Stochastic Call Center Staffing with Uncertain Arrival, Service and Abandonment Rates: A Bayesian Perspective

Tevfik Aktekin
Department of Decision Sciences
Peter T. Paul College of Business and Economics, University of New Hampshire

Tahir Ekin
Department of Computer Information Systems and Quantitative Methods
McCoy College of Business, Texas State University

October 13, 2016

Abstract

In this paper, we introduce staffing strategies for the Erlang-A queuing system in call center operations with uncertain arrival, service and abandonment rates. In doing so, we model the system rates using gamma distributions that create randomness in operating characteristics used in the optimization formulation. We divide the day into discrete time intervals where a simulation based stochastic programming method is used to determine staffing levels. More specifically, we develop a model to select the optimal number of agents required for a given time interval by minimizing an expected cost function, which consists of agent and abandonment (opportunity) costs, while considering the service quality requirements such as the delay probability. The objective function as well as the constraints in our formulation are random variables. The novelty of our approach is to introduce a solution method for the staffing of an operation where all three system rates (arrival, service and abandonment) are random variables. We illustrate the use of the proposed model using both real and simulated call center data. In addition, we provide solution comparisons across different formulations, consider a dynamic extension, and discuss sensitivity implications of changing constraint upper bounds as well as prior hyper-parameters.

1 Introduction

In this paper, we introduce a novel staffing method for call center operations with uncertain arrival, service and abandonment rates using stochastic programming, more specifically an augmented probability simulation (APS) based optimization method. Our optimization formulation differs from the rest of the literature by its treatment of both the objective function and the constraint as random variables due to the uncertainty in all system rates. These models are also referred to as doubly stochastic processes. Finding the optimal solution to such a formulation is not straightforward and to the best of our knowledge had not been considered in the call center staffing literature previously.

A call center operation can simply be described as a system where customers receive service through agents (servers) over the phone. One of the major challenges that call center management faces is to find the correct balance between system efficiency and customer satisfaction. For instance, a call center with a small number of agents might be inexpensive to operate but might lead to high levels of customer dissatisfaction and low customer retention rates. These issues have made call centers a fertile area for academic research in service engineering and operations research; see Gans et al. [23] and Aksin et al. [1] for detailed reviews.

In general, call center operations can be classified into two categories; inbound and outbound. Inbound call centers receive calls from customers (calling a bank, customer service for help, etc.) whereas in outbound call centers, agents reach out to customers (collection companies and telemarketers that are promoting new products and features). The focus of this study will be on operations that deal with inbound call centers. Several queuing models have been used in call center research including the M(t)/M/s(t)+M (Erlang-A with time varying arrivals and agents) with exponentially distributed non-homogeneous arrival, service and abandonment times. This model assumes that arrivals occur according to a time varying arrival rate of $\lambda(t)$, a constant service rate of μ and a constant abandonment rate of θ. To make use of such queuing models, one needs to estimate $\lambda(t), \mu$ and θ using appropriate models and historical data based on which point estimates are obtained. However, as pointed out by Brown et al. [14], research from a statistical inference perspective is scarce in the call center literature. Mandelbaum and Zeltyn [38] also raise the need for models which can capture parameter uncertainty. Since typically the model rates are assumed to be fixed, their uncertainty is rarely attempted to be captured with the exception of that of the arrival rate for which there is a broad literature in forecasting; see Weinberg et al. [53], Shen and Huang [51], Landon et al. [36], Aktekin and Soyer [3]. The only known study in the literature where all system

rates are random in a queuing setting is by Aktekin and Soyer [4].

As pointed out by Aksin et al. [1], roughly sixty to eighty percent of a call center operation consists of labor costs. Thus, call center staffing is a rich area in optimization and queuing research. In a call center setting, staffing can be either long term (total number of people to hire for the whole month) or short term (number of agents required to be available to answer calls during finite periods of time on a given day). In the remainder of the paper, we focus on the latter. One of the commonly used methods is the so called square-root staffing rule which is straightforward to implement (Whitt [55]). However, there are several studies that consider different aspects of staffing in call center operations. Atlason et al. [7] utilize an analytic cutting plane method for staffing and use simulation to determine service quality without using any of the traditional queuing results. Gurvich et al. [29] study large call centers with multiple customer classes and consider the issue of control in staffing. Feldman et al. [21] propose methods of staffing for call center queues with time-varying arrival rates using an iterative simulation based method. Mandelbaum and Zeltyn [39] consider the staffing of many server queues with impatient customers using a constraint satisfaction approach. Gurvich et al. [30] use a fluid approximation to solve a chance constrained formulation, which, in turn, provides staffing and routing policies for multi-class multi-type call centers with arrival rate uncertainty.

Another important aspect of call center operations is the modeling of random abandonment times. Since call center customers can abandon virtual queues that can affect the quality of operations, a significant amount of research has been dedicated to this issue. Abandonment is usually captured via time to event type models from the literature of survival and reliability analyses. These models assume that customers have a random patience level which makes them abandon the virtual queue when they exceed a certain threshold. Palm [46] is the first known study where abandonment is modeled as a random quantity. Bacelli and Hebuterne [9] bring the idea of a random abandonment distribution into the queuing literature and call it the Erlang-A, the well known M/M/s+M model. Mandelbaum and Shimkin [37] use utility functions to determine the abandonment times for customers and propose strategies from a decision theoretic point of view. Zohar et al. [57] discuss the adaptive behavior of impatient customers in call center settings and provide empirical support in its favor. Furthermore, Garnett et al. [24] argue and investigate how staffing decisions are affected if abandonment is ignored. Jouini et al. [33] propose and study models for anticipated delays within call centers considering abandonment. Aksin et al. [2] model the decision process of customers as an optimal stopping problem in a call center setting with impatient

customers.

Our approach differs from the current call center literature mainly in two ways. First is its treatment of the uncertainty in $\lambda(t), \mu$ and θ which creates doubly stochastic arrival, service and abandonment processes. In traditional queuing models, these input rates are assumed to be fixed. In our setup, the uncertainty in the rates are modeled with independent gamma random variables. The modeling approach was first considered in Aktekin and Soyer [4] for the Bayesian inference of the M/M/s+M queues. Second, we propose the use of an augmented probability simulation technique to solve staffing decision problems with randomness in both the objective function and constraints that none of the previous work in the call center literature has been able to solve. In our setup, we consider minimizing an expected cost function by taking advantage of the trade-off between the cost of employment and the cost of abandonment (a lost opportunity cost). In the calculation of the total cost, there are two sources of randomness, the arrival rate and the probability of abandonment. In addition, our formulation also includes a threshold on the delay probability. Such constraints can be set by the call center management or could be based on agreements between the company and its customers as in Gurvich et al. [29]. The use of the total cost with abandonment as an objective function in staffing has been proposed by Mandelbaum and Zeltyn [38] and the use of delay probabilities as constraints has been considered in Gurvich et al. [29] and as targets to achieve in Feldman et al. [21]. However, none of these studies consider randomness in both the objective function and the constraints as well as in all the system rates. Our approach can best be described as a pointwise stationary approximation (PSA) of the Erlang-A model with time varying arrival rates (i.e. M(t)/M/s(t)+M). The PSA idea considered by Green and Kolesar [28] is based on splitting a given time horizon into smaller time periods where each short interval is treated to be in steady state. The PSA approach is also considered in Bassamboo et al. [10] in fluid models for stochastic networks in the context of call center staffing and routing. As pointed out by Garnett et al. [24], PSA is a common and reasonable approach in the analysis of the Erlang-A model with time varying arrival rates as the system will always be stationary due to its birth and death properties (Mandelbaum and Zeltyn [38]). The use of the Erlang-A model considered in this study is justified by the sufficiency of the exponential time to abandonment distribution for practical applications in call centers (Brown et al. [14]). In addition, we also note here that our proposed solution method can be applied to any form in the objective function other than the ones considered here. We discuss an extension of our main approach using various objective functions in our numerical study.

Even though our study borrows the inference strategies from Aktekin and Soyer [4], the two studies differ in several significant ways. The main goal in Aktekin and Soyer [4] is to model system rates as random variables and develop inference for the overall queuing system with implications on operating characteristics. Our goal is to solely develop a staffing strategy using a stochastic optimization model formulation where both the objective function and the constraints are random which is a direct effect of all the system rates being random variables. To the best of our knowledge, such an approach had not been considered in the call center staffing literature previously where most previous work focus on fixed system rates or treat only the arrival rates as random variables. In doing so, we borrow one of the modeling strategies considered in Aktekin and Soyer [4] prior to solving our optimization problem. In addition, the staffing exercise discussed there is not based on any optimization techniques and can at best be considered to be an ad-hoc method for staffing where different values of the number of servers are plugged in the objective function sequentially until a minimum total cost is achieved. Such an approach may not be very efficient and practical for large scale call center operations. In addition, their basic staffing model is not able to consider any constraints and/or dynamic settings with coupled decisions that may be crucial for call center operations with varying operating strategies. In addition, we provide numerical comparisons between our proposed model and others and discuss the implications of using constraints in a staffing model.

The paper is organized as follows. Section 2 summarizes the operating characteristics needed for the implementation of the optimization model and Bayesian inference methods for the system rates. Section 3 describes the stochastic optimization method used to solve the proposed staffing model. Section 4 covers a real life application for a moderate size call center and a simulated example for a relatively large scale call center. We also present insights from sensitivity analysis and consider extensions including a dynamic model. Section 5 concludes with a summary of findings and directions for future work.

2 Queuing Model and Its Bayesian Inference

In this section, we summarize the details of the queuing model whose operating characteristics are used in our staffing method. We consider a call center queuing system where $s(t)$ agents are required at each time interval t (typically for discrete intervals of length 30-60 minutes), with time-varying inter-arrival times with rates $\lambda(t)$, service times with rate μ and abandonment times with rate θ

where all times are exponentially distributed. The system is referred to as the Erlang-A model with time-varying arrivals. The following is a summary of relevant results from Bacelli and Hebuterne [9], Mandelbaum and Zeltyn [38] and Aktekin and Soyer [4].

2.1 Properties of the Erlang-A Model

All subsequent results use the function $A(x,y) = \frac{xe^y}{y^x}\gamma(x,y)$ with $x, y > 0$ and $\gamma(x,y)$ represents an incomplete gamma function, which can be obtained as $\gamma(x,y) = \int_0^y t^{x-1}e^{-t}dt$. For notational convenience, we suppress the dependence on t for all measures such as $\lambda(t)$ and $s(t)$, when we show computations of operating characteristics. Unlike traditional calculations of operating characteristics, all measures are written conditional on the rates, λ, μ and θ as in $f(.|\lambda,\mu,\theta)$. We note here that the Erlang-A model is always considered to be stationary and theoretically is always in steady state unlike other queuing models without abandonment (Mandelbaum and Zeltyn [38]). The first measure of interest is the probability of having n customers in the system given the system rates, $P_n = Pr(N = n|\lambda,\mu,\theta)$, which can be computed as

$$P_n = Pr(N = n|\lambda,\mu,\theta) = \begin{cases} P_s \frac{s!}{n! r^{s-n}} & \text{for} \quad 0 \leq n \leq s, \\ P_s \frac{(\frac{\lambda}{\theta})^{n-s}}{\prod_{i=1}^{n-s}(\frac{s\mu}{\theta}+i)} & \text{for} \quad n \geq s+1, \end{cases} \tag{2.1}$$

and

$$P_s = \frac{E_{1,s}}{1 + [A(\frac{s\mu}{\theta}, \frac{\lambda}{\theta}) - 1]E_{1,s}} \text{ and } E_{1,s} = \frac{\frac{r^s}{s!}}{\sum_{i=0}^{s} \frac{r^i}{i!}}, \tag{2.2}$$

where $r = \frac{\lambda}{\mu}$. Another important operating characteristic is the probability that a customer will wait in the queue (also referred to as delay probability), typically used as a proxy for measuring the service quality of a call center operation. The delay probability, $Pr(T_q > 0|\lambda,\mu,\theta)$, whose unconditional distribution is used in our stochastic optimization model can be computed as

$$Pr(T_q > 0|\lambda,\mu,\theta) = \sum_{n=s}^{\infty} Pr(N = n|\lambda,\mu,\theta), \tag{2.3}$$

where $Pr(N = n|\lambda,\mu,\theta)$ for $n = s, \cdots, \infty$ can be obtained from (2.1). The delay probability has been used to quantify the quality of service as discussed in Garnett et al. [24] where for $Pr(T_q > 0|\lambda,\mu,\theta) \to 0$, the operational regime is called the *quality driven regime* and for $Pr(T_q > 0|\lambda,\mu,\theta) \to 1$ the *efficiency driven regime*.

Another quantity of interest, which is also used to measure the quality of service in a call center

operation, is the probability of abandonment. To obtain this, we need two measures: first is the delay probability (2.3) and second is the abandonment probability given that a customer is not served upon arrival, which is given by

$$Pr(Ab|T_q > 0, \lambda, \mu, \theta) = \frac{1}{\rho A(\frac{s\mu}{\theta}, \frac{\lambda}{\theta})} + 1 - \frac{1}{\rho}, \tag{2.4}$$

where $\rho = \frac{\lambda}{s\mu}$ and is referred to as the offered load. This operating characteristic (2.4) is also directly related to the perceived quality of the call center operation from the perspective of the customer. Finally, the probability of abandonment can be obtained as the product of (2.3) and (2.4) as

$$Pr(Ab|\lambda, \mu, \theta) = \left(\sum_{n=s}^{\infty} Pr(N = n|\lambda, \mu, \theta) \right) \left(\frac{1}{\rho A(\frac{s\mu}{\theta}, \frac{\lambda}{\theta})} + 1 - \frac{1}{\rho} \right). \tag{2.5}$$

(2.5) will be used as part of the objective function of our stochastic optimization scheme to determine optimal staffing policies. One of the challenges in call center management is to find the right balance between the service quality and the operation efficiency when determining staffing. To quantify such a trade-off, Mandelbaum and Zeltyn [38] define a measure of average operational cost (per unit of time) which we use as the objective function of our optimization model. The same objective function is also considered in Aktekin and Soyer [4]. Similar to the operating characteristics, the cost is also a function of the random system rates, λ, μ, θ. Therefore, it is a random variable itself. The conditional operational cost with s servers, $C(s|\lambda, \mu, \theta)$, can be defined as

$$C(s|\lambda, \mu, \theta) = cs + a\lambda Pr(Ab|\lambda, \mu, \theta), \tag{2.6}$$

where c is the staffing cost per unit of time and a is the abandonment cost per unit of time (opportunity cost). Since the rates are random variables, neither evaluating $C(s|\lambda, \mu, \theta)$ nor using it as an objective function in an optimization setting is straightforward. To minimize the expected value of $C(s|\lambda, \mu, \theta)$ where expectation is taken with respect to the joint distribution of λ, μ and θ, we need to evaluate the following form

$$E_{\lambda,\mu,\theta}\{C(s|\lambda, \mu, \theta)\} = \int_{\lambda} \int_{\mu} \int_{\theta} \{cs + a\lambda Pr(Ab|\lambda, \mu, \theta)\} p(\lambda, \mu, \theta) d\lambda \, d\mu \, d\theta. \tag{2.7}$$

The above formulation is one of the main differences between our proposed approach and the rest of the traditional queuing literature where the cost function will be typically computed as

$cs + a\lambda Pr(Ab)$ since the system rates are all assumed to be fixed inputs. These rates are all random variables in our proposed methodology, and we are able to quantify the uncertainty in these system rates which is a feature not shared with traditional queuing models. From a statistical inference perspective, we would need to rewrite the above joint probability distribution when we observe arrival, service and abandonment data as in $p(\lambda, \mu, \theta|Data)$. In other words, we need to have a mechanism to describe and update our uncertainty about system rates in the light of new call center data which we discuss next.

2.2 Bayesian Inference

To describe our uncertainty in the system rates, we use Bayesian inference which nicely handles hierarchical structures and right-censored abandonment data as in our setup. Our goal is to obtain the joint probability distribution of system rates given the data and our approach is similar to one of the models described in Aktekin and Soyer [4] who considered Bayesian queuing models for various scenarios. For various applications of queuing systems using the Bayesian point of view we refer the reader to McGrath et al. [41], McGrath and Singpurwalla [40], Armero and Bayarri [5], Armero and Bayarri [6], Wiper [56], Rios Insua et al. [48], and Rios Insua et al. [49].

Let n_x be the number of inter-arrival times, n_y the number of service times, n_z the number of abandonment times and n'_z the number of waiting times for customers who did not abandon. Let $(x_1, \cdots, x_{n_x})$ be a sample from the exponential inter-arrival times distribution with rate λ, $(y_1, \cdots, y_{n_y})$ be a sample from the exponential service times distribution with rate μ, and $(z_1, \cdots, z_{n_z}, z'_1, \cdots, z'_{n'_z})$ be a collection of $n_z + n'_z$ samples from the exponential abandonment times distribution with rate θ. We also define $Data = (x_1, \cdots, x_{n_x}, y_1, \cdots, y_{n_y}, z_1, \cdots, z_{n_z}, z'_1, \cdots, z'_{n'_z})$. Thus, the likelihood function for independent and exponentially distributed arrival, service and abandonment times can be computed via

$$L(\lambda, \mu, \theta; Data) \propto \lambda^{n_x} e^{-\lambda \sum_{i=1}^{n_x} x_i} \mu^{n_y} e^{-\mu \sum_{i=1}^{n_y} y_i} \theta^{n_z} e^{-\theta \sum_{i=1}^{n_z} z_i} e^{-\theta \sum_{i=1}^{n'_z} z'_i}, \tag{2.8}$$

where the term $e^{-\theta \sum_{i=1}^{n_z} z_i} e^{-\theta \sum_{i=1}^{n'_z} z'_i}$ is due to the survival function of the exponential distribution which describes the contribution of customers who received service and did not abandon the queue (right-censored data). One of the attractive features of Bayesian methods is that once the likelihood is set, handling inference with right-censored data is immediate, especially with exponential

likelihoods. Prior to observing data, we assume gamma distributions as

$$\lambda(t) \sim G(a_0, b_0)\ \forall t,\ \mu \sim G(c_0, d_0) \text{ and } \theta \sim G(e_0, f_0), \tag{2.9}$$

where we include the dependence on t to show that each non-homogeneous arrival rate follows an independent gamma random variable a priori. The hyper-parameters, a_0, b_0, c_0, d_0, e_0 and f_0, are chosen that our prior uncertainty about the system rates is at best vague. Typically, they are all set at 0.001 to obtain uniform prior uncertainty. Given the likelihood (2.8) and the independent priors (2.9), the posterior distributions of λ, μ and θ (suppressing the dependence on t for λs for notational convenience) can be obtained as gamma densities given by

$$\lambda|Data \sim G(a_1, b_1), \mu|Data \sim G(c_1, d_1) \text{ and } \theta|Data \sim G(e_1, f_1), \tag{2.10}$$

where $a_1 = a_0 + n_x$, $b_1 = b_0 + (\sum_{i=1}^{n_x} x_i)$, $c_1 = c_0 + n_y$, $d_1 = d_0 + (\sum_{i=1}^{n_y} y_i)$, $e_1 = e_0 + n_z$, $f_1 = f_0 + (\sum_{i=1}^{n_z} z_i) + (\sum_{i=1}^{n'_z} z'_i)$. Therefore, the joint probability distribution of system rates, $p(\lambda, \mu, \theta|Data)$, would be the product of the independent gamma distributions from (2.10). Given this joint probability distribution, it is possible to compute the delay probability and the cost function. For instance, the unconditional delay probability can be computed as

$$Pr(T_q > 0) = \int_{\lambda} \int_{\mu} \int_{\theta} Pr(T_q > 0|\lambda, \mu, \theta) p(\lambda, \mu, \theta|Data) d\lambda\, d\mu\, d\theta. \tag{2.11}$$

Since (2.11) will not be available in closed form, we can use a Monte Carlo approximation as in

$$\frac{1}{G} \sum_{i=1}^{G} Pr(T_q > 0|\lambda^{(i)}, \mu^{(i)}, \theta^{(i)}), \tag{2.12}$$

where G represents the number of samples generated using Monte Carlo sampling, and $\lambda^{(i)}$, $\mu^{(i)}$, $\theta^{(i)}$ are the samples generated from the posterior distributions from (2.10). We can use these unconditional operating characteristics in our optimal staffing formulation. Next, we propose a stochastic optimization technique for a given time interval t where this delay probability (2.11) as well as the cost function (2.7) are used as a constraint and an objective function.

3 Augmented Probability Simulation for Call Center Staffing

There is a variety of stochastic programming methods used to solve decision making problems under uncertainty; see Birge and Louveaux [12] for a review. Our model is based on expectation based stochastic programs, which require the estimation of the expectation function as well as its optimization. A closed form analytically available solution cannot be obtained for our model, therefore requiring the use of a simulation based method. Simulation based stochastic programming approaches are typically based on variants of Monte Carlo sampling, see Homem-de Mello and Bayraksan [31] and Rubinstein and Kroese [50]. Models for decision making under uncertainty generally assume that uncertainty is independent of previous decisions. Consideration of dependence between decisions and random variables increases computational complexity as pointed out by Dyer and Stougie [19], mainly because of the larger size of the state space; see Goel and Grossmann [27] for a review.

The complex functional relationship and dependence between the probability of abandonment and the number of servers prevents our model to be solved by crude Monte Carlo based methods in an efficient manner. The decision dependency structure of our model is not direct, specifically the probability distributions of the model parameters are not dependent on the decision variables. However, the probability of abandonment and probability of delay are functions of both model parameters and decision variables, thus we need to investigate the joint decision and state space to find the optimal solution. We solve our optimization problem with decision dependent uncertainty using the augmented probability simulation (APS) based optimization approach of Ekin et al. [20]. This approach estimates and optimizes for the objective function by directly using a function of decisions and random values, without an exact evaluation of the objective function. It allows us to deal with conditional dependence between the state and the decision spaces. In addition, we can utilize the posterior distributions from Bayesian inference within a Markov chain Monte Carlo (MCMC) simulation setting. Bayesian frameworks are also considered in Morton and Popova [43] and Kirschenmann et al. [35] for sampling and estimation purposes prior to the solution of the optimization problem. Different from these approaches, APS involves transformation of the optimization problem into a grand simulation by treating the decision variable as random for computational purposes. In what follows, we provide details about the APS method and show its implementation for the proposed call center staffing problem.

3.1 Augmented Probability Simulation (APS)

The APS idea was initially proposed by Bielza et al. [11] and further extended by Müller et al. [45] to maximize expected utilities. Recently, the method was extended to solve stochastic programming problems with constraints by Ekin et al. [20], particularly two stage stochastic programs with recourse. In this paper, we extend it to solve a one-stage stochastic program with constraints and decision dependency.

APS is based on treating the decision variable as random and constructing a state space on both decision variables and uncertainty shocks. Simulating from the joint distribution of decision and random variables solves for the expectation of the objective function and the optimal decision simultaneously. This algorithm is general and can accommodate arbitrary probability models and utility functions.

This paper particularly focuses on solving a one-stage stochastic program with constraints,

$$\begin{aligned} \max_{s} \quad & E_{\xi}[Q(s,\xi)] \\ \text{subject to} \quad & f(s) \leqslant \alpha \end{aligned} \tag{3.1}$$

by maximizing the expected value of the objective function $Q(s,\xi)$ with respect to the decision variable s and random variable ξ, while satisfying the constraint $f(s) \leqslant \alpha$ where $f(s)$ is a function of the decision variable and α is a deterministic parameter. Maximization can be achieved by simulation from an augmented probability distribution $\pi(s,\xi)$ of s and ξ. We will treat s to be random with a uniform distribution measure $U(s)$ and assume the probability distribution, $p(\xi)$ to be known. We define an auxiliary distribution

$$\pi(s,\xi) \propto Q(s,\xi)p(\xi)U(s)\mathbb{1}(f(s) \leqslant \alpha), \tag{3.2}$$

and obtain the marginal distribution of s as

$$\pi(s) \propto \int_{\xi} \pi(s,\xi)d\xi\, \mathbb{1}(f(s) \leqslant \alpha),$$

as proportional to the expected value of $Q(s,\xi)$ where $\mathbb{1}$ is the indicator function. Then the optimal decision, s^*, is given by the mode of the marginal distribution $\pi(s)$. In the above, we assume that $Q(s,\xi)$ is non-negative to obtain a proper probability density.

We simulate from the augmented probability model $\pi(s,\xi)$ using MCMC methods which can generate samples from non-standard distributions, see Gamerman and Lopes [22] for a review. Particularly, we utilize Gibbs sampling based on the idea that iterative samples from the conditional distributions result in samples from the joint distribution in the limit (Casella and George [15]). This will require simulation from the conditional distribution $\pi(\xi|s)$ which is a 'tilted' version of the distribution $p(\xi)$ as $\pi(\xi|s) \propto Q(s,\xi)p(\xi)$. In addition, we need to be able to simulate from the conditional distribution, $\pi(s|\xi) \propto Q(s,\xi)\mathbb{1}(f(s) \leqslant \alpha)$. As a consequence, the mode of the marginal samples of s provides the approximately optimal solution when the Markov chain converges.

A shortcoming of the APS arises with flat and high dimensional expected utility (objective function) surfaces where the modes may not be very distinct. We address this issue by using the "J copies" trick proposed by Müller [44]. This is based on replacing the expected utility surface by a J^{th} power transformation that uses a more peaked surface without changing the mode. This property is similar to simulated annealing (Kirkpatrick et al. [34]) that powers up a function to be maximized in order to find its optimum. Simulated annealing assumes availability of $E_\xi[Q(s,\xi)]$ for direct evaluation, whereas APS does not.

Using J independent draws of the random variable; $\boldsymbol{\xi}_J = \{\xi_1, \ldots, \xi_j, \ldots, \xi_J\}$, we modify our augmented distribution as

$$\pi_J(s, \boldsymbol{\xi}_J) \propto \prod_{j=1}^{J} \Big[Q(s,\xi_j)p(\xi_j)\mathbb{1}(f(s) \leqslant \alpha) \Big], \tag{3.3}$$

where we assume that $Q(s,\xi_j)$ is a non-negative function. We generate J copies of ξ_j for each s and ensure that $Q(s,\xi_j)$ is feasible for each of the J copies. Since the ξ_js are independent and identically distributed draws, the marginal distribution of s is given by

$$\pi_J(s) \propto E_\xi \Big[Q(s,\xi)\mathbb{1}(f(s) \leqslant \alpha) \Big]^J, \text{ where } E_\xi[Q(s,\xi)]^J = \Big[\int Q(s,\xi)p(\xi)d\xi \Big]^J.$$

One key feature is that the mode of the samples from the marginal distribution of decision variable, $\pi_J(s)$ converges to the optimum, and therefore this algorithm results with an approximately optimal decision s^*, for a sufficiently large value of J and MCMC convergence (Müller et al. [45]).

We use MCMC to sample from $\pi_J(s, \boldsymbol{\xi}_J)$. A Gibbs sampler is constructed using the conditional distributions $\pi(s|\boldsymbol{\xi}_J)$ and $\pi(\boldsymbol{\xi}_J|s)$. Samples from the joint conditional distribution $\pi(\boldsymbol{\xi}_J|s)$ can be obtained by independent draws from the conditionals $\pi(\xi_j|s)$ for $j = 1, ..., J$. More specifically, we

use

$$\pi(s|\boldsymbol{\xi}_J) \propto \prod_{j=1}^{J} Q(s,\xi_j)\mathbb{I}(f(s) \leqslant \alpha) \tag{3.4}$$

$$\pi(\xi_j|s) \propto Q(s,\xi_j)p(\xi_j); \forall j = 1,2,..,J$$

We start the chain at the initial feasible values of $(s^{(0)}, \boldsymbol{\xi}_J^{(0)})$ and sample through the conditional distributions. Therefore, we define a Markov chain $(s^{(g)}, \boldsymbol{\xi}_J^{(g)}) \xrightarrow{D} (s^{(\infty)}, \boldsymbol{\xi}_J^{(\infty)})$ as $g \to \infty$, with the pair $(s^{(\infty)}, \boldsymbol{\xi}_J^{(\infty)})$ having distribution $\pi_J(s, \boldsymbol{\xi}_J)$ where $\xrightarrow{D}$ represents convergence in distribution. Therefore, these marginal draws of s are from $\pi_J(s)$. Finally, we can use the mode of our draws to have the approximately optimal decision, s^*.

Non-standard full conditional distributions may require the use of Metropolis-Hastings algorithm (Chib and Greenberg [16]). The choice of the proposal distribution requires that we have an irreducible and aperiodic Markov chain. A standard choice is a symmetric distribution such as normal or heavy tailed t with appropriate scaling. Bielza et al. [11] provides a further discussion regarding the specification of the proposal distribution within APS algorithms.

3.1.1 Convergence Properties

This subsection presents the asymptotic properties that show the convergence of the approximately optimal decision to the optimal decision. We are interested in convergence from three perspectives. First, we describe the convergence properties of the Markov chain as a function of G, the length of the chain. We utilize MCMC methods to sample from non-standard distributions. For a fixed value of J, Pincus [47] and Tierney [52] presented the theoretical MCMC convergence results in number of iterations, G. It is a standard result that $(s, \boldsymbol{\xi}_J^g)_{g=1}^G$ become draws from the joint distribution, $\pi_J(s, \boldsymbol{\xi}_J)$ as $G \to \infty$ (Gamerman and Lopes [22]). In order to assess practical convergence, Brooks-Gelman-Rubin (BGR) statistics (Brooks and Roberts [13]) are reported.

Second, the choice of the augmentation parameter, J, requires careful consideration. Müller et al. [45] showed that a time non-homogeneous Markov chain, whose draws converge to $\pi_\infty(s)$ with a Dirac measure on the desired optimum, can be constructed by increasing $J^{(g)}$ at a logarithmic rate. Jacquier et al. [32] present the proof that the standard error of estimating s^* goes to zero as J becomes large. Then, by gradually increasing J at a log-rate we can guarantee finding the optimal decision. For practical purposes, instead of a formal implementation, we recommend a diagnostic approach that increases $J^{(g)}$ until MCMC draws stabilize (Gelman and Rubin [26]). Then, it is

enough to sample from the joint density and to estimate the mode along the chain for a value of J on a given schedule, which depends on the flatness of the objective function surface, see Jacquier et al. [32].

Last, we are interested in the convergence of Monte Carlo simulation used to estimate the probability of delay, $Pr(T_q > 0) = E_{\lambda,\mu,\theta}\{Pr(T_q > 0|\lambda, \mu, \theta)\}$ within the APS framework. By setting the number of Monte Carlo draws within each MCMC iteration to J, we use the same set of J draws at each MCMC iteration. Convergence in number of Monte Carlo draws is guaranteed by the law of large numbers, $\hat{E}_\xi[Pr(T_q > 0)] = \frac{1}{J}\sum_{j=1}^{J} Pr(T_q > 0|\lambda^{(j)}, \mu^{(j)}, \theta^{(j)}) \to E_\xi[Pr(T_q > 0)]$, as $J \to \infty$ provided that $\mathbb{E}_\xi[|Pr(T_q > 0|] < \infty$; see Dempster et al. [18].

Our approach provides a number of computational advantages compared to other simulation methods. For instance, one can simply use a crude Monte Carlo or even MCMC simulation method, then check for the constraints for simple cases. There are a number challenges with this. First, one needs to compute the expected objective function value for each point in the target decision space. Conversely, the proposed APS algorithm draws samples of the decision variable, s, from the regions of the decision space with high objective function values and the draws of ξ_j are tilted away from the probability density $p(\xi_j)$ toward $Q(s, \xi_j)p(\xi_j)$. The algorithm samples "smart" values of ξ_j where the importance function is the objective function that tightens around the optimal decision, s^*, with the convergence. This reduces the Monte Carlo error since no optimization effort is wasted in parts of the parameter space with low objective function values and results in reduced sample variance leading to performance improvement (Ekin et al. [20]).

3.2 Proposed APS Formulation

We now discuss the details of the stochastic optimization method where the delay probability (2.12) is used as a constraint and the cost function (2.7) is used as the objective function to be minimized while satisfying the staffing requirements for a given time interval. Using the PSA approach of Green and Kolesar [28] which is common in the analysis of the Erlang-A model with time varying arrival rates, we split the day into finite time intervals where the system is stationary. We solve the staffing problem sequentially for each time interval. For convenience, the formulation below does not include an index on t. To summarize, given the vector of random variables $\xi = \{\lambda, \mu, \theta\}$, the

formulation of our stochastic optimization model will be

$$\begin{aligned} \min_{s} \quad & E_{\xi}[cs + a\lambda Pr(Ab|\xi)] \\ \text{subject to} \quad & Pr(T_q > 0) \leq \alpha. \end{aligned} \tag{3.5}$$

where α is a pre-specified delay threshold and s is the number of agents. $Pr(Ab|s, \lambda, \mu, \theta)$ and $Pr(T_q > 0)$ are both functions of arrival, service and abandonment rates as well as the number of agents. Therefore, there is a certain level of decision dependent randomness, albeit indirectly, within this formulation, which is not trivial to solve. We deal with the above problem for several discrete time intervals for a given day where λ is simply replaced by $\lambda(t)$ of the relevant time interval t. To solve the problem, we restructure (3.5) as a maximization problem and add a constant value, M, to the objective function in order to shift it to the positive region without changing the structure of the probability density function. We estimate $Pr(T_q > 0)$ by using J *iid* copies of the random variables and use

$$Pr(T_q > 0) \approx \frac{1}{J} \sum_{j=1}^{J} Pr(T_q > 0|\lambda^{(j)}, \mu^{(j)}, \theta^{(j)}).$$

The joint distribution is constructed as

$$\pi_J(s, \boldsymbol{\xi}_J) \propto \prod_{j=1}^{J} \left(M - cs - a\lambda Pr(Ab|\xi_j) p(\xi_j) \right) \mathbb{1}(Pr(T_q > 0) \leq \alpha),$$

where $\boldsymbol{\xi}_J = \{\lambda_1, \mu_1, \theta_1, \lambda_2, \mu_2, \theta_2, ..., \lambda_J, \mu_J, \theta_J\}$ and $\xi_j = \{\lambda_j, \mu_j, \theta_j\}$.

The mode of simulated decision variables provides the approximately optimal solution with minimum expected cost. We resort to MCMC methods to sample from this joint distribution. In a Metropolis within Gibbs setting, the full conditional distributions are

$$\pi(s|\boldsymbol{\xi}_J) \propto \prod_{j=1}^{J} \left(M - cs - a\lambda Pr(Ab|\xi_j) \right) \mathbb{1}(Pr(T_q > 0) \leq \alpha),$$

$$\pi(\xi_j|s) \propto \left(M - cs - a\lambda Pr(Ab|\xi_j) \right) p(\xi_j), \ \forall \, j = 1, 2, .., J.$$

We draw the number of servers, s, from the maximizing decision space given the constraints. We use a Normal proposal density within the random walk Metropolis Hastings algorithm. We draw J *iid* copies of the random vector $\xi = \{\lambda, \mu, \theta\}$ from the tilted probability distribution $\pi(\xi_j|s)$. The

probability of abandonment, $Pr(Ab|\xi_j)$, and probability of delay, $Pr(T_q > 0)$ are dependent on both s and ξ, showing the existence of decision dependence.

Our algorithm is general in that we can either draw the random variable, ξ_j from a multivariate distribution, exploiting the potential dependence, or we can utilize the marginal distributions to sample from each distribution. Below, we present the implementation for the case that the random variables $\xi = \{\lambda, \mu, \theta\}$ have independent posterior densities and we sample from their marginal full conditional distributions that are given by

$$\pi(\lambda_j|\mu_j, \theta_j, s) \propto \Big(M - cs - a\lambda Pr(Ab|\xi_j)\Big)p(\lambda_j),$$

$$\pi(\mu_j|\lambda_j, \theta_j, s) \propto \Big(M - cs - a\lambda Pr(Ab|\xi_j)\Big)p(\mu_j),$$

$$\pi(\theta_j|\lambda_j, \mu_j, s) \propto \Big(M - cs - a\lambda Pr(Ab|\xi_j)\Big)p(\theta_j).$$

We use Metropolis Hastings to sample from each of these distributions. For instance, after drawing the candidate λ_j from the proposal density $g(\lambda_j)$, we multiply the density of this draw by the objective function value. We use the distribution of $p(\lambda_j)$ as the proposal density for λ_j. This choice only affects the rate of convergence and results in an objective function-tilted blanket. We employ a Metropolis Hastings algorithm, with the acceptance probability

$$\alpha(\lambda_j, \lambda_{*j}) = min\left[\frac{(M - cs - a\lambda_{*j}Pr(Ab|\xi_j)g(\lambda_{*j})}{(M - cs - a\lambda_j Pr(Ab|\xi_j)g(\lambda_j)}, 1\right],$$

where λ_{*j} denotes the j^{th} candidate draw for λ. We accept the candidate draw and set the value of λ_j to λ_{*j} if the uniform draw $U(0, 1)$ is smaller than $\alpha(\lambda, \lambda_*)$, otherwise we do not change the value of λ_j. This is repeated for a number of iterations until convergence of the Markov chain is detected. The sampling process for $\pi(\mu_j|\lambda_j, \theta_j, s)$ and $\pi(\theta_j|\lambda_j, \mu_j, s)$ are similar and are therefore omitted from the discussion to preserve space.

4 Numerical Illustrations

In order to show the implementation of the proposed model, we present two cases; the first one uses real call center data for a small-medium size operation and the second one uses a simulated example for a slightly larger scale operation. For both datasets, we focus on the short-term staffing problem where the number of agents required for each finite time interval is determined prior to

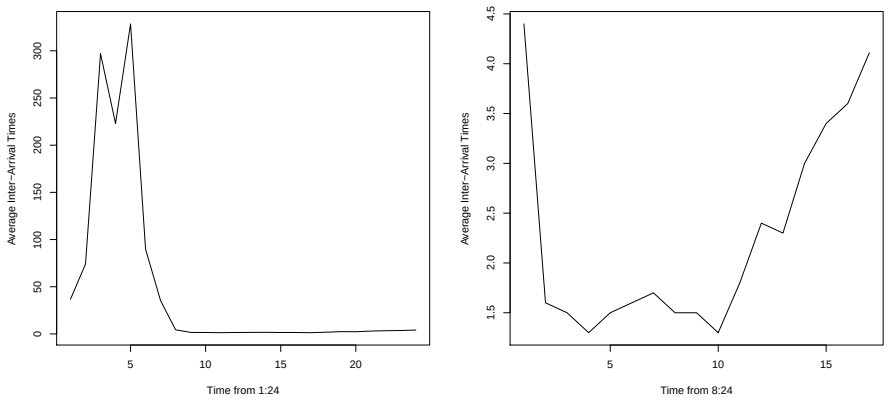

Figure 1: Average Inter-Arrival Times Between 1:24 (left) and Between 8:24 (right)

the start of the operation for a given day. For instance, given that we have observed past data for various weeks, at the beginning of the following week, we identify the staffing requirements for each time interval for Monday. In our example, we only focus on one day of the week, however we also note that it would be straightforward to implement our proposed model sequentially for each day of the week. For instance, as we collect arrival, service and abandonment data for Monday, we can update the uncertainty about our respective system rates and set the new staffing requirements for Tuesday.

4.1 Real Call Center Data Example

The real call center data considered in the sequel is from an anonymous bank operation and is publicly available at Data [17]. The whole data has daily information for a full given year. For illustrative purposes, we only focus on the month of January which had a total of 31,360 observations for arrival, service and abandonment times. We note here that we removed observations that were designated as "PHANTOM" calls. In what follows, we first briefly describe the data at hand, followed by our implementation.

As pointed out by Feldman et al. [21], typically staffing is held constant over fixed intervals (say 30-60 minutes) in call center operations. Therefore, we decided to split the day into 24 intervals of 60 minutes. The plot of average inter-arrival times is shown in the left panel of Figure 1 where time=1 represents the interval 12-1AM and time=24 represents the interval 11 PM-12 AM. Based on the average inter-arrival estimates, the operation between 12-7AM has very small arrival rates (i.e. very high inter-arrival times) as opposed to the rest of the day. Thus, we decided to focus

our attention on operations between 7AM until the end of the day (time=8, ... , 24). In addition,

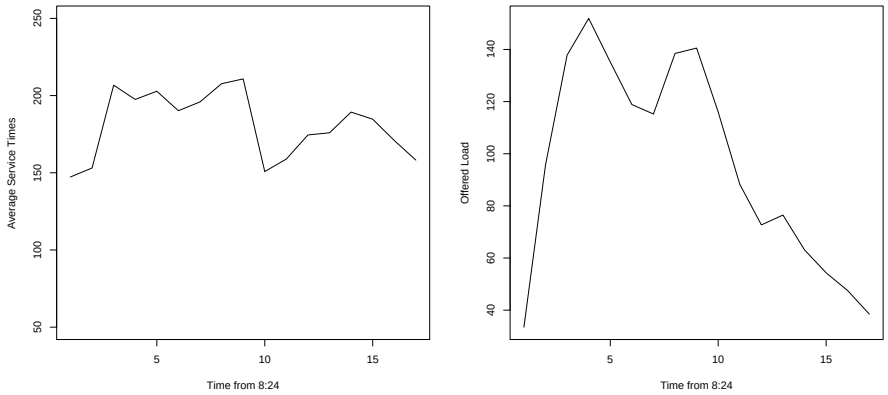

Figure 2: Average Service Times Between 8:24 (left) and Offered Load Between 8:24 (right)

Figure 2 shows the average service times (left panel) as well as the offered load (right panel) between the intervals 8 and 24. We note that the offered load between 9AM and 2PM (time periods 9-14) is higher with respect to the rest of the day and would be an example of a center with heavy traffic. As our proposed model is general, it can provide solutions to staffing in call center settings that range from low to heavy traffic levels. The average service times vary between 150 and 200 minutes. Such a variation can easily be captured by the random service rate structure we are imposing in our models using the gamma prior distributions as in (2.9). Alternatively, we could make the rates time dependent and estimate their stochastic evolution via $\mu(t)$, where each rate follows a separate gamma distribution. For illustrative purposes, we assume that the service rate is not time dependent in our implementation.

4.1.1 Optimization Results and Implications

Prior to observing data, we need to specify the hyper-parameters for the system rates, $\lambda(t), \mu$ and θ as in (2.9). To quantify our uncertainty about these rates, we assumed $\lambda(t) \sim G(0.001, 0.001) \forall t, \mu \sim G(0.001, 0.001)$ and $\theta \sim G(0.001, 0.001)$, which implies that our uncertainty about arrival, service and abandonment times is at best vague. To run our optimization model, in (3.5) we assumed that $\alpha = 0.9$ making the system an efficiency driven system, in the terminology of Garnett et al. [24]. In addition, we set c=50 and a=10,000 in the cost function for illustrative purposes. These two measures can be set by the call center management depending on the actual costs of staffing and perceived opportunity costs of the operation. We note that convergence was attained for the

real call center data example and provide a more detailed discussion of convergence for only the simulated example to preserve space in the narrative.

The optimal number of servers for each time period as a result of our approach are shown in the left panel of Figure 3 where the first time period coincides with the time period 7-8AM. The results indicate that, in the morning, the call center should start with 35 servers and increase the staffing up to a maximum of 72 servers during the day. The maximum number also gives an idea to the call center management for an upper bound of staffing on a given day. For instance, in this case the management can have 72 servers handy during the day where they are assigned to different tasks, say in the morning where only 35 of them are needed to answer calls.

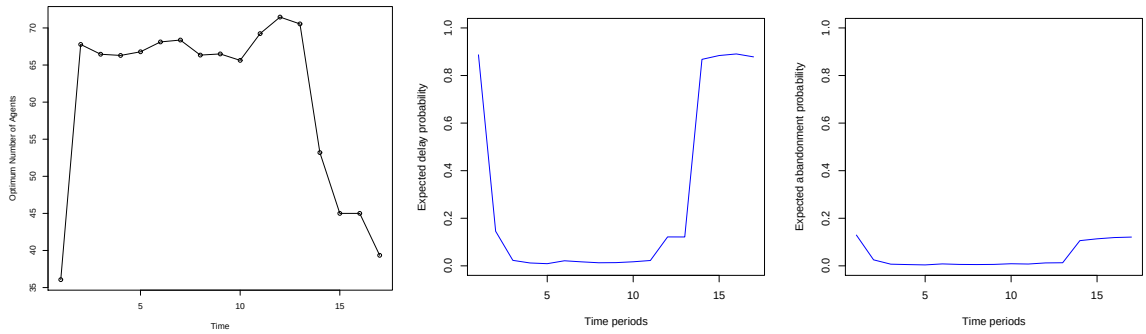

Figure 3: Optimal Number of Servers (left) Probability of Delay, $Pr(T_q > 0)$ (middle), and Probability of Abandonment, $Pr(Ab)$ (right)

Other operating characteristics of interest for call center management are the steady state fraction of number of customers abandoning and the probability of delay, whose estimates are in Figure 3. One of the constraints in our optimization model was $Pr(T_q > 0) \leq \alpha$, which is satisfied for all time periods in the middle panel of Figure 3. The estimates were between 0.009 and 0.890. The abandonment probabilities range between 0.004 and 0.129 during the day as shown in the right panel of Figure 3. As expected, when there is a steep increase in the optimal number of agents both delay and abandonment probabilities go down and vice-versa.

A measure widely used in call center staffing applications is the value of β, the so called service level. The well known square root staffing for operations with no abandonment is given by $s^* = r + \beta\sqrt{r}$, where s^* is the appropriate number of servers and $r = \frac{\lambda}{\mu}$ (Whitt [54]). For M/M/s+M queues, Garnett et al. [24] extended the staffing rule to account for customer abandonment and allowed β to take negative values. For limiting values of $Pr(Ab)$ and $Pr(T_q > 0)$, Garnett et al. [24] introduce three operating regimes for designing a call center and refer to them as quality driven,

efficiency driven and quality efficiency driven regimes. For values between -1 and 2, the regime is quality efficiency driven, for larger values of β, it is quality driven, and for smaller values it is efficiency driven. Even though, we do not explicitly use the staffing rule in identifying the optimal number of servers, it is possible to infer what kind of operational regime our optimal results imply by simply calculating the following implied service level (expected value of β) via

$$E(\beta) = \frac{s^* - E(r)}{E(r)}, \tag{4.1}$$

where $E(r)$ represents the expected value of the offered load which is a random variable in our approach. We used (4.1) to show the implied service level for our numerical example in the left panel of Figure 4, where $E(\beta)$ is mostly less than -1, implying an efficiency driven regime for this specific scenario, except for the first and the last six time periods where it can be defined as quality efficient driven in the terminology of Garnett et al. [24]. Since $\alpha = 0.9$ is the same for all time intervals, $E(\beta)$ stays in similar regions. We discuss the implications of changing α in our simulated example. The right panel of Figure 4 shows the behavior of the average operating cost over time periods for given values of c and a. As expected, it is cheaper to operate the center when the offered load is the least as can also be observed from the right panel of Figure 2.

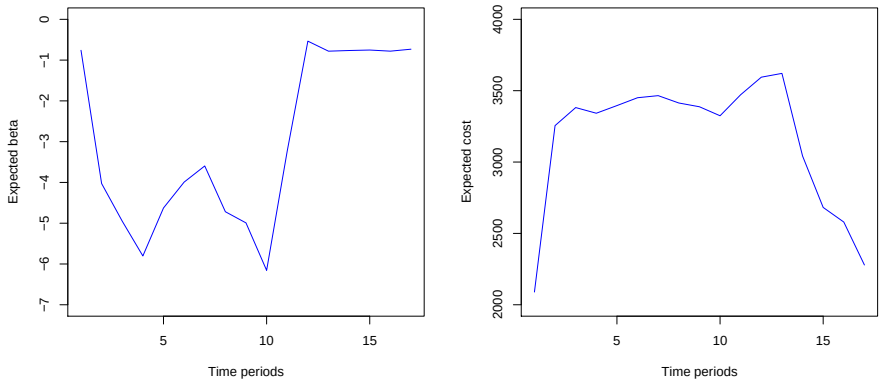

Figure 4: Expected Implied Level of Service, $E[\beta]$ and Expected Total Cost

4.2 Simulated Data Example

To discuss further properties of the proposed model, we simulated the operation of a larger call center that operates, say between 9AM to 8PM (a total of 11 time periods on a given day) where the arrival, service and abandonment processes are exponentially distributed with rates $\lambda(t), \mu$ and

θ where $\lambda(1) \sim G(10, 20), \lambda(t) = \lambda(t-1)+10t$ for $t = 2, \ldots, 6, \lambda(t) = \lambda(t-1)-10t$ for $t = 7, \ldots, 11$, $\mu \sim G(5, 1000)$ and $\theta \sim G(10, 1000)$. We chose these specific rate distributions so that our simulated data example can show the implications of using various delay thresholds on staffing. This was not possible with the real data example due to inherently already low agent utilization rates.

4.2.1 Optimization Results and Implications

We set $c = 10$ and $a = 10,000$ in our implementation. In order to show the performance of the model for different operating regimes, we used minimum delay thresholds of 0.1, 0.5 and 0.9. As before, we assumed flat priors for the system rates as $\lambda(t) \sim G(0.001, 0.001)\ \forall t, \mu \sim G(0.001, 0.001)$ and $\theta \sim G(0.001, 0.001)$. Even though the prior distributions for all the arrival rates are assumed to be the same, their posterior distributions will be drastically different. The observed arrival data during each time period will significantly influence the final posterior distribution estimates. We discuss the implications of using different priors in our sensitivity analysis section below. Figure 5 shows the optimal number of servers for different regimes. As expected, the quality driven regime (where $\alpha = 0.1$) requires a larger number of servers for all points in time.

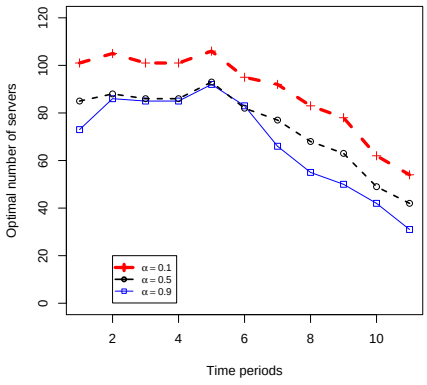

Figure 5: Optimal Number of Servers

Table 1 shows a summary of the optimal number of servers, expected implied levels of service, delay probabilities, abandonment probabilities and the expected operating costs for the quality driven regime. This regime, which has high quality of service expectations, resulted with lower delay and abandonment probabilities and the expected cost is mostly dominated by the total number of servers. The estimated abandonment probability is around 1%, which is in line with current practice in call center operations that typically require a threshold of 4%. However, for

other regimes, when α is 0.5 or 0.9, the abandonment probability is never less than 4%, which can be observed from the left panel of Figure 6. Furthermore, delay probabilities are estimated to be quite high (more than 0.40) for both efficient and quality-efficiency driven regimes as seen in the right panel of Figure 6. In other words, even a balanced operation with $\alpha = 0.5$ does not come close to what a quality driven operation can achieve in terms of service requirements.

Time	s^*	$E(\beta)$	$Pr(T_q > 0)$	$Pr(Ab)$	$E[Cost]$
1	101	1.217	0.098	0.012	1022.374
2	105	1.293	0.093	0.010	1063.691
3	101	1.244	0.098	0.010	1027.999
4	101	1.279	0.098	0.010	1025.784
5	106	0.962	0.090	0.009	1062.539
6	95	1.308	0.099	0.009	964.782
7	92	1.132	0.093	0.012	909.489
8	83	1.234	0.095	0.013	840.509
9	78	1.265	0.091	0.012	789.538
10	62	1.305	0.088	0.013	628.228
11	54	1.269	0.091	0.015	547.653

Table 1: Results for the Constraint with $\alpha = 0.1$

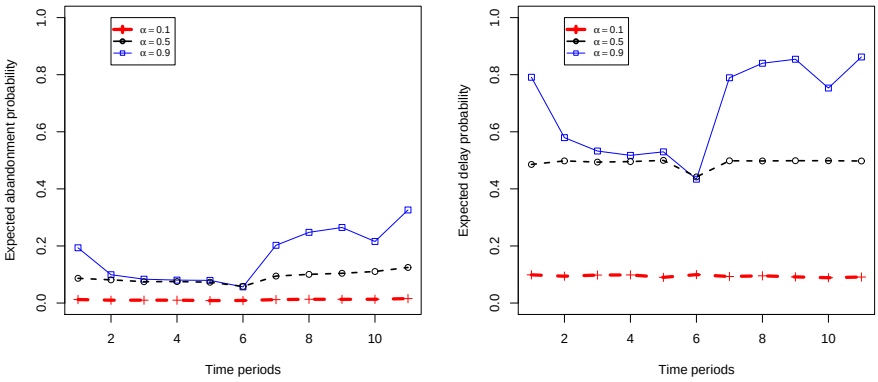

Figure 6: Probability of Abandonment, $Pr(Ab)$, and Probability of Delay, $Pr(T_q > 0)$

Figure 7 shows the behavior of the expected implied service level (left) and the expected total cost (right) over different time periods in a given day for the optimal number of servers of the three operating regimes. The expected implied service level is mostly below zero for both efficiency and quality efficiency driven regimes, which is in line with what we had observed with the delay probabilities in Figure 6. The expected implied service level for the quality driven case is around 1, significantly higher than the rest. An interesting observation is that a decrease from 0.5 to 0.1

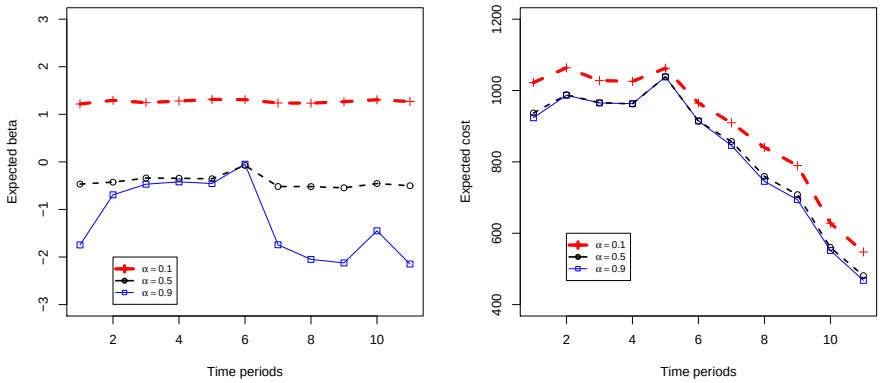

Figure 7: Expected Implied Level of Service, $E[\beta]$ and Expected Total Cost

in the delay threshold creates a much bigger quality of service difference than a decrease from 0.9 to 0.5. In addition, the cost estimates for both efficiency and quality efficiency driven regimes are quite close to each other since their respective optimal number of servers are similar as opposed to that of the quality driven regime. We believe that such insights would be valuable for the call center management in identifying quality of service requirements in setting staffing levels. These will be investigated in the next subsection in detail.

After $G = 1,000$ replications of the MCMC with $J = 100$ where the initial 100 burn-in samples are discarded, MCMC convergence is practically supported by BGR diagnostics. For instance, when the minimum delay threshold is 0.9, BGR values for 3 chains and 11 time intervals are found to be between 1.01 and 1.11, that are reasonable (Gelman and Hill [25]). The interplay between J and G is crucial to reach fast convergence. In our setting, J is also important in estimating $Pr(T_q > 0)$. The computational experiment of Ekin et al. [20] points out, a good strategy would be to increase the value of J when MCMC convergence is reached.

Figure 8 presents the trace plots for the number of servers in period 6 (as an example) for different values of J. It can be seen that the draws converge around the optimal number of servers more quickly for increasing values of J. The standard deviations of the number of servers are 2.083, 0.765, 0.597 and 0.401 for J values of 2, 10, 25 and 100, respectively. Increasing J results in a smaller standard deviation for the decision variable and faster convergence to the optimal decision.

Next, we conducted a simple comparison between our method and a basic simulation based approach. The reported results above use $G * J = 1,000 * 100$ draws. For a fair comparison, we ran

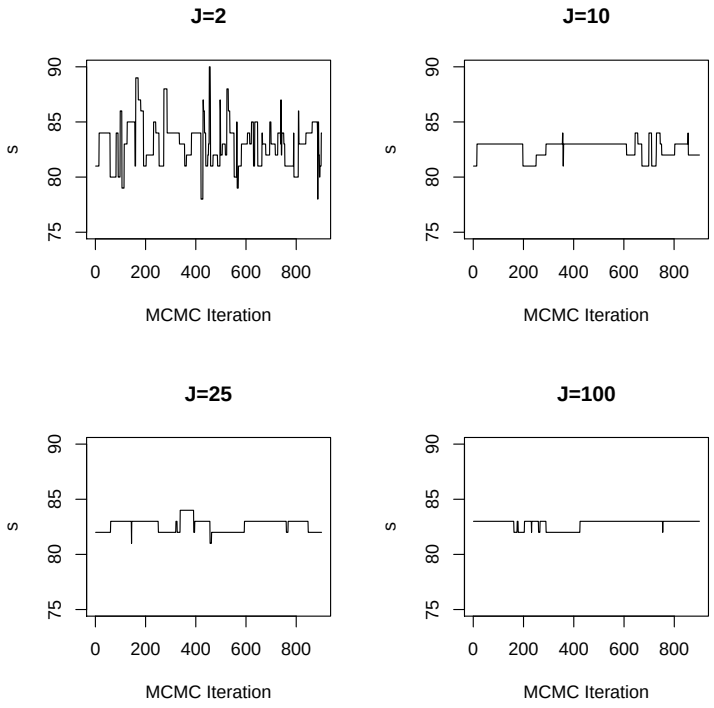

Figure 8: Convergence of the server level for period 6 for different values of J

a Monte Carlo simulation with the same number of draws. We use 500 draws for each decision point $(1, 2, .., 200)$ in the decision space of $(1, 200)$. We computed the expected costs and the probability of delay, and manually checked if the constraint is satisfied. We found similar optimal decisions albeit with higher standard deviations. The efficiency gains of our approach can be explained by drawing "smart" samples of the random variables. It can be argued that one can use MCMC simulation or an importance-sampling based approach to conduct smart sampling. For instance, Morales et al. [42] use a similar idea in Bayesian design of queues with an application in aeronautic maintenance. They consider a lower bound on the probability of having at least a certain level of operational components in an airplane using the idea of Monte Carlo approximation. However, such an approach still require the expectation computation for all the decision points, whereas our proposed method concentrates on the "maximizing" portions of the decision space. It should also be noted that these basic simulation/numerical analysis approaches cannot be utilized for dynamic formulations where decisions are coupled. Furthermore, using these methods will be challenging in the case of tight constraints due to high rejection rates and can result in additional levels of computational burden. Thus, efficiency gains due to smarter sampling would become important especially with flat objective functions.

4.3 Sensitivity Analysis

One of the important features of Bayesian inference is the use of prior distributions for model parameters. In our setup, all system rates have prior distributions as introduced in (2.9). In all of our previous analyses, we have been setting the hyper-parameters so that all the prior distributions were uninformative (flat). Such an approach might be favored when there is no previous information about the call center. It is also possible to set the hyper-parameters such that the prior information about the system rates are stronger. This could be the option favored by most call center operations with previous history on arrivals, service or abandonment (or a combination of the three). An important question is how sensitive our results are to changes in the hyper-parameters of the prior distributions. To assess the sensitivity of our solution method for the simulated case, we considered four cases where the hyper-parameters are set to different values. The case for the prior of λ_1 is shown in Figure 9 for illustrative purposes. For the first three cases, the prior uncertainty is slightly getting stronger but stays still fairly uninformative. In the fourth case, the prior information is strong and is centered around the posterior distribution. To find the values of the hyper-parameters, we use the properties of the gamma distribution where the mean is a_0/b_0 and the variance is a_0/b_0^2. Based on the posterior mean and the variance, it was straightforward to solve for a_0 and b_0 and to obtain the values, 1000 and 9810, as in Figure 9. We altered the hyper-parameters of all system rates in a similar manner. For the proposed model, the optimal staffing results were identical for all four cases which shows evidence in favor of no sensitivity to prior values for our model. This can most probably be attributed to the abundance of data at hand as the effect of the prior distribution becomes less influential for inference given data.

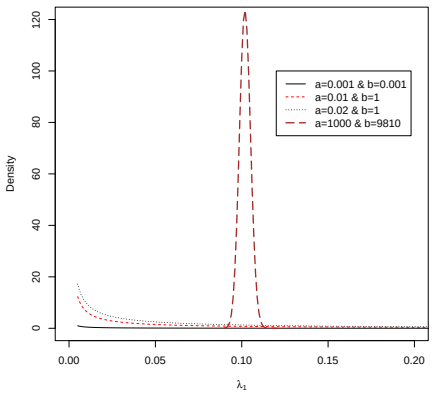

Figure 9: Sensitivity to Prior Distributions

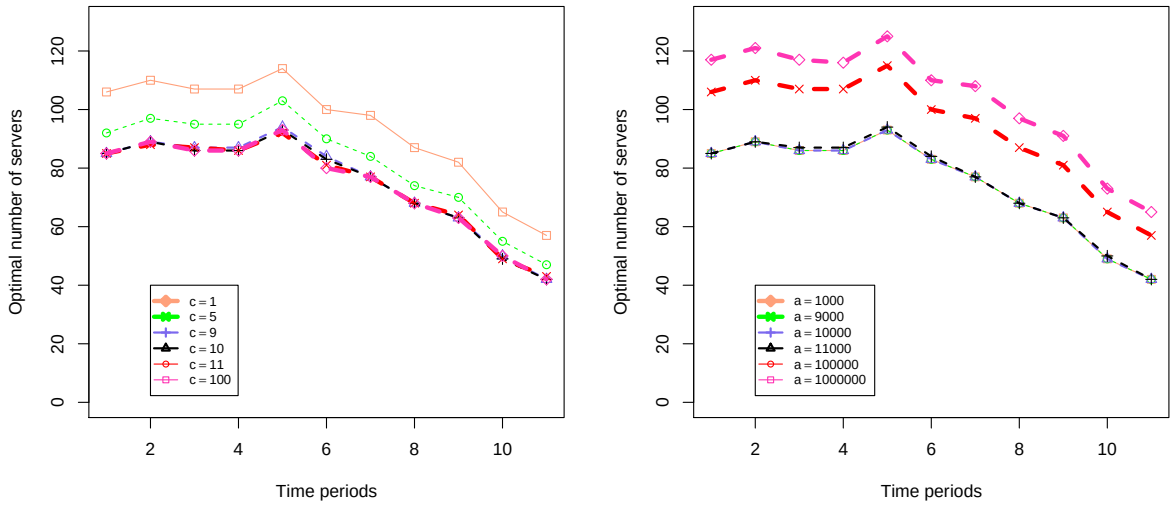

Figure 10: Optimal Number of Servers for different values of c (left) and for different values of a (right)

Next, we investigate the sensitivity to the objective function coefficients to unit staffing and abandonment costs. This can help the decision maker evaluate the effect of salary changes or the impact of abandonment on operations. Our initial parameter values were arbitrarily set at $c = 10$ and $a = 10,000$. We have investigated the sensitivity of the optimal number of servers to both small and large changes in parameters. The left panel of Figure 10 presents the changes in the optimal number of servers for different values of c. For small changes in the amount of 10%, the optimal decision does not change significantly, whereas the number of servers increase when c becomes relatively very low. The sensitivity results for different values of a can be seen in the right panel of Figure 10. Similarly, for small changes in a, the optimal decision remains stable. However, when the unit cost of abandonment increases by large amounts, the optimal number of servers increase at a higher rate.

Finally, we conduct sensitivity analysis for α, the delay threshold. Based on the optimal number of servers, service levels, and the abandonment probability estimates given in Figures 5, 6 and 7, we can observe the changes in those measures as well as in optimal staffing as a function of α. This raises the question of whether it is possible to find a threshold level of α, say $\alpha^\star$, such that the optimal solution or the operating measures are not significantly different across various threshold levels. Given the complexity of the formulations, where both the objective function and the constraints are random quantities, a closed form solution to this problem is not possible.

However, we can numerically investigate the changes in the optimal solution which in turn affect the operating measures. To do so, we obtain the optimal solution for several values of α for our proposed model and show the results in Table 2 between 0.1 and 0.9 with increments of 0.1 units.

Based on the estimates from Table 2, the difference in staffing becomes less significant after $\alpha = 0.5$ for time slots 2-6 as can be observed from Table 2. The time slots 2-6 coincide with the busiest times of the day. This can be explained by the imposed service level beyond 0.5. In this case, the operation is described as efficiency driven, requiring a small number of servers for relatively non-busy time slots such as 1 and 7-11. For lower values of α, we can see apparent differences across the solutions for each time slot of the day regardless of how busy the system is. This type of sensitivity analysis can be conducted regularly by the call center management to assess the effect of the delay threshold and possibly to quantify the effect of such changes on the actual operation cost.

	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$
Time	s	s	s	s	s	s	s	s	s
1	101	96	92	88	85	81	78	73	73
2	105	99	95	92	89	86	86	86	86
3	101	96	93	90	87	85	85	85	85
4	101	96	92	89	87	85	85	85	85
5	106	103	100	96	93	92	92	92	92
6	95	90	86	83	82	83	82	82	82
7	92	88	84	80	78	74	70	66	66
8	83	78	75	72	68	65	61	57	55
9	78	72	69	66	64	60	57	53	50
10	62	57	55	52	50	47	43	42	42
11	54	50	47	45	43	40	37	34	31

Table 2: Sensitivity to α

4.4 Modeling Extensions

Our proposed approach is general and can handle any form of objective function and constraints. This subsection presents a number of formulations, including a dynamic model, provides comparisons and further discusses the implementation of the proposed approach.

4.4.1 Dynamic Staffing

First, we consider a dynamic formulation with an objective function that includes an additional cost to change the level of staffing between time periods. This type of formulation would be

important when changes in the staffing level lead to a trade-off type cost. It may take time and resources to assign agents to other tasks, or to recall them from other tasks when there is a need for additional agents. As staffing decisions are not decoupled as before, this dynamic formulation can not be solved with ad-hoc simulation methods or numerical analysis. This further underlines the importance of our approach. The formulation of this dynamic model is

$$\begin{aligned} \min_{s(t)} \quad & E_{\xi}\Big[cs(t) + k|s(t) - s^*(t-1)| + a\lambda(t)Pr(Ab|\xi)\Big] \\ \text{subject to} \quad & Pr(T_q > 0) \leq \alpha. \end{aligned} \tag{4.2}$$

where $s(t)$ is the number of servers to be determined for period t, $s^*(t-1)$ represents the optimal number of servers from the previous period $t-1$ and k denotes the unit cost of changing the staffing level between two consecutive time periods.

We solve the above formulation sequentially (one period at a time) where the solution from a given period is dependent upon that of the previous period. That the abandonment and delay probabilities are revised sequentially for every time period as before. Finding the optimal solution dynamically did not create any significant additional computational burden for our approach. Our experiments suggest that the dynamic formulation takes up to 9.2 % more computational time with respect to the initial static formulation. For instance, the static formulation with $J = 25$ takes 803.7 seconds whereas the dynamic formulation with the same parameters takes 871.4 seconds. This provides additional support that our method can accommodate different formulations of objective function and constraints.

Figure 11 shows the optimal number of servers using the dynamic formulation (4.2) for $\alpha = 0.5$ and some arbitrary values of k . When $k = 0$, the formulation in (4.2) reduces to our initial model. In the first five time periods, the staffing level is roughly the same for all values of k with very small deviations. However, for $k = 0$, the optimal staffing level significantly changes from one period to another during periods 6-11. This results with significant changes among these different scenarios. As k gets larger (i.e. when it is more costly to switch the staffing levels), the model penalizes changes. After a certain level of k (larger than 10), the staffing levels become more or less stable as the cost of the switching levels becomes too large for the model to favor changes.

In addition, several other formulations can be considered using our model, depending on the end user's objectives and resources. For instance, it is possible to assume a changing cost structure for negative versus positive changes in the staffing level. Alternatively, if the call center management

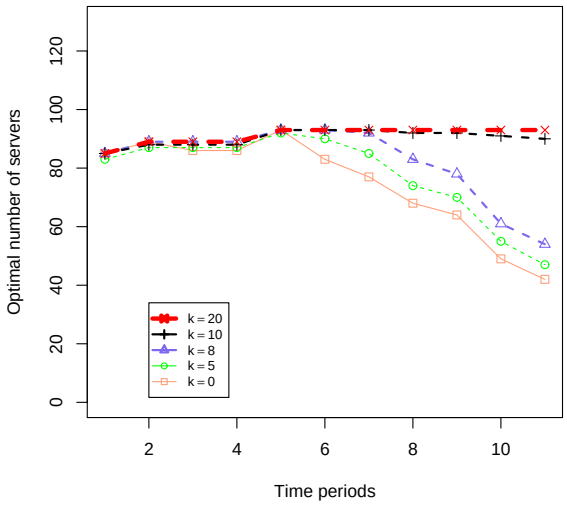

Figure 11: Optimal Solutions for the Dynamic Extension when $k = 0, 5, 8, 10, 20$

has limited number of agents for different time periods, this can be handled by adding an upper bound constraint on the number of servers for each period or on the change in the staffing level with respect to the previous period. Another possibility is to consider a cost function which consists of a combination of several system characteristics: the number of busy servers, empty servers, served customers, customers waiting in the queue and waiting time in the queue as considered by Ausín et al. [8].

4.4.2 Other Formulations

Next, we provide comparisons across four different formulations. Model 1 is the proposed formulation in (3.5). Model 2 is the formulation of Aktekin and Soyer [4], an unconstrained model with the same objective function. We also consider formulations that can help explore the implications of using abandonment and delay probabilities as constraints, instead of as part of the objective function. Therefore, we construct Models 3 and 4 which have the same objective function (with only the expected staffing cost) but with constraints on only delay and abandonment probabilities. For Models 1 and 3, α was set to 0.5. In order to provide a fair comparison, the threshold for the abandonment probability in Model 4 is set to 0.1 which is close to the average abandonment probability when α is 0.5. The results of these comparisons are shown in Tables 3, 4 and 5.

The impact of imposing a constraint can be seen clearly from Table 3. The differences between

	Model 1			Model 2		
Time	s^*	$Pr(T_q > 0)$	$Pr(Ab)$	s^*	$Pr(T_q > 0)$	$Pr(Ab)$
1	85	0.486	0.087	73	0.702	0.203
2	88	0.498	0.081	86	0.543	0.123
3	86	0.494	0.075	85	0.494	0.105
4	86	0.496	0.075	85	0.501	0.106
5	93	0.500	0.073	92	0.499	0.102
6	82	0.442	0.058	82	0.443	0.090
7	77	0.498	0.094	65	0.693	0.216
8	68	0.498	0.100	55	0.725	0.257
9	63	0.499	0.104	50	0.728	0.278
10	49	0.499	0.111	42	0.644	0.256
11	42	0.498	0.125	31	0.688	0.325

Table 3: Comparison of Results for Models 1 and 2

the optimal number of servers are more significant in time periods 1 and 7-11. In busier time periods 2-6, the difference is not as evident. For Model 1, all delay probabilities are close to the threshold of $\alpha = 0.5$, whereas for Model 2 they are in the range between 0.443 and 0.728. The abandonment probabilities for Model 1, are between 0.058 and 0.125 in the vicinity of 0.100, whereas Model 2 results in larger abandonment probability estimates, as high as 0.325. This justifies the importance of considering constraints and underlines the contribution of the proposed optimization model to the literature with respect to the simulation based evaluation of Aktekin and Soyer [4].

	Model 1			Model 3		
Time	s^*	$Pr(T_q > 0)$	$Pr(Ab)$	s^*	$Pr(T_q > 0)$	$Pr(Ab)$
1	85	0.486	0.087	84	0.491	0.088
2	88	0.498	0.081	89	0.483	0.075
3	86	0.494	0.075	86	0.483	0.072
4	86	0.496	0.075	86	0.479	0.072
5	93	0.500	0.073	93	0.491	0.071
6	82	0.442	0.058	80	0.471	0.065
7	77	0.498	0.094	77	0.486	0.091
8	68	0.498	0.100	68	0.490	0.098
9	63	0.499	0.104	64	0.476	0.097
10	49	0.499	0.111	49	0.492	0.101
11	42	0.498	0.125	42	0.489	0.122

Table 4: Comparison of Results for Models 1 and 3

Next, we explore the impact of considering the delay probability within the objective function and as a constraint versus only as a constraint. Since the expected abandonment cost is a function of the delay probability, it could be argued that the delay probability is over-penalized within

the first formulation. Models 1 and 3 provide similar results with small variations in the optimal number of servers, where the difference range is between 0 (minimum) and 2 (maximum) servers for a given time interval with an absolute average of 0.45. This may be attributed to the slightly different delay and abandonment probability estimates.

	Model 2			Model 4		
Time	s^*	$Pr(T_q > 0)$	$Pr(Ab)$	s^*	$Pr(T_q > 0)$	$Pr(Ab)$
1	73	0.702	0.203	83	0.529	0.098
2	86	0.543	0.123	86	0.570	0.097
3	85	0.494	0.105	83	0.590	0.097
4	85	0.501	0.106	83	0.577	0.095
5	92	0.499	0.102	89	0.605	0.097
6	82	0.443	0.090	77	0.617	0.096
7	65	0.693	0.216	76	0.507	0.097
8	55	0.725	0.257	68	0.474	0.094
9	50	0.728	0.278	64	0.467	0.094
10	42	0.644	0.256	50	0.454	0.097
11	31	0.688	0.325	44	0.405	0.094

Table 5: Comparison of Results for Models 2 and 4

Lastly, we compare Models 2 and 4 where the abandonment probability appears as a constraint. As expected, the results are significantly different. The lack of a constraint in Model 2 results in a lower number of servers, albeit with higher delay and abandonment probability estimates. For instance, the estimates for Model 2 range between 0.090 and 0.325 whereas for Model 4 this range is between 0.094 and 0.097 due to the imposed minimum threshold. This also justifies the use of constraints in order to impose the service level criteria in a more accurate fashion. Depending on the goal of the call center management, one of these formulations can be preferred for staffing purposes. If the management is concerned for both the delay and the abandonment probabilities then Model 1 would be a more suitable option.

5 Conclusion and Directions for Future Work

In this study, we considered an augmented probability simulation technique for staffing call centers for the Erlang-A queuing system with discrete time intervals (using the PSA argument) where the arrival, service and abandonment rates are uncertain. In modeling the system rates, we used Bayesian inference by assuming gamma distributed prior distributions. The treatment of the system rates as uncertain quantities creates an extra layer of randomness in key operating characteristics

such as the delay and abandonment probabilities as well as the cost function. These models are also known as doubly stochastic processes. Traditional queuing methods only utilize point estimates and provide relevant measures as deterministic functions, whereas our model provides functions that are random variables. Our optimization formulation is based on the idea of selecting the optimal number of agents required for a given time interval by minimizing an expected cost function which consists of agent and abandonment costs that are constrained by a pre-specified minimum delay threshold. To illustrate the use of our proposed approach, we used both real and simulated call center data, investigated implementation issues, provided comparisons to other formulations, considered a dynamic extension and discussed implications of prior and constraint sensitivity.

To summarize, the novelty of our proposed model comes from solving the staffing decision problem where all system rates are random. Also, our method is general and can be applied to any form of the objective function as evidenced by the dynamic extension considered in our numerical study. The solution technique we consider for staffing, the augmented probability simulation, has not been previously used in the call center literature. In addition, our proposed methodology extends the APS approach by solving a one stage stochastic program with decision dependent uncertainty where both the objective function and constraints are random. Furthermore, to the best of our knowledge, this extension is a methodological novelty of our study and can be utilized for solving similar decision problems with varying cost functions. For instance, our proposed model might be applied in the setting of emergency room staffing in hospitals where abandonment is treated as the termination of a patient while waiting for treatment. In this case, obviously having a constraint on the abandonment rate or the delay probability would be crucial. It also makes sense to treat the arrival, service and abandonment rates of certain groups of patients as random since there might be a considerable uncertainty especially on abandonment and service rates. Other notable application areas would be network systems, bank operations and chat systems, where a queuing system with potential abandonment is typically observed and staffing is of interest.

One of the limitations of our approach is the assumption about exponentially distributed arrival, service and abandonment processes of the Erlang-A model. Even though our optimization method is very general and can be applied to many different formulations, the operating characteristic functions used in the formulation are dependent upon the Markovian assumption. Thus, deviations from exponentiality may skew optimal results. Brown et al. [14] indicate that the Erlang-A model is considered to be robust when the exponential abandonment assumption is violated, we believe that our proposed model can be extended to account for general abandonment given that some

of the well known computational barriers for the M/M/s+G model can be overcome, especially with random rates. In cases where the arrival rates exhibit an identifiable form of within-day cyclical behavior, it may be possible to set the log prior mean of the $\lambda(t)$s to some known cyclical function (e.g. a linear combination of sine and cosine terms) using our approach. Another potential extension would be to allow for dependence rather than assuming independence between system rates of arrival, service and abandonment via their joint distributions. Since traditional queuing methods assume that system rates are fixed, it will not be possible to consider such a feature. However, the use of Bayesian inference can help us estimate the joint posterior distribution of system rates in a straightforward manner. The most important issue is, once again, computational since our augmented simulation method requires that we jointly generate the decision variables with the posterior distribution of system rates. We believe this would be a worthwhile extension to consider for future work and the feasibility of the idea is currently under consideration.

References

[1] Z. Aksin, M. Armony, and V. Mehrotra. The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6): 665–688, 2007.

[2] Z. Aksin, B. Ata, S. M. Emadi, and C. L. Su. Structural estimation of callers' delay sensitivity in call centers. *Management Science*, 59(12):2727–2746, 2013.

[3] T. Aktekin and R. Soyer. Call center arrival modeling: A Bayesian state-space approach. *Naval Research Logistics*, 58(1):28–42, 2011.

[4] T. Aktekin and R. Soyer. Bayesian analysis of queues with impatient customers: Applications to call centers. *Naval Research Logistics*, 59(2):441–456, 2012.

[5] C. Armero and M. Bayarri. Prior assessment for prediction in queues. *The Statistician*, 43(1): 139–153, 1994.

[6] C. Armero and M. Bayarri. Bayesian prediction in M/M/1 queues. *Queuing Systems*, 15: 401–417, 1994.

[7] J. Atlason, M. A. Epelman, and S. G. Henderson. Optimizing call center staffing using simulation and analytic center cutting-plane methods. *Management Science*, 54(2):295–309, 2008.

[8] M. C. Ausín, R. E. Lillo, and M. P. Wiper. Bayesian control of the number of servers in a GI/M/c queueing system. *Journal of Statistical Planning and inference*, 137(10):3043–3057, 2007.

[9] F. Bacelli and G. Hebuterne. On queues with impatient customers. *Performance'81*, (2): 159–179, 1981.

[10] A. Bassamboo, J. M. Harrison, and A. Zeevi. Pointwise stationary fluid models for stochastic processing networks. *Manufacturing & Service Operations Management*, 11(1):70–89, 2009.

[11] C. Bielza, P. Müller, and D. Rios Insua. Decision analysis by augmented probability simulation. *Management Science*, 45(7):995–1007, 1999.

[12] J. Birge and F. Louveaux. *Introduction to Stochastic Programming, Second Edition*. Springer-Verlag, Berlin, 2011.

[13] S. P. Brooks and G. O. Roberts. Assessing convergence of Markov chain Monte Carlo algorithms. *Statistics and Computing*, 8(4):319–335, 1998.

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

[15] G. Casella and E. I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3): 167–174, 1992.

[16] S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

[17] Data. Technion, Israel Institute of Technology, 2000. Available at http://iew3.technion.ac.il /serveng /callcenterdata/.

[18] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society,Series B*, 339:11738, 1977.

[19] M. Dyer and L. Stougie. Computational complexity of stochastic programming problems. *Mathematical Programming*, 106(3):423–432, 2006.

[20] T. Ekin, N. G. Polson, and R. Soyer. Augmented Markov chain Monte Carlo simulation for two-stage stochastic programs with recourse. *Decision Analysis*, 11(4):250–264, 2014.

[21] Z. Feldman, Z. Mandelbaum, W. Massey, and W. Whitt. Staffing of time-varying queues to achieve time-stable performance. *Management Science*, 54(2):324–338, 2008.

[22] D. Gamerman and H. Lopes. *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference(2nd ed.)*. Chapman & Hall, 2006.

[23] N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, 2003.

[24] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4:208–227, 2002.

[25] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

[26] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

[27] V. Goel and I. E. Grossmann. A class of stochastic programs with decision dependent uncertainty. *Mathematical Programming*, 108(2-3):355–394, 2006.

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

[29] I. Gurvich, M. Armony, and A. Mandelbaum. Service-level differentiation in call centers with fully flexible servers. *Management Science*, 54(2):279–294, 2008.

[30] I. Gurvich, J. Luedtke, and T. Tezcan. Staffing call centers with uncertain demand forecasts: A chance-constrained optimization approach. *Management Science*, 56(7):1093–1115, 2010.

[31] T. Homem-de Mello and G. Bayraksan. Monte carlo sampling-based methods for stochastic optimization. *Surveys in Operations Research and Management Science*, 19(1):56–85, 2014.

[32] E. Jacquier, M. Johannes, and N. Polson. MCMC maximum likelihood for latent state models. *Journal of Econometrics*, 137:615–640, 2007.

[33] O. Jouini, Z. Aksin, and Y. Dallery. Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management*, 13(4):534–548, 2011.

[34] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, et al. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[35] T. Kirschenmann, E. Popova, P. Damien, and T. Hanson. Decision dependent stochastic processes. *European Journal of Operational Research*, 234(3):731–742, 2014.

[36] J. Landon, F. Ruggeri, M. Tarimcilar, and R. Soyer. Modeling latent sources in call center arrival data. *European Journal of Operational Research*, 204(3):597–603, 2010.

[37] A. Mandelbaum and N. Shimkin. A model for rational abandonments from invisible queues. *Queueing Systems*, 36:141–173, 2000.

[38] A. Mandelbaum and S. Zeltyn. *Advances in Service Innovations*, chapter Service Engineering in Action: The Palm/Erlang-A Queue, with Applications to Call Centers, pages 17–48. Springer, 2007.

[39] A. Mandelbaum and S. Zeltyn. Staffing many-server queues with impatient customers: constraint satisfaction in call centers. *Operations Research*, 57(5):1189–1205, 2009.

[40] M. F. McGrath and N. D. Singpurwalla. A subjective Bayesian approach to the theory of queues II - inference and information in M/M/1 queues. *Queuing Systems*, 1:335–353, 1987.

[41] M. F. McGrath, D. Gross, and N. D. Singpurwalla. A subjective Bayesian approach to the theory of queues I - modeling. *Queuing Systems*, 1:317–333, 1987.

[42] J. Morales, M. E. Castellanos, A. M. Mayoral, R. Fried, and C. Armero. Bayesian design in queues: An application to aeronautic maintenance. *Journal of statistical planning and inference*, 137(10):3058–3067, 2007.

[43] D. P. Morton and E. Popova. A bayesian stochastic programming approach to an employee scheduling problem. *IIE Transactions*, 36:155–167, 2004.

[44] P. Müller. Simulation-based optimal design. In J. Bernardo and A. Berger J.O., A.P.Dawid, editors, *Bayesian Statistics 6*, pages 459–474. Oxford University Press, 1999.

[45] P. Müller, B. Sansó, and M. De Iorio. Optimal Bayesian design by inhomogeneous Markov chain simulation. *Journal of the American Statistical Association*, 99(467):788–798, 2004.

[46] C. Palm. Research on telephone traffic carried by full availability groups. *Tele*, (1):107, 1957.

[47] M. Pincus. A Monte Carlo method for the approximate solution of certain types of constrained optimization problems. *Operations Research*, 18(6):1225–1228, 1970.

[48] D. Rios Insua, M. Wiper, and F. Ruggeri. Bayesian analysis of M/Er/1 and M/h_k/1 queues. *Queuing Systems*, 30:289–308, 1998.

[49] D. Rios Insua, F. Ruggeri, and M. Wiper. *Bayesian Analysis of Stochastic Process Models*. Wiley New York, 2012.

[50] R. Y. Rubinstein and D. P. Kroese. *Simulation and the Monte Carlo method*, volume 707. John Wiley & Sons, 2011.

[51] H. Shen and J. Z. Huang. Interday forecasting and intraday updating of call center arrivals. *Manufacturing & Service Operations Management*, 10(3):391–410, 2008.

[52] L. Tierney. Markov chains for exploring posterior distributions (with discussion). *The Annals of Statistics*, 22(4):1701–1762, 1994.

[53] J. Weinberg, L. D. Brown, and J. R. Stroud. Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *Journal of the American Statistical Association*, 102(480):1185–1198, 2007.

[54] W. Whitt. Understanding the efficiency of multi-server service systems. *Management Science*, 38(5):708–723, 1992.

[55] W. Whitt. What you should know about queueing models to set staffing requirements in service systems. *Naval Research Logistics*, 171(5):476–484, 2007.

[56] M. Wiper. Bayesian analysis of Er/M/1 and Er/M/c queues. *Journal of Statistical Planning and Inference*, 69:65–79, 1998.

[57] E. Zohar, A. Mandelbaum, and N. Shimkin. Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Science*, 48(4):566–583, 2002.