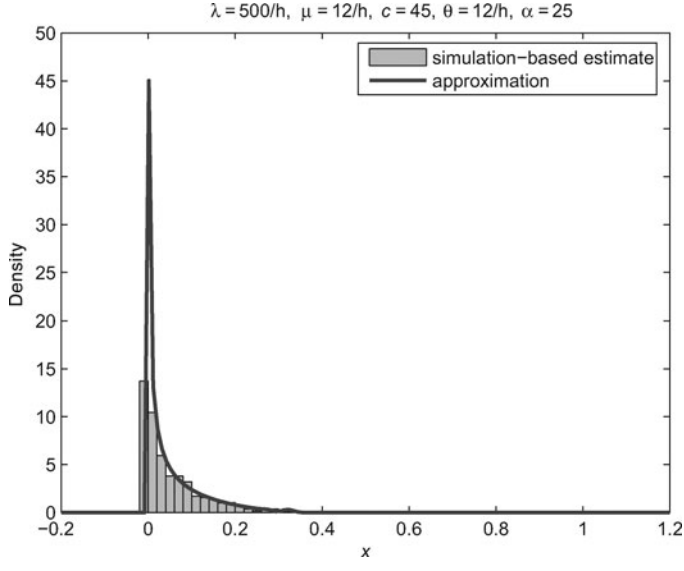

FIGURE 5. Plot of the distribution estimate for the fraction of calls that abandon when $\lambda = 500$, $\mu = 12$, $c = 45$, $\theta = 12$, and $\alpha = 25$.

where $N(s;\gamma)$ is a Poisson random variable, with mean γ's giving the number of arrivals in $[0,s]$.

The strong Markov property for $X(\cdot;\gamma)$ ensures that $(Z_i(\gamma) : i \geq 1)$ is a Markov chain. We can then apply a CLT (e.g., Meyn and Tweedie [17, Chap. 17]) to assert that under appropriate conditions,

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} I(Z_i(\gamma) \in A) - f(\gamma)\right] \Rightarrow \sigma(\gamma)\mathcal{N}(0,1),$$

as $n \to \infty$, where $\sigma^2(\gamma)$ is a variance constant. Again, under appropriate conditions a random-time change gives

$$N^{1/2}(s;\gamma)\left[\frac{1}{N(s;\gamma)}\sum_{i=1}^{N(s;\gamma)} I(Z_i(\gamma) \in A) - f(\gamma)\right] \Rightarrow \sigma(\gamma)\mathcal{N}(0,1)$$

as $s \to \infty$. A converging-together argument then ensures that

$$(\gamma s)^{1/2}\left[\frac{1}{N(s;\gamma)}\sum_{i=1}^{N(s;\gamma)} I(Z_i(\gamma) \in A) - f(\gamma)\right] \Rightarrow \sigma(\gamma)\mathcal{N}(0,1). \qquad \textbf{(12)}$$

The limit result (12) then ensures that so long as a period is "long enough," the ratio $S(1)/N(1)$ is approximately normally distributed. Therefore, when γ is chosen to equal $\lambda B(1)$ independent of all else, the distribution of $S(1)/N(1)$ can be approximated by the normal mixture

$$\mathcal{N}\left(f(\lambda B(1)), \frac{\sigma^2(\lambda B(1))}{\lambda B(1)}\right).$$

It is likely very difficult to compute $f(\cdot)$ and $\sigma^2(\cdot)$ for complex models of service systems, but these quantities could certainly be estimated through simulation, and then the normal-mixture approximation is easily constructed numerically.

5. DISCUSSION

For most service systems, the determination of the staffing levels needed to meet service objectives is a very challenging problem. Traditional methods assume that the number of calls arriving within a given time period can be forecasted reasonably accurately and thus model these call arrivals as a Poisson process with a fixed parameter. However, the magnitude of forecasting errors that we have observed in our dataset suggests that the assumption of accurate forecasts is often invalid, and that the arrival rates themselves might be better modeled as random variables. Moreover, we have

shown that ignoring this arrival rate variability can have a significant negative impact on system performance.

In addition, the results presented in this article suggest several areas for future research; for example, the performance measures presented in Sections 3 and 4 suggest alternative methods for determining staffing levels, which will be higher than those suggested by the typical approaches. By accounting explicitly for arrival rate variability, such methods will lead to (probabilistically) shorter customer waiting times and lower call abandonment rates—but at a cost of increased staffing levels. It would be worthwhile to investigate techniques for choosing staffing levels while accounting for the trade-off between service quality and staffing costs in the context of arrival rate variability.

A related result from our analysis is that even when one chooses staffing levels to ensure that the long-run fraction of customers who experience reasonable waiting times is high, there can still be periods when the customer experience is very poor. This happens because of a mismatch between the *realized* arrival rate and the number of servers available to handle the customer load (although we do not explicitly consider it in this article, absenteeism could also contribute to such a mismatch).

A simple way to avoid such problems is to hire extra staff to ensure that the system can handle a larger-than-foreseen arrival rate without excessive customer waiting times and abandonment. However, hiring additional staff as "insurance" for such high traffic days is very expensive while also resulting in excessive staff on the days when the realized arrival rate is lower than average. This suggests that contracting call center staff, also commonly referred to as "outsourcing," is an important area for additional research. Because of the rapid growth in the call center outsourcing industry, contracting has recently been explored by several researchers, including Gans and Zhou [11] and Milner and Lennon-Olsen [18]. Modeling arrival rate variability and exploring contract structures in this context is of both theoretical and practical interest.

In practice, call volume forecasts are often updated on an intraday basis, as discussed by Avramidis et al. [1], Weinberg et al. [25], and Shen and Huang [22]. In addition, staffing levels can be modified in some manner (overtime shifts, voluntary time off, etc.) to account for the updated forecast as explored by Mehrotra et al. [16]. However, such intraday updating of forecasts and staffing levels is typically not captured in optimal staff scheduling models, partly due to the need for mathematical simplicity, but also because such decisions are usually made in a somewhat ad hoc way. This staff scheduling optimization problem is, in the language of stochastic programming, a classical "recourse" problem and is a natural extension of the research presented in this article.

Finally, our approach for estimating forecast errors was chosen for its mathematical tractability and for clarity of exposition, rather than for accurately capturing the features of the data. We view it as a "first-order" model that captures the primary aspects of forecast errors, but for more accurate predictions of both long and short-run performance, one might try to extend our analysis to more realistic models of forecast errors.

Acknowledgments

We would like to thank an anonymous referee for suggestions that improved the exposition. This work was supported in part by National Science Foundation grant DMI-0400287.

References

1. Avramidis, A.N., Deslauriers, A., & L'Ecuyer, P. (2004). Modeling daily arrivals to a telephone call center. *Management Science* 50(7): 896–908.
2. Bassamboo, A., Harrison, J., & Zeevi, A. (2006). Design and control of a large call center: Asymptotic analysis of an LP-based method. *Operations Research* 54: 419–435.
3. Billingsley, P. (1968). *Convergence of probability measures*. New York: Wiley.
4. Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., & Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association* 100: 36–50.
5. Brown, L.D., Zhang, R., & Zhao, L. (2001). Root un-root methodology for non parametric density estimation. Technical report, Department of Statistics, University of Pennsylvania, Philadelphia.
6. Brown, L.D. & Zhao, L.H. (2002). A test for the Poisson distribution. *Sankhya* 64 (A-3): 611–625.
7. Chen, B.P.K. & Henderson, S.G. (2001). Two issues in setting call center staffing levels. *Annals of Operations Research* 108: 175–192.
8. Cinlar, E. (1972). Superposition of point processes. In P.A.W. Lewis (ed.), *Stochastic point processes: Statistical analysis, theory, and applications*. New York: Wiley Interscience.
9. Deslauriers, A., L'Ecuyer, P., Pichitlamken, J., Ingolfsson, A., & Avramidis, A.N. (2007). Markov chain models of a telephone call center in blend mode. *Computers and Operations Research* 34(6): 1616–1645.
10. Gans, N., Koole, G., & Mandelbaum, A. (2003). Telephone call centers: Tutorial, review and research prospects. *Manufacturing Service Operations Management* 5: 79–141.
11. Gans, N. & Zhou, Y.-P. (2007). Call-routing schemes for call-center outsourcing. *Manufacturing & Service Operations Management* 9(1): 33–50.
12. Grassmann, W.K. (1988). Finding the right number of servers in real-world queuing systems. *Interfaces* 18(2): 94–104.
13. Harrison, J.M. & Zeevi, A. (2005). A method for staffing large call centers using stochastic fluid models. *Manufacturing & Service Operations Management* 7: 20–36.
14. Jongbloed, G. & Koole, G. (2001). Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry* 17: 307–318.
15. Liptser, R.S. & Shiryayev, A.N. (1989). *Theory of Martingales*. Boston: Kluwer Academic.
16. Mehrotra, V., Ozluk, O., & Saltzman, R. (2009). Intelligent procedures for intra-day updating of call center agent schedules. *Production and Operations Management*.
17. Meyn, S.P. & Tweedie, R.L. (1993). *Markov chains and stochastic stability*. London: Springer-Verlag.
18. Milner, J. & Lennon-Olsen, T.M. (2007). Service level agreements in call centers: Perils and prescriptions. *Management Science* 54(2): 238–252.
19. Robbins, T.R., Medeiros, D.J., & Dum, P. (2006). Evaluating arrival rate uncertainty in call centers. In L.F. Perrone, F.P. Wieland, J. Liu, B.G. Lawson, D.M. Nicol, & R.M. Fujimoto (eds.), *Proceedings of the 2006 winter simulation conference*. Piscataway NJ: IEEE, pp. 2180–2187.
20. Ross, A.M. (2001). Queueing systems with daily cycles and stochastic demand with uncertain parameters. Ph.D. dissertation, University of California, Berkeley, Berkeley.
21. Serfling, R.J. (1980). *Approximation theorems of mathematical statistics*. New York: Wiley.
22. Shen, H. & Huang, J.Z. (2008). Interday forecasting and intraday updating of call center arrivals. *Manufacturing & Service Operations Management* 10(3): 391–410.
23. Steckley, S.G., Henderson, S.G., & Mehrotra, V. (2005). Performance measures for service systems with a random arrival rate. In M.E. Kuhl, N.M. Steiger, F.B. Armstrong, & J.A. Joines (eds.), *Proceedings of the 2005 winter simulation conference*. Piscataway, NJ: IEEE, pp. 566–575.

24. Thompson, G.M. (1999). Server staffing levels in pure service environments when the true mean daily customer arrival rate is a normal random variate. Unpublished manuscript.
25. Weinberg, J., Brown, L.D., & Stroud, J.R. (2006). Bayesian forecasting of an inhomogeneous poisson process with applications to call center data. *Journal of the American Statistical Association* 102(480): 1185–1198.
26. Whitt, W. (1976). Bivariate distributions with given marginals. *Annals of Statistics* 4: 1280–1289.
27. Whitt, W. (1999). Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters* 24: 205–212.
28. Whitt, W. (2002). *Stochastic-process limits*. Springer Series in Operations Research. New York: Springer.
29. Whitt, W. (2006). Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management* 15: 88–102.
30. Wolff, R.W. (1989). *Stochastic modeling and the theory of queues*. Englewood Cliffs, NJ: Prentice Hall.

APPENDIX A

Proofs of Selected Results

PROOF OF PROPOSITION 2.1: If $\alpha = \infty$, then, in what follows, interpret $c/\alpha = 0$ for any finite c. We have that $\hat{\alpha}_n \to \alpha$ a.s. if and only if $(s_n^2 - 1)/\bar{\lambda}_n - 1/\alpha \to 0$ a.s., and this occurs if and only if

$$\frac{\sum_{k=1}^{n}(Z(k)^2 - (1 + \lambda(k)/\alpha))}{n\bar{\lambda}_n} - \frac{\bar{Z}_n^2}{\bar{\lambda}_n} \to 0 \quad \text{a.s.} \tag{A.1}$$

The proof follows if we show that each of the terms in (A.1) converges to 0 a.s. as $n \to \infty$. To do so we use the following martingale strong law of large numbers from Liptser and Shiryayev [15, p. 144] for each term separately. Let $(\mathcal{F}_n : n \geq 0)$ be a filtration, i.e., an increasing sequence of sigma fields. ■

THEOREM A.1 (Liptser and Shiryayev [15]): *Let $(M_n, \mathcal{F}_n : n \geq 0)$ be a square-integrable martingale with $M_0 = 0$. Let $(L_n : n \geq 0)$ be nondecreasing in n with $L_n \in \mathcal{F}_n$ for all n. Define*

$$V_n = \sum_{k=1}^{n} \mathbb{E}((M_k - M_{k-1})^2 | \mathcal{F}_{k-1})$$

and assume that

$$\sum_{n=1}^{\infty} \frac{V_{n+1} - V_n}{(1 + L_n)^2} < \infty \quad a.s. \tag{2}$$

and $\mathbb{P}(L_\infty = \infty) = 1$, where $L_\infty = \lim_{n\to\infty} L_n$. Then $M_n/L_n \to 0$ a.s.

For $n \geq 0$, define $\mathcal{F}_n = \sigma\{N(1), \ldots, N(n), \lambda(1), \ldots, \lambda(n+1)\}$.

First, we show that $\bar{Z}_n^2/\bar{\lambda}_n \to 0$ a.s., as $n \to \infty$. Define

$$M_n = \sum_{k=1}^{n} Z(k)$$

and

$$L_n = n\sqrt{\bar{\lambda}_n},$$

so that

$$\begin{aligned} V_n - V_{n-1} &= \mathbb{E}[Z(n)^2|\mathcal{F}_{n-1}] \\ &= 1 + \lambda(n)/\alpha \\ &\leq 1 + \lambda^*/\alpha. \end{aligned}$$

Hence,

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{V_{n+1} - V_n}{(1+L_n)^2} &\leq \sum_{n=1}^{\infty} \frac{1 + \lambda^*/\alpha}{n^2\bar{\lambda}_n} \\ &< \infty. \end{aligned}$$

We conclude that $M_n/L_n \to 0$ a.s. as $n \to \infty$, and so $\bar{Z}_n^2/\bar{\lambda}_n \to 0$ a.s. as $n \to \infty$.

To show that the other term in (A.1) converges to zero, define

$$M_n = \sum_{k=1}^{n} [Z(k)^2 - (1 + \lambda(k)/\alpha)]$$

and

$$L_n = n\bar{\lambda}_n,$$

so that

$$\begin{aligned} V_n - V_{n-1} &= \mathbb{E}\left[(Z(n)^2 - (1 + \lambda(n)/\alpha))^2|\mathcal{F}_{n-1}\right] \\ &= \mathbb{E}\left[(Z(n)^2 - (1 + \lambda(n)/\alpha))^2|\lambda(n)\right]. \end{aligned} \tag{3}$$

Direct calculation shows that (3) equals $\sum_{i=-1}^{4} a_i\lambda(n)^i$ for certain coefficients a_i that are functions of α. We assumed that $\lambda(n)$ was bounded away from zero and infinity, and hence (3) is uniformly bounded in n, by b say. It immediately follows that

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{V_{n+1} - V_n}{(1+L_n)^2} &\leq \sum_{n=1}^{\infty} \frac{b}{(n\bar{\lambda}_n)^2} \\ &\leq \sum_{n=1}^{\infty} \frac{b}{n^2\lambda_*^2} \\ &< \infty, \end{aligned}$$

so that (2) holds and the proof is complete. ■

PROOF OF PROPOSITION 3.3: For notational simplicity we suppress the dependence on γ. For $u \geq 0$, let $U(u) = I(X(u) \in A)$, where $I(\cdot)$ is the indicator function that is 1 if its argument is true and 0 otherwise. Then U is left-continuous and has right limits because the same is true of X. Let $N = (N(u) : u \geq 0)$ be the Poisson arrival process (with rate γ). For arbitrary $v \geq 0$, $(N(v+u) - N(v) : u \geq 0)$ is independent of $(U(u) : 0 \leq u \leq v)$ and $(N(u) : 0 \leq u \leq v)$. Then $s(\gamma) = \gamma\, \mathbb{E}_v \int_0^1 U(u)\, du$ by the PASTA result (e.g., Wolff [30, §5.16]). By Fubini's theorem, for arbitrary $v \geq 0$, $\mathbb{E}_v \int_0^v U(u)\, du = \int_0^v \mathbb{E}_v\, U(u)\, du$. Therefore,

$$\mathbb{E}_v \int_0^v U(u)\, du = \int_0^v \mathbb{P}_v(X(u) \in A)\, du. \tag{A.4}$$

Taking $v = 1$, it follows that $s(\gamma) = \gamma \int_0^1 \mathbb{P}_v(X(u) \in A)\, du$.

For the second result, the system is in steady state at time 0 so that $v = \pi$. However, $\mathbb{P}_\pi(X(u) \in A) = \mathbb{P}_\pi(X(0) \in A)$ for all $u \geq 0$. Defining $f(\gamma) = \mathbb{P}_\pi(X(0) \in A)$, we see from (A.4) that

$$s(\gamma) = \gamma\, \mathbb{E}_\pi \int_0^1 U(u)\, du = \gamma f(\gamma). \tag{A.5}$$

To see that $f(\gamma)$ can be interpreted as the long-run fraction of customers that receive satisfactory service, define the stochastic process $R = (R(v) : v \geq 0)$, where $R(v) = \int_0^v U(u)\, \mathrm{d}N(u)$. Then the fraction of customers that have received satisfactory service by time v is given by $R(v)/N(v)$. Now, it is assumed that $R(v)/N(v)$ converges to some constant p as $v \to \infty$ a.s., where p is the long-run fraction of customers that receive satisfactory service. We show that $p = f(\gamma)$. Since $R(v)/N(v)$ converges to p, it follows that $\int_0^v U(u)\, \mathrm{d}u/v$ also converges to p as $v \to \infty$, from the PASTA result (e.g., Wolff [30, §5.16]). But $p = \mathbb{E}_v\, p = \mathbb{E}_v \lim_{v\to\infty} (1/v) \int_0^v U(u)\, \mathrm{d}u$. The bounded convergence theorem establishes that we can exchange the limit and expectation, and this, together with (A.5), completes the proof. ■

APPENDIX B

Computing the Variance Constants in the Central Limit Theorem

Let $\delta(i)$ denote the rate at which the CTMC X leaves state i and let π_X and π_Y denote the steady-state distributions associated with X and Y, respectively. Since

$$\pi_X(y) = \frac{\pi_Y(y)/\delta(y)}{\sum_z \pi_Y(z)/\delta(z)},$$

it follows that

$$\beta = \sum_{y=0}^{d} \pi_X(y)\delta(y) = \left(\sum_{z=0}^{d} \pi_Y(z)/\delta(z) \right)^{-1}.$$

Note that π_X or π_Y are easily computed and therefore so is β.

We also need to compute u and v. These are given by

$$u = \sum_{i=0}^{c-1} \pi_Y(i) P_Y(i, i+1)$$

and

$$v = \sum_{i=0}^{d-1} \pi_Y(i) P_Y(i, i+1),$$

where P_Y is the transition matrix of Y.

Finally, recall that for $1 \le i, j \le 2$,

$$\begin{aligned}\Sigma_{ij} &= \mathbb{E}_{\tilde{\pi}}[\tilde{g}_i(\tilde{Y}_0)\tilde{h}_j(\tilde{Y}_0) + \tilde{h}_i(\tilde{Y}_0)\tilde{g}_j(\tilde{Y}_0) - \tilde{h}_i(\tilde{Y}_0)\tilde{h}_j(\tilde{Y}_0)] \\ &= \sum_{x,y} \pi_Y(x) P_Y(x,y)[\tilde{g}_i(x,y)\tilde{h}_j(x,y) + \tilde{h}_i(x,y)\tilde{g}_j(x,y) - \tilde{h}_i(x,y)\tilde{h}_j(x,y)].\end{aligned}$$

It remains to specify how to compute $\tilde{g}_i(x,y)$. Define

$$h_i(x) = \mathbb{E}_x \tilde{h}_i(x, Y_1) = \sum_{y=0}^{d} \tilde{h}_i(x,y) P_Y(x,y)$$

to be the "smoothed" version of $\tilde{h}_i$, for $i = 1, 2$ and $x = 0, \ldots, d$. There are multiple solutions to the equations defining $\tilde{g}_i$, all of which differ by an additive constant. In what follows we use one such solution for $\tilde{g}_i$, which is

$$\begin{aligned}\tilde{g}_i(x,y) &= \sum_{k=0}^{\infty} \mathbb{E}_{(x,y)} \tilde{h}_i(Y_k, Y_{k+1}) \\ &= \tilde{h}_i(x,y) + \sum_{k=1}^{\infty} \mathbb{E}_{(x,y)} \tilde{h}_i(Y_k, Y_{k+1}) \\ &= \tilde{h}_i(x,y) + \sum_{k=1}^{\infty} \mathbb{E}_{(x,y)} h_i(Y_k) \\ &= \tilde{h}_i(x,y) + g_i(y),\end{aligned}$$

where

$$g_i(y) = \sum_{k=0}^{\infty} \mathbb{E}_y h_i(Y_k)$$

solves $(P_Y - I)g_i(y) = -h_i(y)$ for all y, and has the property that $\pi_Y g_i = 0$. It is therefore possible to compute g_i from these latter relations and then substitute back to obtain $\tilde{g}_i$.