Operations Research Letters 46 (2018) 379–384

Contents lists available at ScienceDirect

Operations Research Letters

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

Staffing call centers under arrival-rate uncertainty with Bayesian updates

Jing Zan [a], John J. Hasenbein [b,*], David P. Morton [c], Vijay Mehrotra [d]

[a] Uber Technologies Inc., 555 Market Street, San Francisco, CA 94105, United States
[b] Graduate Program in Operations Research and Industrial Engineering, Department of Mechanical Engineering, University of Texas at Austin, Austin, TX 78712, United States
[c] Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208, United States
[d] University of San Francisco, Department of Business Analytics and Information Systems, School of Management, 2130 Fulton Street, San Francisco CA 94117, United States

ARTICLE INFO

Article history:
Received 21 September 2017
Received in revised form 9 March 2018
Accepted 11 April 2018
Available online 1 May 2018

Keywords:
Call center staffing
Quality-of-service
Erlang-C
Stochastic programming
Bayesian update

ABSTRACT

We consider the problem of staffing service centers with quality-of-service constraints. We focus on the case where the arrival rates are uncertain. We introduce formulations that handle staffing decisions made over two decision periods, minimizing the staffing costs over the stages while satisfying a service quality constraint on the second stage operation. A Bayesian update is used to obtain the second-stage arrival-rate distribution based on the first stage prior arrival-rate distribution and the observations in the first stage.

1. Introduction

In this paper we develop a methodology for optimally updating call forecasts and staffing levels in a later period based on information about customer demand in earlier periods, and in turn to optimally set initial staffing levels based on the opportunity for future updates. In particular, this methodology is designed to explicitly account for the correlation between call volumes across time intervals, the cost of initially staffing agents, and the costs of (and opportunities for) later updates to staffing levels.

For call centers, the process of determining staffing levels and agent schedules is typically based on forecasted call volumes. The traditional approach to call center staffing and agent scheduling has been to use call arrival forecasts as estimates for the mean arrival rate for a non-homogeneous Poisson process (NHPP) with piecewise constant arrival rates over specific time intervals that are typically 15-, 30- or 60-minutes long. However, as several researchers (including Brown et al. [3], Steckley et al. [11], and Taylor [12]) have noted, this level of forecast accuracy is difficult to achieve. In particular, after analyzing a great deal of historical data, Brown et al. [3] conclude that call center arrival rates should be modeled as random variables rather than deterministic point forecasts. Similar empirical findings are presented by several researchers including Avramidis et al. [1] and Steckley et al. [11], which in turn has led these and many others (including Jongbloed and Koole [7], Bassamboo et al. [2], Harrison and Zeevi [5], Whitt [13], Robbins and Harrison [10], Liao et al. [8]) to model call arrival rates as random variables.

Finally there are two other closely related papers that deal with call center staffing. Gans et al. [4] develop a parametric forecasting model along with a stochastic programming formulation with recourse for optimally determining agent schedules in the presence of random call arrival rates that are correlated across intervals. Mehrotra et al. [9] present a method for optimal intraday recourse actions to adjust initial staffing levels in response to statistically significant deviations from arrival-rate forecasts. These two papers, and ours, examine staffing problems with arrival rate updating. Our forecast updating scheme is much simpler than the method presented in [4]. Mehrotra et al. [9] do not present any particular updating method (leaving the decision to the modeler). We also address a simpler call center framework, with just one type of caller. Due to these two differences, we can provide what are effectively closed-form solutions to the two-stage staffing problem for various performance measures (utilization is the only metric discussed in detail herein, due to space constraints, but further results appear in [14]). In contrast, the models in [4] and [9] require solution of more complex optimization problems. For simpler call center settings, our model is useful in that the explicit solutions

* Corresponding author.
E-mail address: jhas@mail.utexas.edu (J.J. Hasenbein).

https://doi.org/10.1016/j.orl.2018.04.003
0167-6377/

can be plugged more easily into larger optimization models that may be modeling, say, company-wide costs and revenue. Which of these three models is most useful depends on the particular application.

Several researchers have also pointed out the need for a more effective methodology for setting initial staffing levels given the high level of uncertainty in call forecasts, the correlation in arrival rates across periods within the same day, and the opportunities for updating staffing levels in future periods based on early period call volumes. In this paper, we offer a solution to address this gap in the existing literature. Specifically, in this paper, we seek to optimize staffing decisions over two stages while taking into account random arrival rates, cross-period arrival rate correlations, and opportunities for intra-day staffing adjustments.

2. Two-stage staffing problem with given first-stage staffing decision

We consider the problem of staffing a call center with a single class of customers and a pool of homogeneous service agents. The system manager operates under a quality-of-service (QoS) constraint, which can be quite general. The queueing model we use to represent such a service staffing problem is an $M/M/c$ model. Hence we assume Poisson arrivals and i.i.d. exponential service times. We further assume the system we study has a stochastic *arrival rate*. That is, we assume that arrivals to the system occur according to a doubly stochastic Poisson process.

The most general problem we solve is framed as follows. Consider operating a call center over two time periods, and assume that: (i) the distribution of the arrival rate for the first stage is known or has been previously estimated; (ii) the staffing level for the first-stage, x_1, is selected at the beginning of the first stage; and, (iii) the number of customers who arrive during the first stage, n, is observed. We update the distribution of the arrival rate for the second stage based on n and then pick the staffing level, x_2, for stage two based on the updated distribution. In Section 3 we provide a way to choose both x_1 and x_2. However, in this section we consider a simpler version of the problem. In particular, we change (ii) above and instead assume that the staffing level x_1 is given.

The call center's manager has two competing concerns. First the manager is concerned with the staffing cost for the second stage (we do not consider the cost for the first stage here, since the staffing level for the first stage is given), and hence would tend to hire as few servers in the second stage as possible. Second, the manager is concerned with service quality, which will be poor if an insufficient number of servers are hired. In this section, we use the function $\alpha(x_2, \lambda)$ to represent any quality-of-service metric which depends on x_2 and λ. For example, this function could be the probability that a customer must wait, under a second period staffing level x_2 given arrival rate λ. To streamline mathematical analysis we use continuous extensions of such metrics. In particular, the staffing level is allowed to be continuous. As long as the metric is monotone in the number of servers, an optimal solution to the single-variable problem with a discrete decision variable follows directly from the continuous version of the problem. We use Λ to denote the arrival rate as a random variable, and we use λ to denote a specific realization. Without loss of generality we assume that each server has a service rate of 1.

Let c be the unit staffing cost, c^+ be the unit staffing cost for additional service agents, and c^- be the unit salvage cost for sending unneeded service agents home early. We make the natural assumption that $c^+ > c > c^- > 0$. The QoS constraint is parameterized by ϵ which indicates the required level of service. Let $F_\Lambda(\cdot)$ be the cdf of the random arrival rate Λ. In applied settings, F_Λ is estimated from historical data and represents our initial belief regarding the arrival rate. After observing arrivals in the first stage, this belief (distribution) is updated using the usual Bayesian framework. To facilitate analysis and forecast updating, in the sequel we assume that this prior for the call rate distribution is gamma, as described in Section 3.

The optimization model that minimizes staffing costs subject to the QoS constraint is then:

$$\min_{x_2 \geq 0} \quad cx_1 + c^+(x_2 - x_1)^+ - c^-(x_1 - x_2)^+ \tag{1a}$$

$$\text{s.t.} \quad \int_0^\infty \alpha(x_2, \lambda) dF_\Lambda(\lambda) \leq \epsilon. \tag{1b}$$

The integral in the QoS constraint in (1) simply gives the unconditional value of this QoS metric. This formulation assumes that the QoS metric is a quantity for which smaller values are better (e.g., probability of waiting for service, average waiting time). Of course, other types of QoS metrics can be transformed into this representation.

In order to precisely state our first results, we need some generic, relatively mild assumptions on the QoS metric of interest. First, let $A = \{(x, \lambda) \in \mathbb{R}^2_+ \cap \{x > \lambda > 0\}\}$. In particular, the set A characterizes the stability region of the $M/M/c$ model (recall that each server is assumed to have a service rate of 1). We use A' to denote the complement of A. The following conditions on $\alpha(x, \lambda)$ and F_Λ are used in the sequel:

- **(A1)** $\alpha(x, \lambda) \geq 0$ on $\mathbb{R}^2_+$ and is a continuous function on A. For all $(\bar{x}, \bar{\lambda}) \in A' \cap \mathbb{R}^2_+$, $\alpha(\bar{x}, \bar{\lambda}) = \alpha_{max} \geq 1$. For all $\lambda > 0$, $\lim_{x \searrow \lambda} \alpha(x, \lambda) = \alpha_{max}$.
- **(A2)** On A, $\alpha(x, \lambda)$ is strictly decreasing in x for all $\lambda > 0$ and strictly increasing in λ for all $x > 0$.
- **(A3)** $\lim_{x \to \infty} \int_0^\infty \alpha(x, \lambda) dF_\Lambda(\lambda) = 0$.
- **(A4)** On A, $\alpha(x, \lambda)$ is differentiable in λ and $\frac{\partial \alpha(x,\lambda)}{\partial \lambda}$ is strictly decreasing in x for all $\lambda > 0$.
- **(A5)** $\lim_{x \to \infty} \alpha(x, \lambda) = 0$ for all $\lambda > 0$, and $\lim_{\lambda \to 0} \alpha(x, \lambda) = 0$, for all $x > 0$.

The function $\alpha(x, \lambda)$ represents a QoS metric at arrival rate λ when we have x service agents. We associate higher values of the metric with worse states of the system. Assumption **(A1)** implies that the worst value of the metric is assigned when the system is unstable, allowing for the possibility that this worst value is infinite. It also ensures that stable systems can possess arbitrarily poor metric values. Assumption **(A2)** ensures that the metric improves when either the arrival rate is reduced or the number of servers is increased. These first two assumptions hold for a variety of common metrics including probability of delay, utilization, and average queue-length. Assumption **(A3)** is essentially required to ensure that model (1) has a solution for positive values of ϵ. If F_Λ has bounded support, then this condition holds by applying **(A5)** and dominated convergence. In fact, in the case of the average queue-length metric, bounded support is necessary. If F_Λ is unbounded with finite mean, as is the case in most of our examples, then $\alpha(x, \lambda) \leq K \cdot \max(1, \lambda)$ and **(A5)** ensure that **(A3)** holds, again by dominated convergence. This Lipschitz-style inequality holds for utilization and metrics which are probabilities.

Although the conditions above were stated in the framework of the $M/M/c$ model, our results extend to other models with a slight modification of the conditions. For example, the $M/M/c+M$ model is stable for all positive values of λ and μ, due to abandonments from the system. In this case, we redefine A to be $\{(x, \lambda) \in \mathbb{R}^2_+\}$ and require that **(A1)**–**(A5)** hold with the redefined set A, except for the third part of **(A1)**, which is no longer necessary. So, for the $M/M/c + M$ queue, the metric of the long-run proportion of customers who abandon the system satisfies these conditions and the results and algorithms in the remainder of the paper also work for this model and metric. Computational results for the $M/M/c + M$ model are summarized in Zan [14].