European Journal of Operational Research 207 (2010) 1608–1619

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

A stochastic programming model for scheduling call centers with global Service Level Agreements

Thomas R. Robbins *, Terry P. Harrison

Department of Marketing and Supply Chain, College of Business, East Carolina University, Greenville, NC, United States
Department of Supply Chain and Information Systems, Smeal College of Business, Penn State University, University Park, PA, United States

ARTICLE INFO

Article history:
Received 9 June 2009
Accepted 11 June 2010
Available online 19 June 2010

Keywords:
Stochastic programming
Scheduling
OR in manpower planning
Call centers

ABSTRACT

We consider the issue of call center scheduling in an environment where arrivals rates are highly variable, aggregate volumes are uncertain, and the call center is subject to a global service level constraint. This paper is motivated by work with a provider of outsourced technical support services where call volumes exhibit significant variability and uncertainty. The outsourcing contract specifies a Service Level Agreement that must be satisfied over an extended period of a week or month. We formulate the problem as a mixed-integer stochastic program. Our model has two distinctive features. Firstly, we combine the server sizing and staff scheduling steps into a single optimization program. Secondly, we explicitly recognize the uncertainty in period-by-period arrival rates. We show that the stochastic formulation, in general, calculates a higher cost optimal schedule than a model which ignores variability, but that the expected cost of this schedule is lower. We conduct extensive experimentation to compare the solutions of the stochastic program with the deterministic programs, based on mean valued arrivals. We find that, in general, the stochastic model provides a significant reduction in the expected cost of operation. The stochastic model also allows the manager to make informed risk management decisions by evaluating the probability that the Service Level Agreement will be achieved.

1. Introduction

A call center is a facility designed to support the delivery of some interactive service via telephone communications; typically an office space with multiple workstations manned by agents who place and receive calls (Gans et al., 2003). Call centers are a large and growing component of the US and world economy, and are estimated to employ over 2 million call center agents (Aksin et al., 2007). Large scale call centers are technically and managerially sophisticated operations and have been the subject of substantial academic research. Call center applications include telemarketing, customer service, help desk support, and emergency dispatch.

Staffing is a critical issue in call center management, as direct labor costs often account for 60–80% of the total operating budget of a call center (Aksin et al., 2007). This paper addresses the scheduling problem in a call center with highly variable and uncertain arrival rates. The work is directly related to a research project with a provider of outsourced technical support delivered via globally distributed call centers. This operation involves providing help desk support to large corporate and government entities. While the scope of services varies from account to account, many accounts require 24×7 support and virtually all accounts are subject to some form of Service Level Agreement (SLA). There are multiple types of SLAs, but the most common specifies a minimum level of the Telephone Service Factor (TSF). A TSF SLA specifies the proportion of calls that must be answered within a specified time. For example, an 80/120 SLA specifies that 80% of calls must be answered within 120 seconds. A very important point is that the service level applies to an extended period, typically a week or month. Therefore, the desk is often staffed so that at some times the service level is underachieved, sometimes overachieved, and is on target for the entire month. The key challenge involved with staffing this call center is meeting a fixed SLA with a variable and uncertain arrival rate pattern.

Throughout this analysis, we will evaluate the models using three test problems based on specific outsourcing projects. Project J is a corporate help desk for a large industrial company averaging about 750 calls a day, where the volatility of call volume is relatively low. Project S is a help desk that provides support to workers in a large national retail chain. Call volume on this desk is about 2000 calls a day. Because this desk supports users in retail stores, as opposed to a corporate office, the daily seasonality of call volume is quite different.

* Corresponding author. Tel./fax: +1 252 328 6588.
E-mail address: robbinst@ecu.edu (T.R. Robbins).

0377-2217/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2010.06.013

This company is making major changes in its IT infrastructure and therefore call volume is very volatile and difficult to forecast. Project O is a help desk that provides support to corporate and retail site users of another retail chain. This is a smaller desk with about 500 calls a day, where call volume is fairly volatile and shocks are relatively common. We also examine various scheduling options. At one extreme, we only allow workers to be assigned to five 8-hour shifts per week. At the opposite extreme, we allow a wide range of part time schedules. We allow for a total of five different flexibility options (A-E), which are summarized in the Appendix in Tables A.1 and A.2.

2. Literature

There is a large body of literature addressing call center issues. Gans et al. (2003) provides a detailed and comprehensive review of the literature. Aksin et al. (2007) is another more recent review of the call center literature. Call center papers cover a wide range of topics and encompass a number of OR methodologies, including queuing theory, optimization, and simulation.

The primary problem we address in this paper is shift scheduling. The basic approach to this problem was first outlined in a paper by Dantzig (1954), which addressed scheduling toll booth operators. Dantzig formulated his model as a weighted set covering problem with known staffing requirements; the objective being to find the minimal cost covering from a set of available schedules. In the weighted set covering approach, the staffing levels in each time period are calculated exogenously and are defined as hard constraints that must be satisfied in any feasible schedule. Segal (1974) showed that without considering breaks the problem could be solved as a network flow problem in polynomial time. However, when breaks are scheduled explicitly the problem becomes NP Hard (Garey and Johnson, 1979). Due to the large number of potential schedules, especially when breaks are explicitly scheduled, much of the early research focused on solution algorithms.

Many early papers focused on heuristic algorithms. Henderson and Berry (1976) apply two types of heuristics. The first heuristic reduces the number of shift types, scheduling against only a reduced set of schedules referred to as the *working subset*. The second approximation is the scheduling algorithm, where the authors use three different scheduling heuristics. Another stream of research attacks the problem using an implicit scheduling approach. Implicit scheduling models use two sets of decision variables; one to assign breakless shifts, another to fit breaks. Implicit scheduling approaches are addressed in Bechtold and Jacobs (1990), Thompson (1995) and Aykin (1996). Several other papers address related problems (Brusco and Johns, 1996; Brusco and Jacobs, 1998, 2000). A succinct overview of a two-stage approach to scheduling in a call center environment is provided in Section 12.7 of Pinedo (2005).

Customer service is an important consideration in call centers, and many centers are subject to SLAs. Milner and Olsen (2008) examine contract structures in call centers with SLAs. Baron and Milner (2006) examine optimal staffing under various SLAs. These papers classify SLAs as Individual Based (IB), Period Based (PB), or Horizon Based (HB). IB-SLAs assess a financial penalty for every customer not served within the specified service level. The PB-SLA specifies penalties for each time period in which the service level target is not achieved. Periods are defined as intervals over which the arrival rate can be considered constant – typically 15 or 30 minute intervals. The HB-SLA specifies penalties for service level shortcomings over an extended period such as a week or month. In this paper we examine scenarios where a HB-SLA has been specified with the horizon specified as one week.

Most call center scheduling models in the literature implement a hard constraint for service level on a period-by-period basis – a PB-SLA. Scheduling for a PB-SLA is straightforward using the Stationary Independent Period by Period (SIPP) approach. The SIPP approach is described in detail in Green et al. (2001), but essentially the day is divided into short periods, typically 15 or 30 minutes. In each period, the arrival rate is assumed to be constant and performance is assumed to be independent of the performance in other periods. In each period, a queuing model, often the Erlang C model, is used to calculate the staffing level required to achieve the service level requirement. A set covering integer program is then used to schedule shifts. This two phased approach splits the task into a server sizing task, based on queuing models, and a staff scheduling task, based on discrete optimization.

A few models are formulated to solve a global service level requirement, i.e. an HB-SLA. It is our experience that outsourcing contracts often specify an HB-SLA, and all of the projects we examined were subject to this type of SLA. Koole and van der Sluis (2003) attempt to develop a staffing model that optimizes a global objective based on an HB-SLA. Their model uses a local search algorithm, and to ensure convergence to a global optimum they require agent schedules with no breaks, and assume no abandonment. Their model also assumes a time varying, but known, arrival rate. Cezik and L'Ecuyer (2007) solve a global service level problem using simulation and integer programming. They use simulation to estimate service level attainment and integer programming to generate the schedule. The IP model generates cuts via subgradient estimation calculated via simulation. The model solves the sample average problem and therefore ignores arrival rate uncertainty, but it does allow for multiple skills. This model is an extension of the model presented in Atlason et al. (2004). In a related paper Avramidis et al. (2007a) use a local search algorithm to solve the same problem. A related model is presented in Avramidis et al. (2007b). Fukunaga et al. (2002) describe a commercial scheduling application widely used for call center scheduling. Global service level targets are modeled as soft constraints while certain staffing restrictions are modeled as hard constraints. The algorithm uses an artificial intelligence based search heuristic. Atlason et al. (2008) develop an algorithm that combines server sizing and staff scheduling into a single optimization problem. This model focuses on the impact that staffing in one time period can have on performance in the subsequent period, a fact ignored in SIPP models. The algorithm utilizes discrete event simulation to calculate service levels under candidate staffing models, and a discrete cutting plane algorithm to search for improving solutions.

Each of these models either assumes that the per-period arrival rate is known or schedules against the expected arrival rate. The issue of arrival rate uncertainty has been addressed in several recent papers. Both major call center reviews (Gans et al., 2003; Aksin et al., 2007) have sections devoted to arrival rate uncertainty. Brown et al. (2005) perform a detailed empirical analysis of call center data. While they find that a time-inhomogeneous Poisson process fits their data, they also find that arrival rate is difficult to predict and suggest that the arrival rate should be modeled as a stochastic process. Many authors argue that call center arrivals follow a doubly stochastic process, a Poisson process where the arrival rate is itself a random variable (Chen and Henderson, 2001; Whitt, 2006; Aksin et al., 2007). Arrival rate uncertainty may exist for multiple reasons. Arrivals may exhibit randomness greater than that predicted by the Poisson process due to unobserved variables; the weather may have an impact on emergency calls (Chen and Henderson, 2001), the state of an organization's IT infrastructure may have an impact on support center calls (Robbins, 2007), and TV advertising may have an impact on inbound volume to a sales center (Andrews and Cunningham, 1995). Call volume is highly seasonal over the course of a day, week, month and year (Andrews and Cunningham, 1995; Gans et al., 2003; Robbins, 2007). Call center managers attempt to account for these factors when they develop forecasts, yet forecasts are subject

to significant error. Robbins (2007) compared four months of weekday forecasts to actual call volume for 11 call center projects. He found that the average forecast error exceeds 10% for 8 of 11 projects, and 25% for 4 of 11 projects. The standard deviation of the daily forecast to actual ratio exceeds 10% for all 11 projects. Steckley et al. (2009) compared forecasted and actual volumes for nine weeks of data taken from four call centers. They showed that the forecasting errors are large and modeling arrivals as a Poisson process with the forecasted call volume as the arrival rate can introduce significant error. Robbins et al. (2006) used simulation analysis to evaluate the impact of forecast error on performance measures, demonstrating the significant impact forecast error can have on system performance.

Some recent papers address staffing requirements when arrival rates are uncertain. Bassamboo et al. (2005) develop a model that attempts to minimize the cost of staffing plus an imputed cost for customer abandonment for a call center with multiple customer and server types when arrival rates are variable and uncertain. They solve the staffing and routing problems using an LP based method that is asymptotically optimal. Harrison and Zeevi (2005) use a fluid approximation to solve the sizing problem for call centers with multiple call types, multiple agent types, and uncertain arrivals. Their model seeks to minimize a deterministic staffing cost function along with a penalty cost associated with abandonment. Their approach models the staffing problem as a multidimensional newsvendor model and solves it through a combination of linear programming and simulation. Whitt (2006) allows for arrival rate uncertainty as well as uncertain staffing, i.e. absenteeism when calculating staffing requirements. Steckley et al. (2004) examine the type of performance measures to use when staffing under arrival rate uncertainty. Each of these models incorporate arrival rate uncertainty into the server sizing step, but do not explicitly address the staff scheduling step.

The model presented in our paper seeks to allow for arrival rate uncertainty while simultaneously integrating the server sizing and staff scheduling steps. We do this through a model formulated as a stochastic integer program. The theory of stochastic programming is well defined. Birge and Louveaux (1997) is a classic text that reviews both the theory of stochastic programming and numerous solution algorithms. A standard method of solving stochastic programs is to solve the sample path problem, solving the optimization problem against a discrete set of samples referred to as scenarios. Mak et al. (1999) discuss important statistical properties associated with sample path optimization.

3. Problem formulation and solution approach

In this model, we attempt to find a minimal cost staffing plan that satisfies a global service level requirement. Our model estimates the number of calls that meet the service level requirement in each period by making a piecewise linear approximation to the TSF curve; the curve that relates the number of agents to a given service level for a given arrival rate. In this section, we first present our formulation of the model, including our approach for estimating service levels. We then outline a process for solving large-scale integer program that results. Finally, we present a post-optimization approach to assess the quality of the resulting solution, an important consideration in stochastic program.

3.1. Formulation

We formulate the model as a two stage, mixed-integer stochastic program. In the first stage, staffing decisions are made and in the second stage, call volume is realized and we calculate SLA attainment. We formulate a model with the following definitions:

Sets

I	time periods
J	possible schedules
K	scenarios
H	points in a linear approximation

Deterministic parameters

c_j	cost of schedule j
a_{ij}	indicates if schedule j is staffed in time period i
g	global SLA goal
m_{ikh}	slope of piecewise TSF approximation h in period i of scenario k
b_{ikh}	intercept of piecewise TSF approximation h in period i of scenario k
p_k	probability of scenario k
μ_i	minimum number of agents in period i
d_j	maximum number of agents available for schedule j
r	per point penalty cost of TSF shortfall

Decision variables

x_j	number of resources assigned to schedule j

State variables

y_{ik}	number of calls in period i of scenario k answered within service level
S_k	proportional TSF shortfall in scenario k

Stochastic parameters

n_{ik}	number of calls in period i of scenario k

$$\min \quad \sum_{j\in J} c_j x_j + \sum_{k\in K} p_k r S_k \tag{3.1}$$

$$\text{subject to } y_{ik} \leqslant n_{ik}\left(m_{ikh}\sum_{j\in J} a_{ij}x_j + b_{ikh}\right) \quad \forall i \in I,\ k \in K,\ h \in H, \tag{3.2}$$

$$\sum_{i\in I} n_{ik}S_k \geqslant \sum_{i\in I}(gn_{ik} - y_{ik}) \quad \forall k \in K, \tag{3.3}$$

$$y_{ik} \leqslant n_{ik} \quad \forall i \in I,\ k \in K, \tag{3.4}$$

$$\sum_{j\in J} a_{ij}x_j \geqslant \mu_i \quad \forall i \in I, \tag{3.5}$$

$$x_j \leqslant d_j \quad \forall j \in J, \tag{3.6}$$

$$x_j \in \mathbb{Z}^+, \quad y_{ik} \in \mathbb{R}^+, \quad S_k \in \mathbb{R}^+ \quad \forall i \in I,\ k \in K,\ j \in J. \tag{3.7}$$

The objective of this model (3.1) is to minimize the sum of the total cost of staffing and the expected penalty cost associated with failure to achieve the desired service level. The optimization occurs over a set K of sample realizations of call arrivals. Constraint (3.2) defines the variable y_{ik} as the number of calls answered within the SLA target in period i of scenario k based on a convex linear approximation of the TSF curve shown in Fig. 3.1. Constraint (3.3) calculates the TSF proportional shortfall, S_k: the maximum of either the percentage point difference between the target TSF and achieved TSF or zero. Constraint (3.4) limits the calls answered within the SLA target to the total calls received in the period. Constraint (3.5) defines the minimum number of agents in any period. The minimum agent level is set to the maximum of the global minimum number of agents required by policy, typically two agents, and the staffing level required to achieve a minimum service level at expected call volumes. In our test cases, the parameter d_j is set to the maximum of two, and the number of agents that results in a service level of at least 50% at the average volume for the period. Constraint (3.6) sets an upper limit on the number of agents assigned to each schedule. The purpose of this constraint is to limit the number of agents assigned to a schedule based on agent availability or willingness to work. In practice, this constraint also allows the call center manager to turn off certain schedules as he sees fit. Constraint (3.7) defines the non-negativity and integer conditions for program variables.

For a given planning horizon and scheduling interval, the size of the model, and therefore the computation effort required to solve it, is driven in large part by two factors; the number of potential schedules (J) and the number of scenarios (K). The number of integer variables is equal to the number of schedules, while the number of continuous variables is equal to the product of the number of scenarios and the number of time periods, plus the number of scenarios. A common planning horizon is one week, and a common interval is 30 minutes. Increasing the planning horizon or decreasing scheduling interval will directly increase the number of time periods (I) and indirectly increase the number of schedules (J) and therefore the resources required to solve the problem.

In this analysis, we are creating schedules for a week (with explicit breaks between shifts, but not within shifts.) In simple cases, where we allow only 5 day a week, 8-hour shifts, the number of possible schedules is 576. In more complex cases, where we have a wider range of full and part time schedule options we have 3696 schedules. (Details are presented in Table A.2.) We investigate the number of scenarios required in the next section, but 50 scenarios are not unreasonable. This implies the requirement to solve models with 3696 integer variables and over 16,000 continuous variables.

This program (3.1)–(3.7) is solved over some set of sample outcomes from the statistical model of call arrival patterns. Multiple approaches are available for generating simulated arrival patterns. A thorough analysis is provided in Avramidis et al. (2004). For our test problems, we use a simple two-stage algorithm similar to the model in Weinberg et al. (2007). We use a multiphase, multiplicative model where the arrival rate is the product of a daily number of calls and the proportion of daily calls received in that time period; both of which

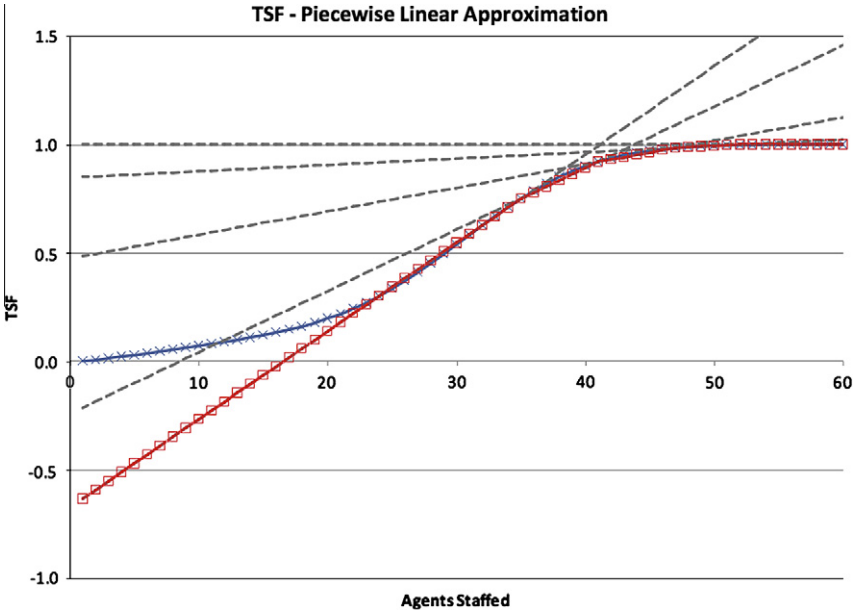

Fig. 3.1. Piecewise approximation of TSF.

are random. Details of the algorithm are presented in the Appendix in Fig. A.1, but it should be noted that the scheduling algorithm is in no way dependent on the model of arrivals.

3.2. TSF approximation

The objective of our optimization model is to find the lowest cost staffing plan giving a service level constraint based on the TSF. In order to satisfy service level constraints, the model must estimate the service level that will be achieved for a given staffing model and call pattern. We perform this estimate by using a piecewise linear approximation of the TSF curve based on the Erlang A queuing model.

The Erlang A model is a widely accepted model for call center systems with a non-negligible abandonment rate. Erlang A assumes calls arrive via a Poisson process with rate λ and are served by a set of homogeneous agents with an exponentially distributed service time with mean $1/\mu$. If no agent is available when the call arrives, it is placed in an infinite capacity queue where it waits for the next available agent. Each caller has a patience level, which are iid draws from an exponential distribution with mean $1/\theta$. If a caller is not served by the time her patience expires, she hangs up. The call center is also assumed to have infinite capacity, so no calls are blocked. For each test project, we utilize estimates derived from actual call center data. In steady state, the staffing decision then involves forecasting the arrival rate λ_i and setting the staff level based on the Erlang A approximation. The result is a nonlinear S-shaped curve that, for a fixed arrival rate, relates the achieved service level to the number of agents staffed.

The TSF curve is neither convex nor concave over the full range of staffing. For very low staffing levels, where performance is very poor, the curve is convex, and we experience increasing efficiency from incremental staffing. For higher staffing levels, the curve becomes concave, and the impact of incremental staffing becomes decreasing. Note that the area of convexity corresponds to very poor system performance, an area where we do not plan to operate. In addition, embedding this function in our optimization model would create a non-convex optimization problem. To address this problem we create a piecewise linear approximation to the TSF curve as shown in Fig. 3.1.

In this graph, the straight lines represent the individual constraints, and the piecewise linear function is our approximation of the non-linear curve. This graph has five linear segments, including a horizontal segment at a service level of 100%. The optimization model requires that the TSF is less than each line segment. The piecewise linear approximation and the true TSF curve are very close for service levels above 25%. For very low staffing levels, the linear approximation will overly penalize performance, potentially calculating a negative TSF level. The optimization process will force these constraints to be binding and will force the TSF to be non-negative. Our assumption is that we are almost always operating in the higher performance region. In all our test cases, we constrain the problem so that expected performance in any period is greater than 50% via constraint (3.5).

3.3. Solution algorithm

Our model is formulated with a finite number of call arrival patterns, and can therefore be expressed as a deterministic equivalent mixed integer program and as such can be solved by an implicit enumeration (branch and bound) algorithm. Algorithms such as branch and bound, which ignore the special structure of a stochastic program, tend to become quite inefficient for large-scale stochastic programs (Birge and Louveaux, 1997). A common approach for solving stochastic programs is to exploit the structure of the program through a decomposition algorithm (Birge and Louveaux, 1997). We implemented a version of the L-Shaped decomposition algorithm adapted for a discrete first stage. We decompose the problem into a master problem where the staffing decision is made, and a series of sub-problems where the TSF shortfall is calculated for each scenario.

Let v denote the major iterations of the algorithm. Also let E^v_{ik} and e^v_{ik} denote the coefficients of the cut generated in iteration k. The master problem is then defined as

$$\min \quad \sum_{j\in J} c_j x_j + \theta^v \tag{3.8}$$

$$\text{subject to } \theta^v \geqslant \sum_{k\in K} p_k E^v_{ik} \sum_{j\in J} a_{ij}x_j + e^v_{ik} \quad \forall i \in I, v, \tag{3.9}$$

$$\sum_{j\in J} a_{ij}x_j \geqslant \mu_i \quad \forall i \in I, \tag{3.10}$$

$$x_j \leqslant d_j \quad \forall j \in J, \tag{3.11}$$

$$x_j \in \mathbb{Z}^+, \quad \boldsymbol{\theta}^v \in \mathbb{R}^+ \quad \forall j \in J. \tag{3.12}$$

In this problem, θ^v represents an estimate of the TSF shortfall penalty term. Let (x^v, θ^v) be an optimal solution. For each realization of the random vector $k = 1, \ldots, K$, we then solve the following subproblem

$$\min \quad rS_k \tag{3.13}$$

$$\text{subject to } y_{ik} \leqslant n_{ik}\left(m_{ikh}\sum_{j\in J} a_{ij}x^v_j + b_{ikh}\right) \quad \forall i \in I,\ k \in K,\ h \in H, \tag{3.14}$$

$$\sum_{i\in I} n_{ik}S_k \geqslant \sum_{i\in I}(gn_{ik} - y_{ik}) \quad k \in K, \tag{3.15}$$

$$y_{ik} \leqslant n_{ik} \quad \forall i \in I,\ k \in K, \tag{3.16}$$

$$x_j \in \mathbb{Z}^+, \quad y_{ik} \in \mathbb{R}^+, \quad S_k \in \mathbb{R}^+ \quad \forall i \in I,\ j \in J,\ k \in K. \tag{3.17}$$

We use the dual variables from the solution of the set of sub-problems to improve the approximation of the penalty term. Let $\pi 1^v_{ikh}$ be the dual variables associated with (3.14), $\pi 2^v_k$ the dual variables associated with (3.15), and $\pi 3^v_{ik}$ the dual variables associated with (3.16). We then calculate the following parameters used for cut generation:

$$E_{ik}^{v+1} = \sum_{i\in I}\sum_{h\in H}\pi 1_{ikh}^{v} m_{ikh}\sum_{j\in J} a_{ij}x_j^{v},$$

$$e_k^{v+1} = \sum_{i\in I}\left[\pi 3_{ik}^{v} n_{ik} + \sum_{h\in H}\pi 1_{ikh}^{v} b_{ikh} n_{ik}\right] - \pi 2_k^{v} g\sum_{i\in I} n_{ik}.$$

We use these values to generate a constraint of the form (3.9). Set $v = v + 1$, add the constraint to the master program and iterate. The algorithm solves the master program and then solves each subprogram for the fixed staffing level defined in the master solution. Based on the solution of the sub-problems, each iteration adds a single cut to the master problem. These cuts create an outer linearization of the penalty function (Geoffrion, 1970).

The solution of the master problem provides a lower bound on the optimal solution, while the average of the subproblem solutions provides an upper bound (Birge and Louveaux, 1997). In our implementation, we solve the LP relaxation of the master until an initial tolerance level on the optimality gap is achieved, and we then reapply the integrality constraints. We continue to iterate between the master MIP and the subprogram LPs until a final tolerance gap is achieved. Whereas the branch and bound approach solves a single large MIP, the decomposition solves a large number of relatively small LPs and a small number of moderately sized MIPs and MIP relaxations. A representative instance with 100 scenarios required 30 major iterations, thereby requiring the solution of the master problem 30 times and the subproblem 3000 times. The master was solved as an LP relaxation 26 times and as a MIP four times. (Fig. 10-8 in the supplementary material illustrates the convergence of the L-Shaped decomposition algorithm for a particular instance with 384 schedules and 100 scenarios.) As is the case with a branch and bound algorithm, relatively good bounds are found in the first few iterations. Convergence then slows as each successive iteration cuts a smaller area from the feasible region of (3.8)–(3.12).

3.4. Post-optimization analysis

The solution to the simple path formulation of a stochastic program is an approximation of the solution to the true optimization problem in which parameters are random variables (Mak et al., 1999). A well-developed theory exists for assessing the quality of simple path approximations based on Monte Carlo sampling techniques (Birge and Louveaux, 1997; Mak et al., 1999; Bayraksan and Morton, 2009). In this section, we outline a process whereby we used this method to test the quality of the solution we obtain when evaluating against the sample of 25 arrival patterns.

The solution of (3.1)–(3.7) is the optimal solution of the sample path problem. We denote the objective value of this solution as z_n^*, where n is the number of scenarios used to calculate the solution. This is a biased estimate of the solution to the true problem; that is, the problem evaluated against the continuous distribution of arrival rates. We denote the objective of the true solution as z^*. Mak et al. (1999) show that the expected bias in the solution is decreasing in sample size

$$E[z_n^*] \leqslant E[z_{n+1}^*] \leqslant z^*.$$

From a practical perspective, a key decision is determining the number of scenarios to use in our optimization. As we increase the number of scenarios, the solution becomes a better approximation of the true solution, but the computational cost of finding that solution increases.

To aid in this process, we perform a post-optimization evaluation of the candidate solution using a Monte Carlo bounding process described in Mak et al. (1999). Denote the solution to the sample problem as $\hat{x}$. We then solve the subprogram (3.13)–(3.17) using $\hat{x}$ as the candidate solution, to obtain the expected cost of implementing this solution. In this analysis, we solve the subprogram with n_u equal 500 scenarios generated independently from the scenarios used in the optimization. The solution to the subprogram gives us an upper bound on the true solution ($\overline{U}(n_u)$), while the solution to the original problem, z_n^*, is a lower bound ($\overline{L}(n_l)$).

To obtain better bounds on the true optimal solution, we may choose to solve the original problem multiple times, each with independently generated scenarios. Denote the number of batches (sets of scenarios) used to solve the original problem as n_ℓ and the sample variance of the objective as $s_\ell(n_\ell)$. Similarly, we calculated the sample variance of the expected outcome of the candidate solution against the n_u evaluation scenarios. We can then define the following standard errors

$$\tilde{\varepsilon}_u = \frac{t_{n_u-1,\alpha} s_u(n_u)}{\sqrt{n_u}},$$

$$\tilde{\varepsilon}_\ell = \frac{t_{n_\ell-1,\alpha}, s_\ell(n_\ell)}{\sqrt{n_\ell}},$$

where $t_{n_u-1,\alpha}$ is a standard t-statistic, i.e. $P\{T_n \leqslant t_{n_u-1,\alpha}\} = 1-\alpha$. We can now define an approximate $(1-2\alpha)$ confidence interval on the optimality gap as

$$\left[0, \left[\overline{U}(n_u) - \overline{L}(n_\ell)\right]^+ + \tilde{\varepsilon}_u + \tilde{\varepsilon}_\ell\right]. \tag{3.18}$$

Note that we take the positive portion of the difference between the upper and lower bounds because it is possible, due to sampling error, that this difference is negative. This procedure allows us to generate a statistical bound on the quality of our solution. (A graphical analysis of the optimality gap is shown in Figs. 10-13 and 10-14 of the supplementary material.) In an optimization problem with 25 scenarios we achieved a gap of $50 on a schedule with a cost in excess of $11,000, a gap of less than 0.5%. Based on this analysis, we concluded that solving the stochastic program with 25 scenarios would provide near optimal solutions. In our test cases, we used 50 scenarios unless noted otherwise.

4. Cost and service level tradeoffs

In our model we control the certainty with which the target service level is achieved by assigning a financial penalty to a service level shortfall. By adjusting the performance penalty factor, r, we adjust the preferred degree of certainty associated with meeting the target.

While the penalty rate r may be set based on the contractual penalty for failing to achieve the service level, there is an additional implicit cost associated with the perception of poor quality. Put another way, managers typically wish to provide a higher probability of achieving the service level than implied by the explicit penalty rate. We now analyze the relationship between the penalty rate, the cost of service delivery, and the confidence associated with the performance target, i.e. the probability that the service level target is achieved.

In a deterministic optimization approach to call center scheduling, we set a performance target for some metric and then find the minimal cost schedule that satisfies that constraint; i.e. we implement the service level requirement as a *hard* constraint. In a stochastic setting, the call volume, and therefore the service level, is random, and the performance target can only be expressed in probabilistic terms. Given the nature of arrival variability, it is neither practical nor desirable to generate a schedule that will always achieve the service level target as this schedule would be prohibitively expensive. Therefore, we wish to implement the service level requirement as a *soft* constraint.

In Tables 4.1–4.3, we show the result of an experiment evaluating the impact of various penalty rates. For each project, we test eight design points, (DPs), each with a different penalty rate. (The same data is shown graphically in Fig. 10-9 of the supplementary material.) The purpose of this experiment is to determine the penalty rate that should be used for each project to achieve a desired confidence of achieving the service level target. In each case, we solve the stochastic problem five times, each with an independent batch of 50 scenarios. We then evaluate each solution against an independently generated set of 500 scenarios to estimate the expected outcome of implementing the candidate solution. The model is solved with the constraint that all schedules are full-time (40 hours), using schedule B defined in Table A.2.

In all cases, low penalties result in a zero confidence and an expected TSF near 60%. As the penalty rate increases, the expected TSF begins to increase as additional staffing is added to offset shortfall penalties. Both factors increase rapidly and then level off as it becomes increasingly expensive to meet the service levels in the tail of the arrival rate distribution. It is interesting to note that each project requires a different penalty rate to achieve a desired confidence level. Project S, which has the largest staff levels and a high degree of variability, requires penalty rates in the range of $200,000 ($2000 per percentage point shortfall) to schedule with greater than 80% confidence. Project O, a smaller project with moderate variability, plateaus with penalty rates around 100,000. Project J, a stable project, stabilizes with penalty rates at or above 75,000. The call center manager seeks to minimize the cost of staffing, while maximizing the probability of achieving the target service level. These two goals are clearly in conflict and the manager must decide how to balance cost and risk: a decision that is obscured in a deterministic optimization approach.

Table 4.1
Cost and service level tradeoffs – Project J.

DP	Penalty rate	Average				Standard deviation			
		Labor cost	Expected outcome	Average TSF (%)	Confidence (%)	Labor cost	Expected outcome	Average TSF (%)	Confidence (%)
1	0	8800	8800	60.5	0.0	0	0	0.00	0.00
2	25,000	10,800	11,008	80.6	61.6	0	18	0.16	2.73
3	50,000	10,880	11,249	81.0	65.7	179	40	1.16	12.71
4	75,000	11,120	11,332	82.6	82.9	179	28	1.11	11.35
5	100,000	11,120	11,419	82.7	83.1	179	127	1.11	11.74
6	150,000	11,200	11,458	83.1	87.9	0	36	0.30	2.74
7	200,000	11,200	11,504	83.1	88.8	0	56	0.23	2.36
8	250,000	11,200	11,597	83.1	89.0	0	72	0.31	2.30

Table 4.2
Cost and service level tradeoffs – Project S.

DP	Penalty rate	Average				Standard deviation			
		Labor cost	Expected outcome	Average TSF (%)	Confidence (%)	Labor cost	Expected outcome	Average TSF (%)	Confidence (%)
1	0	20,880	20,880	52.5	0.0	179	179	0.82	0.00
2	25,000	22,880	26,869	64.1	1.9	179	23	0.71	1.00
3	50,000	26,160	29,280	75.2	41.1	358	31	1.07	7.26
4	75,000	26,800	30,677	77.0	53.2	283	59	0.71	4.76
5	100,000	27,920	31,801	79.5	67.3	769	118	1.42	6.45
6	150,000	29,040	33,554	81.5	76.1	1152	89	1.72	5.03
7	200,000	30,480	34,801	83.7	80.9	1481	343	2.20	6.47
8	250,000	31,920	35,662	85.7	84.4	1559	392	2.26	4.23

Table 4.3
Cost and service level tradeoffs – Project O.

DP	Penalty rate	Average				Standard deviation			
		Labor cost	Expected outcome	Average TSF (%)	Confidence (%)	Labor cost	Expected outcome	Average TSF (%)	Confidence (%)
1	0	8240	8240	54.2	0.0	219	219	1.49	0.00
2	25,000	10,800	11,705	76.8	27.2	0	37	0.17	1.52
3	50,000	11,360	12,294	79.9	62.0	219	37	0.97	11.80
4	75,000	11,600	12,736	80.6	71.6	0	58	0.33	3.72
5	100,000	11,600	13,022	80.9	74.2	0	46	0.21	1.89
6	150,000	12,000	13,595	82.5	86.2	0	21	0.17	2.49
7	200,000	12,000	14,127	82.4	86.0	0	112	0.36	3.40
8	250,000	12,320	14,591	83.1	89.3	179	72	0.71	2.30

The managerial implications here are important. When making day-to-day staffing decisions managers must consider how much risk of missing the service level target they are willing to tolerate. Conversely, they also decide how much insurance to buy in the form of excess capacity. In most situations, managers must make these decision based on intuition. Our model operationalizes this decision by assigning a financial penalty to the possibility of failing to meet the service level target.

5. The impact of variability and VSS

5.1. Overview

The solution of the mean value program generates a biased estimate of the true cost of implementing the proposed solution. Solving a stochastic program reduces that bias, and the bias declines with the number of scenarios, going to zero as the number of scenarios goes to infinity (Mak et al., 1999). The expected cost of implementing the stochastic solution is lower than the cost of implementing the mean value solution, or stated differently we can lower the expected cost of operating the system by explicitly considering variability in our optimization problem. This reduction in cost is known as the Value of the Stochastic Solution (VSS). It is easily shown that VSS is a non-negative quantity (Birge, 1982; Birge and Louveaux, 1997). (Fig. 10-11 in the supplementary material graphically depicts the relationship of the various costs.) We calculate the VSS to determine if there is benefit from solving the stochastic version of the problem. In this section, we calculate the VSS to demonstrate that mean value solutions are optimistically biased, but unlikely to achieve the desired service level, for each of our three test projects.

5.2. VSS and solution convergence

In this section, we estimate the bias and the VSS for the same three test projects previously analyzed for various scenario levels. At each scenario level, we generate five independent batches and solve the program once for each batch. The expected outcome is found by evaluating that solution against 500 evaluation scenarios. Table 5.1 summarizes the results.

In each case we find substantial bias in the Mean Value Solution and find substantial value from implementing the stochastic solution. On the moderately variable project J, the stochastic program reduces expected cost by 13%. On the more variable projects S and O, the stochastic solution reduces cost by over 20%. Also note that the stochastic solution provides a higher confidence that the performance target will be achieved.

For each project listed in Table 5.1 the stochastic program lowers overall expected cost by increasing direct labor. It is somewhat paradoxical that stochastic programs provide better results by calculating worse objective functions. The intuition is, however, straightforward; *deterministic optimization programs assume away uncertainty and therefore do not adequately hedge for variability*; incremental staffing is added in periods with relatively high volumes and high variability.

In Section 3, we showed that the average solution to the stochastic program provides a point estimate on the lower bound of the true optimal solution, while the average expected outcome of the candidate solution forms a point estimate of the upper bound of the true optimal. (Fig. 10-12 in the supplementary material plots the point estimate of the upper and lower solution bounds. Fig. 10-13 plots the 90% confidence interval on the magnitude of the optimality gap.) These graphs show that the mean value problem exhibits significant bias, but that even with a moderate number of scenarios, and a few batches, we are able to generate fairly tight bounds on the true optimal value. The data suggests that solving the problem with as few as 25 scenarios provides reasonably good results, while a 50 or 100 scenario model gives us a tighter bound, which may be useful when trying to make detailed comparisons between alternatives.

6. Comparative analysis

6.1. Introduction

Throughout this paper, we have analyzed a model that includes abandonment and arrival rate uncertainty. Neither of these conditions is included in many industry standard models. As noted in Gans et al. (2003), "common practice uses the M/M/N (Erlang C) queuing model to

Table 5.1
Solution bias and VSS.

Project	Scenarios	Direct cost	Calculated optimum	Expected outcome	Solution bias	VSS	VSS (%)	Confidence level (%)
Project J	MV	10,020	10,081	12,838	2758			1.6
	10	10,824	10,959	11,253	295	1585	12.3	63.5
	25	10,848	11,044	11,146	121	1693	13.2	70.6
	50	10,868	11,044	11,108	64	1730	13.5	74.4
	100	10,884	11,075	11,092	36	1747	13.6	76.8
Project S	MV	23,200	23,240	34,860	11,620			14.0
	10	25,400	25,710	28,663	2953	6197	17.8	56.2
	25	26,720	27,376	27,540	193	7320	21.0	84.6
	50	26,440	27,280	27,496	303	7364	21.1	81.2
	100	26,260	27,069	27,337	304	7523	21.6	81.5
Project O	MV	8820	8820	13,855	5035			69.9
	10	10,488	10,717	11,079	361	2776	20.0	80.2
	25	10,500	10,844	11,009	199	2846	20.5	80.5
	50	10,388	10,872	10,993	125	2862	20.7	80.1
	100	10,520	10,879	10,956	77	2899	20.9	80.8

estimate the stationary system performance of short – half hour or hour – interval" p. 92. Fukunaga et al. (2002) describe a commercial system deployed at over 800 call centers in which "agent requirements are computed by applying the well-known Erlang-C formula." Furthermore, standard industry practice is to make staffing decisions based on a period-by-period (local) service level requirement, "each half hour interval's forecasted λ_i and μ_i give rise to a target staffing level for the period. ... determination of an optimal set of schedules can then be described as the solution to an integer program" (Gans et al., 2003), p. 93. In Section 5.2, we showed that ignoring arrival rate uncertainty leads to verifiably more expensive solutions, on an expected cost basis, than models that account for variability. In this section, we compare the stochastic Erlang A model to the commonly applied mean value arrival rate Erlang C model.

The standard approach described above generates a set of fixed staffing requirements in each period, and then attempts to find the lowest cost schedule to satisfy these requirements. The resulting integer program is a standard weighted set covering problem, which can be expressed as

$$\begin{aligned} \min \quad & \sum_{j\in J} c_j x_j \\ \text{subject to} \quad & \sum_{j\in J} a_{ij} x_j \geqslant b_i, \quad \forall i \in I, \\ & x_{ij} \in \mathbb{Z}^+, \end{aligned}$$

where c_j is the cost of the schedule j, x_j is the number of resources assigned to the jth schedule, and a_{ij} is the mapping of schedules to time periods.

6.2. Locally constrained Erlang C model

We refer to the standard approach described in Gans et al. (2003) as the locally constrained Erlang C model because it uses Erlang C to generate a hard constraint in each period. The general problem with this approach is the constraint created by the per-period service level requirement, coupled with the requirement to schedule resources in shifts. The peak staffing level is set by the peak arrival period and, depending on the length of the arrival peak and the length of the flexibility of the staffing model, a substantial amount of excess capacity may be created in other periods due to shift constraints. The magnitude of the excess capacity will be a factor of the flexibility of the available set of schedules. With more flexible staffing options, the weighted set covering algorithm can match the requirement more closely.

To quantify the impact, we run a locally constrained Erlang C model for each of the three test projects for each of the five schedule sets. The per-period constraints are set so that the service level with expected volumes is at least 80% in every 30 minute period, thus ensuring the global SLA of 80% is met. In Table 6.1, we compare the results of this analysis with the results generated from solving the stochastic program. We solve each test project for each of the five levels of staffing flexibility defined in Table A.2.

The data confirms that the excess staffing is high for 5×8 staffing, but decreases quickly with more flexible scheduling options. It also shows that this is a more significant problem for project J, which has a strong seasonality pattern, than for either Project S or O. The set covering approach tends to overstaff the project and achieves expected service levels higher than those achieved in the stochastic model. However, because the set covering model considers only the expected value and not the variance of arrivals, it is less effective at hedging than the stochastic model. Consider the case of schedule D for project S. The deterministic model has an expected service level of 86.1%, versus the goal of 80%, but still has an expected penalty cost of $4820. The stochastic model, on the other hand, has an expected service level of 83.5%, 2.6% lower, but an expected penalty only slightly higher at $4493.

In all cases, the stochastic model yields a lower direct labor cost and a lower expected cost of operation. The benefit of using the stochastic model is most significant when arrivals have a strong seasonal pattern, as in Project J, or when workforce flexibility is low. With 5×8 only staffing, the stochastic model provides at least 10.8% reduction in operating costs.

6.3. Globally constrained Erlang C model

In the previous section, we showed that the stochastic model based on the Erlang A model provides lower cost solutions than the locally constrained Erlang C model discussed in the literature. An alternative approach is to use a deterministic Erlang C model, ignoring abandonment and uncertainty as in the previous model, but optimizing to global versus local constraints. While this approach is not presented in the literature as far as we know, it is a natural simplification of the stochastic model we have analyzed so far. Because the model is deterministic, it assumes arrival rates are known, and, it will, in general, be easier to solve than the stochastic model. Ignoring abandonment will tend to increase recommended staffing, but ignoring uncertainty will tend to decrease staffing. It may be the case that under some circumstances these errors will cancel each other out, and we can achieve good solutions at a lower computational cost.

The method for formulating and solving these problems is a straightforward implementation of the model (3.1)–(3.7). We solve a mean value version of the problem. The major change is that the coefficients for constraints (3.3) and (3.5) are calculated based on the Erlang C model. We still require a minimum of two agents staffed at all times and a minimum service level at expected volume in every period of at least 50%.

We solve this version of the problem for each of the three projects and for each scheduling option. Since the model is deterministic, there is no need to solve multiple batches. To evaluate the expected cost of implementing the solution, we continue to evaluate the resulting schedule against the stochastic Erlang A model. We assume that the Erlang A model with uncertain arrivals is the correct model and the objective of this analysis is to determine the error introduced by using a Globally Constrained Erlang C model. The results of this analysis are shown in Table 6.2.

This analysis leads to several interesting insights. First, the stochastic model outperforms the global Erlang C model in all cases; in some cases this improvement is large and in others it is small. Given that both models are scheduling to a global objective, the difference is due to a better hedging strategy. Sometimes the stochastic model schedules fewer hours, other times more.

The second insight is that the Mean Value Globally Constrained Erlang C (GCEC) model does much better than the Mean Value Globally Constrained Erlang A (GCEA) model, even under the assumption that the Erlang A model is correct. The GCEC model makes two simplifying

Table 6.1
Comparing the stochastic and local Erlang C schedules.

	Locally constrained Erlang C						SCCS – Erlang A							
	Direct labor	Expected penalty	Expected outcome	Average TSF (%)	Excess cap.	Excess cap. (%)	Direct labor	Expected penalty	Expected outcome	Average TSF (%)	Direct labor savings		Expected savings	
Project J														
Sched A	16,000	0	16,000	91.8	4055	34	11,280	380	11,660	81.1	4720	29.5%	4340	27.1%
Sched B	13,200	0	13,200	91.0	1255	11	10,800	439	11,239	80.4	2400	18.2%	1961	14.9%
Sched C	12,880	0	12880	90.4	935	8	10,944	291	11,235	81.3	1936	15.0%	1645	12.8%
Sched D	12,500	0	12500	89.5	555	5	10,844	259	11,103	81.5	1656	13.2%	1397	11.2%
Sched E	12,300	0	12300	89.2	355	3	10,720	299	11,019	81.3	1580	12.8%	1281	10.4%
Project S														
Sched A	38,000	1565	39,565	91.6	8340	28	30,960	4345	35,305	83.2	7040	18.5%	4260	10.8%
Sched B	32,800	3847	36,647	88.0	3140	11	30,320	4408	34,728	83.7	2480	7.6%	1919	5.2%
Sched C	32,320	4184	36,504	87.4	2660	9	30,384	4349	34,733	83.6	1936	6.0%	1772	4.9%
Sched D	30,900	4820	35,720	86.1	1240	4	30,092	4493	34,585	83.5	808	2.6%	1135	3.2%
Sched E	30,980	4796	35,776	86.2	1320	4	30,096	4499	34,595	83.5	884	2.9%	1181	3.3%
Project O														
Sched A	13,600	384	13,984	85.7	2180	19	11,600	843	12,443	80.2	2000	14.7%	1542	11.0%
Sched B	12,400	514	12,914	83.4	980	9	11,360	897	12,257	80.1	1040	8.4%	656	5.1%
Sched C	12,160	544	12,704	83.0	740	6	11,296	982	12,278	79.5	864	7.1%	426	3.4%
Sched D	11,980	592	12,572	82.4	560	5	11,352	858	12,210	80.2	628	5.2%	362	2.9%
Sched E	11,880	624	12,504	82.1	460	4	11,316	910	12,226	79.9	564	4.7%	278	2.2%

Table 6.2
Comparing the stochastic and global Erlang C schedules.

	Globally constrained Erlang C				SCCS – Erlang A							
	Direct labor	Expected penalty	Expected outcome	Average TSF (%)	Direct labor	Expected penalty	Expected outcome	Average TSF (%)	Direct labor savings		Expected savings	
Project J												
Sched A	14,000	20	14,020	88.6	11,280	380	11,660	81.1	2720	19.4%	2360	16.8%
Sched B	12,000	2	12,002	87.1	10,800	439	11,239	80.4	1200	10.0%	763	6.4%
Sched C	11,760	5	11,765	86.3	10,944	291	11,235	81.3	816	6.9%	530	4.5%
Sched D	11,600	7	11,607	86.3	10,844	259	11,103	81.5	756	6.5%	504	4.3%
Sched E	11,580	26	11,606	85.8	10,720	299	11,019	81.3	860	7.4%	587	5.1%
Project S												
Sched A	35,200	953	36,153	87.3	30,960	4345	35,305	83.2	4240	12.0%	848	2.3%
Sched B	30,400	5412	35,812	84.8	30,320	4408	34,728	83.7	80	0.3%	1084	3.0%
Sched C	30,160	5426	35,586	84.7	30,384	4349	34,733	83.6	−224	−0.7%	854	2.4%
Sched D	29,340	6080	35,420	83.6	30,092	4493	34,585	83.5	−752	−2.6%	835	2.4%
Sched E	29,320	6050	35,370	83.7	30,096	4499	34,595	83.5	−776	−2.6%	775	2.2%
Project O												
Sched A	11,600	976	12,576	79.9	11,600	843	12,443	80.2	0	0.0%	133	1.1%
Sched B	11,200	1305	12,505	78.5	11,360	897	12,257	80.1	−160	−1.4%	247	2.0%
Sched C	11,120	1394	12,514	78.3	11,296	982	12,278	79.5	−176	−1.6%	236	1.9%
Sched D	10,960	1442	12,402	78.0	11,352	858	12,210	80.2	−392	−3.6%	192	1.5%
Sched E	11,080	1421	12,501	78.1	11,316	910	12,226	79.9	−236	−2.1%	276	2.2%

assumptions. First, it assumes away abandonment, which causes the model to be overstaffed. The model also assumes away arrival rate uncertainty, which leads to understaffing. These two effects tend to counterbalance each other, implying it may not be wise to introduce abandonment unless arrival rate uncertainty is also considered.

7. Conclusions and future research

In this paper, we examined the issue of short term shift scheduling for call centers for which it is important to meet a service level commitment over an extended horizon. While the analysis focused exclusively on a TSF based SLA, the model could easily be adapted to support other forms of an SLA; such as abandonment rate or average speed to answer. The model was designed to recognize the uncertainty in arrival rates and was formulated as a mixed-integer two-stage stochastic program. Although difficult to solve, we showed the model is tractable and can be solved in a reasonable amount of time. We also showed that uncertainty is highly relevant in call centers, and that it has a real impact on scheduling decisions.

In Section 5.2, we showed the Value of the Stochastic Solution for this model is substantial; ranging from 12.3% to over 21%. The clear implication is that, for this model formulation, ignoring variability is a costly decision. However, most models in practice ignore both uncertainty and abandonment. The implication is that one should not introduce abandonment into the model without also considering uncertainty. In Section 6.2, we compared this model with the common practice of scheduling to a local Erlang C constraint; that is, scheduling based on a model that ignores abandonment and uncertainty but requires the service level target is achieved in every period. Comparing our model to this common practice, we again found our model achieves lower cost results, ranging from 2.4% to 27%. The basic

implication here is that the Erlang C model sometime achieves good results, likely because the abandonment and uncertainty assumptions create counter balancing errors. However, the stochastic model always achieves a better solution, and in many practical cases the results are substantially better. This is particularly true when the flexibility of the workforce is limited to full- or near-full-time shifts and the set covering approach introduces considerable slack in the schedule.

Finally, we compared this model to a Globally Constrained Erlang C model. This model gives superior results as compared to the local constrained Erlang C, but again our stochastic model outperforms this model in every case, by as little as 1% but by as much as 16%. The overall conclusion is that, compared to the alternative methods analyzed here, the stochastic model gives a lower cost of operation schedule, and sometimes this difference can be substantial. This is a basic property of stochastic programming in general, but in this analysis we have shown that the difference is significant in real world cases.

In addition to providing a lower cost solution, the model presented in this paper addresses the scheduling problem from a fundamentally different perspective. In the standard set covering approach, the service level constraint is a hard constraint, it must be satisfied and any candidate schedule either achieves the service level requirement or does not. But, in reality, the service level is a random variable and we will achieve the SLA target with some probability. Our analysis examines this explicitly and addresses the tradeoffs that managers must make in terms of cost and the confidence of achieving the service level. Our analysis shows that the cost of operation increases nonlinearly with the desired confidence level. This tradeoff is obscured in the deterministic setting.

In future research, this model can be easily extended to use different queuing assumptions, for example, that relax the requirement for exponential service times. The tradeoff of solution precision and computational effort is also an area for future research, examining the impact of changing the convergence parameters discussed in Section 3.1. We will also investigate the implicit scheduling of breaks.

Appendix A. Algorithms and shift patterns

See Figs. A.1–A.4 and Tables A.1 and A.2.

1. Generate a call volume for each day of the week using the mean and standard deviation specified for the day.
2. For each time period in each day generate a random proportion of call volume based on the specified mean and standard deviation for the time period.
3. Normalize the time period proportions so that they sum to 1 for each day.
4. Calculate the per-period call volume by multiplying the daily total by the time period proportion.

Fig. A.1. Simulated call generation algorithm.

1. Generate a week of call volume using the algorithm shown in Fig. A.1 and calculate the associated per-period arrival rate.
2. For a given call volume, select $h + 1$ probability levels for estimating points on the TSF curve. (In practice we use values of .3, .72, .9, .98, and .995 for all periods with call volumes of at least 5. Different values are used for lower call volumes to maintain a concave approximation.)
3. Calculate the staff level required to achieve the target probabilities defined in Step 2.
4. Recalculate the TSF for the integral staffing level calculated in Step 3. We now have $h + 1$ staff level probability pairs on the TSF curve.
5. Calculate the slope (m_{ikh}) and intercept (b_{ikh}) for each pair of adjacent points found in Step 5.
6. Generate a scenario that includes the per-period call volumes (n_{ik}) and h pairs of slope and intercept parameters for each period in the planning horizon.

Fig. A.2. Scenario based TSF approximation approach.

1. Define w, the worst-case acceptable expected service level, and n_{min}, the overall minimum number of agents to be staffed at any time.
2. Determine the expected call arrival rate.
3. Calculate the staff level, n_w, required to achieve the worst-case expected service level defined in Step 1.
4. Calculate $\mu_i = \lceil \min(n_w, n_{\min}) \rceil$, the minimum agents to staff in period i.
5. Write out μ_i in a GAMS compatible format.
6. Repeat Steps 2–6 for each period i.

Fig. A.3. Minimum staff level constraint generation.

1. Calculate the average volume in each 30 minute period of the week.
2. Using the volumes calculated in Step 1, determine the number of agents required to achieve the target service level in each 30 minute period by performing a search.
3. Set the period staffing requirement to the maximum of the number calculated in Step 2 and the global minimal staffing requirement.
4. Use the resulting vector of staffing requirements as the requirement parameter b_i in the IP constraint (3.2).

Fig. A.4. Local constraint generation.

Table A.1
Shift patterns.

Pattern	Description
5×8	5 days a week, 8 hours a day (40 hours week)
4×10	4 days a week, 10 hours a day (40 hours week)
4×8	4 days a week, 8 hours a day (32 hours week)
5×6	5 days a week, 6 hours a day (30 hours week)
5×4	5 days a week, 4 hours a day (20 hours week)

Table A.2
Scheduling patterns.

Pattern	Schedule types included	Feasible schedules
A	5×8 only	336
B	5×8, 4×10	1680
C	5×8, 4×10, 4×8	3024
D	5×8, 4×10, 4×8, 5×6	3360
E	5×8, 4×10, 4×8, 5×6, 5×4	3696

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2010.06.013.

References

Aksin, Z., Armony, M., Mehrotra, V., 2007. The modern call-center: A multi-disciplinary perspective on operations management research. Production and Operations Management 16 (6), 665–668.
Andrews, B.H., Cunningham, S.M., 1995. L.L. Bean improves call-center forecasting. Interfaces 25 (6), 1–13.
Atlason, J., Epelman, M.A., Henderson, S.G., 2004. Call center staffing with simulation and cutting plane methods. Annals of Operations Research 127, 333–358.
Atlason, J., Epelman, M.A., Henderson, S.G., 2008. Optimizing call center staffing using simulation and analytic center cutting-plane methods. Management Science 54 (2), 295–309.
Avramidis, A.N., Deslauriers, A., L'Ecuyer, P., 2004. Modeling daily arrivals to a telephone call center. Management Science 50 (7), 896–908.
Avramidis, A.N., Chan, W., L'Ecuyer, P., 2007a. Staffing Multi-skill Call Centers via Search Methods and a Performance Approximation. University of Montreal.
Avramidis, A.N., Gendreau, M., L'Ecuyer, P., Pisacane, O., 2007b. Simulation-based optimization of agent scheduling in multiskill call centers. In: Fifth Annual International Industrial Simulation Conference (ISC-2007), Delft, The Netherlands.
Aykin, T., 1996. Optimal shift scheduling with multiple break windows. Management Science 42 (4), 591–602.
Baron, O., Milner, J.M., 2006. Staffing to maximize profit for call centers with alternate service level agreements. Operations Research 57 (3), 685–700.
Bassamboo, A., Harrison, J.M., Zeevi, A., 2005. Design and control of a large call center: Asymptotic analysis of an LP-based method. Operations Research 54 (3), 419–435.
Bayraksan, G., Morton, D.P., 2009. Assessing solution quality in stochastic programs via sampling. In: Informs 2009 Tutorials in Operations Research, pp. 102–122.
Bechtold, S.E., Jacobs, L.W., 1990. Implicit modeling of flexible break assignments in optimal shift scheduling. Management Science 36 (11), 1339–1351.
Birge, J.R., 1982. The value of the stochastic solution in stochastic linear programs, with fixed recourse. Mathematical Programming 24, 314–325.
Birge, J.R., Louveaux, F., 1997. Introduction to Stochastic Programming. Springer, New York.
Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Haipeng, S., Zeltyn, S., Zhao, L., 2005. Statistical analysis of a telephone call center: A queueing-science perspective. Journal of the American Statistical Association 100 (469), 36–50.
Brusco, M.J., Jacobs, L.W., 1998. Personnel tour scheduling when starting-time restrictions are present. Management Science 44 (4), 534–547.
Brusco, M.J., Jacobs, L.W., 2000. Optimal models for meal-break and start-time flexibility in continuous tour scheduling. Management Science 46 (12), 1630–1641.
Brusco, M.J., Johns, T.R., 1996. A sequential integer programming method for discontinuous labor tour scheduling. European Journal of Operational Research 95 (3), 537–548.
Cezik, M., L'Ecuyer, P., 2007. Staffing multiskill call centers via linear programming and simulation. Management Science 54 (2), 310–323.
Chen, B.P.K., Henderson, S.G., 2001. Two issues in setting call centre staffing levels. Annals of Operations Research 108 (1), 175–192.
Dantzig, G.B., 1954. A comment on Edie's "Traffic delays at toll booths". Journal of the Operations Research Society of America 2 (3), 339–341.
Fukunaga, A., Hamilton, E., Fama, J., Andre, D., Matan, O., Nourbakhsh, I., 2002. Staff scheduling for inbound call centers and customer contact centers. In: Eighteenth National Conference on Artificial Intelligence, Edmonton, Alberta, Canada.
Gans, N., Koole, G., Mandelbaum, A., 2003. Telephone call centers: Tutorial, review, and research prospects. Manufacturing and Service Operations Management 5 (2), 79–141.
Garey, M.R., Johnson, D.S., 1979. Computers and Intractability: A Guide to the Theory of NP-completeness. W.H. Freeman, San Francisco.
Geoffrion, A.M., 1970. Elements of large-scale mathematical programming: Part I: Concepts. Management Science 16 (11), 652–675 (Theory Series).
Green, L.V., Kolesar, P.J., Soares, J., 2001. Improving the SIPP approach for staffing service systems that have cyclic demands. Operations Research 49 (4), 549–564.
Harrison, J.M., Zeevi, A., 2005. A method for staffing large call centers based on stochastic fluid models. Manufacturing and Service Operations Management 7 (1), 20–36.
Henderson, W.B., Berry, W.L., 1976. Heuristic methods for telephone operator shift scheduling: An experimental analysis. Management Science 22 (12), 1372–1380.
Koole, G., van der Sluis, E., 2003. Optimal shift scheduling with a global service level constraint. IIE Transactions 35, 1049–1055.
Mak, W.-K., Morton, D.P., Wood, R.K., 1999. Monte Carlo bounding techniques for determining solution quality in stochastic programs. Operations Research Letters 24 (1–2), 47–56.
Milner, J.M., Olsen, T.L., 2008. Service-level agreements in call centers: Perils and prescriptions. Management Science 54 (2), 238–252.
Pinedo, M., 2005. Planning and Scheduling in Manufacturing and Services. Springer, New York, NY.
Robbins, T.R., 2007. Managing Service Capacity Under Uncertainty. Unpublished PhD Dissertation, Pennsylvania State University, 240p. <http://personal.ecu.edu/robbinst/> (accessed 01.04.10).
Robbins, T.R., Medeiros, D.J., Dum, P., 2006. Evaluating arrival rate uncertainty in call centers. In: Proceedings of the 2006 Winter Simulation Conference, Monterey, CA.
Segal, M., 1974. The operator-scheduling problem: A network-flow approach. Operations Research 22 (4), 808–823.
Steckley, S.G., Henderson, W.B., Mehrotra, V., 2004. Service System Planning in the Presence of a Random Arrival Rate. Cornell University.
Steckley, S.G., Henderson, S.G., Mehrotra, V., 2009. Forecast errors in service systems. Probability in the Engineering and Informational Sciences (23), 305–332.
Thompson, G.M., 1995. Improved implicit optimal modeling of the labor shift scheduling problem. Management Science 41 (4), 595–607.
Weinberg, J., Brown, L., Stroud, J.R., 2007. Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. Journal of the American Statistical Association 102 (480), 1185–1198.
Whitt, W., 2006. Staffing a call center with uncertain arrival rate and absenteeism. Production and Operations Management 15 (1), 88–102.