Markov chain models of a telephone call center with call blending

Alexandre Deslauriers • Pierre L'Ecuyer • Juta Pichitlamken* •
Armann Ingolfsson† • Athanassios N. Avramidis

GERAD and
Département d'Informatique et de Recherche Opérationnelle
Université de Montréal, C.P. 6128, Succ. Centre-Ville
Montréal, H3C 3J7, CANADA

**Department of Industrial Engineering, Faculty of Engineering*
Kasetsart University, Bangkok, THAILAND

†*School of Business, University of Alberta, Edmonton, Alberta, T6G 2R6, CANADA*

deslauriers.Alexandre@hydro.qc.ca • lecuyer@iro.umontreal.ca •
fengjtp@ku.ac.th • armann.ingolfsson@ualberta.ca • avramidi@iro.umontreal.ca

Motivated by a Bell Canada call center operating in blend mode, we consider a system with two types of traffic and two types of agents. Outbound calls are served only by blend agents, whereas inbound calls can be served by either inbound-only or blend agents. Inbound callers may balk or abandon. There are several performance measures of interest, including the rate of outbound calls and the proportion of inbound calls waiting more than some fixed number of seconds. We present a collection of continuous-time Markov chain (CTMC) models which capture many real-world characteristics while maintaining parsimony that results in fast computation. We discuss and explore the tradeoffs between model fidelity and efficacy and compare our different CTMC models with a realistic simulation model of a Bell Canada call center, used as a benchmark.

1. Introduction

Telephone call centers are an important part of customer service of many organizations. Managing their operations more efficiently attracts much interest as exemplified by a growing body of academic work in various disciplines (see Gans et al. 2003 and Mandelbaum 2003 for extensive overviews). From the operational perspective, most call centers face common challenges such as uncertainties in call arrivals and service times while having to respect certain quality-of-service constraints.

In this paper, we consider a telephone call center with two types of traffic, *inbound* and *outbound*, and two types of agents, *inbound-only* and *blend*. Inbound calls arrive randomly, according to some stochastic process. When traffic is too high, new inbound calls must wait

in a queue. For inbound traffic, we consider *abandonment*, i.e., some customers may not stay in the queue once learning that they are put on hold, or they may leave after spending some time waiting in the queue. When the inbound traffic is low and some blend agents are idle, an automatic *dialer* composes multiple outbound calls in parallel (trying to reach potential customers, e.g., for marketing or direct sales), in order to increase the productivity of the center. *Mismatches* occur when more customers are reached by outbound calls than the number of idle agents. The outbound calls are served only by blend agents, whereas inbound calls can be served by either type.

Managers are interested in performance measures such as agent utilization, abandonment rate, rate of outbound calls, rate of mismatches, fraction of calls waiting more than τ seconds for some constant τ, etc., in the long run. They often want to determine a minimal *staffing* (the number of agents of each kind in the center as a function of time) under certain (stochastic) constraints on the quality of service and the volume of outbound calls completed. Ultimately, they also need to find a daily or weekly *schedule* for a certain number of individual agents. This imposes additional constraints which imply that not all staffings can be realized exactly. For example, each agent must work a minimum number of hours during the day, these hours must be contiguous with a lunch break near the middle, etc. Minimal-cost scheduling is generally more difficult than minimal-cost staffing. Both can be formulated as stochastic integer-programming problems after an appropriate model of the system is defined.

Realistic models of call centers are generally so complex that they can only be handled via stochastic simulation. Typically, the inbound calls do not arrive according to a stationary Poisson process, the call durations are not exponential random variables and their distribution may depend on the time of the day, the number of agents of each type in the center varies from day to day and within each day, and so on. However, running simulation models to determine the staffing and/or the agent schedules in a call center is sometimes too slow. Simplified models that can be solved more quickly, either by analytic formulas or numerically, can be more convenient and appropriate when a fast response is needed. These models must rely on several unrealistic assumptions, so the answers they provide are only rough-cut approximations. But these approximations are often much more useful than precise answers that come too late. For this reason and because of their simplicity, such approximations are widely used in the case of inbound-only call centers (e.g., the Erlang-C formula and the "square root" rule).

A call center can be naturally viewed as a queueing system. With drastic simplifications, one may obtain queueing formulas for the performance measures of interest (see Koole and Mandelbaum 2002 for a recent overview of queueing models in call center applications). The so-called *Erlang-C formula* and its variations have traditionally been used to model call centers with inbound traffic only. In that context, the center is modelled as an $M/M/n$

queue, with Poisson arrivals, exponential service times (time spent by a customer with an agent), identical servers (agents), and no customer abandoning the queue or receiving a busy signal. The $M/M/n$ model is appealing because the number of calls in the system as a function of time is then a continuous-time Markov chain (CTMC) whose steady-state (long-run) probabilities are easily determined. From this probability distribution, the long-run performance measures of interest can be conveniently computed.

The $M/M/n$ model has been modified to accommodate features such as abandonment, blocking, time non-stationarity, and outbound traffic. We first describe some earlier work in this area before explaining our enhancements.

Brandt and Brandt (1999a) allow customer abandonment via the concept of an *impatient customer* who has a generally-distributed *maximal patience time* beyond which he abandons the queue. In addition, arrival and service rates may be dependent on the number of customers in the system. Brandt and Brandt (1999b) also model a secondary queue (e.g., call-back service queue) of lower priority than the inbound traffic queue. Brown et al. (2005) fit the $M/M/n$ model that is augmented with the exponential patience time (a.k.a. "Erlang-A" with A for abandonment) to actual call center data. They find that the Erlang-A model provides a useful approximation to performance measures such as the average waiting time and the fraction of customers experiencing positive waiting times.

The aim of this paper is to develop practical CTMC models for call centers operating in *blend mode*, i.e., outbound calls are initiated when the inbound traffic is low. Operating in blend mode is appealing because it improves agent productivity. For this reason, it has become popular in modern call centers, but it also increases the complexity of the system. No single model can be the most appropriate solution for all situations, because certain simplifying assumptions are reasonable in some cases and totally unrealistic in other cases. For example, in the model, we may have to distinguish the inbound and outbound calls being served if their service time distributions differ significantly, and not if they are similar. For this reason, it is appropriate to define a collection of models, as we do in this paper. The simplest model could be the right tool for one call center while a more detailed model might be needed for another center.

We propose five CTMC models with varying degree of complexity. In the simplest model, M_1, all agents are identical blend agents, and the inbound and outbound service times are i.i.d. exponential (therefore, from the agents' point of view, inbound and outbound calls are indistinguishable). Outbound calls are made only one at a time, so mismatches never occur. Model M_2 differs from M_1 only in that M_2 allows parallel outbound call dialing, which sometimes causes mismatches. In Model M_3, inbound and outbound calls are differentiated, and inbound-only and blend agents are distinguished.

Being the richest model, M_3 is also the most costly to compute; therefore, we provide two special cases of M_3 that are less demanding in computation. In M_4, the expected inbound

service time is equal to that of the outbound service time, and inbound and blend agents are differentiated. Complementary to M_4 is M_5, where every agent is blend, but the expected inbound service time is allowed to be different from the expected outbound service time.

All five models are time-stationary; however, we explain how to extend them to more realistic non-stationary and doubly stochastic arrival processes, and how to use M_1, M_2 and M_4 as approximations when the inbound and outbound service times have different means. All these models being CTMCs, all interarrival times, service times, patience times, etc., are exponential. Non-exponential times could be considered in principle via phase-type distributions, but this would enlarge the state space and make the models much slower to solve.

In our models, the dialer automatically determines when to make outbound calls and how many, as a function of the current state of the system, using a *threshold-type* policy: the number of outbound calls to attempt is a non-decreasing function of the number of idle blend agents. This is motivated by the results of Bhulai and Koole (2003). Essentially under the assumptions of the $M/M/n$ model with outbound calls in the background, and if the objective is to maximize the rate of outbound calls subject to a steady-state mean delay constraint for inbound calls, these authors showed that a threshold-type policy for initiating outbound calls is optimal when the inbound and outbound calls have the same expected service times, and is close to optimal otherwise.

In the case where service times (call durations) have a small coefficient of variation, one could think of using the *elapsed times* of calls to form predictors of their *residual times* and define a *predictive dialing policy* based on that information. Such a policy would initiate dialing whenever the estimated residual time of an on-going call becomes small enough (e.g., near the average time for reaching a customer). This type of strategy is discussed by Samuelson (1999). In CTMC models, however, service times are always assumed to be exponential, which means that the elapsed time gives no information on the residual service time, because of the memoryless property of the exponential distribution. We have also observed that in real-life call centers, the service times actually have *more variability* than for the exponential distribution and an increasing hazard rate function. In that case, a longer elapsed time means a longer expected residual time.

Our contributions are: First, we define and study CTMC models simple enough to allow fast computation of their steady-state probabilities, especially for M_1, M_2, and M_4, while capturing many real-world characteristics (e.g., to our knowledge, mismatched and failed outbound calls have not been incorporated in CTMC models of call centers before). We also develop methods to compute various call center performance measures with these models. The models can provide an approximation of the number of agents needed to satisfy the waiting time requirement. Second, we provide further approximation techniques to handle non-stationary and doubly stochastic arrival processes. Third, we explore the tradeoff

between model fidelity and efficacy through an empirical study where we use a realistic simulation model of a Bell Canada call center as a benchmark.

The remainder of the paper is organized as follows: In the next section, we specify the CTMC models. We use the steady-state probabilities for each model to obtain call center performance measures, as discussed in Section 3. In Section 4, we develop a heuristic approach for relaxing the assumption of equal average service times of M_1, M_2, and M_4, and we address the non-stationarity of actual call centers and a doubly stochastic version of the Poisson arrival process. In Section 5, we compare the performance of the CTMC models and their agreement with the simulation results for an example of a realistic call center. We also explore the sensitivity of our results to selected assumptions. Based on these comparisons, we provide suggestions on when each model is appropriate. Section 6 briefly outlines how the CTMC models might be used for optimal staffing and scheduling. We conclude in Section 7 with a summary and future research directions.

2. CTMC models

We present the CTMC models and explain how to compute their steady-state probabilities in Sections 2.1–2.3. Then we use the steady-state probabilities to compute performance measures that are relevant to call center applications in Section 3.

2.1 Model M_1: all blend agents and no mismatches

First, we describe modelling assumptions underlying M_1: Our call center consists of n identical blend agents with a single FIFO waiting queue of finite capacity c. Inbound calls arrive according to a Poisson process with rate λ. Service times for inbound and outbound calls are i.i.d. exponentially distributed with rate μ. Customers that are not served immediately hang up with probability $1-\gamma$; otherwise, they join the queue from which they will abandon if their waiting time is greater than a maximal *patience time*. The patience times are exponentially distributed with mean $1/\eta$ and are independent for different customers.

The automatic dialer of M_1 uses a threshold-type policy to schedule outbound calls: The dialer attempts to make an outbound call if there are $\dot{n}$ or fewer busy agents, where the pre-determined threshold $\dot{n}$ satisfies $1 \leq \dot{n} \leq n$. The time from when the dialer dispatches a call until it registers a successful or failed attempt is exponentially distributed with mean $1/\nu$. Each outbound call is answered by a customer with probability κ. All of this is modeled via the state transition rates of the CTMC. As a consequence, if an arrival occurs while the dialer waits for a customer to answer, and if the number of busy agents then exceeds $\dot{n}$, then the dialing in progress is simply stopped (this is implicit in the definition of λ_k below).

The state variable $X_1(t)$ is the total number of calls—inbound and outbound—in the system at time t. Under M_1 assumptions, $\{X_1(t); t \geq 0\}$ is a CTMC with state space

$S_1 = \{0, 1, 2, ..., n + c\}$. Because $X_1(t) = k$ can change only to $k \pm 1$, it is a birth-and-death process, where the birth rates λ_k and death rates μ_k are state-dependent as follows:

$$\lambda_k = \begin{cases} \lambda + \kappa\nu, & k = 0, 1, \ldots, \dot{n} \\ \lambda, & k = \dot{n} + 1, \ldots, n - 1 \\ \gamma\lambda, & k = n, \ldots, n + c - 1 \\ 0, & \text{otherwise} \end{cases}$$

$$\mu_k = \begin{cases} k\mu, & k = 1, 2, .., n - 1 \\ n\mu + (k - n)\eta, & k = n, \ldots, n + c \\ 0, & \text{otherwise.} \end{cases}$$

The stationary probabilities $\{\pi_0, \pi_1, \ldots, \pi_{n+c}\}$ can be determined recursively from the birth and death rates (for example, see Ross 1983 or Taylor and Karlin 1998). They are given in Appendix A.1.

2.2 Model M_2: all blend agents with parallel dialing and mismatches

One of the limitations of M_1 is that mismatched calls are neglected. In practice, call center managers regard mismatches as highly undesirable, and they control mismatches (on an operational basis) by manipulating the dialer policy, which dictates the number of parallel outbound calls to attempt, given the state of the system. In this section, we modify the dialer of M_1 to allow for the possibility of mismatches.

At each end-of-service epoch, the dialer of M_2 attempts to make multiple outbound calls in parallel. The dialer composes outbound calls only when there are $\dot{n}$ or fewer busy agents. The number of outbound calls dialed is $v(I) \geq 0$, where $v(I)$ is a pre-determined function of the number of idle agents I. We assume that the dialer recognizes whether a call is answered *instantaneously* as soon as a call is dispatched. This simplifying assumption is to keep the state space unidimensional. Because multiple outbound calls are made in parallel, mismatched calls can occur under M_2. Specifically, when there are k calls in the system, and z outbound calls are answered, then $\max(0, k + z - n)$ of the outbound calls are mismatched, and lost. Because each call is answered with probability κ, the number of answered outbound calls Z is a binomial random variable, with probability mass function:

$$\phi_I(z) = \binom{v(I)}{z} \kappa^z (1 - \kappa)^{v(I)-z} \qquad \text{for } 0 \leq z \leq v(I). \tag{1}$$

Another possibility would be to assume the following. In states for which there are no more than $\dot{n}$ busy agents and where $v(I) > 0$, a *dialer-reaching-customers* event occurs at

some constant rate r. When such an event occurs, Z outbound calls are answered where Z is a binomial random variable with mass function $\phi_I(z)$ defined in (1). In other words, the dialer-reaching-customers events would occur according to a stationary Poisson process with rate r, and the number of customers reached at such an event would be a binomial with parameter $v(I)$ that depend on the current state at the time of that event. In some states $v(I)$ would be zero. It would not be difficult to modify the transition probabilities and construct the infinitesimal generator for this variant of our model.

On the other hand, modeling nonzero dial resolution delays that are *independent* across customers would require an extra state variable (the number of pending dials). This would make the CTMC model more complicated and more costly to solve.

Aside from the outbound calling process, the other assumptions of Model M_1 still hold under M_2. Model M_2 involves the following transition types:

1. **Inbound arrival:** State changes from k to $k+1$, for $k < n+c$, at rate λ for $k < n$ and at rate $\lambda\gamma$ for $n \le k < n+c$.

2. **Abandonment:** State changes from k to $k-1$, for $k > n$, at rate $(k-n)^+\eta$.

3. **Service completion without subsequent outbound dialing:** If the current state is k, then the number of busy servers immediately after the service completion will be $\min(k,n)-1$. If $\min(k,n)-1 > \dot{n}$, then no outbound dialing will occur and the state will change to $k-1$, at rate $\min(k,n)\mu$.

4. **Service completion followed by z outbound calls that are answered:** This transition is possible when the current state k satisfies $k-1 \le \dot{n}$, resulting in $v(n-k+1)$ dialed outbound calls. Note that $m = \min(z, n-k+1)$ of the z answered calls will begin service and $z-m$ answered calls will result in mismatches. This transition occurs at the rate $k\mu\phi_{n-k+1}(z)$. Here, we have a family of transitions types that correspond to all z such that $0 \le z \le v(n-k+1)$.

The state variable for M_2 is $X_2(t)$, the total number of calls in the system at time t. The process $\{X_2(t),\, t \ge 0\}$ is a CTMC whose inifinitesimal generator Q_2 can be constructed from the transition types listed above. The state space S_2 for M_2 is the same as for M_1. The steady state probability vector π can then be found by solving $\pi Q_2 = 0$ and $\sum_{k\in S_2} \pi_k = 1$.

2.3 Model M_3: two types of agents

Model M_3 keeps the dialer of M_2, but it distinguishes between inbound and outbound agents, and the service times of inbound and outbound calls may have different means. There are n_1 inbound-only agents who serve only inbound calls and n_2 blend agents who can process both inbound and outbound calls. The total number of agents is thus $n = n_1 + n_2$. Service

times for inbound and outbound calls are i.i.d. exponentially distributed with mean $1/\mu_1$ and $1/\mu_2$, respectively. The outbound calling process of M_3 is almost identical to that of M_2 except that the parallel outbound calls are initiated only when there are at most $\dot{n}$ busy agents (of any type), and at least one blend agent is idle. Given that these conditions are satisfied, the number of attempted outbound calls is $v(I_2)$, a pre-specified function of the number of idle blend agents I_2. Note that this function can be zero for small values of I_2, thus implementing a threshold on the number of idle blend agents. If an incoming call arrives when both inbound agents and blend agents are available, it is serviced by an inbound agent.

The following processes describe various aspects of model M_3:

$$\begin{aligned}
B_1(t) &= \text{number of busy inbound agents} \\
I_1(t) &= n_1 - B_1(t) = \text{number of idle inbound agents} \\
Q(t) &= \text{number of waiting inbound calls} \\
B_{21}(t) &= \text{number of blend agents serving inbound calls} \\
B_{22}(t) &= \text{number of blend agents serving outbound calls} \\
B_2(t) &= \text{number busy blend agents} \\
I_2(t) &= n_2 - B_2(t) = \text{number of idle blend agents} \\
B(t) &= B_1(t) + B_2(t) = \text{total number of busy agents}
\end{aligned}$$

We will view $X_3(t) = (B_1(t), B_{21}(t), B_{22}(t), Q(t))$ as the state variable for M_3. The "supplementary variables" $I_1(t)$, $B_2(t)$, $I_2(t)$, and $B(t)$ are uniquely determined by the value of the state variable. Using lowercase letters to denote the values of processes at a point $s = (b_1, b_{21}, b_{22}, q)$ in the sample space, we can express the state space as

$$\begin{aligned}
S_3 &= \{s : b_1, b_{21}, b_{22} \geq 0, b_1 \leq n_1, b_{21} + b_{22} \leq n_2, q = 0\} \\
&\quad \cup \{s : b_1 = n_1, b_{21} + b_{22} = n_2, 0 < q \leq c\},
\end{aligned}$$

with cardinality $(n_1 + 1)(n_2 + 2)(n_2 + 1)/2 + (n_2 + 1)c$. As the expression for the state space shows, S_3 is three-dimensional, although we use four state variables, for notational convenience. The state variables $B_1(t)$ and $Q(t)$ could be replaced by their sum, but this would complicate some of the expressions later in this subsection.

We define the process in terms of the following eight transition types, defined by the destination state from a generic origin state s, the conditions that the origin state must satisfy, and the transition rate. This information suffices to construct the infinitesimal generator matrix Q_3 and to calculate stationary probabilities. We use the supplementary variables i_1, b_2, i_2, and b to simplify some of the expressions, and we use I to denote the indicator function.

1. **Inbound arrival:**

- Destination state: $(b_1 + I\{b_1 < n_1\}, b_{21} + I\{b_1 = n_1, b_2 < n_2\}, b_{22}, q + I\{b = n\})$.
- Condition: $q < c$.
- Rate: $\lambda(1 - (1 - \gamma)I\{b = n\})$.

2. **Abandonment:**
   - Destination state: $(b_1, b_{21}, b_{22}, q - 1)$.
   - Condition: $q > 0$.
   - Rate: $q\eta$.

3. **Inbound agent completes inbound call, no outbound dialing:**
   - Destination state: $(b_1 - I\{q = 0\}, b_{21}, b_{22}, q - I\{q > 0\})$.
   - Condition: $b_1 > 0$ and $(b - 1 > \dot{n}$ or $i_2 = 0)$.
   - Rate: $b_1\mu_1$.

4. **Blend agent completes inbound call, no outbound dialing:**
   - Destination state: $(b_1, b_{21} - I\{q = 0\}, b_{22}, q - I\{q > 0\})$.
   - Condition: $b_{21} > 0$, $b - 1 > \dot{n}$.
   - Rate: $b_{21}\mu_1$.

5. **Blend agent completes outbound call, no outbound dialing:**
   - Destination state: $(b_1, b_{21} + I\{q > 0\}, b_{22} - 1, q - I\{q > 0\})$.
   - Condition: $b_{22} > 0$, $b - 1 > \dot{n}$.
   - Rate: $b_{22}\mu_2$.

6. **Inbound agent completes inbound call, followed by z outbound calls that are answered:** Note that some of the z answered calls may result in mismatches. Specifically, $m = \min(z, i_2)$ of the outbound calls will be connected to blend agents. This transition type represents a family of transitions, for all z such that $0 \leq z \leq v(i_2)$.
   - Destination state: $(b_1 - 1, b_{21}, b_{22} + m, 0)$.
   - Condition: $b_1 > 0, b - 1 \leq \dot{n}$ and $i_2 \geq 1$.
   - Rate: $b_1\mu_1\phi_{i_2}(z)$.

7. **Blend agent completes inbound call, followed by z outbound calls that are answered:** Now, $m = \min(z, i_2+1)$ of the z outbound calls will be connected to blend agents and z can range from 0 to $v(i_2 + 1)$.

- Destination state: $(b_1, b_{21} - 1, b_{22} + m, 0)$.
- Condition: $b_{21} > 0, b - 1 \leq \dot{n}$.
- Rate: $b_{21}\mu_1\phi_{i_2+1}(z)$.

8. **Blend agent completes outbound call, followed by z outbound calls that are answered:** As with the previous transition type, $m = \min(z, i_2+1)$ of the z outbound calls will be connected to blend agents and z can range from 0 to $v(i_2+1)$.

   - Destination state: $(b_1, b_{21}, b_{22} - 1 + m, 0)$.
   - Condition: $b_{22} > 0, b - 1 \leq \dot{n}$.
   - Rate: $b_{22}\mu_2\phi_{i_2+1}(z)$.

The last two models, M_4 and M_5, are simplifications of M_3 that have fewer states and thus require less computational time than M_3.

2.4 Model M_4: two types of agents and $\mu_1 = \mu_2$

When the mean service times of inbound and outbound calls are the same ($\mu_1 = \mu_2 = \mu$), we can model the system with a simplified state space compared to M_3: we do not need to distinguish between inbound and outbound calls processed by a blend agent. Maintaining the definitions of $B_1(t)$, $B_2(t)$, and $Q(t)$ from M_3, the state variable for the system is $X_4(t) = (B_1(t), B_2(t), Q(t))$. Denoting a point in the state space by $s = (b_1, b_2, q)$, the state space becomes

$$\begin{aligned} S_4 &= \{s : b_1, b_2, \geq 0, b_1 \leq n_1, b_2 \leq n_2, q = 0\} \\ &\quad \cup \{s : b_1 = n_1, b_2 = n_2, 0 < q \leq c\} \end{aligned}$$

with cardinality $(n_1+1)(n_2+1)+c$. It is important that as the number of agents increases, the state space of M_4 becomes considerably smaller than that of M_3. For example, with $n_1 = n_2 = c = 20$, M_3 has 5,271 states while M_4 has 461 states. As n_1 and n_2 grow larger, the M_4 "savings" compared to M_3 is roughly a *factor* of $n_2/2$. The transition types for M_4 are similar to the ones for M_3 and they are given in detail in Appendix A.2.

2.5 Model M_5: all blend agents and $\mu_1 \neq \mu_2$

In M_5, we assume all agents are blend, so there is no need for distinction between agent types. The mean service times are allowed to differ between inbound and outbound calls. The state variable is $X_5(t) = (B_1(t), B_2(t), Q(t))$, with $B_1(t)$ being the number of agents busy with inbound calls and $B_2(t)$ being the number of agents busy with outbound calls. The cardinality of the state space S_5 is $(n+2)(n+1)/2 + (n+1)c$. Appendix A.3 lists the transition types for M_5.

3. Performance measures

In this section, we show how to use the steady-state probabilities to determine the waiting time distribution and other performance measures of call centers. In order to unify the calculation of various measures across our five models, individual components of the vector state variable in a generic state s are denoted (when convenient) by the lowercase symbol associated to the model's state variable. Thus $b(s)$ denotes the number of busy agents in state s for any model, $i_2(s)$ denotes the number of idle blend agents in state s for models 3 and 4, etc.

3.1 Waiting time distribution

Let $W_q(\ell)$ denote the waiting time in the queue experienced by the ℓth call that enters the system. The end of waiting time is triggered by either service completion or abandonment. We will compute the limiting distribution

$$\bar{F}(\tau) = \lim_{\ell \to \infty} \Pr\{W_q(\ell) > \tau\}.$$

Let $\bar{F}_a(\tau)$ denote the limiting distribution $\bar{F}(\tau)$ for Model M_a, for $a = 1, 2, \ldots, 5$. In each case, this limiting distribution exists because the model is regenerative and aperiodic (since it has a finite state space and there is always a nonzero probability of returning to the empty state after a bounded number of steps). Conditioning on the system state $\widetilde{X}_a(\ell)$ upon arrival of the ℓth call, and using the PASTA property (see Wolff 1989), we get

$$\begin{aligned} \bar{F}_a(\tau) &= \lim_{\ell \to \infty} \sum_{s \in S_a} \Pr\left\{ W_q(\ell) > \tau \,\middle|\, \widetilde{X}_a(\ell) = s \right\} \Pr\{\widetilde{X}_a(\ell) = s\} \\ &= \sum_{s \in S_a} \Pr\left\{ W_q(\ell) > \tau \,\middle|\, \widetilde{X}_a(\ell) = s \right\} \pi_s. \end{aligned} \tag{2}$$

Note that the conditional probabilities in the expression above do not depend on ℓ because the process $\widetilde{X}_a(\cdot)$ is Markovian and stationary. Accordingly, we simplify notation below by dropping "ℓ" with the understanding that conditioning is with respect to the stationary, random system state.

If customers never abandon the queue (i.e., have infinite patience times), an incoming customer must wait until all of the customers who are already in the system (if there are any) finish. Thus, when the service times are i.i.d. exponential, the distribution of positive waiting time, conditional on the customers who are already in the system, is simply gamma (a.k.a. Erlang). In our case, we have abandonments: a customer leaves the queue once his waiting time reaches his patience time. Assuming the patience times are i.i.d. exponential, the following lemma provides the conditional probabilities needed in (2) for models M_1, M_2, and M_4. Here, $\widetilde{B}$ denotes the stationary random number of busy servers and $\widetilde{Q}$ denotes the

stationary random number of waiting inbound calls (these are components of the stationary system state $\widetilde{X}_a$). Notice that Equation (3) bears resemblance to the Erlang distribution.

Lemma 1 (Riordan 1962, pages 110-111) *Suppose the maximal patience times are i.i.d. exponential with mean $1/\eta$. Under Model M_1, M_2, and M_4, we have that for all all $j \geq 0$,*

$$\begin{aligned} f(\tau;\mu) &\stackrel{\text{def}}{=} \Pr\left\{W_q > \tau \,\middle|\, \widetilde{B} = n, \widetilde{Q} = q\right\} \\ &= e^{-\eta\tau}\frac{\psi(\psi+1)\cdots(\psi+q)}{q!}\sum_{j=0}^{q}(-1)^j\binom{q}{j}\frac{e^{-\eta(\psi+j)\tau}}{\psi+j} \\ &= e^{-\eta\tau(1+\psi)}\sum_{j=0}^{q}\frac{(\psi)_j(1-e^{-\eta\tau})^j}{j!}, \end{aligned} \tag{3}$$

where $\psi = n\mu/\eta$, $(\psi)_0 = 1$ and $(\psi)_j = (\psi)(\psi+1)\cdots(\psi+j-1)$ for $j \geq 1$.

We recommend using the last expression of (3) in computations because all terms in the series have the same sign. This expression was developed for us by Richard Simard. With the intermediate expression, given by Riordan (1962), cancellation errors between terms of opposite sign can lead to totally wrong results. This expression, combined with the steady state probabilities π_s, allows us to calculate $\bar{F}_a(\tau)$ for models 1, 2, and 4, because in these models, service times of inbound and outbound calls are i.i.d. An alternative definition of QoS could use instead the *virtual wait time* W_q, defined as the stationary wait time for an infinite-patience customer; to get the tail probability of this random variable, one simply divides the right-hand side of (3) by $e^{-\eta\tau}$.

In models M_3 and M_5, the mean service times differ across inbound and outbound calls. We do not have closed-form expressions for the conditional probabilities required in (2) for these models. In principle, one could develop recursive formulas for these probabilities via a methodology similar to that of Koole (2004), but the analysis is very messy and would not lead (as far as we can see) to a simple closed-form formula. We consider two alternative approaches; one approximate and one exact.

The approximation is based on the natural idea of exploiting the result (3) via a pooling of the two different means for each state in question. More precisely, given $s \in S_a$, the conditional probability in (2) is approximated by $f(\tau;\mu_a(s))$ as in (3), where

$$\mu_a(s) = \begin{cases} \left(\frac{(b_1(s)+b_{21}(s))\mu_1^{-1}+b_{22}(s)\mu_2^{-1}}{b(s)}\right)^{-1}, & s \in S_a,\ a = 3 \\ \left(\frac{b_1(s)\mu_1^{-1}+b_2(s)\mu_2^{-1}}{b(s)}\right)^{-1}, & s \in S_a,\ a = 5 \end{cases} \tag{4}$$

is the *state-dependent pooled mean* service time of inbound and outbound calls. A different approximation approach would be to pool the mean service times right from the start and use model M_2 or M_4 (see Section 4.1). The difference is that here (for M_5 and M_3), we do

not pool the means in the CTMC model itself but pool them separately for each state. This should lead to a better approximation especially when μ_1 and μ_2 differ significantly.

Alternatively, one can use the following exact approach (discussed, for example, in Grassmann 1977b and Gross and Miller 1984) to calculate $\bar{F}_a(\tau)$ for $a = 1, 2, \ldots, 5$. The state space for each model can be partitioned into states where some agents are idle and states where all agents are busy, i.e.,

$$\begin{aligned} S_a^1 &= \{s \in S_a : b(s) < n\} \\ S_a^2 &= \{s \in S_a : b(s) = n\}. \end{aligned}$$

The probability that $W_q > \tau$ is the same as the probability that the first passage time to the set S_a^1 is greater than τ in a modified version of the original CTMC. To obtain the modified process, we collapse the state space into $S_a' = \{s^1\} \cup S_a^2$, where state s^1 corresponds to an aggregation of S_a^1. We discard transitions corresponding to inbound arrivals and outbound dialing and we direct all transitions whose original destination state was in S_a^1 to the aggregated state s^1. State s^1 thus becomes absorbing, and reaching this state corresponds to the ℓth customer entering service.

Denote the transient probability of the modified process being in state s at time t by $\phi_s(t)$. We set the initial state probabilities for the modified process equal to the stationary probabilities for the original process, i.e.,

$$\begin{aligned} \phi_{s^1}(0) &= \sum_{s \in S_a^1} \pi_s \qquad \text{and} \\ \phi_s(0) &= \pi_s \text{ for } s \in S_a^2. \end{aligned}$$

Then, we calculate the transient probabilities for the modified process at time τ, using the uniformization method (see, e.g., Grassmann 1977a or Gross and Miller 1984). We then have that $\bar{F}_a(\tau) = 1 - \phi_{s^1}(\tau)$.

Although this exact approach requires the calculation of transient probabilities, the computation time is typically considerably less than that required to calculate the stationary probabilities for the original process, for two reasons. First, the modified process typically has far fewer states than the original process. For example, the modified state space for M_3 can be represented as

$$S_3' = \{s^1\} \cup \{s \in S_3 : b_1 = n_1, b_{21} + b_{22} = n_2, 0 \le q \le c\}$$

with cardinality $1 + (n_2 + 1)(c + 1)$. With $n_1 = n_2 = c = 20$, the original process has 5,271 states while the modified process has only 442 states. Second, the waiting time standard τ is typically very small (often it is 20 seconds), and the uniformization method requires $O(\nu\tau)$ matrix multiplications (see Grassmann 1977a), each of which requires $O(n_2 c)$ operations, where ν is the maximum transition rate out of any state in the modified process. This can

be compared to the computational effort required to calculate the stationary probabilities for M_3. This computational effort will depend on the method used to solve the linear system of equations, but it must be *at least* proportional to the number of states, i.e., larger or equal to $K(n_1 n_2^2 + n_2 c)$ for some constant K. For M_3, $\nu = n\max(\mu_1, \mu_2) + c\eta$. If one assumes that $c = O(\sqrt{n})$, $n_1 = O(n)$, and $n_2 = O(n)$, then uniformization has overall complexity $O(\tau \max(\mu_1, \mu_2) n^{2.5})$ while solving for the steady state probabilities is at least Kn^3 for some constant K. In our numerical examples, the CPU time to compute the service level by the exact method was always smaller than the time required to compute the steady-state probabilities, usually by a factor of 10 or more. But we cannot qualify this time as negligible. Moreover, implementing this exact method requires more work than just using the pooled means approximation.

We conducted a small set of experiments to assess the effect of this approximation error on the *service level*, defined as $1 - \bar{F}(\tau)$. We compare the exact values under M_3, the M_3 approximation with the pooled mean (4), and the approximation via M_4. We are mainly interested in the effect of the ratio $\zeta = \mu_2/\mu_1$ on model error. In Table 1, rows 1-3 correspond to our call center for time periods 13, 16, and 25, respectively (see Section 5.1 for the values of all parameters). Rows 4-6 and 7-9 correspond to the same periods, where we keep the original values for μ_1 and adjust μ_2 so that $\zeta = 2$ and 0.5, respectively.

Table 1: Computed service level $1 - \bar{F}(\tau)$ under alternative models ($\tau = 20$ sec).

Period	ζ	M_3 exact	M_3 pooled	M_4
13	1.307	0.9106	0.9117	0.9255
16	1.155	0.9611	0.9615	0.9735
25	1.155	0.9429	0.9593	0.9613
13	2	0.9257	0.9259	0.9495
16	2	0.9794	0.9779	0.9939
25	2	0.9673	0.9746	0.9840
13	0.5	0.8715	0.8712	0.8041
16	0.5	0.8931	0.8897	0.8229
25	0.5	0.8928	0.9192	0.9083

These and further experiments suggest that the M_3 pooled-mean approximation is very good for a wide range of the ratio ζ, whereas the M_4 approximation is considerably less robust to deviations of ζ from one.

3.2 Other performance measures

Other long-run performance measures can be obtained using the steady-state probabilities, as outlined in this subsection.

Agent utilization

Let $B(t)$ denote the number of busy agents at time t, and $u_a = \lim_{t\to\infty} \int_0^t B(\zeta)d\zeta/nt$ be the steady-state agent utilization fraction under Model M_a, for $a = 1, 2, \ldots, 5$. If $b(s)$ is the value of $B(t)$ for state $s \in S_a$, then we have that

$$u_a \;=\; \frac{1}{n}\sum_{s\in S_a} b(s)\pi_s. \tag{5}$$

The following performance measures are concerned with the number of calls per unit time.

Rate of lost inbound calls

Let R_a^ℓ be the rate at which inbound calls are lost for M_a, $a = 1, 2, \ldots, 5$. We have that

$$R_a^\ell = \lambda(1-\gamma) \sum_{s\in S_a: b(s)=n, q(s)<c} \pi_s + \eta \sum_{s\in S_a} q(s)\pi_s + \lambda \sum_{s\in S_a: q(s)=c} \pi_s$$

for $a = 1, 2, \ldots, 5$.

Rate of calls served

The steady-state rate of calls served of any type, denoted R_a^{t} for model a, is

$$R_a^{\mathrm{t}} = \sum_{s\in S_a} b(s)\mu\pi_s, \quad a = 1, 2, 4, \tag{6}$$

$$R_3^{\mathrm{t}} = \sum_{s\in S_3} \left((b_1(s) + b_{21}(s))\mu_1 + b_{22}(s)\mu_2\right)\pi_s, \tag{7}$$

and

$$R_5^{\mathrm{t}} = \sum_{s\in S_5} \left((b_1(s)\mu_1 + b_2(s)\mu_2\right)\pi_s. \tag{8}$$

The rate of inbound calls served equals, by inbound-call conservation, the rate of inbound calls accepted into the system: $\lambda - R_a^\ell$. Now the rate of outbound calls served can be calculated as

$$R_a^{\mathrm{o}} = R_a^{\mathrm{t}} - (\lambda - R_a^\ell), \quad a = 1, 2, .., 5. \tag{9}$$

Rate of mismatches

We denote the steady-state rate of mismatches by R_a^{m} for M_a, for $a = 1, 2, \ldots, 5$. Recall that R_1^{m} is zero. For the other four models, the steady-state rate of calls answered by outbound customers (not all are necessarily connected) is:

$$R_2^{\mathrm{r}} = \kappa \sum_{b=1}^{\dot{n}+1} b\mu v(n-b+1)\pi_b, \tag{10}$$

$$R_3^{\mathrm{r}} = \kappa \sum_{s\in S_3: b(s)-1\le \dot{n}, i_2(s)>0} b_1(s)\mu_1 v(i_2(s))\pi_s + \kappa \sum_{s\in S_3: b(s)-1\le \dot{n}} \left(b_{21}(s)\mu_1 + b_{22}(s)\mu_2\right) v(i_2(s)+1)\pi_s, \tag{11}$$

$$R_4^{\mathrm{r}} = \kappa \sum_{s\in S_4: b(s)-1\le \dot{n}, i_2(s)>0} b_1(s)\mu v(i_2(s))\pi_s + \kappa \sum_{s\in S_4: b(s)-1\le \dot{n}} b_2(s)\mu v(i_2(s)+1)\pi_s, \tag{12}$$

$$R_5^{\mathrm{r}} = \kappa \sum_{s\in S_5: b(s)-1\le \dot{n}} \left(b_1(s)\mu_1 + b_2(s)\mu_2\right) v(n-b(s)+1)\pi_s, \tag{13}$$

where we used that the expected number of calls answered by outbound customers, conditional on dialing at a state with i idle blend agents, is $\kappa v(i)$. Thus the rate of mismatches is

$$R_a^{\mathrm{m}} = R_a^{\mathrm{r}} - R_a^{\mathrm{o}}, \quad a = 1, 2, ..., 5. \tag{14}$$

We will compare the above long-run performance measures of M_1–M_5 with the results of a more realistic simulation model in Section 5. Before that, we discuss some extensions of M_1–M_5.

4. Refinements

In this section, we consider "relaxing" the assumptions of equal average service times in M_1, M_2 and M_4, and of deterministic time-stationary arrival rates in all models.

4.1 Different inbound and outbound service times: adapting M_1, M_2 and M_4

The analysis of M_1, M_2 and M_4 was made under the simplifying assumption that inbound and outbound service times are i.i.d. with rate μ. In practice, however, inbound and outbound service rates are usually different. In this section, we propose a method for choosing μ, so that these models can provide reasonable approximations in such cases.

Suppose that the inbound and outbound service times are i.i.d. with means $1/\mu_1$ and $1/\mu_2$, respectively. Then it seems intuitive to approximate the mean service time $1/\mu$ with the weighted average $p/\mu_1 + (1-p)/\mu_2$, where p is the long-run proportion of calls that are inbound. However, this p actually depends on μ. Denoting it by $p(\mu)$, we thus have the relationship

$$\frac{1}{\mu} = \frac{p(\mu)}{\mu_1} + \frac{1-p(\mu)}{\mu_2}. \tag{15}$$

We will show that this equation has a solution μ_{E} which we call the *effective service rate*. (This notion of effective service rate parallels the definition in Wolff 1989, page 266, and we will characterize μ_{E} so that it can be numerically determined; see Lemma 2).

For Model M_a, $a = 1, 2, 4$, the function $p(\cdot)$ is

$$p(\mu) \;=\; 1 - \frac{R_a^{\mathrm{o}}(\mu)}{R_a^{\mathrm{t}}(\mu)}, \tag{16}$$

where $R_a^{\mathrm{o}}(\mu)$ and $R_a^{\mathrm{t}}(\mu)$ are calculated in (9) and (6), respectively.

A solution to (15) is equivalent to a root of the function

$$h(\mu) = \frac{p(\mu)}{\mu_1} + \frac{1 - p(\mu)}{\mu_2} - \frac{1}{\mu}. \tag{17}$$

The following lemma characterizes this function and proves the existence of a solution to (15). It also implies a simple algorithmic solution approach.

Lemma 2 *The CTMC corresponding to each model of interest (M_1, M_2, or M_4) and induced by any μ in the interval $\mathcal{I} \stackrel{\text{def}}{=} (\min(\mu_1, \mu_2), \max(\mu_1, \mu_2))$ is positive recurrent. The function $h(\cdot)$ is continuous, with $h(\mu_1) < 0 < h(\mu_2)$ if $\mu_1 < \mu_2$, and $h(\mu_2) < 0 < h(\mu_1)$ if $\mu_1 > \mu_2$. Thus, Equation (15) has at least one solution in $\mathcal{I}$.*

Proof. For any relevant model M_a, $a = 1, 2, 4$, and for any $\mu \in \mathcal{I}$, the associated Markov chain is easily seen to be irreducible, aperiodic and, in view of the finiteness of the state space, positive recurrent. This establishes the existence of a stationary distribution and that the stationary probabilities $\pi_i(\mu)$, $i \in S_a$ are all positive. The transition rates are continuous functions of μ, which implies that the stationary probabilities are also continuous functions of μ. The continuity of $h(\cdot)$ follows from the continuity of $R_a^{\mathrm{o}}(\mu)$ and $R_a^{\mathrm{t}}(\mu)$, for $a = 1, 2, 4$ (see (9) and (6)). Now we consider the case $\mu_1 < \mu_2$ and M_1. Since the stationary probabilities $\pi_i(\mu)$, $i = 1, 2, \ldots, n + c$, are all positive, we have $p(\mu_1) < 1$, implying $h(\mu_1) < 0$, and $p(\mu_2) > 0$, implying $h(\mu_2) > 0$. The remaining cases are handled analogously. □

In light of Lemma 2, a solution to (15) for M_1, M_2 and M_4 can be found via root-bracketing methods. Bracketing algorithms begin with a closed interval known to contain a root. The size of this interval is iteratively reduced until it encloses the root to within a desired tolerance. The width of the search interval when it terminates provides an error estimate for the location of the root. In our implementation (see Section 5.2), we use the Brent-Dekker method (Brent 1971 and Bus and Dekker 1975), which combines an interpolation strategy with a bisection algorithm. From our experience, this algorithm is fast and robust. We note that under M_1, a potentially more efficient way of solving (15) is via algorithms that exploit derivative information, e.g., Newton-Raphson. The birth and death rates for that model can be used to express $\pi_k(\mu)$ in closed form (see Ross 1983 or Taylor and Karlin 1998) and the resulting expressions can then be differentiated.

4.2 Non-stationary and doubly stochastic Poisson arrival processes

In practice, the arrival process to a call center is *not* time-stationary; the arrival rate varies from day to day and within each day. The usual modelling approach in such a context is to partition the time period of interest into subintervals over which the arrival rate is assumed to be constant. A CTMC model can be used by adopting the simplifying assumption that the system is in steady-state during each subinterval (Deslauriers 2003, Green et al. 2001, Jongbloed and Koole 2001). After computing the performance measure for each subinterval, the overall performance is simply the appropriate weighted average. The planning intervals are usually 15 or 30-minute long for the reason that the available data in call centers is typically aggregated into averages over 15 or 30-minute periods. Of course, such a model is only a rough-cut approximation; in reality, the arrival rate is not piecewise constant over fixed-length periods, and the system is never in steady-state. We assess the amount of error introduced by this approximation for a particular example in Section 5.

Besides the non-stationarity, it has been observed that the arrivals to a call center are not realistically modeled by a process with a *deterministic* time-varying arrival rate (Avramidis et al. 2004, Jongbloed and Koole 2001). This has led to models with piecewise constant arrival rates, whose actual values are random variables. Brown et al. (2005) developed a model that induces dependence between arrival rates across days and across time intervals within a day, while Avramidis et al. (2004) developed models that focus on the dependence of arival rates across time intervals within a day. Specifically, suppose the time period of interest, $[t_{\text{begin}}, t_{\text{fin}}]$, is partitioned into ℓ subintervals defined by the time cut points $t_{\text{begin}} = t_0 < t_1 < \ldots < t_\ell = t_{\text{fin}}$. We assume that over the time period $[t_{i-1}, t_i]$, arrivals occur according to a Poisson process whose arrival rate Λ_i is a continuous random variable with density g_i, for $i = 1, 2, \ldots, \ell$, and where the densities g_i are allowed to be conditional densities given past (relevant) information, e.g., Avramidis et al. (2004) or Brown et al. (2005). In the analysis below, we develop approximate performance measure formulæ for a Poisson arrival model with a generally distributed rate parameter.

Suppose there is an infinite stream of independent realizations of the time period $[t_{\text{begin}}, t_{\text{fin}}]$, and we are interested in performance measures defined by averages over these realizations. In our context, the time period $[t_{\text{begin}}, t_{\text{fin}}]$ will correspond to one day of operation of the call center. Finally, in order to be able to compute the quantities of interest via the formulæ developed for our CTMC models, we make the important simplifying assumption that the initial state of the CTMC at the beginning of each subinterval, conditional on $\Lambda_i = \lambda$, obeys the steady-state probabilities that correspond to an arrival rate of λ for that subinterval.

Some performance measures are time-averages (e.g., agent utilization and rate of outbound calls completed), while others are averages per call (e.g., fraction of calls whose waiting time exceeds a given threshold). To see how to compute these averages, we concentrate on some subinterval i and first consider the case of an infinite-horizon *average per call*, for the

successive realizations of subinterval i only. For example, suppose we want to compute the long-run fraction of inbound calls whose waiting time exceeds a given constant τ. If W_q is the waiting time of a *random* inbound call in steady-state, in subinterval i, then this fraction is $\bar{F}_a^{(i)}(\tau) \stackrel{\text{def}}{=} \Pr\{W_q > \tau\}$. When the arrival rate is fixed at λ, we denote this fraction as $\bar{F}_{a,\lambda}(\tau)$ for Model M_a, given in (2). For a given subinterval $[t_{i-1},\, t_i]$, let D be the number of inbound call arrivals, A_τ of which wait for longer than τ, during this subinterval. Across successive days, we have i.i.d. realizations of the pair (D, A_τ). Because the system is assumed to be always in steady-state within a given subinterval, we have that $\mathrm{E}[A_\tau] = \mathrm{E}[D]\Pr\{W_q > \tau\}$, and

$$\begin{aligned}
\bar{F}_a^{(i)}(\tau) &= \Pr\{W_q > \tau\} = \frac{\mathrm{E}[A_\tau]}{\mathrm{E}[D]} = \frac{\mathrm{E}[A_\tau]}{(t_i - t_{i-1})\mathrm{E}[\Lambda_i]} \\
&= \frac{1}{(t_i - t_{i-1})\mathrm{E}[\Lambda_i]} \int_0^\infty \mathrm{E}[A_\tau | \Lambda_i = \lambda] g_i(\lambda) d\lambda \\
&= \frac{1}{\mathrm{E}[\Lambda_i]} \int_0^\infty \lambda \bar{F}_{a,\lambda}(\tau) g_i(\lambda) d\lambda \qquad (18)
\end{aligned}$$

for Model M_a, $a = 1, 2, \ldots, 5$. The quantities $\bar{F}_a^{(i)}(\tau)$ and $\bar{F}_{a,\lambda}(\tau)$ can be interpreted as average costs per call if we define the cost for a call as equal to 1 when its waiting time exceeds τ and 0 otherwise. Equation (18) can be generalized to an arbitrary cost function by replacing A_τ by the total cost of all calls over the given subinterval and the quantities $\bar{F}_a^{(i)}(\tau)$ and $\bar{F}_{a,\lambda}(\tau)$ by the appropriate infinite-horizon (or steady-state) average costs per call. For Model M_a, we denote these averages by $\nu_a^{(i)}$ and $\nu_{a,\lambda}$ to obtain

$$\nu_a^{(i)} = \frac{1}{\mathrm{E}[\Lambda_i]} \int_0^\infty \lambda \nu_{a,\lambda} g_i(\lambda) d\lambda. \qquad (19)$$

The overall average per call over $[t_{\text{begin}},\, t_{\text{fin}}]$ is the weighted average $\nu_a = \sum_{i=1}^{\ell} w_i \nu_a^{(i)}$, where

$$w_i = \frac{(t_i - t_{i-1})E[\Lambda_i]}{\sum_{j=1}^{\ell}(t_j - t_{j-1})E[\Lambda_j]}$$

represents the fraction of inbound calls arriving in subinterval i in the long run.

In the case of *time-average* performance measures $\vartheta_a^{(i)}$ and $\vartheta_{a,\lambda}$, the analog of (19) is

$$\vartheta_a^{(i)} = \int_0^\infty \vartheta_{a,\lambda} g_i(\lambda) d\lambda, \qquad (20)$$

and the overall average is $\vartheta_a = \sum_{i=1}^{\ell}(t_i - t_{i-1})\vartheta_a^{(i)}/(t_\ell - t_0)$.

Now we consider the distribution of Λ_i. Jongbloed and Koole (2001) model Λ_i as a gamma random variable, with density

$$g_i(\lambda) = \frac{\beta_i^{-\alpha_i}}{\Gamma(\alpha_i)} \lambda^{\alpha_i - 1} e^{-\lambda/\beta_i}, \qquad (21)$$

for $\lambda > 0$, where $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$ is the gamma function. They assume that the Λ_i's are mutually independent. This model, which we call the *Poisson-gamma* arrival process model, is appealing because it is flexible and mathematically tractable; under it, the number of arrivals in a given subinterval has the negative binomial distribution. Alternatives to (21) include a multivariate generalization of the above model in Avramidis et al. (2004) that allows dependence over different time intervals within a day, or an auto-regressive model of Brown et al. (2005) that assumes dependent and normally distributed Λ_i's.

5. Case study: Bell Canada call center

In this section, we compare the CTMC performance measures with the results of a realistic simulation model of a Bell Canada call center developed in Deslauriers (2003) and also described in Pichitlamken, Deslauriers, L'Ecuyer, and Avramidis (2003). That model was made as realistic as possible given the information we had and was calibrated so that its behavior (in terms of performance measures) was very close to that of the real system. Here we actually use a slightly simplified version of the simulation model, but the simplifications have little impact on the model's behavior. First, we describe the experimental setup.

5.1 Simulation model

The call center modeled in Deslauriers (2003) operates from 8:00 to 20:30, i.e., 8:30 PM. Agents receive only inbound calls before 14:00. After that, some of the agents are in blend mode, and there are also outbound calls. All of the available data is aggregated as *averages* over half-hour periods. That is, for each half hour, we have the number of arrivals (inbound), the number of outbound calls, the sum of service times for the inbound calls served and similarly for the outbound calls, but not the call-by-call arrival times or service times (with one exception; see below). Therefore, we assume that the model parameters (e.g., arrival rate and service time distributions) are constant over each half hour; with the notation of Section 4.2, we have $\ell = 25$ and $t_i = 8 + i/2$ for $i = 0, 1, \ldots, 25$.

In the simulation model, the arrival process is doubly stochastic, with random arrival rate Λ_i that is constant within period i. Moreover, for a given day, these random variables Λ_i are *dependent* and are distributed as

$$\Lambda_i \;=\; W\lambda_i, \tag{22}$$

where the λ_i's are parameters, and W is a gamma random variable with parameters (α', β'), and $\mathrm{E}[W] = 1$ (see Avramidis et al. 2004 and Deslauriers 2003 for further discussion of this model). The idea of the random factor W is to account for the day-to-day traffic variation. The parameter values estimated from the data were $\alpha' = 29.7$, $\beta' = 0.0336$ (the coefficient of variation of W is 5.45), and the λ_i's can be found in Table 2, along with all other

relevant model parameters. Because the parameter values depend on the day of the week (see Pichitlamken et al. 2003 for further discussion), we simulate each day of the week separately; the values given here are for Monday.

The inbound service times are gamma distributed with parameters (α_i, β_i) in period i. For the outbound service times (only), we happened to have over 50 thousand individual observations which we use to estimate the service time density via a kernel density estimation method. We generate the service times from that density, using the UNURAND package (Leydold and Hörmann 2002).

The other aspects of the simulation model not mentioned here are the same as in M_3. The patience time is exponentially distributed with mean $1/\eta_i$ in period i. The probability $1-\gamma$ of a customer immediately leaving the queue once realizing that he is put on hold is 0.005. When the total number of idle agents is at least 4 and $I_2 \geq 1$ of them are blend agents, the system dials $v(I_2) = 2I_2$ outbound calls in parallel. We thus have $\dot{n} = n - 4$. We do not claim that this policy is close to optimal in any sense; it is just an approximation of the rule that was used in the center. The *dial resolution delay*, i.e., the time required for the dialer to either start the call or recognize that the attempt is not successful, is exponentially distributed with a mean of 2 seconds (also for M_1). The simulation results are essentially unchanged when we set the delay to zero, which is what we have in Models M_2—M_5. The success probability is κ_i in period i.

The staffing used in this particular center (last two columns of Table 2) may look a bit strange (for example, compare the mix of inbound-only and blend agents in period 13 and near the end of the day). We neither claim nor believe that this staffing is optimal in any sense. It was determined by the managers under restrictive constraints on the shifts of individual agents, due to union agreements (for example, no agent can be put in blend mode for more than 4 hours a day, agents must work for a certain number of consecutive hours, etc.). Determining an "optimal" staffing without taking these constraints into account would most certainly not give a feasible solution to the scheduling problem.

5.2 Implementation details for models M_1 to M_5

For the CTMC models in this case study, we adopted the Poisson-gamma arrival process model detailed in Section 4.2. The parameters of these models were chosen to match the corresponding values in the simulation model over each period i, whenever possible. All models had the same mean service times for inbound calls as well as for outbound calls, and the same mean for each Λ_i. The input parameters of the CTMC models that differ from those of the simulation model can be found in Table 3. The number of agents for M_1, M_2 and M_5 is $n_1 + n_2$. The average service time of outbound calls is 440.2 seconds.

The service rate μ in M_1, M_2 and M_4 is the effective service rate μ_E defined in Section 4.1. We find a root of the function $h(\cdot)$ in (17), where each evaluation of $h(\cdot)$ requires solving

Table 2: Input parameters of the simulation model.

Period i	Start time (hour)	Arrival rate λ_i per half hour	Outbound success prob. κ_i	Mean patience time $1/\eta_i$ (sec)	Inbound serv. time (sec)		Inbound agents	Blend agents
					α_i	β_i	n_1	n_2
1	8.0	32.11	0	400	0.729	817.0	12	0
2	8.5	45.96	0	400	0.729	817.0	18	0
3	9.0	58.48	0	400	0.729	817.0	22	0
4	9.5	66.50	0	700	0.729	817.0	25	0
5	10.0	73.44	0	700	0.729	817.0	27	0
6	10.5	72.87	0	600	0.729	817.0	26	0
7	11.0	74.13	0	600	0.729	817.0	26	0
8	11.5	71.40	0	600	0.729	817.0	24	0
9	12.0	68.32	0	600	0.620	927.6	22	0
10	12.5	68.04	0	600	0.620	927.6	23	0
11	13.0	71.55	0	500	0.620	927.6	28	0
12	13.5	70.11	0	500	0.620	927.6	25	0
13	14.0	68.50	0.27	500	0.620	927.6	25	5
14	14.5	67.71	0.27	500	0.620	927.6	23	11
15	15.0	68.45	0.28	500	0.755	753.8	23	16
16	15.5	72.93	0.29	500	0.755	753.8	23	18
17	16.0	71.92	0.29	500	0.755	753.8	21	16
18	16.5	66.15	0.30	500	0.553	996.9	17	14
19	17.0	49.50	0.33	500	0.553	996.9	15	11
20	17.5	48.45	0.37	500	0.553	996.9	10	16
21	18.0	39.00	0.40	500	0.553	996.9	4	16
22	18.5	34.97	0.38	500	0.518	981.6	3	16
23	19.0	30.80	0.41	500	0.518	981.6	3	16
24	19.5	28.26	0.41	100	0.518	981.6	3	17
25	20.0	20.68	0.41	50	0.518	981.6	3	16

the system of linear equations identifying the steady-state probabilities. For the root-finding problem, we use the GSL (Galassi et al. 2002) implementation of the Brent-Dekker root-bracketing method. The algorithm starts with the interval given in Lemma 2 and stops with an error smaller than 10^{-7} on $1/\mu_{\mathrm{E}}$.

For a given μ, we obtain the steady-state probabilities by solving the system of linear equations, using the LU decomposition method implemented in LAPACK (Anderson et al. 1999). To compute the steady-state performance measures of interest via (19) and (20), we perform numerical integration with the adaptive 15-point Gaussian quadrature method called the Gauss-Kronrod rule (Piessens et al. 1983). The integration terminates when the estimated relative error is deemed sufficiently small. In our example, the error was small enough so that all the digits given in Table 4 are significant.

5.3 Comparison between the simulation and CTMC models

With the parameter settings of Tables 2 and 3, we now compare the performance measures provided by the CTMC models to those obtained from simulation. If we assume that the simulation model is realistic, this amounts to assessing the approximation error made by the CTMC models.

Table 4 summarizes the performance measures averaged (a) when being in inbound mode, and (b) in blend mode. Column *Simulation* (§*5.1*) contains the simulation results whose underlying distributions are as described in Section 5.1, whereas column *Simulation* (*exp*) presents the simulation results when the arrival and service time distributions are exponential. (The values are given with 95% confidence intervals. The symbol ϵ represents a value smaller than 0.1.) The *quality of service* (QoS) is defined as the probability that an inbound call waits in the queue for less than 20 seconds. The other performance measures were defined in Section 3. Note that during inbound mode, there is only one set of results for M_1–M_5, because M_2–M_5 reduce to M_1 when there are no outbound calls. There are no mismatches for M_1 because this model does not allow them. As a result of input modelling, the expected total number of inbound calls arriving to the call center, which is the sum of the number of inbound calls served and lost, is identical across all the CTMC models and the simulation model. We see greater discrepancies between our CTMC models and the simulation in the inbound mode (Table 4a) than in the blend mode (Table 4b), mostly because the CTMC models wrongly assume steady-state conditions at all times and this has a larger impact at the beginning of the day. Indeed, in the simulation model, the day begins with an empty system, and the arrival rate and number of agents change significantly during the first five half hours; see Table 2.

Other than the steady-state condition, M_1–M_5 differ from the simulation model in the following ways: (a) the arrival rates Λ_i are dependent within the same day in the simulation (see (22)) and independent in the CTMC models; (b) the service times in the CTMC models

Table 3: Input parameters of M_1–M_5. The average service times are in seconds.

Period	Arriv. proc. param. (/0.5 hr)		Avg. svc. time			
i	α_i	β_i	($1/\mu_1$ for M_3 and M_5, $1/\mu_E$ for M_1, M_2 and M_4)			
1	16.9	1.9	595.6			
2	38.3	1.2	595.6			
3	13.6	4.3	595.6			
4	26.6	2.5	595.6			
5	21.6	3.4	595.6			
6	34.7	2.1	595.6			
7	35.3	2.1	595.6			
8	23.8	3.0	595.6			
9	24.4	2.8	575.1			
10	24.3	2.8	575.1			
11	15.9	4.5	575.1			
12	17.1	4.1	575.1			
Period	Arriv. proc. param. (/0.5 hr)		Avg. in. svc. time $1/\mu_1$	Avg. svc. time $1/\mu_E$		
i	α_i	β_i	M_3 and M_5	M_1	M_2	M_4
13	27.4	2.5	575.1	534.9	530.6	554.4
14	18.3	3.7	569.1	518.2	514.5	529.5
15	18.5	3.7	569.1	508.3	505.0	516.3
16	22.1	3.3	569.1	509.1	505.9	515.0
17	24.8	2.9	569.1	516.1	512.5	520.9
18	24.5	2.7	551.3	511.1	507.6	513.7
19	16.5	3.0	551.3	504.0	500.3	509.3
20	28.5	1.7	551.3	502.8	498.9	502.1
21	19.5	2.0	551.3	506.9	502.0	503.3
22	26.9	1.3	508.5	477.6	474.8	475.4
23	11.0	2.8	508.5	473.2	470.5	471.1
24	31.4	0.9	508.5	468.7	465.6	466.4
25	51.7	0.4	508.5	462.4	459.4	460.5

do not have the same distributions as those in the simulation; and (c) the dialer operates differently across the different models. Despite these differences, M_3 yields results that are close to the simulation values on most measures when the call center is in the blend mode (see Table 4b). It is not surprising that M_3 outperforms our other CTMC models since it is the most faithful in details to the simulation model. What is interesting is that M_4 also performs very well in this problem setting. This is because μ_1 and μ_2 are not drastically different; therefore, the pooled service time, $1/\mu_E$, is not too far apart from either $1/\mu_1$ or $1/\mu_2$ (see Tables 2–3). On the other hand, M_5 performs poorly in the current setting because it assumes that every agent is blend, but the fraction of blend agents varies from 15% to 85%.

Let us recall that we propose M_4 and M_5 as simplified versions of M_3 to speed up the computation because the number of states in the CTMCs is significantly reduced when we assume either that $\mu_1 = \mu_2$ (as in M_4) or that every agent is blend (as in M_5). For example, for half hour 16, there are 4940 states in M_3, compared to only 476 states in M_4 and 1743 states in M_5 (the number of states as a function of n_1, n_2, and c is given for models M_3-M_5 in section 2). We expect M_4 to provide a good approximation of M_3 when μ_1 and μ_2 are close, whereas M_5 should yield results close to M_3 when most agents are blend.

To evaluate how much of the difference between M_1–M_5 and the simulation model is due to the choices of distributions, we performed simulations using the same distributions as in the CTMC models. The results are given in column *Simulation* (*exp.*) of Table 4. We see that matching the distributions has little effect on the simulation results. The other two factors, the steady-state condition and the dialer's policy, appear to be more important sources of difference between the simulation and CTMC models. The dialer's policy certainly has a significant impact on the mismatch rate and the QoS (compare M_1–M_5 in Table 4). Further sensitivity analysis is pursued in the next subsection.

Figure 1 shows the evolution of the QoS by half-hour (the confidence intervals are not shown as they are smaller than the size of the series symbols). In the blend mode, the QoS of M_2 and M_5 are lower than that of the simulation, except for half-hour 21, when the QoS drops dramatically. (This is due to the sudden decrease in the total number of agents available that occurs at the same time as the reduction in the fraction of inbound-only agents; see Table 2.) On the other hand, the QoS of M_3 and M_4 closely follows the QoS of the simulation model.

Figure 1: Comparison of the QoS obtained from the CTMC and simulation models for the call center in blend mode.

Table 4: Comparison of the CTMC and the simulation models. The number of calls reported is per day.

(b) **Inbound mode**

Daily measures	M_1—M_5	Simulation	
		(§5.1)	(exp)
QoS (%)	63.33	69.5 $\pm$ ϵ	67.8 $\pm$ ϵ
Agent utilization (%)	83.3	82 $\pm$ 0.1	82.7 $\pm$ ϵ
# inbound calls served	709.80	721$\pm$ 4	725 $\pm$ 1.1
# of lost inbound calls	62.5	47$\pm$ 1.4	47 $\pm$ 0.5

(c) **Blend mode**

Daily measures	M_1	M_2	M_3	M_4	M_5	Simulation	
						(§5.1)	(exp)
QoS (%)	89.81	80.44	87.46	89.22	79.30	87.3 $\pm$ ϵ	86.4 $\pm$ ϵ
Agent utilization (%)	91.5	95.1	89.1	87.9	95.2	89 $\pm$ 0.1	89 $\pm$ ϵ
# inbound calls served	657.5	647.2	653.3	655.5	645.6	648 $\pm$ 4	651 $\pm$ 1.1
# of lost inbound calls	9.8	20.0	13.9	11.8	21.7	11.6 $\pm$1.4	12.1 $\pm$ 0.5
# outbound calls served	523.2	590.3	492.5	472.3	592.5	480.4 $\pm$ 2	484.7 $\pm$ 0.9
# mismatches	0	27.7	50.6	50.6	27.7	60.9 $\pm$ 0.3	62.0 $\pm$ 0.2

5.4 Sensitivity to selected model assumptions

In this section, we study empirically the effect of the dial resolution delay and non-exponential service and patience times on both true system performance and the accuracy of the CTMC models. We focus on steady-state call center performance measures, which we estimate to high accuracy via simulation and compare to results from the CTMC models to assess model errors. We will use two half-hour periods from the Bell call center to illustrate how the appropriateness of the different models depends on model parameters. We begin by assuming a Poisson arrival process and then repeat the analysis using the the Poisson-Gamma model of Section 5.1. Dial resolution delays are i.i.d exponentially distributed with mean δ.

Tables 5 to 7 show results for period 16 (high staffing, a balanced mix of agents). Simulation point estimates are accompanied by 95% confidence intervals; ϵ denotes entries less than 0.1; when point estimates are small and there is high simulation accuracy, we show interval half-widths as percentages of the point estimate.

Table 5 shows sensitivity to mean dial resolution delay δ and compares to the CTMC values (for M_1, we set $\nu = 1/\delta$). We see that with the exception of mismatch volume, all performance measures are insensitive as we move from $\delta = 0.001$ (indicated as '0' in tables) to $\delta = 10$ seconds. Table 6 shows sensitivity of the system to distributional assumptions by substituting a lognormal distribution (L) with the same mean as the assumed exponential (E) and coefficient of variation (CV, standard deviation divided by the mean) equal to two; this yields more variable times than the exponential case (CV is one). Here, all performance measures are virtually constant across the four pairs of distributions for service time and patience. With respect to CTMC model errors in Tables 5 and 6, M_3 and M_4 are overall very good—relative errors are small for all performance measures, with the exception of mismatches. The other models do much less well, and their weakness is easily explained by the fact that the all-agents-are-blend assumption is strongly violated; M_2 and M_5 overestimate the outbound volume and agent utilization and underestimate QoS; M_1 is sensitive to the mean delay δ but is not accurate in either case.

Table 7 shows sensitivity to dial resolution delay and then to non-exponential service and patience time distributions for a Poisson-Gamma arrival process; here, naturally, CTMC models use the refinement in Section 4.2. Remarkably, performance measures (simulated and CTMC-based) are very close to the Poisson case, with a notable exception being the inbound loss rate, which is more than doubled. The CTMC model errors behave qualitatively the same as in the Poisson case.

Table 8 summarizes results for period 21 (low staffing, agent majority is blend type). While the overall picture is mostly in line with period 16, a difference is that M_2 and M_5 do much better than in period 16; this is expected, since the all-blend-agents assumption is closer to reality.

While extrapolating from limited empirical results is risky, these results suggest: (1) the

assumption of exponential service and patience times is robust against a lognormal, higher-variance alternative; this holds for all performance measures; (2) the assumption of zero dial resolution delays is robust against the alternative of reasonable dial resolution delays; this holds for all performance measures except mismatches; (3) our models' weakest point is estimating mismatches because the dialing process is not modeled well; (4) violations of the all-blend-agents assumption are costly in terms of CTMC model accuracy; and (5) when service rates for inbound and outbound calls differ by a reasonable amount, the heuristic of section 4.1 makes it possible to use simpler models with small loss of accuracy (evidenced by the closeness of M_4 to M_3 and the closeness of M_2 to M_5).

Table 5: CTMC and simulation results showing sensitivity of simulation to outbound resolution delay δ. Period 16, arrival process is Poisson, service and patience are exponential. Call volumes are per half hour.

Perf. measure	M_1		M_2	M_3	M_4	M_5	Simulation	
	$\delta = 0$	$\delta = 10$					$\delta = 0$	$\delta = 10$
QoS (%)	94.8	98.9	88.5	96.1	97.3	88.2	96.7±ϵ	96.4±ϵ
Agent util.(%)	94.9	81.7	96.5	90.4	88.4	96.5	89.5±ϵ	89.2±ϵ
Inb. calls	72.5	72.8	71.9	72.6	72.7	71.9	72.65±ϵ	72.62±ϵ
Lost inb. calls	0.45	0.09	1.00	0.33	0.23	1.02	0.29±1.4%	0.31±1.1%
Outb. calls	65.4	42.7	68.8	57.7	54.2	68.8	55.95±ϵ	55.64±ϵ
Mismatches	0	0	1.82	5.64	5.74	1.82	3.77±ϵ	7.37±ϵ

Table 6: Sensitivity of simulation to some service-patience distribution pairs; E/L denotes exponential service and lognormal patience, and so on. Period 16, arrival process is Poisson, mean outbound resolution is 10 seconds. Call volumes are per half hour.

Perf. measure	E/E	E/L	L/E	L/L
QoS (%)	96.4±ϵ	96.4±ϵ	96.3±ϵ	96.4±ϵ
Agent util.(%)	89.2±ϵ	89.2±ϵ	89.2±ϵ	89.2±ϵ
Inb. calls	72.62±ϵ	72.62±ϵ	72.61±ϵ	72.60±ϵ
Lost inb. calls	0.32±1.1%	0.32±1.6%	0.33±1.5%	0.34±1.5%
Outb. calls	55.64±ϵ	55.66±ϵ	55.57±ϵ	55.56±ϵ
Mismatches	7.37±ϵ	7.38±ϵ	7.33±ϵ	7.31±ϵ

6. Optimization

As mentioned in the introduction, a primary use of the CTMC models developed here would be to determine an appropriate staffing or work schedule for a call center. Optimization

Table 7: CTMC and simulation results showing sensitivity of simulation to outbound resolution delay and then to non-exponential service and patience: exponential (E) versus lognormal (L). Period 16, arrival process is Poisson-Gamma. Call volumes are per half hour.

Perf. measure	M_1		M_2	M_3	M_4	M_5	Simulation		
	$\delta = 0$	$\delta = 10$					$\delta = 0$, E	$\delta = 10$, E	$\delta = 10$, L
QoS (%)	92.6	97.1	86.3	92.8	94.7	85.0	94.6±.7	94.6±.7	94.5±.6
Agent util.(%)	94.9	80.9	96.5	89.2	87.1	96.6	88.2±.1	89.2±.5	88±.5
Inb. calls	72.3	72.7	71.7	72.3	72.4	71.6	72.5 ±3	72.3±1.2	72.2±1.4
Lost inb. calls	0.65	0.26	1.21	0.66	0.47	1.34	0.64±.2	0.65±.1	0.73±.1
Outb. calls	65.7	41.6	69.1	56.1	52.4	69.3	54.1 ±2	54.1±1.1	54.1±1.1
Mismatches	0	0	1.83	5.55	5.55	1.83	3.75±.2	7.37±.18	7.29±.17

Table 8: CTMC and simulation results showing sensitivity of simulation to outbound resolution delay and then to non-exponential service and patience: exponential (E) versus lognormal (L). Period 21, arrival process is Poisson. Call volumes are per half hour.

Perf. measure	M_1		M_2	M_3	M_4	M_5	Simulation		
	$\delta = 0$	$\delta = 10$					$\delta = 0$, E	$\delta = 10$, E	$\delta = 10$, L
QoS (%)	90.2	92.4	77.3	80.5	82.0	76.4	81.6±ϵ	80.9±ϵ	81.3±ϵ
Agent util.(%)	90.1	85.9	93.9	93.0	92.8	94.0	92.7±ϵ	92.6±ϵ	92.5±ϵ
Inb. calls	38.5	38.6	37.8	37.9	38.0	37.7	38.0±ϵ	37.9±ϵ	37.7±ϵ
Lost inb. calls	0.53	0.42	1.23	1.08	0.98	1.30	1.06±ϵ	1.10±ϵ	1.33±ϵ
Outb. calls	27.1	21.9	30.6	28.6	29.5	29.7	28.2±ϵ	28.2±ϵ	28.4±ϵ
Mismatches	0	0	2.39	2.82	2.97	2.32	2.60±ϵ	4.79±ϵ	4.81±ϵ

in that context raises another set of non-trivial challenges that we plan to address in a subsequent paper. We now briefly discuss these issues and how they can be handled.

For inbound-only call centers, it is still common practice to use the Erlang-C or Erlang-A formulæ as crude approximations to determine the minimal staffing, for each half-hour period, under a given set of constraints on the quality of service. Our CTMC models can be used in a similar way for call centers operating in blend mode.

In the simplest case of a steady-state model for a single period and where the decision variables are only the number of agents of each type, for example, we may want to minimize $c_1 n_1 + c_2 n_2$ under the constraints that $\Pr\{W_q \le \tau\} \ge b_q$, $R^\ell \le b_\ell$, and $R^o \ge b_o$, for some constants c_1, c_2, b_q, b_ℓ, and b_o, where W_q is the waiting time of a random call, R^ℓ is the expected number of abandonments, and R^o is the expected number of successful outbound calls.

Due to the nature of the constraints, the set of feasible solutions in this context is normally an increasing set, i.e., if (n_1, n_2) is a feasible solution and $(\tilde{n}_1, \tilde{n}_2) \ge (n_1, n_2)$, then $(\tilde{n}_1, \tilde{n}_2)$ is also feasible. Let us assume that this is the case. For models with a single type of agent (all agents are blend), the minimal number of agents that satisfies all the constraints can be found easily via binary search, for example. For models with two types of agents, for each integer $n_1 \ge 0$, let us define $n_2^*(n_1)$ as the minimal value of n_2 such that (n_1, n_2) is a feasible solution. The optimal solution (if it exists) must lie on the boundary defined by the function n_2^*. Then it suffices to determine this boundary over some reasonable range of values of n_1 and evaluate the cost $c_1 n_1 + c_2 n_2$ of all solutions on the boundary to find the optimal one. To determine the boundary, start at a given n_1 and find $n_2^*(n_1)$ by binary search. Then search for $n_2^*(n_1+1)$ by starting at $n_2 = n_2^*(n_1)$ and decreasing n_2 by 1 until the boundary is found. Repeat with $n_1 + 2$, and so on. In the other direction, search for $n_2^*(n_1 - 1)$ by starting at $n_2 = n_2^*(n_1)$ and increasing n_2 by 1 until the boundary is found, and so on. This procedure exploits the fact that $n_2^*(n_1)$ and $n_2^*(n_1 \pm 1)$ are usually close to each other.

More generally, the decision variables in the optimization model may also include the parameters of the dialer's policy. For instance, one could define $v(I) = \text{round}((\psi_1 + \psi_2 I)^+)$ and optimize the values of $\dot{n}$, ψ_1, and ψ_2. This makes the optimization a bit more challenging. In the case where the constraints are on daily averages (e.g., the fraction of customers waiting less than τ in the long run must be at least b_q, or the expected total number of outbound calls per day must be at least b_o) then we have an optimization problem where the decision variables are the values of $(n_1, n_2, \dot{n}, \psi_1, \psi_2)$ for each period of the day. So if there are 25 periods, we have a nonlinear integer programming problem with 125 variables.

All of this is for the staffing problem. The scheduling problem gives rise to a yet more complicated integer program where the decision variables n_1 and n_2 for each period are replaced by the number x_i of agents having the daily working schedule i, for each possible working schedule. This challenging problem is the subject of our on-going investigation.

7. Conclusion and directions for future work

We have studied CTMC models of a call center at different levels of detail. The models were first developed under a time-stationarity assumption and then extended to cover the case of a piecewise-constant doubly-stochastic arrival rate. We compared performance measure estimates obtained by these models with those obtained by a more detailed simulation model of a real-life call center. In the blend environment, the discrepancy between the results of the CTMC models and the simulation was less than 1% for important performance measures such as the fraction of inbound calls with a response time less than 20 seconds (the QoS), but was higher for other measures such as the rate of mismatches and the rate of abandonments. The latter measures are strongly influenced by how we model the dialer's policy.

The practical usefulness of the CTMC models, as "quick" alternatives to simulation, depends on what performance measure(s) we are mostly interested in and how much error we are ready to accept. Even with detailed simulation models, the error on certain performance measures may exceed 10% because of the uncertainty in model parameters (estimated from the data), monthly and yearly variations and trends, and several modeling assumptions (e.g., occasionally, the list of customers to be reached by outbound calls might be empty, or may contain mostly customers who are rarely at home, whereas we assume that this list is infinite, etc.).

An important utility of these CTMC models is for the staffing and scheduling problems. Much like the Erlang-C or Erlang-A formulæ for the case of inbound-only traffic, they provide *rough-cut* approximations of certain performances measures of the system that can feed an optimization algorithm. Solving these optimization problems is non-trivial but feasible. Eventually, the solutions can be refined in a second stage by using a more detailed simulation model combined with simulation-based optimization techniques. Studying how to do that effectively is the subject of current investigation.

Appendix

A.1 Steady-state probabilities for M_1

The steady-state probabilities for model M_1 are

$$\begin{aligned}
\theta_j &= \begin{cases}
\frac{1}{j!}\left(\frac{\lambda+\kappa\nu}{\mu}\right)^j, & j = 1,\ldots,\dot{n} \\
\frac{1}{j!}\left(\frac{\lambda+\kappa\nu}{\mu}\right)^{\dot{n}+1}\left(\frac{\lambda}{\mu}\right)^{j-\dot{n}-1}, & j = \dot{n}+1,\ldots,n-1 \\
\frac{1}{n!}\left(\frac{\lambda+\kappa\nu}{\mu}\right)^{\dot{n}+1}\left(\frac{\lambda}{\mu}\right)^{n-\dot{n}-1}\prod_{l=1}^{j-n}\frac{\gamma\lambda}{n\mu+l\eta}, & j = n,\ldots,n+c \\
0, & \text{otherwise}
\end{cases} \\
\pi_k &= \begin{cases}
\left[\sum_{j=0}^{\infty}\theta_j\right]^{-1}, & k = 0 \\
\theta_k\pi_0, & k = 1,2,\ldots,n+c \\
0, & \text{otherwise.}
\end{cases}
\end{aligned} \tag{A1}$$

A.2 M_4 transition types

1. **Inbound arrival:**
   - Destination state: $(b_1 + I\{b_1 < n_1\}, b_2 + I\{b_1 = n_1, b_2 < n_2\}, q + I\{b = n\})$.
   - Condition: $q < c$.
   - Rate: $\lambda(1 - (1-\gamma)I\{b = n\})$.
2. **Abandonment:**
   - Destination state: $(b_1, b_2, q-1)$.
   - Condition: $q > 0$.
   - Rate: $q\eta$.
3. **Inbound agent completes call, no outbound dialing:**
   - Destination state: $(b_1 - I\{q = 0\}, b_2, q - I\{q > 0\})$.
   - Condition: $b_1 > 0$ and $(b - 1 > \dot{n}$ or $i_2 = 0)$.
   - Rate: $b_1\mu$.
4. **Blend agent completes call, no outbound dialing:**
   - Destination state: $(b_1, b_2 - I\{q = 0\}, q - I\{q > 0\})$.

- Condition: $b_2 > 0$ and $b - 1 > \dot{n}$.
- Rate: $b_2\mu$.

5. **Inbound agent completes call, followed by z outbound calls that are answered:** Of the z answered calls, $m = \min(z, i_2)$ calls will be connected to blend agents. This is a family of transitions generated by all z such that $0 \leq z \leq v(i_2)$.
   - Destination state: $(b_1 - 1, b_2 + m, 0)$.
   - Condition: $b_1 > 0$ and $b - 1 \leq \dot{n}$ and $i_2 \geq 1$.
   - Rate: $b_1\mu\phi_{i_2}(z)$.

6. **Blend agent completes call, followed by z outbound calls that are answered:** Of the z answered calls, $m = \min(z, i_2 + 1)$ calls will be connected to blend agents. This is a family of transitions generated by all z such that $0 \leq z \leq v(i_2 + 1)$.
   - Destination state: $(b_1, b_2 - 1 + m, 0)$.
   - Condition: $b_2 > 0$ and $b - 1 \leq \dot{n}$.
   - Rate: $b_2\mu\phi_{i_2+1}(z)$.

A.3 M_5 transition types

1. **Inbound arrival:**
   - Destination state: $(b_1 + I\{b < n\}, b_2, q + I\{b = n\})$.
   - Condition: $q < c$.
   - Rate: $\lambda(1 - (1 - \gamma)I\{b = n\})$.

2. **Abandonment:**
   - Destination state: $(b_1, b_2, q - 1)$.
   - Condition: $q > 0$.
   - Rate: $q\eta$.

3. **Inbound call service completion, no outbound dialing:**
   - Destination state: $(b_1 - I\{q = 0\}, b_2, q - I\{q > 0\})$.
   - Condition: $b_1 > 0$ and $b - 1 > \dot{n}$.
   - Rate: $b_1\mu_1$.

4. **Outbound call service completion, no outbound dialing:**

- Destination state: $(b_1 + I\{q > 0\}, b_2 - 1, q - I\{q > 0\})$.
- Condition: $b_2 > 0$, $b - 1 > \dot{n}$.
- Rate: $b_2\mu_2$.

5. **Inbound call service completion, followed by z outbound calls that are answered:** Of the z answered calls, $m = \min(z, n - b + 1)$ calls will be connected. This is a family of transitions generated by all z such that $0 \leq z \leq v(n - b + 1)$.
   - Destination state: $(b_1 - 1, b_2 + m, 0)$.
   - Condition: $b_1 > 0, b - 1 \leq \dot{n}$.
   - Rate: $b_1\mu_1\phi_{n-b+1}(z)$.
6. **Outbound call service completion, followed by z outbound calls that are answered:** Of the z answered calls, $m = \min(z, n - b + 1)$ calls will be connected. This is a family of transitions generated by all z such that $0 \leq z \leq v(n - b + 1)$.
   - Destination state: $(b_1, b_2 - 1 + m, 0)$.
   - Condition: $b_2 > 0, b - 1 \leq \dot{n}$.
   - Rate: $b_2\mu_2\phi_{n-b+1}(z)$.

Acknowledgments

This research was supported by grants number OGP-0110050, CRDPJ-251320, and 203534 from the Natural Sciences and Engineering Research Council of Canada (NSERC), a grant from Bell Canada via the Bell University Laboratories, and grant number 00ER3218 from NATEQ-Québec, and a Canada Research Chair to the third author. The work of the first author was supported by an NSERC-Canada scholarship. We thank Bell Canada for providing us with the data and their support, and Richard Simard for his help in doing the computations.

REFERENCES

Anderson, E., Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. 1999. *LAPACK users' guide*. Third ed. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Avramidis, A. N., A. Deslauriers, and P. L'Ecuyer. 2004. Modeling daily arrivals to a telephone call center. *Management Science* 50 (7): 896–908.

Bhulai, S., and G. Koole. 2003. A queueing model for call blending in call centers. *IEEE Transactions on Automatic Control* 48:1434–1438.

Brandt, A., and M. Brandt. 1999a. On the $M(n)/M(n)/s$ queue with impatient calls. *Performance Evaluation* 35 (1-2): 1–18.

Brandt, A., and M. Brandt. 1999b. A two-queue priority system with impatience and its application to a call center. *Methodology and Computing in Applied Probability* 1:191–210.

Brent, R. P. 1971. An algorithm with guaranteed convergence for finding a zero of a function. *Computer Journal* 14:422–425.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association* 100 (469): 36–50.

Bus, J. C. P., and T. J. Dekker. 1975. Two efficient algorithms with guaranteed convergence for finding a zero of a function. *ACM Transactions of Mathematical Software* 1 (4): 330–345.

Deslauriers, A. 2003. Modélisation et simulation d'un centre d'appels téléphoniques dans un environnement mixte. Master's thesis, Department of Computer Science and Operations Research, University of Montreal, Montreal, Canada.

Galassi, M., J. Davies, J. Theiler, B. Gough, G. Jungman, M. Booth, and F. Rossi. 2002, December. GNU scientific library: Reference manual. Downloadable from `http://sources.redhat.com/gsl`.

Gans, N., G. Koole, and A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management* 5:79–141.

Grassmann, W. K. 1977a. Transient solutions in Markovian queueing systems. *Computers and Operations Research* 4:47–53.

Grassmann, W. K. 1977b. Transient solutions in Markovian queues. *European Journal of Operational Research* 1:396–402.

Green, L. V., P. J. Kolesar, and J. Soares. 2001. Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research* 49 (4): 549–564.

Gross, D., and D. R. Miller. 1984. The randomization technique as a modeling tool and solution procedure for transient Markov processes. *Operations Research* 32 (2): 343–361.

Jongbloed, G., and G. Koole. 2001. Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry* 17:307–318.

Koole, G. 2004. A formula for tail probabilities of Cox distributions. *Journal of Applied Probability* 41 (3): 935–938.

Koole, G., and A. Mandelbaum. 2002. Queueing models of call centers: An introduction. *Annals of Operations Research* 113:41–59.

Leydold, J., and W. Hörmann. 2002. *UNURAN—a library for universal non-uniform random number generators*. Available at `http://statistik.wu-wien.ac.at/unuran`.

Mandelbaum, A. 2003. Call centers (centres): Research bibliography with abstracts. Down-

loadable from ⟨ie.technion.ac.il/~serveng/References/ccbib.pdf⟩.

Pichitlamken, J., A. Deslauriers, P. L'Ecuyer, and A. N. Avramidis. 2003. Modeling and simulation of a telephone call center. In *Proceedings of the 2003 Winter Simulation Conference*, 1805–1812: IEEE Press.

Piessens, R., E. de Doncker-Kapenga, C. W. Uberhuber, and D. K. Kahaner. 1983. *QUADPACK a subroutine package for automatic integration*. Springer Verlag.

Riordan, J. 1962. *Stochastic service systems*. New York: John Wiley & Sons Inc. The SIAM series in applied mathematics.

Ross, S. M. 1983. *Stochastic processes*. Wiley Series in Probability and Mathematical Statistics.

Samuelson, D. A. 1999. Call attempt pacing for outbound telephone dialing systems. *Interfaces* 29 (5): 66–81.

Taylor, H. M., and S. Karlin. 1998. *An introduction to stochastic modeling*. third ed. San Diego: Academic Press.

Wolff, R. W. 1989. *Stochastic modeling and the theory of queues*. New York: Prentice-Hall.