Proceedings of the 2003 Winter Simulation Conference
S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, eds.

MODELLING AND SIMULATION OF A TELEPHONE CALL CENTER

Juta Pichitlamken, Alexandre Deslauriers, Pierre L'Ecuyer, and Athanassios N. Avramidis

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

ABSTRACT

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. Our objective is to allocate a number of agents such that some service requirement is satisfied. We have taken two approaches in analyzing this staffing problem: We developed a simulation model of the call center, which allows us to do a what-if analysis, as well as continuous-time Markov chain (CTMC) queueing models, which provide approximations of system performance measures. We describe the simulation model in this paper.

1 INTRODUCTION

We consider a telephone call center with two types of traffic, *inbound* and *outbound*, and two types of agents, *inbound-only* and *blend*. The number of agents of each type can vary from day to day and within each day. The inbound calls arrive according to a Poisson process whose rate may itself evolve as a 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.

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. We are primarily interested in finding the number of agents such that at least $p\%$ of customers should have delay time less than s seconds, for arbitrary p and s. Other performance measures of interest are agent utilization, abandonment rate, and rate of outbound calls.

We have taken two approaches in analyzing the staffing problem: We developed a simulation model of the call center and continuous-time Markov chain (CTMC) queueing models. Each method has its own appeal: The simulation approach is highly flexible, e.g., it can be tailored to specific details and is easy to modify. The simulation model also allows us to do a what-if analysis and learn additional information that may otherwise not be available, e.g., times that customers are willing to wait before abandoning. On the other hand, the CTMC models are insightful, sometimes faster computationally, and relatively easier to construct than a simulation model. Moreover, a call center can be naturally viewed as a queueing system, e.g., the simplest CTMC model for an inbound call center is a $M/M/s$ queueing model (see Gans, Koole, and Mandelbaum 2002 and Koole and Mandelbaum 2002 for an overview of queueing models in call center applications). In this paper, we will only describe our simulation model of call centers. Our CTMC development can be found in Deslauriers et al. (2003).

This paper is organized as follows: Section 2 discusses some difficulties that we encountered in modelling the call centers. We describe our data analysis in Section 3. Section 4 explains how we construct our simulation model and how it is validated. Supposing that the simulation model reproduces performance of a real call center, we explore how other management policies affect the call center performance in Section 5.

2 DIFFICULTIES ENCOUNTERED

Although the call center staffing problem poses many real-world issues that require us to make seemingly simplistic assumptions, the resulting simulation model is reasonably good at emulating the performance of a real call center (see Section 5) while maintaining its parsimony. We describe the issues that we have encountered

in developing the call center simulation model in this section.

The types of data that are traditionally available at call centers poses many challenges, one of which is due to the aggregation of data over some period, typically 30 minutes. That is, for each half hour, we have the number of (inbound call) arrivals, the sum of service times of the inbound calls served, and similarly for the outbound calls, but not the arrival times or service times of the individual calls (with the exception on outbound calls in our case). The lack of call-by-call information complicates the data analysis because standard parameter estimation methods generally do not apply. In addition, the stochastic nature of call centers add difficulties to the data analysis, e.g., the arrival rates varies from day to day and within each day.

It has been observed that the arrivals to a call center are not realistically modelled by a process with a deterministic time-varying arrival rate (Avramidis, Deslauriers, and L'Ecuyer 2003, Jongbloed and Koole 2001 and Brown et al. 2002). From empirical study, call center arrivals are known to have a variance that is considerably higher than implied by Poisson arrival (Jongbloed and Koole 2001 and Deslauriers 2003) and strong positive association between the arrivals in different time periods (Tanir and Booth 1999 and Brown et al. 2002).

Moreover, some relevant information is simply not available. For example, in an ideal world, we would use the distribution of time that a customer is willing to wait before abandoning the queue (called the *patience time*) to model the abandonment process. Instead, what we have is the number of customers abandoning and a rough histogram of distribution of the waiting times before they hang up. In other words, we have a problem of highly *censored* data; we only observe the maximal patience times of those customers who abandon, but we have no information about these times for customers who are served.

Another piece of missing information is how the dialer works (i.e., the algorithm it uses to activate outbound calls and how many outbound calls it makes) for it is a proprietary knowledge of a software vendor. We gather from Bell staff that the dialer considers the number of idle agents and some measures of the quality of service, e.g., the fraction of inbound calls that waits for longer than some threshold averaged over some previous time interval. However, we do not know how the dialer actually uses this information. Because the dialer is key to the call center performance in blend environment, the lack of knowledge on the dialer makes it more difficult to validate our simulation model. Specifically, when we compare the simulated performance measures to the empirical values, we cannot be sure if the discrepancies we observe are due to our modelling assumptions or due to our lack of knowledge on the dialer.

Human aspects of call center operations also complicate model validation. We observe that the empirical quality of service (QoS, defined as the fraction of inbound customers whose waiting time is smaller than 20 seconds) is better than the target (80%) most of the time. From the discussion with Bell staff, we speculate that this is partly because call center managers respond "too quickly" when they observe short-term poor QoS by manipulating the dialer aggressiveness parameter that controls how often the dialer makes outbound calls and how many calls it attempts at a time. We do not know how the managers control the dialer in real time or if they do so in a systematic fashion. In essence, the manager's control of a dialer coupled with the algorithms inside the dialer constitute a black box which we regard as the dialer in our model.

Another human factor comes from the call center agents themselves. The time that they are available to take a call is very likely to be less than the time for which they are scheduled, because of coffee breaks, trips to restrooms, absenteeism, etc. At this moment, due to the lack of information and for the sake of model simplicity, these factors are taken into account globally by reducing the number of agents by some fixed percentage (see Section 4).

3 INPUT MODELLING

The call center 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. Because all the available data is aggregated as *averages* over half-hour periods, it is natural to assume that the model parameters (e.g., arrival rate, service time distributions) are constant over each half hour, and we proceed as such. That is, the planning horizon is partitioned into half-hour time periods; period 1 is 8.00-8.30, period 2 is 8.30-9.00, and so on. We experiment with the empirical data from Bell Canada to find the fitting distribution for each process, but we will discuss only the arrival and the service processes in this paper (see Deslauriers 2003 for additional details on the model). For those who are interested in data analysis for call centers, Brown et al. (2002) offer extensive study of call-by-call data as well as investigation of how well conventional queueing models perform in such cases (inbound-only call centers).

3.1 Arrival Processes

After verifying that a Poisson process with a *deterministic* time-varying arrival rate cannot realistically model

the call center arrivals, we consider a Poisson process with *stochastic* rate. This choice is partially supported by the empirical evidence in Brown et al. (2002) where they concludes that the arrival processes of call center are well modelled as an inhomogeneous Poisson process. Let X_i be the number of inbound call arrivals in half hour i, with the probability mass function:

$$\Pr\{X_i = x\} \quad = \quad e^{-\Lambda_i}\frac{\Lambda_i^x}{x!}. \tag{1}$$

Jongbloed and Koole (2001) model Λ_i as a gamma random variable with density:

$$g_i(\lambda) = \frac{\beta_i^{-\alpha_i}}{\Gamma(\alpha_i)}\lambda^{\alpha_i-1}e^{-\lambda/\beta_i}, \tag{2}$$

for $\lambda > 0$, where $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$. 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 time interval has the negative binomial distribution.

To estimate the parameters in (2), we use the maximum likelihood estimation (MLE) method where we estimate the parameters of the negative binomial distribution. Let r be the number of days in our data, d be the total number of half-hour periods in a day (in our case, it is 25), $X_{i,j}$ be the number of arrivals in half hour i of day j,

$$\begin{aligned}
\bar{X}_i &= \sum_{j=1}^{r} X_{i,j}/r \\
M_i &= \max_{1\le j\le r} X_{i,j} \\
f_i(k) &= \sum_{j=1}^{r} \mathcal{I}\{X_{i,j} \ge k\},
\end{aligned}$$

where $\mathcal{I}\{\epsilon\}$ is 1 if ϵ is true and 0 otherwise. The log-likelihood function of observing $\{X_{i,j}, 1 \le i \le d, 1 \le j \le r\}$ under (1)–(2) is

$$\begin{aligned}
\widetilde{\phi}_i(\alpha_i) \quad = \quad & \sum_{k=1}^{M_i} f_i(k)\ln(\alpha_i + k - 1) \\
& + r\alpha_i \ln(\alpha_i/(\alpha_i + \bar{X}_i)) \\
& + r\bar{X}_i \ln(\bar{X}_i/(\alpha_i + \bar{X}_i)).
\end{aligned}$$

We will denote an estimator of a parameter a as $\widehat{a}$. The desired $\widehat{\alpha}_i$ is the the value at which $\widetilde{\phi}_i(\alpha_i)$ is maximized with respect to α_i, i.e., the root of the first derivative of $\widetilde{\phi}_i(\alpha_i)$ with respect to α_i. The parameter $\widehat{\beta}_i$ can be obtained by first solving for the negative binomial parameter $\widehat{\varphi}_i = \widehat{\alpha}_i/(\widehat{\alpha}_i + \bar{X}_i)$ then $\widehat{\beta}_i$ is simply $(1 - \widehat{\varphi}_i)/\widehat{\varphi}_i$.

We test the goodness of fit via the Kolmogorov-Smirnov (KS) test statistic

$$D_i \quad \stackrel{\text{def}}{=} \quad \sup_x \left|\widetilde{F}_i(x) - \widehat{F}_i(x)\right|, \tag{3}$$

where $\widetilde{F}_i(x)$ is the empirical distribution and $\widehat{F}_i(x)$ is the estimated distribution of X_i for half hour i. Because the empirical data are already used for estimating the distribution parameters, the distribution of D_i under the null hypothesis is complicated and unknown. We estimate it via a bootstrapping technique (Ross 1997) as follows. Using the parameterized distribution $F_{\hat{\alpha}_i,\hat{\beta}_i}$, we simulate a new sample path for the same length of time as the empirical data. From this realization, we again estimate the distribution parameters in (2) via the MLE method so that we can compute the bootstrapped $D_i{}^*$ via (3). By repeating this process, say, B, times we can estimate the p-value—the probability that we observe D_i conditional on the hypothesis that the parameterized distribution is the true underlying distribution—for half hour i as:

$$p_i \quad \approx \quad \frac{1}{B}\sum_{k=1}^{B} \mathcal{I}\{D_{i,k}^* \ge D_i\}.$$

A drawback of the Poisson-gamma process is that the number of arrivals in one time period is independent of those in all other periods. This assumption rarely holds in practice (Tanir and Booth 1999). Avramidis, Deslauriers, and L'Ecuyer (2003) model the Λ_i in (1) as *dependent* random variables where

$$\Lambda_i \quad = \quad W\lambda_i, \tag{4}$$

the λ_i's are constants to be estimated, and W is a gamma random variable with parameters (α', β') and $\mathrm{E}[W] = 1$. (See Brown et al. 2002 for a Poisson model with auto-regressive rate parameters across successive days.) The idea is to let the random factor W account for the day-to-day traffic variation. The main advantages of this model are (a) mathematical tractability: under (1) and (4), the distribution of X_i is negative multinomial distribution; and (b) the possibility of time dependence, which improves modelling realism, e.g., a time period with a heavy inbound traffic would likely be followed by a high call volume in the next time period (however, the random variable W induces a positive correlation between the number of arrivals in successive half hours). In addition, the arrival process (4) has fewer parameters than (2), i.e., 26 vs. 50 parameters for 25 time periods under study.

We estimate parameters in (4) via MLE. Let

$$o_l \quad = \quad \sum_{j=1}^{r} \mathcal{I}\left\{\sum_{i=1}^{d} X_{i,j} \ge l\right\}$$

Table 1: Parameter Estimates for the Poisson Arrival Process with a Gamma-Distributed Correlation Factor for Tuesday (The Number of Arrivals is Per Half Hour)

	Value		Value
$\widehat{\alpha}'$	36.0	$\widehat{\lambda}_{13}$	57.3
$\widehat{\beta}'$	0.0278	$\widehat{\lambda}_{14}$	54.8
$\widehat{\lambda}_1$	26.5	$\widehat{\lambda}_{15}$	57.8
$\widehat{\lambda}_2$	38.4	$\widehat{\lambda}_{16}$	58.8
$\widehat{\lambda}_3$	52.4	$\widehat{\lambda}_{17}$	60.1
$\widehat{\lambda}_4$	61.1	$\widehat{\lambda}_{18}$	54.3
$\widehat{\lambda}_5$	63.8	$\widehat{\lambda}_{19}$	46.3
$\widehat{\lambda}_6$	62.2	$\widehat{\lambda}_{20}$	40.5
$\widehat{\lambda}_7$	66.2	$\widehat{\lambda}_{21}$	35.0
$\widehat{\lambda}_8$	59.8	$\widehat{\lambda}_{22}$	31.2
$\widehat{\lambda}_9$	58.8	$\widehat{\lambda}_{23}$	27.1
$\widehat{\lambda}_{10}$	57.4	$\widehat{\lambda}_{24}$	24.7
$\widehat{\lambda}_{11}$	58.0	$\widehat{\lambda}_{25}$	17.9
$\widehat{\lambda}_{12}$	57.8		

$$q = \max_{1\le j\le r} \sum_{i=1}^{d} X_{i,j},$$

and Δ be a constant that is independent to the parameters we want to estimate. For $\mathrm{E}[W] = 1$, we have that $\beta' = 1/\alpha'$, and the log likelihood function is:

$$\begin{aligned}\widetilde{\phi}(\alpha', \lambda_1, \ldots, \lambda_d) &= \sum_{l=1}^{q} o_l \log(\alpha' + l - 1) \\ &+ \Delta + r\alpha' \log\left(\frac{\alpha'}{\sum_{k=1}^{d}\lambda_k + \alpha'}\right) \\ &+ \sum_{j=1}^{r}\left(\sum_{i=1}^{d} X_{i,j}\log\left(\frac{\lambda_i}{\sum_{k=1}^{d}\lambda_k + \alpha'}\right)\right).\end{aligned}$$

We get a better goodness of fit when we assume that the arrival process is time-of-the-day and day-of-the-week dependent. Table 1 shows the estimated parameters for Tuesdays. We observe the arrival rates that are time-of-the-day dependent. We will use model (4) in our simulation model.

3.2 Service Time Distributions

We tried to fit an inbound service time S_1 as an exponential random variable with rates that are piecewise-constant over each half-hour. Recall that we only have sum of service times and not call-by-call service times. Let $X'_{i,j}$ be the number of inbound calls served in half-hour i on day j, and $Y_{i,j}$ is the sum of service times of these calls. The maximum likelihood estimator of the service rate for half hour i is

$$\widehat{\mu}_{1,i} = \frac{\sum_{j=1}^{r} {X'}_{i,j}}{\sum_{j=1}^{r} Y_{i,j}},$$

We assess the goodness of fit by the bootstrapping KS test described in Section 3.1. Because we do not have individual service times, we use the property of the exponential distribution that, for $Z_1, Z_2, \ldots, Z_\alpha$ i.i.d. exponential random variables with mean β, $\sum_{i=1}^{\alpha} Z_i$ is a gamma random variable with parameters α and β. We do the bootstrapping KS test on the *sum* of service times. We find that the exponential distribution does not provide a satisfactory fit to our data. Note that we simply test if the sum of the service times, conditional on the number of calls served, follows a gamma distribution. This test is weaker than testing if individual service times are exponential, yet the null hypothesis is rejected by a wide margin (the estimated p-values are ≤ 0.001 for 11 of the 25 half-hour periods), so it seems that the service times are far from exponential.

As is also suggested in Brown et al. (2002), we have experimented with the lognormal distribution whose density is

$$\frac{1}{x\sqrt{2\pi\sigma^2}} \exp\frac{-(\ln x - \mu)^2}{2\sigma^2} \text{ for } x > 0. \tag{5}$$

We use the method of moments to get the estimators. Let the kth moment of S_1 be $m_k = \mathrm{E}(S_1^k)$, and the average inbound service time for half hour i be $\bar{Y}_i$. The first two moments are simply

$$\mu = 2\ln m_1 - \frac{1}{2}\ln m_2,\ \sigma^2 = \ln m_2 - 2\ln m_1, \tag{6}$$

and the m_1 estimate is

$$\widehat{m}_1 = \frac{\sum_{k=1}^{d} X'_k \bar{Y}_k}{\sum_{k=1}^{d} X'_k}. \tag{7}$$

We obtain m_2 via the relationship $m_2 = \mathrm{Var}(S_1) + m_1^2$, and the unbiased estimator of $\mathrm{Var}(S_1)$ (see Deslauriers 2003 for the proof):

$$\widehat{\mathrm{Var}}(S_1)^2 = \frac{1}{d-1}\sum_{k=1}^{d} X'_k(\bar{Y}_k - \widehat{m}_1)^2. \tag{8}$$

For S_1 in seconds, we get $\hat{\mu} = 5.874$ and $\hat{\sigma}^2 = 0.948$ in (5).

Another distribution we have explored is the gamma. The method of moments yields the gamma parameter estimates in Equation (2) as:

$$\widehat{\beta} = \frac{\widehat{\mathrm{Var}}(S_1)^2}{\widehat{m}_1} \text{ and } \widehat{\alpha} = \frac{\widehat{m}_1^2}{\widehat{\mathrm{Var}}(S_1)^2}$$

We considered the case where the service times are time-of-the-day independent and dependent, and we got a better fit with the latter. Table 2 shows the parameter estimates and its bootstrapped p-values. We decided to use the gamma distribution in our simulation model in Section 4 because it was easier to test the goodness of fit for this distribution than for the lognormal, and the fit was reasonably good.

Table 2: Estimated Parameters of Service Times of Inbound Calls Under the Gamma Model (The Service Times are in Seconds)

i	$\widehat{\alpha}_i$	$\widehat{\beta}_i$	p_i	i	$\widehat{\alpha}_i$	$\widehat{\beta}_i$	p_i
1	1.374	357.4	0.55	14	0.702	838.4	0.34
2	0.924	608.9	0.78	15	0.782	727.9	0.89
3	0.956	634.1	0.50	16	0.776	725.2	0.73
4	0.807	750.5	0.65	17	0.821	678.3	0.93
5	0.735	811.3	0.21	18	0.565	992.3	0.76
6	0.706	843.6	0.61	19	0.583	970.4	0.73
7	0.700	888.9	0.60	20	0.583	917.0	0.66
8	0.533	1126.2	0.58	21	0.496	1080.2	0.16
9	0.664	863.9	0.37	22	0.487	1089.0	0.44
10	0.740	770.0	0.89	23	0.506	1021.4	0.35
11	0.465	1218.0	0.38	24	0.536	944.1	0.79
12	0.615	935.0	0.93	25	0.505	986.7	0.34
13	0.697	844.9	0.56				

Unlike inbound service times, we do have call-by-call outbound service times. We first explore modelling the outbound service times with parameterized distributions such as exponential, gamma and lognormal. The lognormal appears to be a good choice if we assume that the outbound service times are half-hour dependent. Nevertheless, because we have a large amount of data, the KS goodness-of-fit test rejects all the distributions we tried. In the simulation, we generate the service times with a density obtained via a kernel density estimation method, using the UNURAND package (Leydold and Hörmann 2002).

4 OTHER ASPECTS OF THE SIMULATION MODEL

In our simulation model of the call center, there are $n_{i,1}$ identical inbound agents and $n_{i,2}$ identical blend agents during period i. These integers are parameters of the model. There is a single FIFO waiting queue for inbound calls. A customer who is not served immediately hangs up with probability 0.005; otherwise, he joins the queue from which he will abandon if experiencing a waiting time greater than his *patience time*. We model this patience time as an exponential random variable with mean $1/\eta_i$ for half hour i.

Our dialer model tries to emulate the real dialer in that the decision on when and how many outbound calls to make is based on the current state of the system. When the service of a customer ends, if the number of idle blend agents is N_2, the dialer makes outbound calls if all of the following three conditions are satisfied: (a) $N_2 > 1$; (b) the number of busy agents (of any type) is at most $n_{i,1} + n_{i,2} - 4$; and (c) more than 75% of the inbound calls that arrive over the last 10 minutes wait for less than 20 seconds. The number of calls composed is $2N_2$ if the percentage of mismatches averaged over the last 15 minutes does not exceed 8% of the total outbound calls attempted; otherwise, the number of outbound calls composed is N_2. We do not claim that this heuristic is a good control policy for the dialer. We merely want to reproduce what we have observed in the empirical data and learned by talking to the call center managers.

Each outbound call successfully reaches a customer with probability κ_i during half hour i. The answering time for an outbound call, defined as the time required by the dialer to either reach the customer or recognize that the attempt is not successful, is exponentially distributed with mean 2 seconds.

Under the arrival process and service time distributions described in Section 3 and the parameter values in Table 3, we have validated our simulation model by comparing the simulation results to the empirical data collected form the center. Using the average number of agents from the empirical data, we noticed that our QoS was higher and the agent occupation fraction (defined as the ratio of times agents are busy to the total scheduled times) lower than in the data. We think that this is because our agents are too "efficient," in a sense that they are never absent or take a break. We were able to obtain results much closer to the empirical data by assuming that the inbound agents are available only 90% of time and the blend agents are available only 85% of time. (To implement this, we simply multiplied the number of agents of each type by the appropriate percentage rounded to the next integer. The numbers given in Table 3 are *before* this reduction).

In Section 3, we saw that the arrival rates are day-of-the-week dependent, and so is the number of agents. Thus, we simulate each day of the week separately. Table 4 shows the simulation results for Tuesdays. The half widths of the 95% confidence intervals are obtained by assuming that the simulation outputs are i.i.d. normally distributed. Here and in all other forthcoming tables, ϵ denotes a value less than 0.1. For most performance measures, there is no significant difference between the results of the simulation and the empirical data. The number of abandoned calls in the simulation could better match what we observe in the empirical

Table 3: Input Parameters of the Simulation Model for Tuesdays

Period i	Start time (hr)	End time (hr)	Out. success prob. κ_i	Mean patience time $1/\eta_i$ (sec)	# in. agents n_1	# blend agents n_2
1	8.0	8.5	0	400	11.4	0
2	8.5	9.0	0	400	18.6	0
3	9.0	9.5	0	400	24.3	0
4	9.5	10.0	0	700	27.9	0
5	10.0	10.5	0	700	28.1	0
6	10.5	11.0	0	600	28.1	0
7	11.0	11.5	0	600	27.7	0
8	11.5	12.0	0	600	27.8	0
9	12.0	12.5	0	600	25.8	0
10	12.5	13.0	0	600	25.9	0
11	13.0	13.5	0	500	29.0	0
12	13.5	14.0	0	500	28.9	0
13	14.0	14.5	0.27	500	26.6	6.1
14	14.5	15.0	0.27	500	25.0	14.1
15	15.0	15.5	0.28	500	25.6	19.5
16	15.5	16.0	0.29	500	26.5	21.7
17	16.0	16.5	0.29	500	24.8	20.2
18	16.5	17.0	0.30	500	21.4	18.6
19	17.0	17.5	0.33	500	19.6	14.1
20	17.5	18.0	0.37	500	9.9	21.3
21	18.0	18.5	0.40	500	4.1	21.2
22	18.5	19.0	0.38	500	3.2	20.6
23	19.0	19.5	0.41	500	2.8	19.8
24	19.5	20.0	0.41	100	3.3	21.9
25	20.0	20.5	0.41	50	3.2	20.8

Table 4: Comparison of Daily Performance Measures Averaged from Empirical Data and Those Obtained by Simulation of 100,000 Days

Performance measure	Tuesday	
	Simulated	Empirical
QoS (%)	$88.3 \pm \epsilon$	87.9 ± 2.4
Inbound calls arrived	1230.9 ± 1.3	1228.1 ± 67.1
Abandoned calls	26.9 ± 0.2	28.1 ± 4.8
Outbound calls attempted	1952.9 ± 1.6	1783.7 ± 218.1
Outbound calls served	601.7 ± 0.5	565.3 ± 69.6
Mismatches	44.4 ± 0.1	38.5 ± 6.2
Agent occupation (%)	$71.1 \pm \epsilon$	71.7 ± 2.8

data had we have better information on customer patience time.

We have developed a simulation tool in C for simulating our models of call centers. The software has a modular design, which is practical in that it enables users to understand the structure and relationships between the various aspects of the model without going into much detail. In addition, it allows stability in the general structure of the simulation model as it evolves; the modification of certain details is done inside the corresponding module while leaving other modules unchanged.

The simulation programs are fast. To give an idea, it takes approximately 12 minutes of CPU time on a 2MHz Athlon-XP processor running the RedHat Linux 8 operating system to simulate 100,000 operating days of the call center.

5 NUMERICAL EXPERIMENTS

The goal of this section is to provide some examples of what-if analysis that the simulation model allows us to do. For this purpose, we assume that the simulation model described in Section 4 replicates the performance of a real call center, and we use it to benchmark the changes that we make.

5.1 Improving the Dialer's Operation

The policy of the dialer used in our model (and in the call center) attempts at maintaining the QoS above 80%

every day, and perhaps every hour, by basing its decision on the QoS and number of mismatches observed over the past 10 minutes. It would certainly be less restrictive to respect the QoS requirement only over the long term (say, one month or one year) rather than in the short term. Simulation experiments can give us an idea of how much we can gain by changing the policy in that direction, i.e., adopting a policy that avoids looking at the QoS over the past few minutes or hours and bases its decision only on the current system's state. Additional motivation for looking at this came from the observation that call center managers may have a tendency to modify the dialer's aggressiveness parameters and over-react when they see poor QoS in the last few minutes. This type of behavior degrades the performance of the system in the long run.

To illustrate this, we made simulation experiments with the following simple dialing rule: at the end of a service, if N_2 blend agents are idle, the system dials round(τN_2) numbers in parallel for some fixed constant τ, where "round" means rounding to the nearest integer.

Table 5 gives the results for $\tau = 1.2$, 1.4, 1.6, and 2.0. We see that the volume of outbound calls completed increases significantly, and the agent occupation increases slightly, compared with the original rule described in Section 4 (see Table 4 under column *Simulated*). Of course, these values also increase with τ. The QoS decreases slightly, but still remains well above the 80% limit, even for $\tau = 2$. The number of abandonments is larger than with the original rule and increases slowly with τ. The number of mismatches, on the other hand, increases very rapidly with τ. It is smaller than with the original rule for $\tau \leq 1.4$ and larger for $\tau \geq 1.6$. (Note that with $\tau = 1$, there would be no mismatch.) The appropriate choice of τ would depend on how the call center managers value the increases/decreases in these different performance measures. For instance, we see that the policy with $\tau = 1.6$ achieves a much larger volume of outbound calls (around 11% increase) than the original rule. On the other hand, there are more abandonments and mismatches. It should be left to the managers to decide if the value of the increased volume of outbound calls outweights the "cost" of these additional abandonments and slight QoS decrease.

More refined stationary rules could also be considered and could certainly improve on the simple rules in Table 5. Such rules could take into account the number of idle agents of each type and perhaps the current arrival rate λ_i. Then, one can define an optimization problem by imposing constraints on some of the long-run performance measures and incorporating the others into the objective function. The decision variables of this problem would be the parameters of the dialing rule (for the above simple rule, it is τ). The optimization problem could be solved via optimization-by-simulation methodology.

Table 5: Daily Performance Measures Obtained from the Simulation with a New Dialing Heuristic

Performance measures	τ				Half width
	1.2	1.4	1.6	2.0	
QoS (%)	86.0	85.4	84.5	84.1	ϵ
Abandonments	36.1	38.2	41.6	42.8	0.3
Outbound calls served	639.0	652.6	669.5	677.0	0.4
Mismatches	2.2	15.3	75.0	98.8	0.1
Agent occup. (%)	72.0	72.3	72.8	73.0	ϵ

5.2 Sensitivity of the Performance Measures to the Staffing Level

We now look at how the performance measures are affected by a change in the staffing level. From the simulation experiment, we observe that the QoS is higher than what is required (88.3% vs 80%). The simulation model allows us to assess the call center performance if we lower the staffing level.

Table 6 shows the performance measures when we decrease the number of agents by 5%. The QoS is still comfortably above the requirement, but the number of outbound calls served decreases. These results enable the call center managers to evaluate if the saving of 5% reduction in the number of agents is enough to compensate the loss of revenue resulting from fewer outbound calls and the loss of customer satisfaction as manifested by the increase in the number of abandoned calls. (One possible explanation for the decrease in agent occupancy is that when there are fewer agents, the dialer is triggered to make outbound calls less frequently as the threshold condition is harder to satisfy.)

Table 6: Daily Performance Measures Obtained from the Simulation with 5% Fewer Agents

Performance measures	Tuesday
QoS (%)	85.4 $\pm$ ϵ (-2.9)
Outbound calls served	540.4 $\pm$ 0.5 (-61.3)
Abandoned calls	34.6 $\pm$ 0.3 ($+7.7$)
Mismatches	39.3 $\pm$ 0.1 (-5.1)
Agent occupation (%)	68.8 $\pm$ ϵ (-2.3)

5.3 Sensitivity of the Performance Measures to the Distributions of Stochastic Processes

In the next experiment, we modify the assumptions of our simulation model by changing the distributions of the arrival process and the service times of inbound and outbound calls to resemble those of a M/M/s queueing model which is often used to model call centers (Koole and Mandelbaum 2002). As we previously mentioned, we have also developed CTMC models in parallel to the simulation model, so we are curious to see how the change in the input distributions would affect the call center performance.

We consider the Poisson arrival process with the deterministic time-of-the-day dependent arrival rates (i.e., $W = 1$ in (4)) and exponentially distributed service times (for inbound and outbound calls, with different distribution parameters). Note that these distributions have the same means as their corresponding counterparts that we have chosen for our original simulation model.

Table 7 shows a significant increase in the QoS of the simulation under the new set of distributions compared to the original simulation model. This is not surprising, because assuming deterministic arrival rates reduces the traffic variability. This reduces congestion in the system and improves the QoS. We also observe a significant decrease in the in the volume of outbound calls. Our experiment shows that simply using the assumptions of a M/M/s queueing model can give significant error in performance measures estimates. The significance of these errors depends of course on the other sources of error in the model (e.g., amount and reliability of the data) and also on what the managers find acceptable.

Table 7: Daily Performance Measures Obtained from the Simulation Where the Arrival Process is Poisson with Deterministic Rates and Exponential Service Times

Performance measures	Tuesday
QoS (%)	91.7 $\pm\epsilon$ (+3.4)
Outbound calls served	572.0 $\pm$ 0.7 ($-$29.7)
Abandoned calls	16.5 $\pm$ 0.2 ($-$10.4)
Mismatches	42.2 $\pm$0.1 ($-$2.2)
Agent occupation (%)	70.2 $\pm\epsilon$ ($-$0.9)

REFERENCES

Avramidis, A. N., A. Deslauriers, and P. L'Ecuyer. 2003. Modeling daily arrivals to a telephone call center. Technical report, GERAD and DIRO, University of Montreal. Submitted for publication.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. 2002, November. Statistical analysis of a telephone call center: A queueing-science perspective. Technical report, The Wharton School, University of Pennsylvania, Philadelphia. Preprint.

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.

Deslauriers, A., J. Pichitlamken, P. L'Ecuyer, and A. N. Avramidis. 2003. Markov chain models of a telephone call center in blend mode. Technical report, GERAD and DIRO, University of Montreal. Preprint.

Gans, N., G. Koole, and A. Mandelbaum. 2002. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management*. To appear.

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., 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>.

Ross, S. M. 1997. *Simulation*. Second ed. Academic Press.

Tanir, O., and R. J. Booth. 1999. Call center simulation in Bell Canada. In *Proceedings of the 1999 Winter Simulation Conference*, ed. P. A. Farrington, H. B. Nemhard, D. T. Sturrock, and G. W. Evans, 1640–1647. Piscataway, New Jersey: IEEE Press. Available on line via <www.informs-cs.org>.

ACKNOWLEDGMENTS

This research was supported by grants number OGP-0110050 and CRDPJ-251320 from NSERC-Canada, a grant from Bell Canada via the Bell University Laboratories, and grant number 00ER3218 from NATEQ-Québec to the third author. The work of the second author was supported by an NSERC-Canada scholarship. We thank Bell Canada for providing us the data and their support, and Eric Buist for his help in running the simulations.

AUTHOR BIOGRAPHIES

JUTA PICHITLAMKEN received her Ph.D. from the Department of Industrial Engineering and Manage-

ment Sciences at Northwestern University in 2002. She is currently a Postdoctoral Fellow at the University of Montreal. Her research interests include ranking and selection procedures and simulation optimization. Her e-mail address is <pichitla@iro.umontreal.ca>.

ALEXANDRE DESLAURIERS has completed his M.Sc. in Operations Research at the Université de Montréal in 2003. He worked on the modeling and simulation of call centers, under the supervision of Pierre L'Ecuyer. He is now with Hydro-Québec.

PIERRE L'ECUYER is Professor in the Département d'Informatique et de Recherche Opérationnelle, at the Université de Montréal. His main research interests are random number generation, quasi-Monte Carlo methods, efficiency improvement via variance reduction, sensitivity analysis and optimization of discrete-event stochastic systems, and discrete-event simulation in general. He obtained the prestigious *E. W. R. Steacie* fellowship in 1995-97 and a *Killam* fellowship in 2001-03. His recent research articles are available on-line at <http://www.iro.umontreal.ca/~lecuyer>.

ATHANASSIOS (THANOS) N. AVRAMIDIS is an Invited Researcher at the Département d'Informatique et de Recherche Opérationnelle at the Université de Montréal. He has been on the faculty at Cornell University and a consultant with SABRE Decision Technologies. His primary research interests are Monte Carlo simulation, particularly efficiency improvement, the interface to probability and statistics, and applications in computational finance, service operations. His web page is <http://www.iro.umontreal.ca/~avramidi>.