Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

TWO-STAGE CHANCE-CONSTRAINED STAFFING WITH AGENT RECOURSE FOR MULTI-SKILL CALL CENTERS

Wyean Chan
Thuy Anh Ta
Pierre L'Ecuyer
Fabian Bastin

Department of Computer Science and Operations Research
Université de Montréal
2920 chemin de la Tour
Montreal, Quebec, H3C 3J7, CANADA

ABSTRACT

We consider a stochastic staffing problem with uncertain arrival rates. The objective is to minimize the total cost of agents under some chance constraints, defined over the randomness of the service level in a given time period. In the first stage, an initial staffing must be determined in advance based on imperfect forecast of the arrival rates. At a later time, when the forecast becomes more accurate, this staffing can be corrected with recourse actions, by adding or removing agents at the price of some penalty costs. We present a method that combines simulation, mixed integer programming, and cut generation to solve this problem.

1 INTRODUCTION

Call centers are important service systems that can be found practically anywhere in our society, for example for government services, financial institutions, retail stores, airlines, emergency systems (police, ambulances, etc.), and so on. For many organizations, the call center is the primary medium of interaction with customers. Statistical studies have measured a positive link between customer satisfaction and customer loyalty, which can lead to larger long-term revenue (Heskett et al. 1994). Additionally, workforce salary accounts for roughly 60% to 70% of the operating cost of a call center (Gans, Koole, and Mandelbaum 2003). It is thus important to optimize the workforce in order to control the operating expense while providing good customer satisfaction.

In a multi-skill call center, customers are represented by calls and are categorized by their requested type of service. Telephone agents are divided into groups based on their skill set, where each skill corresponds to a call type. We are only interested in inbound calls that request an interaction with an agent; that is, we ignore calls that are served by or have abandoned in the *interactive voice response* (IVR) unit. Calls are assigned to agents by a router, also called an *automatic call distributor* (ACD), following a chosen routing policy; see Chan, Koole, and L'Ecuyer (2014) for some examples of routing policies. When a new call arrives and the router cannot find any idle agent to serve it, this call is placed in a waiting queue. A customer abandons the queue when her waiting time exceeds her patience time, which is a random variable. Further details on call center operations can be found in Gans, Koole, and Mandelbaum (2003).

In reality, the workforce management (WFM) must optimize multiple stochastic problems that range from days or weeks (scheduling) to months (hiring process) in advance, while facing different levels of uncertainty. Indeed, it is known that arrival rates in call centers are uncertain and depend on multiple factors, such as the day of the week, time of the day, level of busyness, holidays and special events, or publicity

campaigns; see Channouf et al. (2007), Ibrahim and L'Ecuyer (2013), Ibrahim et al. (2016), Oreshkin, Régnard, and L'Ecuyer (2016), and references therein. Most existing studies on WFM for multi-skill call centers simplify their models by assuming that the arrival rates are known perfectly, and they do not consider any staffing recourse. Also, these studies often consider quality-of-service (QoS) targets (constraints) with respect to the long-term expected value, which is an average over an infinite number of days. A drawback of such constraints on the expectations is that a QoS measure for a "particular day", which is a random variable, may often miss the target, even if the solution meets the long-term average target. If a manager really wants to meet the QoS targets for a given (large) proportion of the days, then distributional (chance) constraints have to be imposed and are appropriate for this situation. The aim of this paper is to develop a framework and computational tools to do that.

As a primary contribution, we extend and adapt the simulation-based linear cut algorithm of Ceẓik and L'Ecuyer (2008) to solve a chance-constrainted two-stage staffing with recourse problem for multi-skill call centers. This problem has been studied for single-skill call centers, for which most proposed algorithms approximate the QoS with some analytical formulas (see Section 2 for references). However, we know of no such analytical formula for the multi-skill case. To the best of our knowledge, this paper is the first that studies a two-stage optimization problem for staffing with recourse in a multi-skill call center and proposes a simulation-based solution method to solve that problem. We consider a staffing problem for which some initial staffing decisions must be taken in advance, based on forecast with important stochastic variability. At a later time, when more accurate demand forecast becomes available, the initial staffing may be corrected by adding or removing agents, at the price of some penalty costs. Another difference from the literature is that we consider chance constraints that require the QoS over a day to be satisfied with a minimum probability threshold, instead of constraints defined on the expectations, which are never observed in reality.

The remainder of this paper is structured as follows. A short review of the literature on WFM for call centers is given in Section 2. Section 3 presents the model of a multi-skill call center, the chance constraints, and the two-stage stochastic staffing problem with recourse. We present our extended simulation-based cutting-plane algorithm to solve this stochastic staffing problem in Section 4. We compare the performance of our algorithm with a traditional two-step (TS) approach, described in Section 5. Numerical results are reported in Section 6, followed by a conclusion in Section 7.

2 LITERATURE REVIEW

Most of the work found in the literature focus on the optimization of single-skill call centers, where simple analytical formulas, such as the Erlang formulas (Cooper 1981, Garnett, Mandelbaum, and Reiman 2002) are often used. However, these formulas are not applicable to the multi-skill context in which we are interested, and for which only simulation can provide accurate estimation of the QoS.

A few authors optimize the traditional staffing problem (single stage, without recourse) for multi-skill call centers. For the staffing of a single period, Ceẓik and L'Ecuyer (2008) adapt the simulation-based cutting-plane algorithm of Atlason, Epelman, and Henderson (2008), who adapted it from the cutting-plane method of Kelley Jr. (1960). This method is based on the fact that the service level function for a single queue with abandonment has the shape of a sigmoid function, which is concave above a certain threshold (see Section 4.1); in particular, the function becomes concave as the number of agents increases. This concavity does not necessarily hold in the multi-skill case, so the cutting-plane method then becomes a heuristic, but it nevertheless performs well empirically. Avramidis et al. (2010) extend this cutting-plane algorithm to a scheduling problem with multiple time periods and constraints on work schedules of agents. Neighborhood search algorithms, guided by simulation or approximation formula, are proposed by Wallace and Whitt (2005), Pot, Bhulai, and Koole (2008), and Avramidis, Chan, and L'Ecuyer (2009) for the single period staffing problem. Bhulai, Koole, and Pot (2008) present a two-step scheduling algorithm with temporary group transfers to reduce over-staffing in certain periods. All the above papers assume that the

arrival rates are known perfectly (although simulation-based methods can integrate stochastic arrival rates in the simulation).

Two-stage stochastic staffing and scheduling problem have been considered before; see, e.g., Kim and Mehrotra (2015) and references therein. However, previous studies generally model directly the stochastic staffing requirements into the problem. In call center problems, the main stochastic variables are usually the arrival rates of the calls, which determine the required staffing. Robbins and Harrison (2010) optimize a stochastic scheduling problem (without recourse) for a single-skill call center, by replacing the non-linear Erlang formula by a piecewise-linear function, and they solve a mixed integer problem. They minimize the cost of the agents and the penalty cost for missing the service level target. Robbins and Harrison (2008) propose a variable neighborhood search, combined with simulation, to solve the same problem. Gans et al. (2015) consider a two-stage scheduling problem with recourse (add or remove agents) for single-skill call centers. They assume that a forecast update is made available at midday, and agents can be added or removed to correct the schedules. They consider constraints on the fraction of abandonments, and they approximate the abandonment function of a Markovian queue by a piecewise-linear function, similar to Robbins and Harrison (2010). These methods do not apply to multi-skill problems, because it is unclear if the QoS functions can be linearized in this context, and it may be too time-consuming to do so with simulation.

For the stochastic staffing of multi-skill call centers, Harrison and Zeevi (2005), and Bassamboo, Harrison, and Zeevi (2006) optimize the schedules using a fluid model that approximates the abandonment rate in a call center. The problem is solved as a two-stage stochastic problem where the first-stage variables are the schedules, and the second-stage variables are the work allocations (or routing). Gurvich, Luedtke, and Tezcan (2010) use a similar two-stage stochastic problem framework. They consider chance constraints on the expectation of the fraction of abandonment, for stochastic arrival rates. That is, a solution is feasible if the probability that the *expected steady-state fraction* of abandonment of a day is equal or greater to a given threshold $(1-\delta)$. The parameter δ can represent the level of risk tolerance of the manager. We studied a two-stage stochastic staffing with recourse problem in Chan et al. (2014) using the same type of chance constraints, but with respect to the service levels. However, the algorithm and results were not presented in this extended abstract (for a poster). In this paper, we consider a different type of chance constraint. In practice, some managers are interested on the probability that the observed (realized) QoS of the day meets the constraint. A solution that meets the constraint on the expectation may still fail to meet the target QoS on a single day because of stochastic variance. Instead of the expectation, our chance constraint are defined with respect to the realized value of the QoS.

3 MODEL AND FORMULATION

3.1 Call Center Model

We consider a multi-skill call center with K call types (numbered from 1 to K), and I agent groups (numbered from 1 to I). Agents in group i are assumed to be homogeneous, and they share the same skill set $\mathscr{S}_i \subseteq \{1,...,K\}$, where each skill correspond to a call type. Alternatively, this means a call of type k can be served by an agent who belongs to the set of groups $\mathscr{G}_k = \{i : k \in \mathscr{S}_i\}$. The calls are assigned to agents by a router. For a call of type k, we define $1/\mu_{k,i}$ as the mean service time for an agent of group i to serve this call, and we define $1/\nu_k$ as the mean patience time of this call. The staffing vector is $\boldsymbol{y} = (y_1,\ldots,y_I)^{\mathrm{T}}$, where y_i is the number of agents in group i. In this study, we assume a day with only one period, similarly to Wallace and Whitt (2005) and Cezik and L'Ecuyer (2008).

For a "random" day, the arrival process for call type k is assumed to be time-homogeneous Poisson with rate Λ_k for the entire day, where Λ_k is a random variable, and the Λ_k's are independent across different days. We denote by λ_k a realization of Λ_k. We suppose that several days in advance, in the first stage, Λ_k has a prior distribution, which corresponds to some initial distributional forecast. At a later time (the

second stage), the distributional forecast is updated, which means that Λ_k has a new (posterior) distribution. Typically, the posterior distribution has less uncertainty (smaller variance), but this may not be always true.

To make things more concrete in this paper, we assume that the distribution of Λ_k is parameterized by some parameter θ_k, whose initial distribution determines the prior distribution of Λ_k in the first stage, and whose value is known in the second stage. Typically, knowing θ_k does not tell us the realization of Λ_k, but it may (for example) significantly reduce its variance compared with the prior distribution. In our numerical example, we will consider that $\Lambda_k = \theta_k \beta_k$ is the product of two independent random variables, where β_k is the random busyness factor of a day with mean 1 (see Avramidis, Deslauriers, and L'Ecuyer 2004). In the first stage, both θ_k and β_k are random variables, but in the second stage, only β_k remains random.

3.2 Service Level (SL)

In practice, call centers often measure their QoS according to the *service level* (SL). The SL measures the fraction of calls that are answered within a given time τ, called the *acceptable wait threshold* (AWT). When there are call abandonments, there exist multiple definitions of the SL formula with different ways to account the abandonments. In this study, we use one of the SL formulas implemented in the simulation library *ContactCenters* (Buist and L'Ecuyer 2005); alternative definitions of the SL can be found in Jouini, Koole, and Roubos (2013). Let N be the total number of calls that arrived in a day, $A(\tau, \boldsymbol{y})$ be the number of calls served after waiting at most τ, and $L(\tau, \boldsymbol{y})$ be the number of calls that abandoned after waiting more than τ. Since all these variables are random, the SL of a given time period is also a random variable, and the SL necessarily depends on the number of agents $\boldsymbol{y}$. The SL formula is:

$$S(\tau, \boldsymbol{y}) = \frac{A(\tau, \boldsymbol{y})}{N - L(\tau, \boldsymbol{y})}.$$

Most existing studies on WFM for call centers, especially those that are based on Erlang formulas, consider staffing constraints with respect to an SL defined as a fraction of customers with good QoS over an infinite number of independent and identically distributed (i.i.d.) days. This type of SL is defined as $\bar{S}(\tau, \boldsymbol{y}) = \mathbb{E}[A(\tau, \boldsymbol{y})]/\mathbb{E}[N - L(\tau, \boldsymbol{y})]$.

A manager usually wants to guarantee some minimum level of SL. For example, a manager may want 80% of calls to be answered within 20 seconds of wait time. Multi-skill call centers are much more difficult to optimize under SL constraints than their single-skill counterpart, because the SL can only be estimated reliably by simulation.

3.3 Chance Constraints on the SL

In practice, a manager observes the random variable $S(\tau, \boldsymbol{y})$, not the expected fraction $\bar{S}(\tau, \boldsymbol{y})$. The variable $S(\tau, \boldsymbol{y})$ may have significant stochastic variance, even if the arrival rate is constant. This means that the observed SL of a day may be well below the target, even if the staffing gives an expected SL (over an infinite number of days) above the target. If a manager is interested in satisfying the SL target most of the days, then a distributional chance constraint on $S(\tau, \boldsymbol{y})$ may be appropriate.

We define the following chance constraint on the distribution of $S(\tau, \boldsymbol{y})$: on a random day, the SL target must be met with probability $1 - \delta$ or higher, for a given risk level δ selected by the manager. Given the staffing vector $\boldsymbol{y}$, let $S_k(\tau_k, \boldsymbol{y})$ be the SL of call type k during the day, with AWT τ_k, and let $S(\tau, \boldsymbol{y})$ be the aggregate SL of the day over all calls. All of these are random variables, whose distributions depend on the staffing $\boldsymbol{y}$. The chance constraints are:

$$\begin{aligned} \mathbb{P}[S_k(\tau_k, \boldsymbol{y}) \geq l_k] &\geq 1 - \delta_k, \qquad 1 \leq k \leq K, \\ \mathbb{P}[S(\tau, \boldsymbol{y}) \geq l] &\geq 1 - \delta, \end{aligned}$$

where l_k and l are the SL targets, and δ_k and δ are the given risk thresholds in the interval $(0,1)$. To give an example of chance constraints, setting $\delta_k = \delta = 0.05$, $l_k = l = 0.8$, and $\tau_k = \tau = 20$ seconds means that 80% of calls in a day must be answered within 20 seconds, with at least 95% probability.

3.4 Staffing Problem with Recourse

We now describe our two-stage staffing problem with stochastic arrival rates. In the first stage, the manager must select an initial staffing $\boldsymbol{x} = (x_1, \ldots, x_I)^{\mathrm{T}}$, at the corresponding cost per agent of $\boldsymbol{c} = (c_1, \ldots, c_I)^{\mathrm{T}}$, based on an initial forecast that gives only the prior distributions of the Λ_k's, parametrized by a random parameter θ_k. In the second stage, the manager obtains the realizations of $\theta_1, \ldots, \theta_k$, which provides updated forecasts, and can modify the initial staffing $\boldsymbol{x}$ by adding or removing agents at some penalty costs. Note that even when knowing θ_k, the manager may not know the exact arrival rate λ_k (the realization of Λ_k) in the second stage. And even if she does, she still does not know the SL for the day, so there is still uncertainty. In the special case where we suppose perfect forecast of the arrival rate in the second stage, Λ_k simply has a degenerate posterior distribution (with only one possible realization).

In the second stage, given the posterior distributions of the Λ_k's, the manager can add $r_i^+(\Lambda)$ extra agents to group i at a greater cost of $c_i^+ > c_i$ per agent, or remove $r_i^-(\Lambda) \le x_i$ agents in group i and save c_i^- per agent, where $0 \le c_i^- < c_i$. After the recourse, the new number of agents in group i is $y_i(\Lambda) = x_i + r_i^+(\Lambda) - r_i^-(\Lambda)$. Let $\boldsymbol{c}, \boldsymbol{c}^+$, $\boldsymbol{c}^-$, and $\boldsymbol{y}(\Lambda)$ be the vectors with components c_i, c_i^+, c_i^-, and $y_i(\Lambda)$, respectively. We define the recourse vectors as $\boldsymbol{r}^+(\Lambda) = (r_1^+(\Lambda), \ldots, r_I^+(\Lambda))^{\mathrm{T}}$, and $\boldsymbol{r}^-(\Lambda) = (r_1^-(\Lambda), \ldots, r_I^-(\Lambda))^{\mathrm{T}}$.

Given a staffing $\boldsymbol{y}(\Lambda)$, the SL of call type k and the aggregate SL are random variables $S_k(\tau_k, \boldsymbol{y}(\Lambda))$ and $S(\tau, \boldsymbol{y}(\Lambda))$. We consider the following chance-constrained staffing problem with recourse for multi-skill call centers with arrival rate uncertainty:

$$\min \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} + \mathbb{E}_\theta\left[Q_\theta(\boldsymbol{x}, \theta)\right], \text{ subject to: } \boldsymbol{x} \ge 0 \text{ and integer}, \tag{P1a}$$

where $Q_\theta(\boldsymbol{x}, \theta)$ is the cost of the second-stage problem, given initial staffing $\boldsymbol{x}$ and $\theta = (\theta_1, \ldots, \theta_K)$:

$$Q_\theta(\boldsymbol{x}, \theta) = Q_\Lambda(\boldsymbol{x}, \Lambda) = \min (\boldsymbol{c}^+)^{\mathrm{T}} \boldsymbol{r}^+(\Lambda) - (\boldsymbol{c}^-)^{\mathrm{T}} \boldsymbol{r}^-(\Lambda)$$

subject to:

$$\begin{aligned} \boldsymbol{x} + \boldsymbol{r}^+(\Lambda) - \boldsymbol{r}^-(\Lambda) &= \boldsymbol{y}(\Lambda), \\ g_k(\boldsymbol{y}(\Lambda), \Lambda) := \mathbb{P}[S_k(\tau_k, \boldsymbol{y}(\Lambda)) \ge l_k] &\ge 1 - \delta_k, \qquad \forall k, \\ g(\boldsymbol{y}(\Lambda), \Lambda) := \mathbb{P}[S(\tau, \boldsymbol{y}(\Lambda)) \ge l] &\ge 1 - \delta, \\ 0 \le r_i^-(\Lambda) &\le x_i, \qquad \forall i, \\ \boldsymbol{r}^+(\Lambda), \boldsymbol{r}^-(\Lambda) &\ge 0 \text{ and integer}. \end{aligned} \tag{P1b}$$

The functions g_k and g represent the probabilities that the SL targets of call type k and the aggregate SL are met for a random day, given the staffing vector $\boldsymbol{y}(\Lambda)$ and forecast Λ. One particularity of our model is that the second-stage problem (P1b) is much more difficult than the first-stage problem (P1a) because of the SL constraints. The first stage has only non-negativity and integrality constraints on $\boldsymbol{x}$. Our stochastic problem has complete recourse, because we can add an unlimited number of agents in (P1b) for any Λ. If we were considering a rostering problem with limited number of agents, then the constraints on the SL should probably be relaxed and replaced by some penalty functions in the objective. However, it is not clear how to choose the penalty factors.

3.5 Sample Average Approximation (SAA) Problem

In general, the arrival rates are continuous parameters. Therefore, Λ are continuous random variables, and this makes the staffing problem more difficult to optimize. Instead of solving the two-stage problem (P1a), we solve a sample average approximation (SAA) version, where we generate M scenarios of θ

by Monte Carlo which in turn define the distributions of the Λ_k's in the second-stage problem. Let $\hat{\Lambda}_m = (\hat{\Lambda}_{1,m}, \ldots, \hat{\Lambda}_{K,m})$ be the vector of K distributions of the arrival rates of scenario m with probability $p_m > 0$, and $\sum_{m=1}^{M} p_m = 1$. In our numerical examples, we will assume, without loss of generality, the same probability $p_m = 1/M$ for all m.

Moreover, we do not know how to compute exactly the probability functions g_k and g in (P1b), but we can approximate their empirical values by simulation. Let $\hat{g}_k$ and $\hat{g}$ be the empirical values obtained by simulating N independent and identically distributed days. Since we have a finite number of scenarios in the second stage, we lighten the notation by using indexed variables $\boldsymbol{r}_m^+ = \boldsymbol{r}^+(\hat{\Lambda}_m)$, $\boldsymbol{r}_m^- = \boldsymbol{r}^-(\hat{\Lambda}_m)$ and $\boldsymbol{y}_m = (y_{1,m}, \ldots, y_{I,m})^{\mathrm{T}}$ for scenario m. We approximate (P1a) by the following SAA problem formulated in extensive form:

$$\min \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} + \sum_{m=1}^{M} p_m \left[(\boldsymbol{c}^+)^{\mathrm{T}}\boldsymbol{r}_m^+ - (\boldsymbol{c}^-)^{\mathrm{T}}\boldsymbol{r}_m^-\right]$$

subject to:

$$\begin{aligned}
\boldsymbol{x} + \boldsymbol{r}_m^+ - \boldsymbol{r}_m^- &= \boldsymbol{y}_m, && \forall m, \\
\hat{g}_k(\boldsymbol{y}_m, \hat{\Lambda}_m) := \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}[\hat{S}_k^n(\tau_k, \boldsymbol{y}_m) \geq l_k] &\geq 1 - \delta_k, && \forall k, \forall m, \\
\hat{g}(\boldsymbol{y}_m, \hat{\Lambda}_m) := \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}[\hat{S}^n(\tau, \boldsymbol{y}_m) \geq l] &\geq 1 - \delta, && \forall m, \\
0 \leq \boldsymbol{r}_m^- &\leq \boldsymbol{x}, && \forall m, \\
\boldsymbol{x}, \boldsymbol{r}_m^+, \boldsymbol{r}_m^- &\geq 0 \text{ and integer}, && \forall m,
\end{aligned} \tag{S1}$$

where $\mathbb{I}$ is a 0-1 indicator function, and $\hat{S}_k^n(\tau_k, \boldsymbol{y}_m)$ and $\hat{S}^n(\tau, \boldsymbol{y}_m)$ are the SL of call type k and aggregate SL for the n-th simulated day, given staffing vector $\boldsymbol{y}_m$ and $\hat{\Lambda}_m$.

4 CUTTING-PLANE (CP) ALGORITHM FOR TWO-STAGE STOCHASTIC STAFFING

We present our main contribution of the paper in this section. Ceẓik and L'Ecuyer (2008) propose a simulation-based cutting-plane algorithm to solve the staffing problem of multi-skill call centers. We extend their algorithm to include recourse actions and solve a two-stage stochastic staffing problem with uncertain arrival rates.

4.1 Concavity of the SL Function

In order to use the CP algorithm, Ceẓik and L'Ecuyer (2008) assume the hypothesis that the expected SL functions $\bar{S}$ and $\bar{S}_k$ are concave, or at least concave around the optimal staffing solution. They observe numerically that the functions $\bar{S}$ and $\bar{S}_k$ often display the shape of a sigmoid function (or stretched "S"), where the functions are first convex and become concave as the number of agents increase.

In our problem with chance constraints, we are interested in the probability functions g_k and g, instead of the expected SL functions. Fortunately, our numerical observations show that the simulated functions $\hat{g}_k$ and $\hat{g}$ also have a similar "S" shape, see Figure 1 for an example with one agent group. In this figure, we can apply the CP algorithm if the risk level δ is in the concave region of $\hat{g}$, that is, if $1-\delta$ is above 0.4.

4.2 CP Staffing Problem

One major difficulty in solving (S1) is that it is non-linear because of the chance constraints on the SL. The main idea of the CP method of Ceẓik and L'Ecuyer (2008), and Atlason, Epelman, and Henderson (2008) is to replace the non-linear SL constraints in (S1) by a set of linear constraints, and we obtain an

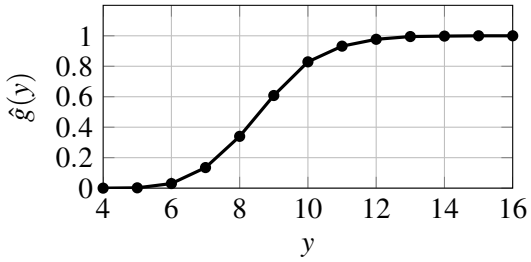

Figure 1: Example of function $\hat{g}(y)$ showing an "S" shaped curve.

equivalent linear problem (S2). Unfortunately, we do not know all the constraints of (S2) beforehand, so we use the CP algorithm to build it.

The CP method is an iterative algorithm that starts at an infeasible solution, and it adds new linear cuts based on the subgradient of $\hat{g}_k$ and $\hat{g}$ until a feasible solution is obtained. Assuming that the cuts do not eliminate any feasible solutions of (S1), then the first feasible solution will also be optimal for (S1). For each scenario m, let the matrix $\boldsymbol{B}_m^v$ and vector $\boldsymbol{b}_m^v$ define the set of linear cut constraints that replace the chance constraints on $\hat{g}_k$ and $\hat{g}$, at iteration v. The constraint parameters $\boldsymbol{B}^1 = \boldsymbol{b}^1 = \emptyset$ are initially empty. To avoid starting the algorithm at a null solution (all-zero solution) or in a non-concave region, we add heuristic linear constraints to cover a fraction α_k of the arrival rate of call type k, as described in Chan (2013). These constraints are also used in the fluid scheduling model of Bassamboo, Harrison, and Zeevi (2006). These heuristic constraints require additional continuous variables $w_{k,i,m} \geq 0$ which define the (fractional) number of agents of group i working on calls of type k in scenario m. The parameters α_k should be selected such that the initial solution is hopefully in a concave region of $\hat{g}_k$ and $\hat{g}$. For example, in Figure 1, α_k should be set such that the starting solution y would be greater than 7. The parameter α_k is generally chosen around 1. The CP problem (S2) at iteration v is:

$$\min \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} + \sum_{m=1}^{M} p_m \left[(\boldsymbol{c}^+)^{\mathrm{T}}\boldsymbol{r}_m^+ - (\boldsymbol{c}^-)^{\mathrm{T}}\boldsymbol{r}_m^-\right]$$

subject to:

$$\begin{aligned}
\boldsymbol{x} + \boldsymbol{r}_m^+ - \boldsymbol{r}_m^- &= \boldsymbol{y}_m^v, && \forall m,\\
\boldsymbol{B}_m^v \boldsymbol{y}_m^v &\geq \boldsymbol{b}_m^v, && \forall m,\\
\sum_{i \in \mathscr{G}_k} \mu_{k,i} w_{k,i,m} &\geq \alpha_k \hat{\Lambda}_{k,m}, && \forall k, \forall m,\\
\sum_{k \in \mathscr{S}_i} w_{k,i,m} &\leq y_{i,m}, && \forall i, \forall m,\\
w_{k,i,m} &\geq 0, && \forall k, \forall i, \forall m,\\
0 \leq \boldsymbol{r}_m^- &\leq \boldsymbol{x}, && \forall m,\\
\boldsymbol{x}, \boldsymbol{r}_m^+, \boldsymbol{r}_m^- &\geq 0 \text{ and integer}, && \forall m.
\end{aligned} \qquad (\mathrm{S2}^v)$$

If there is a large number of scenarios, one could exploit the block angular structure of the cut constraints $\boldsymbol{B}_m^v \boldsymbol{y}_m^v \geq \boldsymbol{b}_m^v$, and use the well-studied L-shaped method (Van Slyke and Wets 1969). However, the main challenge remains that the parameters $\boldsymbol{B}_m^v$ and $\boldsymbol{b}_m^v$ are initially unknown, and these can only be determined by simulation, which will generally be the most time-consuming part of the algorithm.

4.3 Subgradient-Based Linear Cuts

We explain briefly how to generate the constraint parameters $\boldsymbol{B}_m^v$ and $\boldsymbol{b}_m^v$, since the procedure is identical to Ceẓik and L'Ecuyer (2008) with the exception that we replace the expected SL functions $\bar{S}$ and $\bar{S}_k$ by the

chance-constraint functions $\hat{g}$ and $\hat{g}_k$. These subgradient-based linear cuts are generated independently for each scenario. Now consider scenario m with realization $\hat{\Lambda}_m$, and let $\boldsymbol{y}_m^v$ be the current solution at iteration v. We will drop $\hat{\Lambda}_m$ from the notation, since it is fixed. We present the case where the chance constraint $\hat{g}$ for the aggregate SL is not satisfied, but the procedure is the same for the per call type constraint $\hat{g}_k$. Let the vector $\boldsymbol{q}(\boldsymbol{y}_m^v)$ of size I be the subgradient of $\hat{g}$ at point $\boldsymbol{y}_m^v$. We estimate the i-th element $q_i(\boldsymbol{y}_m^v)$ by forward finite difference, with step size d, using simulation:

$$q_i(\boldsymbol{y}_m^v) = [\hat{g}(\boldsymbol{y}_m^v + d\boldsymbol{e}_i) - \hat{g}(\boldsymbol{y}_m^v)]/d,$$

where $\boldsymbol{e}_i$ is a unit vector with 1 at the i-th position and 0 elsewhere. Normally, we set $d = 1$, but when the simulation has a lot of noise (e.g., the number of simulated days N is small), we may set d to 2 or 3. Assuming $\boldsymbol{q}(\boldsymbol{y}_m^v)$ is a subgradient of $\hat{g}$ at point $\boldsymbol{y}_m^v$, we have the following valid inequality $\hat{g}(\boldsymbol{y}_m^v) + \boldsymbol{q}(\boldsymbol{y}_m^v)(\boldsymbol{y}_m - \boldsymbol{y}_m^v) \geq \hat{g}(\boldsymbol{y})$. Since we want to find $\boldsymbol{y}_m$ such that $\hat{g}(\boldsymbol{y}_m) \geq 1 - \delta$, then we can add the following cut to (S2^v) without cutting the optimal solution of (S1):

$$\boldsymbol{q}(\boldsymbol{y}_m^v)\boldsymbol{y}_m \geq 1 - \delta - \hat{g}(\boldsymbol{y}_m^v) + \boldsymbol{q}(\boldsymbol{y}_m^v)\boldsymbol{y}_m^v. \tag{1}$$

Notice that all terms in the right-hand side of (1) are known at iteration v. This subgradient-based cut is added to $\boldsymbol{B}_m^v$ and $\boldsymbol{b}_m^v$, and they become $\boldsymbol{B}_m^{v+1}$ and $\boldsymbol{b}_m^{v+1}$ at the next iteration $v+1$.

4.4 CP Algorithm

We present the framework of the CP algorithm to optimize the two-stage stochastic staffing problem (S1). The algorithm alternates between solving the problem (S2^v) and generating cuts by simulation. We set a limit of $v_{\max}$ on the maximum number of iterations. The algorithm finishes with a simulation-based local search to refine the solution.

Step 0 Initialization: Set $v = 1$. For each scenario m, set $\boldsymbol{B}_m^1 = \boldsymbol{b}_m^1 = \emptyset$ (empty matrix and vector).
Step 1 Solve problem (S2^v).
Step 2 For each scenario m, simulate staffing solution $\boldsymbol{y}_m^v$. If all constraints in (S1) are satisfied, that is, $\hat{g}_k(\boldsymbol{y}_m^v, \hat{\Lambda}_m) \geq 1 - \delta_k$ for all k, and $\hat{g}(\boldsymbol{y}_m^v, \hat{\Lambda}_m) \geq 1 - \delta$, then go to Step 5, else continue.
Step 3 For each scenario m with unsatisfied constraints on $\hat{g}_k$ or $\hat{g}$, generate one or many subgradient-based cuts defined by (1) at point $\boldsymbol{y}_m^v$. Add the new cut(s) to $\boldsymbol{B}_m^v$ and $\boldsymbol{b}_m^v$.
Step 4 Set $v := v + 1$. If $v = v_{\max}$, go to Step 5, else return to Step 1.
Step 5 Apply local search on the solutions $\boldsymbol{x}, \boldsymbol{r}^+$ and $\boldsymbol{r}^-$, by using simulation. Stop the algorithm.

To reduce the execution time of the algorithm, we may want to avoid solving (S2^v) during the early iterations when the number of subgradient-based cuts is small (because the solution will likely be infeasible anyway). Hence, we solve each of the M scenarios separately as a single staffing problem without recourse, then we take the cuts generated in this phase to initialize $\boldsymbol{B}_m^1$ and $\boldsymbol{b}_m^1$. For this phase, we can use bigger risks δ_k and δ to have more conservative cuts, then we reset them to the original risk levels when starting Step 1.

The local search helps reduce the cost, but its role is also to correct the simulation noise of approximated functions $\hat{g}_k$ and $\hat{g}$. A larger N is used in the local search than during the CP algorithm in order to improve the feasibility of the solution for real functions g and g_k. This local search is inspired by the simulation-based neighborhood search of Avramidis, Chan, and L'Ecuyer (2009).

5 TWO-STEP (TS) METHOD

A traditional approach to optimize the staffing is usually to divide the main problem into smaller and simpler problems. In general, such an approach has the advantage of executing faster, but it has the inconvenience

to likely be suboptimal. Nevertheless, it will be interesting to compare numerically the performance of the CP algorithm with a simpler two-step (TS) method, which we will describe here.

As its name implies, the TS method is divided into 2 separate steps. In step 1, we determine the number of agents required in each group for each scenario m, independently from the other $M-1$ scenarios. We use an adapted version of Ceẓik and L'Ecuyer (2008) to solve each scenario m independently. Let $\bar{\boldsymbol{y}}_m = (\bar{y}_{1.m}, \ldots, \bar{y}_{I,m})^{\mathrm{T}}$ be the optimal solution of scenario m in step 1. Step 2 optimizes the initial staffing $\boldsymbol{x}$ and recourse actions by taking $\bar{\boldsymbol{y}}_1, \ldots, \bar{\boldsymbol{y}}_M$ as requirement constraints:

$$\min \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} + \sum_{m=1}^{M} p_m[(\boldsymbol{c}^{+})^{\mathrm{T}}\boldsymbol{r}_m^{+} - (\boldsymbol{c}^{-})^{\mathrm{T}}\boldsymbol{r}_m^{-}]$$

subject to:

$$\begin{aligned} \boldsymbol{x} + \boldsymbol{r}_m^{+} - \boldsymbol{r}_m^{-} &\geq \bar{\boldsymbol{y}}_m, && \forall m, \\ 0 \leq \boldsymbol{r}_m^{-} &\leq \boldsymbol{x}, && \forall m, \\ \boldsymbol{x}, \boldsymbol{r}_m^{+}, \boldsymbol{r}_m^{-} &\geq 0 \text{ and integer}, && \forall m. \end{aligned} \tag{T1}$$

The principal difference with (S2^v) of CP is that (T1) is more restrictive (and simpler), because it only uses the staffing requirements found in step 1 rather than the subgradient-based cuts to delimit the feasible region of solutions, as in (S2^v).

6 NUMERICAL EXPERIMENTS

In this section, we evaluate the performance of the proposed algorithms using different models based on data from a call center of a major utility company in Quebec, Canada. We consider a model with 6 call types and 8 agent groups. We optimize the staffing for a single period (day) of 10 hours. The service times follow log-normal distributions with average service times between 5.1 and 11.3 minutes. The patience times are exponential variables with means between 36 and 52 minutes.

We now describe the uncertainty of the arrival rates of this example. The arrival rate λ_k of type k of a day (this rate stays constant throughout the day) is a random variate (realization) of random variable Λ_k, which is the product of two random variables. That is, $\Lambda_k = \theta_k \beta_k$, where β_k is the busyness factor of the day (see Avramidis, Deslauriers, and L'Ecuyer 2004) and follows a symmetric triangular distribution of mean and mode 1, minimum 0.9, and maximum 1.1. The variable θ_k represents the random "mean" arrival rate of type k. For the different k, we suppose that θ_k is normally distributed with mean from 0.45 to 9.15 calls per minute and 10% standard deviation. In the first stage of the staffing problem, both θ_k and β_k are random variables. In the second stage, θ_k is known, but β_k remains random.

In our numerical test, we consider $M = 20$ and 50 scenarios of θ_k. We choose the parameters as follows: the acceptable waiting times are $\tau_k = \tau = 120$ (seconds), the targets of SLs are $l_k = l = 80\%$, the targets for the probabilities that the SLs are satisfied are $1 - \delta_k = 80\%$ for all call types k, and $1 - \delta = 85\%$ for the aggregated one. The first-stage cost c_i of an agent of group i is defined by

$$c_i = 1 + 0.1(|\mathscr{S}_i| - 1),$$

where $|\mathscr{S}_i|$ is the cardinality of skill set $\mathscr{S}_i$. We define the recourse costs $\boldsymbol{c}^{+}$ and $\boldsymbol{c}^{-}$, for adding and removing agents, proportionally to the initial costs $\boldsymbol{c}$. In practice, there can be many types of recourse actions with diverse penalty costs. For example, the cancellation of a training session may have a low penalty cost, whereas asking on-call agents to come in to work may incur a higher cost. We consider three different cost structures of $\boldsymbol{c}^{+}$ and $\boldsymbol{c}^{-}$, as shown in Table 1.

In Table 2, we report the results obtained by the CP and TS algorithms. During the optimization, we compute $\hat{g}$ and $\hat{g}_k$ by simulating a sample size of $N = 1000$ days. To evaluate the quality of the obtained solutions, we use out-of-sample simulation with sample size of 10000 days. In R1 and R2, the differences

Table 1: Three different recourse cost structures.

	$\boldsymbol{c}^+$	$\boldsymbol{c}^-$
R1	$2\boldsymbol{c}$	$0.5\boldsymbol{c}$
R2	$1.5\boldsymbol{c}$	$0.75\boldsymbol{c}$
R3	$1.1\boldsymbol{c}$	$0.9\boldsymbol{c}$

between $\boldsymbol{c}^+$, $\boldsymbol{c}^-$ and $\boldsymbol{c}$ are high. For these two cases, the final costs and the averaged $\boldsymbol{r}^+$ and $\boldsymbol{r}^-$ values given by the CP algorithm are always smaller than those given by the TS. On the contrary, with R3, when $\boldsymbol{c}^+$, $\boldsymbol{c}^-$ and $\boldsymbol{c}$ are not significantly different, the final costs and the averaged $\boldsymbol{r}^+$ and $\boldsymbol{r}^-$ are closer in value between the two algorithms. The results show that if the recourse actions are expensive, then the CP algorithm generally finds a lower cost solution than TS. This can be explained by the lower number of recourse actions in the solution from CP, which also means that the initial staffing is better chosen. It is important to note that, in the out-of-sample tests, the numbers of infeasible scenarios (i.e., the respective staffing is infeasible) obtained by the CP algorithm are always less than those of TS, as reported under the column "Number infeasible scenarios". Finally, we also report the computational times which show that the CP and TS perform quite similarly in terms of speed. We used a computer server with an Intel(R)-Xeon(R), 16 cores, 2.4GHz and running Linux, and CPLEX 12.6 with LP/IP solvers. The simulations were performed using the Java simulation library *ContactCenters* (Buist and L'Ecuyer 2005).

Table 2: Results of CP and TS for different recourse costs and number of scenarios M.

Recourse costs	M	Algorithm	Total cost	Avg. $\boldsymbol{r}^+$	Avg. $\boldsymbol{r}^-$	Number of infeasible scenarios	CPU time (hours)
R1	20	CP	193.55	1.8	8.3	0	20.9
		TS	200.60	6.0	13.8	4	23.7
	50	CP	193.70	3.0	5.6	7	56.9
		TS	200.32	6.9	13.9	13	54.8
R2	20	CP	190.88	0.8	7.3	1	22.0
		TS	192.41	5.2	13.5	9	27.3
	50	CP	189.69	1.8	7.2	5	53.9
		TS	191.82	6.5	14.6	23	61.9
R3	20	CP	189.83	11.4	7.3	3	20.5
		TS	187.76	12.6	6.1	5	21.1
	50	CP	188.10	10.5	6.1	7	53.0
		TS	186.29	12.3	6.1	13	52.6

7 CONCLUSION

We studied a two-stage stochastic staffing problem with chance constraints for multi-skill call centers with uncertain arrival rates. To solve this problem, we proposed an extension of the simulation-based and linear cut algorithm of Ceẓik and L'Ecuyer (2008). We compared our algorithm with a traditional two-step method on an example inspired by a realistic call center. Numerical results show that our algorithm generally finds better solution than the two-step method, especially when the penalty costs for recourse actions are expensive. We are investigating a more general problem that includes stochastic agent absenteeism.

REFERENCES

Atlason, J., M. A. Epelman, and S. G. Henderson. 2008. "Optimizing Call Center Staffing using Simulation and Analytic Center Cutting Plane Methods". *Management Science* 54 (2): 295–309.

Avramidis, A. N., W. Chan, M. Gendreau, P. L'Ecuyer, and O. Pisacane. 2010. "Optimizing Daily Agent Scheduling in a Multiskill Call Centers". *European Journal of Operational Research* 200 (3): 822–832.

Avramidis, A. N., W. Chan, and P. L'Ecuyer. 2009. "Staffing Multi-Skill Call Centers via Search Methods and a Performance Approximation". *IIE Transactions* 41 (6): 483–497.

Avramidis, A. N., A. Deslauriers, and P. L'Ecuyer. 2004. "Modeling Daily Arrivals to a Telephone Call Center". *Management Science* 50 (7): 896–908.

Bassamboo, A., J. M. Harrison, and A. Zeevi. 2006. "Design and Control of a Large Call Center: Asymptotic Analysis of an LP-Based Method". *Operations Research* 54 (3): 419–435.

Bhulai, S., G. Koole, and A. Pot. 2008. "Simple Methods for Shift Scheduling in Multiskill Call Centers". *Manufacturing & Service Operations Management* 10 (3): 411–420.

Buist, E., and P. L'Ecuyer. 2005. "A Java Library for Simulating Contact Centers". In *Proceedings of the 2005 Winter Simulation Conference*, edited by M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 556–565. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Ceşik, M. T., and P. L'Ecuyer. 2008. "Staffing Multiskill Call Centers via Linear Programming and Simulation". *Management Science* 54 (2): 310–323.

Chan, W. 2013. *Optimisation des Horaires des Agents et du Routage des Appels dans les Centres d'Appels*. Ph. D. thesis, Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Canada.

Chan, W., G. Koole, and P. L'Ecuyer. 2014. "Dynamic Call Center Routing Policies Using Call Waiting and Agent Idle Times". *Manufacturing & Service Operations Management* 16 (4): 544–560.

Chan, W., T. A. Ta, P. L'Ecuyer, and F. Bastin. 2014. "Chance-Constrained Staffing with Recourse for Multi-Skill Call Centers with Arrival-Rate Uncertainty". In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, 4103–4104. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Channouf, N., P. L'Ecuyer, A. Ingolfsson, and A. N. Avramidis. 2007. "The Application of Forecasting Techniques to Modeling Emergency Medical System Calls in Calgary, Alberta". *Health Care Management Science* 10 (1): 25–45.

Cooper, R. B. 1981. *Introduction to Queueing Theory*. second ed. New York, NY: North-Holland.

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., H. Shen, Y.-P. Zhou, N. Korolev, A. McCord, and H. Ristock. 2015. "Parametric Forecasting and Stochastic Programming Models for Call-Center Workforce Scheduling". *Manufacturing & Service Operations Management* 17 (4): 571–588.

Garnett, O., A. Mandelbaum, and M. Reiman. 2002. "Designing a Call Center with Impatient Customers". *Manufacturing and Service Operations Management* 4 (3): 208–227.

Gurvich, I., J. Luedtke, and T. Tezcan. 2010. "Staffing Call Centers with Uncertain Demand Forecasts: A Chance-Constrained Optimization Approach". *Management Science* 56 (7): 1093–1115.

Harrison, J. M., and A. Zeevi. 2005. "A Method for Staffing Large Call Centers Based on Stochastic Fluid Models". *Manufacturing & Service Operations Management* 7 (1): 20–36.

Heskett, J. L., T. O. Jones, G. W. Loveman, W. E. Sasser Jr., and L. A. Schlesinger. 1994, March–April. "Putting the Service-Profit Chain to Work". *Harvard Business Review*:164–174. Reprint 94204.

Ibrahim, R., and P. L'Ecuyer. 2013. "Forecasting Call Center Arrivals: Fixed-Effects, Mixed-Effects, and Bivariate Models". *Manufacturing and Services Operations Management* 15 (1): 72–85.

Ibrahim, R., H. Ye, P. L'Ecuyer, and H. Shen. 2016. "Modeling and Forecasting Call Center Arrivals: A Literature Study and a Case Study". *International Journal of Forecasting* 32 (3): 865–874.

Jouini, O., G. Koole, and A. Roubos. 2013. "Performance Indicators for Call Centers with Impatient Customers". *IIE Transactions* 45 (3): 341–354.

Kelley Jr., J. E. 1960. "The Cutting-Plane Method for Solving Convex Programs". *Journal of the Society for Industrial and Applied Mathematics* 8 (4): 703–712.

Kim, K., and S. Mehrotra. 2015. "A Two-Stage Stochastic Integer Programming Approach to Integrated Staffing and Scheduling with Application to Nurse Management". *Operations Research* 63 (6): 1431–1451.

Oreshkin, B., N. Régnard, and P. L'Ecuyer. 2016. "Rate-Based Daily Arrival Process Models with Application to Call Centers". *Operations Research* 64 (2): 510–527.

Pot, A., S. Bhulai, and G. Koole. 2008. "A Simple Staffing Method for Multiskill Call Centers". *Manufacturing & Service Operations Management* 10 (3): 421–428.

Robbins, T. R., and T. Harrison. 2008. "A Simulation-Based Scheduling Model for Call Centers with Uncertain Arrival Rates". In *Proceedings of the 2008 Winter Simulation Conference*, edited by S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, and J. W. Fowler, 2884–2889. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Robbins, T. R., and T. P. Harrison. 2010. "A Stochastic Programming Model for Scheduling Call Centers with Global Service Level Agreements". *European Journal of Operational Research* 207 (3): 1608–1619.

Van Slyke, R. M., and R. Wets. 1969. "L-Shaped Linear Programs with Applications to Optimal Control and Stochastic Programming". *SIAM Journal on Applied Mathematics* 17 (4): 638–663.

Wallace, R. B., and W. Whitt. 2005. "A Staffing Algorithm for Call Centers with Skill-Based Routing". *Manufacturing & Service Operations Management* 7 (4): 276–294.

AUTHOR BIOGRAPHIES

WYEAN CHAN is a postdoctoral fellow at the Université de Montréal, Canada. He holds a PhD degree in Computer science from the Université de Montréal. His main research projects are in stochastic optimization, simulation, and wait time estimation in call centers. He is a key contributor in the development of simulation and optimization software for call centers available at http://simul.iro.umontreal.ca. His email address is chanwyea@iro.umontreal.ca.

TA THUY ANH is a Ph.D. student at Département d'Informatique et de Recherche Opérationnelle, Université de Montréal. She is student member of CIRRELT and GERAD. Her research concerns models and simulation-based optimization algorithms for problems in contact centers. Her detailed information can be found at http://www-etud.iro.umontreal.ca/~tathuyan/.

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 and an Inria International Chair in Rennes, France. He is a member of the CIRRELT and GERAD research centers. 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 served as Editor-in-Chief for *ACM Transactions on Modeling and Computer Simulation* from 2010 to 2013. He is currently Associate Editor for *ACM Transactions on Mathematical Software*, *Statistics and Computing*, and *International Transactions in Operational Research*. More information can be found on his web page: http://www.iro.umontreal.ca/~lecuyer.

FABIAN BASTIN is Associate Professor in the Département d'Informatique et de Recherche Opérationnelle, at the Université de Montréal, Canada. He is a member of the CIRRELT research center and serves as associate editor for Journal of Industrial and Management Optimization and treasurer of the Montreal Section of the Canadian Operational Research Society. His main research interests are stochastic optimization, nonlinear programming and discrete choice modeling. More information can be found on his web page: http://www.iro.umontreal.ca/~bastin.