Proceedings of the 2005 Winter Simulation Conference
M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.

MODELING AND SIMULATION OF CALL CENTERS

Athanassios N. Avramidis
Pierre L'Ecuyer

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

ABSTRACT

In this review, we introduce key notions and describe the decision problems commonly encountered in call center management. Main themes are the central role of uncertainty throughout the decision hierarchy and the many operational complexities and relationships between decisions. We make connections to analytical models in the literature, emphasizing insights gained and model limitations. The high operational complexity and the prevalent uncertainty suggest that simulation modeling and simulation-based decision-making could have a central role in the management of call centers. We formulate some common decision problems and point to recently developed simulation-based solution techniques. We review recent work that supports modeling the primitive inputs to a call center and highlight call center modeling difficulties.

1 INTRODUCTION

Call centers are an important component of the global economy. Around 3% of the workforce in the United States and Canada works at a call center (Call Center News Service 2001). More people in North America work in call centers than in agriculture. Most of the operating cost of call centers (around 3/4) is labor costs. These call centers handle customer support, phone orders and sales, marketing, governmental information services, emergency services (police, ambulance), etc. A current trend is the extension to a *contact center*, whereby telephone services are enhanced by services in other media such as e-mail, fax, or chat.

In this review, we introduce key notions and describe the decision problems commonly encountered in call center management. The main themes elaborated are: the central role of uncertainty throughout the decision hierarchy; the many operational complexities and relationships between decisions; and a review of work that supports modeling the primitive inputs to a call center. We also make connections to analytical models in the literature, emphasizing insights gained and model limitations. The high operational complexity and the prevalent uncertainty suggest that simulation modeling and simulation-based decision-making could have a central role in the management of call centers. Mehrotra and Fama (2003) also discusses simulation-based decisions for call centers, from an applied point of view. Gans et al. (2003) is an excellent, in-depth tutorial on call centers.

2 KEY NOTIONS

A *call center* is a set of resources (communication equipment, employees, computers, etc.) which enable the delivery of services via the telephone. *Inbound calls* are those initiated by customers calling in to the center. A customer can be *blocked*, i.e., receive a busy signal, if all of the center's phone lines are busy at the time he calls. At first, calls may be connected to an *interactive voice response* (IVR) unit. The latest generation of speech-recognition technology allows IVRs to interpret complex user commands, so customers may be able to "self-serve", i.e., complete the service interaction at the IVR. Otherwise, calls are passed from the IVR to an *automatic call distributor* (ACD). An ACD is a specialized switch designed to route each call to an individual *agent*; if no qualified agent is available, then the call is placed in a queue. Modern ACDs are sophisticated, allowing routing rules based on many criteria. A queued customer may *abandon* without receiving service.

In a *multi-skill* call center, we distinguish various call *types* (or *skills*), and we distinguish agents by their *skill group*, defined as the subset of call types they can handle. *Skill-based routing* (SBR), or simply *routing*, refers to rules (programmed in the ACD) that control in real time the agent-to-call and call-to-agent assignments. There is a trend towards multi-skill centers with SBR (Koole and Mandelbaum 2002); according to Mehrotra and Fama (2003), the multi-skill call center has become ubiquitous.

A *blend center* is one where inbound calls are blended with *outbound calls*; these are initiated by agents calling

customers, usually aided by a *predictive dialer* that tries to anticipate the number of free agents at the time customers are reached. A *mismatch* occurs whenever the called party answers but cannot be served immediately.

Typically, call center managers are interested in many *performance measures*; commonly encountered are: (1) the *service level* (SL); this is the fraction of calls that wait less than an *acceptable wait time* (typically 20 to 30 seconds), usually observed separately by pre-selected target periods (e.g., each hour, day, etc.) and, in multi-skill centers, by call type; (2) the *abandonment ratio*; this is the fraction of calls that abandon; (3) the expected wait time. Additionally, for blend centers: (4) the number of outbound calls completed; (5) the number of mismatches.

3 DECISION PROBLEMS

The hierarchy of call-center decisions can be summarized as follows (loosely adapted from Koole (2005)): *Strategic decisions*: made by upper management, concerning the role of the center in the company, the type of service to be delivered, the budget. *Tactical decisions*: how resources (e.g., budget, human knowledge) should be used; hiring and training of agents. *Planning decisions*: usually, on a weekly basis, new rosters (work schedules for each employee) are made by planners at the call center. *Daily control*: reactions to the current situation, usually taken by shift leaders that monitor service levels and productivity. Typical reactions may be: if the load is less than planned for, then release agents for training or other activities; if the load is more than planned for, then make employees work overtime. *Real-time control*: usually made by the ACD software, sometimes complex; e.g, the call selection and agent selection under SBR; this is the *routing* problem.

Many of these decisions must be made in the face of large uncertainties. At the tactical level, agent hiring and training decisions face uncertain future agent attrition. At the planning level, the staffing and scheduling decisions face uncertainty in future arrival rates (see section 7.1) and also in *realized staffing*, which differs from *planned staffing* due to agent *absenteism.*

The agent hiring and training decisions are part of the broader problem of manpower planning. Bartholomew et al. (1991) review statistical techniques in this field, applying more broadly to sectors beyond call centers. Gans and Zhou (2002) develop a dynamic programming model of long-term hiring and derive optimal policies that are analogs of the inventory literature's "order-up-to" policies. One may envision simulation-based decision-making at this level, but we are not aware of any such work.

In the remainder of this section, we emphasize the decisions at and below the planning level. Consider the decision on how many agents of each skill group to have in the center as a function of time. In a *staffing* problem, the day is divided into periods (e.g., 30 minutes or one hour each) and one simply decides the number of agents of each group for each period, subject to meeting performance constraints, most often on SL, and usually on the abandonment rate. These constraints can be imposed per call type, per period, and/or for aggregations over call types and periods. In a *scheduling* problem, a set of admissible work schedules is first specified, and the decision variables are the number of agents of each skill group in each work schedule. This determines the staffing indirectly, while making sure that it corresponds to a feasible set of work schedules. A yet more constrained version of the problem is when there is a fixed set of available agents to be scheduled for the day or the week, where each agent has a specific set of skills. Then we have a *scheduling and rostering* problem. To issue employee work schedules in a timely manner, these problems must typically be solved several weeks ahead.

These planning problems are closely intertwined with the daily and real-time control problems. The multi-skill and blend capabilities are powerful tools for controlling system performance. In a multi-skill center, the routing may be used as a tool to equalize the SL across classes or enforce desired differences on the SL across classes. The routing may in some cases be subject to technological constraints, and it may also involve objectives that conflict with queueing-system efficiency. As an example of the latter condition, suppose we have call class 1 with a high revenue-generation potential and call class 2 with low or no revenue-generation potential. In addition, agents type A are stronger in selling services, and agents type B are stronger in servicing. Arguably, it is desirable to route calls of class 1 preferentially to type A, and if all type-A agents are busy, only then route to type B. The reverse agent order applies to calls type 2. This "crossed" routing attempts to maximize the rate of assigning "the best agent type for the call type". Similarly, in a blend call center, the outbound capability is a powerful tool for maintaining high agent utilization. Notably, the outbound dialing policy may not be transparent, due to, e.g., a proprietary predictive policy; one such instance is discussed in Deslauriers (2003).

Typically, call center planners solve a single-skill staffing, scheduling, and rostering problem as follows: they ignore (or model very crudely) the uncertainties and *invert* classical formulas such as Erlang-C ($M/M/c$, i.e., without blocking or abandonment) or Erlang-A ($M/M/c+M$, i.e., with abandonment), fed by point forecasts of the arrival, service, and time-to-abandonment rate for the target period, where "inverting" means finding the minimal staffing that meets all target performance constraints. (Encouragingly, Brown et al. (2005) find Erlang-A to work well against empirical data.) The above procedure is applied separately for sub-periods of the day defined so that the arrival rate in each period is deemed near-constant. This

is justified by the Pointwise Stationary (PS) approximation (Green and Kolesar 1991, Whitt 1991). This time-varying staffing is input to a set covering integer programming problem, where decision variables are the counts of selected admissible work schedules and one seeks to minimize staffing costs subject to meeting the target staffing for all periods. A roster is usually created via employee bidding, controlled by a ranking of employees; see Gans et al. (2003).

4 MULTI-SKILL STAFFING AND SCHEDULING: A FORMULATION

As an illustration of a typical real-life call center optimization problem, we adapt from Ceẓik and L'Ecuyer (2004) a Mathematical Programming (MP) formulation of the staffing and scheduling problems in the multi-skill setting. We then briefly review solution approaches.

There are K call types, I skill groups, P time periods, and Q types of work schedules (*shifts*). The *cost vector* is $\mathbf{c} = (c_{1,1}, \ldots, c_{1,Q}, \ldots, c_{I,1}, \ldots, c_{I,Q})^{\mathrm{t}}$, where $c_{i,q}$ is the cost of agent type i having shift q and "t" denotes the vector transpose. The vector of *decision variables* is $\mathbf{x} = (x_{1,1}, \ldots, x_{1,Q}, \ldots, x_{I,1}, \ldots, x_{I,Q})^{\mathrm{t}}$, where $x_{i,q}$ is the number of agents of type i having shift q. We use the vector of *auxiliary variables* $\mathbf{y} = (y_{1,1}, \ldots, y_{1,P}, \ldots, y_{I,1}, \ldots, y_{I,P})^{\mathrm{t}}$ where $y_{i,p}$ is the number of agents of type i in period p. This vector $\mathbf{y}$ satisfies $\mathbf{y} = \mathbf{A}\mathbf{x}$ where $\mathbf{A}$ is a block diagonal matrix with I blocks $\tilde{\mathbf{A}}$, where the element (p, q) of $\tilde{\mathbf{A}}$ is 1 if shift q covers period p, and 0 otherwise. The service level for call type k and period p is

$$g_{k,p}(\mathbf{y}) = \frac{E[\#\text{ calls answered within } s_{k,p} \text{ sec. in period } p]}{E[\#\text{ calls in period } p]}$$

for some constant $s_{k,p}$. Similarly, the aggregate service level over call type k is the expected total number of calls of type k answered within some time limit s_k over the day (say), divided by the expected total number of calls of type k received over the day. We denote by $g_p(\mathbf{y})$, $g_k(\mathbf{y})$ and $g(\mathbf{y})$ the aggregate service levels for period p, call type k, and overall, respectively. The corresponding time limits are s_p, s_k, and s, with minimal service-levels l_p, l_k and l.

A formulation of the *scheduling problem* is

$$\begin{array}{rl} \min & \mathbf{c}^{\mathrm{t}}\mathbf{x} = \sum_{i=1}^{I}\sum_{q=1}^{Q} c_{i,q}x_{i,q} \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{y}, \\ & g_{k,p}(\mathbf{y}) \geq l_{k,p} \quad \text{for all } k,\ p, \\ & g_p(\mathbf{y}) \geq l_p \quad \text{for all } p, \\ & g_k(\mathbf{y}) \geq l_k \quad \text{for all } k, \\ & g(\mathbf{y}) \geq l, \\ & \mathbf{x} \geq 0, \text{ and integer.} \end{array} \tag{P1}$$

The *staffing problem* is a *relaxation* of the scheduling problem where we assume that any staffing $\mathbf{y}$ is admissible. The cost vector is $\mathbf{c} = (c_{1,1}, \ldots, c_{1,P}, \ldots, c_{I,1}, \ldots, c_{I,P})^{\mathrm{t}}$ where $c_{i,p}$ is the cost of an agent of group i in period p. The MP is

$$\begin{array}{rl} \min & \mathbf{c}^{\mathrm{t}}\mathbf{y} = \sum_{i=1}^{I}\sum_{p=1}^{P} c_{i,p}y_{i,p} \\ \text{subject to} & g_{k,p}(\mathbf{y}) \geq l_{k,p} \quad \text{for all } k,\ p, \\ & g_p(\mathbf{y}) \geq l_p \quad \text{for all } p, \\ & g_k(\mathbf{y}) \geq l_k \quad \text{for all } k, \\ & g(\mathbf{y}) \geq l, \\ & \mathbf{y} \geq 0, \text{ and integer.} \end{array} \tag{P2}$$

Simpler instances of (P2) arise by considering a *single period*. Solving the single-period problem in itself should yield practical answers and possibly insights on the joint effect of staffing and routing decisions.

To solve any of these problems, one needs to approximate or estimate the functions $g_\bullet$. Note that $g_{k,p}(\mathbf{y})$ generally depends on the values of $y_{i,j}$ for all i and $j \leq p$, in a very complicated way, and similarly for the other functions $g_\bullet$. For example, the arrival process is generally nonstationary, the service times may have arbitrary distributions, there could be abandonments, routing rules could be complex, etc. Simulation seems to be the only reliable way of estimating the value of these functions for realistic call centers. Ingolfsson et al. (2003) and Atlason et al. (2004) solve multi-period single-skill instances of (P2), and Ceẓik and L'Ecuyer (2004) solve single-period multi-skill instances of (P2). In all three, the solution algorithm involves iterative addition of *cuts* to relaxations of the integer programming problems; the first paper addresses a time-dependent arrival rate and staffing via transient analysis of a continuous-time Markov chain model; the other two papers use simulation to estimate the service levels, and cuts are derived from subgradient estimates for a *sample average approximation* of $g_\bullet$, i.e., a function $\tilde{g}_\bullet$ estimated by simulation. A necessary condition for cut validity is concavity of the functions $g_\bullet$. Atlason et al. (2004) and Ceẓik and L'Ecuyer (2004) document non-concavity of $g_\bullet$, unless the staffing is "sufficiently large", and suggest practical solution heuristics. Avramidis et al. (2005) also solve single-period multi-skill instances of (P2) heuristically, using a randomized search driven by the performance approximation discussed in section 5.3, and, at a final stage, simulation-based, local adjustment.

5 ANALYTICAL MODELS, INSIGHTS AND LIMITATIONS

5.1 Single-skill staffing

Important insights on call-center sizing are available from existing analysis of single-class queueing systems under

limiting conditions. Halfin and Whitt (1981) consider a sequence of $M/M/s$ queues indexed by n with number of servers $s_n = n$, arrival rate λ_n, service rate μ, and *load* $\rho_n = \lambda_n/\mu$. Under the assumptions that $\lambda_n \to \infty$ and $(1 - \rho_n/n)\sqrt{n} \to \beta$ for $0 < \beta < 1$ as $n \to \infty$, they show that $P(W_n > 0) \to \alpha$, where W_n denotes steady-state delay in queue for the nth queue, $\alpha = [1 + \beta\Phi(\beta)/\phi(\beta)]^{-1}$, and Φ and ϕ are the c.d.f. and p.d.f. of a standard normal random variable. This limiting result justifies the *square-root safety staffing* formula (approximation) for achieving a given delay probability α under load ρ: $n = \rho + \delta$, where $\delta = \beta\sqrt{\rho}$ is the "safety staffing" above the load to account for stochastic variability. To obtain this formula, it suffices to multiply the approximation $(1 - \rho/n)\sqrt{n} \approx \beta$ by $\sqrt{n}$ and note that $\sqrt{n}/\sqrt{\rho} \to 1$ as $n \to \infty$. The approximation has a long history, has been extended to more general queues (Whitt 2004a), and is very robust (Borst et al. 2004). Given the target delay probability, the formula shows the load-staffing relationship in simpler terms than the Elang-C formula. An important insight is the economies of scale resulting from increasing system size n. Notably, large n ensures simultaneously high quality of service and high server utilization, which characterize a *quality-and-efficiency driven* (QED) call center.

5.2 Arrival-rate uncertainty and time dependence

Two sources of risk in the recipe for staffing described in section 3 are that future arrival rates are uncertain and time-dependent. Harrison and Zeevi (2005) and Whitt (2004d) demonstrate the importance of arrival-rate uncertainty and show that ignoring this uncertainty typically leads to understaffing. This can be explained by the fact that typical measures of service quality are, in great generality, concave decreasing functions of the arrival rate in the usual region of system loads; see Chen and Henderson (2001). Second, the arrival rate varies considerably within a day (see Section 7.1), so the PS approximation may suffer from large error. Steckley et al. (2004) analyze this error for simple Markovian models.

5.3 Control, performance analysis, and staffing under SBR

Insightful results on good routing policies have been obtained under a limiting regime known as *conventional heavy traffic*: the traffic intensity goes to one (from below) and the fraction of delayed calls goes to one; these conditions characterize an *efficiency-driven* call center. In this limit, the call-to-agent assignment problem disappears (because essentially all calls must wait in queue) and, under certain conditions, *complete resource pooling* occurs; loosely speaking, this means that the agents are coordinated as if they were a generalist "super-server" which serves the workload at the maximum possible rate. In typical models, the incurred cost is $C_i(\tau_i)$ for each call of type i, where C_i is a convex increasing function and τ_i is either queue time or sojourn time; then one derives an *asymptotically optimal* policy, i.e., one whose expected cumulative cost (possibly discounted) is minimal over a large class of routing policies, in this heavy-traffic limit. Such results are usually obtained by analyzing simple *designs*; examples are: (i) an *N design* has two call types and two agent types, a *specialist* type that can handle only one call type and a *generalist* type that can handle both call types; and (ii) a design where all agents are generalists.

Next we describe two cases as above that exemplify different types of (optimal) routing policies that arise. For a multi-skill design with a single generalist agent and convex increasing cost function C_i on sojourn time, the asymptotically optimal policy was found by van Mieghem (1995) and named *generalized $c\mu$ rule*: call type i is assigned the index $\mu_i c_i(a_i(t))$, where μ_i is the class-i service rate, $a_i(t)$ is the time that the oldest class-i call has been waiting at time t, and c_i is the derivative of C_i; the call served is the oldest waiting call of the class with highest index. The optimality result has been extended to the multi-agent, all-generalist design (Mandelbaum and Stolyar 2004). The dependence of the $c\mu$ rule on only the service rates and cost functions means that the rule continues to be correct (optimal) under changes in important factors such as staffing level and arrival rates. Bell and Williams (2001) study the N design with two agents; activity 1 corresponds to processing of class-1 calls by agent 1; for $j = 2, 3$, activity j corresponds to processing of calls of class $j - 1$ by agent 2; the mean of inter-arrival times of class-i calls is $1/\lambda_i$, $i = 1, 2$, and the mean of service times for activity j is $1/\mu_j$, $j = 1, 2, 3$. There is no abandonment, service preemption is allowed, and cost is linear on sojourn time with coefficient c_i for class i. The limiting N design satisfies: (i) $(\lambda_1 - \mu_1)/\mu_2 + \lambda_2/\mu_3 = 1$, and (ii) $\lambda_1 > \mu_1$; that is, in the limit, the sever capacity is just sufficient to process the incoming load, and, moreover, agent 1 needs help from agent 2 to process the load. The authors exhibit an asymptotically optimal policy of *threshold* type: whenever the number of class-1 calls in the system exceeds a threshold, agent 2 gives preemptive-resume priority to class-1 calls over class-2 calls; otherwise, he gives priority to class-2 calls.

Another line of research is on non-asymptotic performance analysis and/or control. Motivated by a desire to simplify the analysis, many authors analyze a call center as a *loss* system, where calls that cannot be served immediately upon arrival are lost. Koole and Talim (2000) develop an approximation of the call-loss process under *overflow routing*, whereby calls overflow downstream along a predetermined list of agent groups until those calls find an agent available, or else they are lost. Franx et al. (2004)

impose severe restrictions on the routing (a crossed routing as in section 3 is not allowed) and develop an approximation of the loss rate for each call type that is claimed to be superior to other known approaches. Bhulai (2004) approximates an optimal routing policy via dynamic programming; one-step policy improvement of a "good" initial policy is proposed as a means to making the procedure practical, given that the state space is very large (high-dimensional) in typical applications. Chevalier et al. (2004) work with loss-type models of a call center that is staffed with a mixture of single-skill and fully-flexible agents. They show that routing calls first to specialists, then (if necessary) to fully-flexible agents, minimizes the loss rate. Further, they adapt Hayward's approximation (see Wolff (1989), pp. 354-355) to support minimum-cost staffing subject to loss-rate performance constraints. The simple rule-of-thumb "80% specialist, 20% fully-flexible agents" is shown to work well in their examples. Avramidis et al. (2005) extend the ideas of Koole and Talim (2000) to model call queueing, allowing abandonment and an arbitrary overflow routing (including the crossed case); they approximate the tail of the distribution of virtual queue time (see section 7.3) for each call type. Such performance approximations may be useful as pure alternatives to simulation or in synergy with simulation, typically to support the staffing and scheduling decisions. In Avramidis et al. (2005), synergy between the analytical performance approximation and simulation was essential to solving efficiently single-period multi-skill staffing problems (see Section 4).

Recent research provides further insights on the advantages of effective coordination of staffing, routing, and skill-set design. Wallace and Whitt (2004) demonstrate by examples (but not theoretically) that endowing agents with two skills, combined with a carefully designed routing, gives a performance (in terms of SL) that is essentially as good as for a system where all agents have all skills. Their routing entails a careful balancing of agents' priorities over different call types. A key insight is that *a little flexibility goes a long way*. Harrison and Zeevi (2005) focus on arrival-rate uncertainty; they assume an optimal routing can be enforced (continually over time), impose staffing and abandonment costs (and no performance constraints), and use fluid approximations of call abandonment. The obtained insight is that the staffing problem can be seen as a *multidimensional newsvendor* problem (van Mieghem 1998). For the small designs they consider, the cost function is nearly flat around the optimum (2-dimensional) staffing.

From the point of view of practical relevance, some of the models discussed above are not satisfactory, for several reasons. First, many call centers of interest normally operate under the QED regime, in which, by definition, a considerable fraction of calls is served immediately, but also considerable is the fraction of calls that experiences some delay. That is, neither conventional heavy-traffic, nor loss-type models are good representations of the QED regime. Second, there is a gap between the simple designs often analyzed and the relative complexity in typical call center designs. Third, the time dependence of arrival rates commonly found in practice is incompatible with the constant arrival rate usually assumed in analytical models; further, the load may temporarily exceed the system processing capacity.

6 SIMULATION ROLE AND MODELING DIFFICULTIES

The discussion in section 3 establishes the central role that uncertainty and complexity play in modern call-center operation and management. Despite the many insights obtained from analytical models discussed in Section 5, the gap between these models and call-center reality is still quite large. In this setting, simulation appears to be the most viable option for accurate performance measurement and subsequent decision support.

Simulation of call centers may involve large, complex models that incorporate some or all of the elements discussed above, notably: (1) uncertainty in many essential primitives, e.g., attrition, absenteism, arrival rates, service times; (2) time-varying arrival patterns; (3) daily control; and (4) real-time control (routing and outbound dialing policies). (Of course, such modeling complexity translates to increased costs.) Such models can be (and already are) useful at various levels in the decision hierarchy. Mehrotra and Fama (2003) give academic examples where simulation is used as a decision-support tool for both staffing and routing decisions in a blend call center. For numerous applications of call center simulation, see Mandelbaum (2003).

The biggest modeling difficulty appears to be the complex daily and real-time control actions. Man-made decisions at these levels may be taken ad-hoc and thus are difficult to model. An outbound dialer with a proprietary (non-transparent) policy is also a considerable modeling difficulty. We are aware of a major call center where the actual SL oscillates many times above and below the target during one day, presumably due to the lack of good coordination between daily and real-time control actions. Properly modeling actions with such effects is difficult, if not futile.

A major possible problem is the lack of detailed, high-quality data. One common difficulty is the lack of connection between call-by-call data stored at the IVR level and downstream, aggregate data, tracked by workforce planning systems, in which the call ID is absent. Collection of high-quality data and subsequent in-depth statistical analysis appear to be important pre-requisites for better understanding of call centers, which in turn is a pre-requisite for advanced simulation modeling.

7 MODELING CALL CENTER PRIMITIVES

We review work relevant to modeling the primitive inputs to a call center, drawing from recent empirical work, primarily Brown et al. (2005), and Jongbloed and Koole (2001), Avramidis et al. (2004), Steckley et al. (2004).

7.1 Arrival process

Properties of call center arrival processes that have emerged in recent studies are:

- P1. The total daily demand (number of calls) has overdispersion relative to the Poisson distribution (the variance is greater than the mean).
- P2. The arrival rate is strongly time-varying within each day.
- P3. There is positive stochastic dependence between arrival rates within each day.
- P4. There is positive stochastic dependence between arrival rates across successive days.

Jongbloed and Koole (2001) analyze data from a Dutch bank, confirm P1, and propose a *doubly stochastic* model under which arrivals follow a Poisson process with a random arrival rate. To model a time-varying arrival rate, they assume independence across successive time periods, thus being inconsistent with P3. Avramidis et al. (2004) propose various models that are consistent with P1-P3, including a multivariate extension of the above model. In a case study of a Bell Canada call center, they show that simulation-based call-center performance measurement is sensitive to the arrival-process model, and more particularly to the presence of correlation within the day.

P4 was observed in several studies. Regressing a day's call volume on the previous day's volume, Brown et al. (2005) explain 50% of the variability. Steckley et al. (2004) report strong call volume correlations between Monday and all remaining days of the same week, usually in the range 40%-90%, and decreasing with time distance. Our own unpublished work confirms this phenomenon. Obviously, P4 implies that an analyst doing a simulation to estimate future performance a few days in advance, should simulate the arrival rate from the conditional distribution given the observed call volume over the recent past (and possibly other covariates).

Summarizing, we have *time-varying, uncertain* arrival rates that are typically positively dependent within a day and across closely-spaced days.

7.2 Service times

Some studies find the exponential distribution provides an adequate fit to empirical data (Kort 1983, Harris et al. 1987). In addition to the exponential, other parametric families that arose in applications include the gamma and the lognormal (Chlebus 1997, Deslauriers 2003, Pichitlamken et al. 2003). Brown et al. (2005) find the lognormal provides an excellent fit to data, especially after excluding short service times. The excellent fit of the lognormal was also present after conditioning: for all types and priorities of customers, for individual agents, for different days of the week, and for all times of the day. A positive implication is that one can apply standard estimation techniques to relate (regress) log(service time) to various *covariates*, i.e, observed information, with obvious modeling benefits.

7.3 Abandonment

The maximal time a customer is willing to wait in queue is his *patience* time, A, also known as *time-to-abandonment*. The time he *must* wait before beginning service is his *virtual queue time*, V. The actual wait time is $W = \min(A, V)$, terminated by either abandonment (whenever $V > W$), or beginning of service ($V = W$).

In heavy traffic, even a small fraction of calls that abandon the queue can have a dramatic effect on system performance (Gans et al. 2003). On the theoretical side, for a many-server queue with abandonment operating under heavy traffic conditions, fluid approximations in Whitt (2004b) show that steady-state performance depends strongly upon the distribution of A beyond its mean. This suggests that modeling abandonment, preferably the distribution of patience (thus going beyond the mean) is important.

With respect to parametric models of patience, the Weibull distribution arises in a theoretical model in Palm (1943) and also in Kort (1983), based on laboratory testing.

How can one estimate the distribution of patience? Typically, the ACD collects data on W and the abandonment-indicator, $\mathbf{1}\{V > W\}$; A cannot be observed. We encounter the classical statistical problem of *censoring*, and techniques from the field of *survival analysis* are applicable. Brown et al. (2005) employ the classical, non-parametric, Kaplan-Meier estimator of the survival function $\Pr\{A > t\}$, for $t > 0$. They observe that the patience hazard rate has two main peaks and explain this phenomenon by observing that both peaks correspond to time points where customers are offered a "please wait" message. We caution, echoing these authors, that the Kaplan-Meier estimator will be biased whenever there is statistical dependence of observations of W and $\mathbf{1}\{V > W\}$; this is likely to happen for observations made successively in time due to highly-dependent covariates, e.g., announcements such as "please wait" or offering expected wait times.

To help prioritize modeling efforts, Whitt (2004c) studies the sensitivity of the Erlang-A model to its parameters and finds, intuitively, that performance is quite sensitive to the arrival and service rate and relatively insensitive to the impatience (time-to-abandonment) rate.

7.4 Retrials

For our purposes, a *retrial* occurs when a customer re-dials into the center after having encountered a busy signal or having abandoned. In most call centers, the majority of retrials is due to customer abandonment, because the bottleneck resource is the agents, not the number of telephone lines. In any case, naively measuring arrival rates leads to overestimation of the volume of *first-time* calls, i.e., net of retrials. Aguir et al. (2004) demonstrate the danger of ignoring retrials; working with Markovian queues, they find that under high-load conditions, the retrial volume can be of the order of first-time calls. Retrial behavior is often modeled by some function that equals the probability of an n-th attempt, given a survival of the customer (no service received) beyond the $(n-1)$-th attempt. Hoffman and Harris (1986) estimate jointly first-call arrival rates and re-trial rates based on ACD data.

8 CONCLUSION

Modern call centers operate under many uncertainties and complexities, notably, uncertain and/or time-varying primitives and complex daily control and routing control actions. These realities stretch the limits of existing analytical models from queueing theory, optimal queueing control, and stochastic programming. Simulation appears to be the most viable option for accurate performance measurement and subsequent decision support.

Major difficulties that await the call-center modeler are to achieve a deep understanding of the daily and real-time control actions and to ensure the availability of high-quality, detailed data. These are pre-requisites to developping realistic models.

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 second author.

REFERENCES

Aguir, M. S., O. Akşin, F. Karaesmen, and Y. Dallery. 2004. On the interaction between retrials and sizing of call centers. Technical report, Department of Industrial Engineering, Koç University.

Atlason, J., M. A. Epelman, and S. G. Henderson. 2004. Call center staffing with simulation and cutting plane methods. *Annals of Operations Research* 127:333–358.

Avramidis, A. N., W. Chan, and P. L'Ecuyer. 2005. Staffing multi-skill call centers using a performance approximation and search methods. manuscript.

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

Bartholomew, D., A. Forbes, and S. McLean. 1991. *Statistical techniques for manpower planning*. 2nd ed. Wiley.

Bell, S., and R. Williams. 2001. Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy. *Annals of Applied Probability* 11:608–649.

Bhulai, S. 2004. Dynamic routing policies in multi-skill call centers. Technical report, Technical report 2004-11, Free University, Amsterdam.

Borst, S., A. Mandelbaum, and M. Reiman. 2004. Dimensioning large call centers. *Operations Research* 52:17–34.

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.

Call Center News Service 2001. Call center statistics. http://www.callcenternews.com/.

Ceşik, M. T., and P. L'Ecuyer. 2004. Staffing multiskill call centers via linear programming and simulation. submitted.

Chen, B., and S. G. Henderson. 2001. Two issues in setting call center staffing levels. *Annals of Operations Research* 108:175–192.

Chevalier, P., R. A. Shumsky, and N. Tabordon. 2004. Routing and staffing in large call centers with specialized and fully flexible servers. Technical report, Simon Graduate School of Business, Uiversity of Rochester.

Chlebus, E. 1997. Empirical validation of call holding time distribution in cellular communications systems. In *Proceedings of the 15th International Teletraffic Congress*, 1179–1188: Elsevier.

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.

Franx, G. J., G. Koole, and A. Pot. 2004, September. Approximating multi-skill blocking systems by hyper-exponential decomposition. Technical report, Vrije Universiteit, The Netherlands, Amsterdam. Preprint.

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

Gans, N., and Y.-P. Zhou. 2002. Managing learning and turnover in employee staffing. *Operations Research* 50:991–1006.

Green, L. V., and P. J. Kolesar. 1991. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science* 37 (1): 84–97.

Halfin, S., and W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations Research* 29:567–588.

Harris, C., K. Hoffman, and P. Saunders. 1987. Modeling the irs telephone taxpayer information system. *Operations Research* 35:504–523.

Harrison, J. M., and A. Zeevi. 2005. A method for staffing large call centers based on stochastic fluid models. *Manufacturing and Service Operations Management*. To appear.

Hoffman, K. L., and C. Harris. 1986. Estimation of a caller retrial rate for a telephone information system. *European Journal of Operational Research* 27 (2): 207–214.

Ingolfsson, A., E. Cabral, and X. Wu. 2003. Combining integer programming and the randomization method to schedule employees. Technical report, School of Business, University of Alberta, Edmonton, Alberta, Canada. Preprint.

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. 2005. Call center mathematics. In preparation.

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

Koole, G., and J. Talim. 2000. Exponential approximation of multi-skill call center architecture. In *Proceedings of QNETs 2000*, 23/1–10. Ilkley (UK).

Kort, B. 1983. Models and methods for evaluating customer acceptance of telephone connections. In *GLOBECOM '83*, 706–714: IEEE.

Mandelbaum, A. 2003. Call centers (centres): Research bibliography with abstracts. Downloadable from <www.ie.technion.ac.il/~serveng/References/ccbib.pdf>.

Mandelbaum, A., and A. Stolyar. 2004. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized c-mu rule. *Operations Research* 52:836–855.

Mehrotra, V., and J. Fama. 2003. Call center simulation modeling: Methods, challenges, and opportunities. In *Proceedings of the 2003 Winter Simulation Conference*, 135–143: IEEE Press.

Palm, C. 1943. Intensitätsschwankungen im fernsprechverkehr. *Ericsson Technics* 44:1–189.

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.

Steckley, S. G., S. G. Henderson, and V. Mehrotra. 2004. Service system planning in the presence of a random arrival rate. submitted.

van Mieghem, J. A. 1995. Dynamic scheduling with convex delay costs: the generalized c*mu* rule. *Annals of Applied Probability* 5:809–833.

van Mieghem, J. A. 1998. Investment strategies for flexible resources. *Management Science* 44:1071–1078.

Wallace, R. B., and W. Whitt. 2004. Resource pooling and staffing in call centers with skill-based routing. manuscript.

Whitt, W. 1991. The pointwise stationary approximation for $M(t)/M(t)/s$ queues is asymptotically correct as the rates increase. *Management Science* 37 (3): 307–314.

Whitt, W. 2004a. A diffusion approximation for the g/gi/n/m queue. *Operations Research* 6:922–941.

Whitt, W. 2004b. Fluid models for many-server queues with abandonments. *Operations Research*. To appear.

Whitt, W. 2004c. Sensitivity of performance in the Erlang a model to changes in the model parameters. manuscript.

Whitt, W. 2004d. Staffing a call center with uncertain arrival rate and absenteeism. manuscript.

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

AUTHOR BIOGRAPHIES

ATHANASSIOS (THANOS) N. AVRAMIDIS is Researcher in the Département d' Informatique et de Recherche Opérationnelle at the Université de Montréal, Canada. 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, call center operations, and transportation. His recent research articles are available on-line from his web page: http://www.iro.umontreal.ca/~avramidi.

PIERRE L'ECUYER is Professor in the Département d'Informatique et de Recherche Opérationnelle, at the Université de Montréal, Canada. He holds the Canada Research Chair in Stochastic Simulation and Optimization. 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* fellow-

ship in 1995-97 and a *Killam* fellowship in 2001-03. His recent research articles are available on-line from his web page: http://www.iro.umontreal.ca/~lecuyer.