Computers & Operations Research 65 (2016) 149–162

Contents lists available at ScienceDirect

Computers & Operations Research

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

A branch-and-bound algorithm for shift scheduling with stochastic nonstationary demand

Mieke Defraeye*, Inneke Van Nieuwenhuyse

Research Center for Operations Management, Department of Decision Sciences and Information Management, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium

ARTICLE INFO

Available online 11 July 2015

Keywords:
Time-varying arrival process
Staffing and scheduling
Personnel planning
Capacity analysis
Optimization

ABSTRACT

Many shift scheduling algorithms presume that the staffing levels, required to ensure a target customer service, are known in advance. Determining these staffing requirements is often not straightforward, particularly in systems where the arrival rate fluctuates over the day. We present a branch-and-bound approach to estimate optimal shift schedules in systems with nonstationary stochastic demand and service level constraints. The algorithm is intended for personnel planning in service systems with limited opening hours (such as small call centers, banks, and retail stores). Our computational experiments show that the algorithm is efficient in avoiding regions of the solution space that cannot contain the optimum; moreover, it requires only a limited number of evaluations to encounter the estimated optimum. The quality of the starting solution is not a decisive factor for the algorithm's performance. Finally, by benchmarking our algorithm against two state-of-the-art algorithms, we show that our algorithm is very competitive, as it succeeds in finding a high-quality solution fast (i.e., with a limited number of simulations required in the search phase).

1. Introduction

Many shift scheduling algorithms presume that the staffing levels, required to ensure a target customer service, are known in advance: the shift scheduling step then boils down to fitting the minimum-cost shift schedule to the requirements. Determining these staffing requirements, however, is often challenging, since in many real-life systems demand is stochastic and nonstationary (i.e., the demand rate varies over time). While some authors have tried to ease this problem by using approximations to estimate the service levels (see, e.g., [3]), it is known that the "two-step" approach suffers from a major flaw, as it may result in a suboptimal schedule [18].

This paper presents an *integrated* approach to the shift scheduling problem with nonstationary stochastic demand: different staffing combinations are explored using implicit enumeration, which allows to efficiently estimate the minimum-cost shift schedule subject to a service level constraint (the probability that the customer waiting time violates a critical level should not exceed a user-defined target). The algorithm is flexible in the sense that it does not rely on any specific methodology to evaluate the customer service implied by a given shift schedule. We opted to use simulation in our experiments, because (1) it requires virtually no restrictions on the assumptions regarding arrival and service process, and (2) it allows us to include real-life complexities of which the impact on customer service cannot easily be estimated analytically, such as customer impatience (abandonments) and the exhaustive service policy (which implies that servers work overtime to finish the customer in service at the time their shift ends). The accuracy of the outcome, however, will depend on the number of replications used.

The algorithm specifically targets service systems with limited opening hours (so-called *terminating* systems, see Law and Kelton [22]). As the computational effort required to run the algorithm to completion tends to increase as the solution space grows, the algorithm is especially suited for systems with a small solution space: for instance, systems that require a limited number of workers (such as banks, retail stores, or small call centers), or have a low to moderate load. Nevertheless, our benchmark results show that it remains competitive with alternative algorithms (as the ones proposed by [18] and [2]) even in settings with higher load, as it is able to find a high-quality solution fast; the effort to verify that this solution is indeed the *optimal* solution in these cases will likely be prohibitive, though.

Our algorithm contributes to the existing literature by proposing straightforward, easy-to-implement rules that efficiently explore the solution space (as opposed to the more complex and

* Corresponding author. Tel.: + 32 0 16326617.
E-mail addresses: mieke.defraeye@kuleuven.be (M. Defraeye), inneke.vannieuwenhuyse@kuleuven.be (I. Van Nieuwenhuyse).

http://dx.doi.org/10.1016/j.cor.2015.06.016
0305-0548/

time-consuming approach of Atlason et al. [1,2]). Moreover, as it is simulation-based, the algorithm can more easily handle a variety of assumptions (such as general service and abandonment processes) than the heuristic of Ingolfsson et al. [18], which uses randomization. As revealed by the benchmark results, the best solution found by the branch-and-bound algorithm tends to outperform the results obtained by the algorithms of Ingolfsson et al. [18] and Atlason et al. [2].

Section 2 gives a brief discussion of the related literature. Section 3 presents the formal problem statement. A detailed description of the algorithm is provided in Section 4. Section 5 discusses the computational experiment, and analyzes the impact of the fathoming rules and the initial solution. Section 6 benchmarks the algorithm against the algorithms of Ingolfsson et al. [18] and Atlason et al. [2]. Finally, concluding remarks are provided in Section 7.

2. Related literature

Shift scheduling for systems with nonstationary arrival rates has received relatively limited attention in the academic literature. The two-step approach, which fits minimum-cost shift schedules to predefined staffing requirements, is by far the most common (see Thompson [26,27], Sinreich and Jabali [25], and Izady and Worthington [20], among others). The main problem, however, is that the staffing levels required to ensure a target customer service level are not straightforward to determine. Moreover, the two-step approach may result in suboptimal shift schedules [14–16], because several staffing solutions might exist that lead to shift schedules with substantially varying costs. Alternatively, shift scheduling can be done directly based on the time-varying arrival rates [5,13,16,21]. These approaches avoid the suboptimality that arises by decomposing the problem into two steps. Yet, including quality of service constraints in the shift optimization is not straightforward, hence authors commonly resort to simplifying assumptions (e.g., exponential service and abandonment times).

Our research is closely related to the work of Ingolfsson et al. [16,18] and Atlason et al. [1,2]. These articles suggest algorithms to determine low-cost shift schedules with a service level constraint on customer waiting time. Ingolfsson et al. [16] evaluate schedule performance by numerical integration of the forward differential equations for $M_t/M/s_t$ queues and apply a genetic algorithm to search for good schedules. Ingolfsson et al. [18] apply a heuristic cutting-plane algorithm and use the randomization method for evaluating schedule performance [10,17,19], which is computationally less expensive but yields similar accuracy [17]. Atlason et al. [1,2] suggest a cutting plane method that uses simulation to evaluate customer service, and add cuts based on the estimated (pseudo)gradients of the service level function (the methodology was later adapted by Cezik et al. [6] to optimize staffing in multiskill call centers, adding heuristics to improve the applicability to large-scale instances). This approach requires substantial computational effort. Atlason et al. [2] show that their algorithm converges towards an optimal solution as the number of replications grows large; in contrast, both Ingolfsson et al. [16] and Ingolfsson et al. [18] are heuristic approaches, that do not guarantee an optimal solution.

The approach developed in this paper is easier to implement than the one proposed in Atlason et al. [1,2], but cannot strictly guarantee the optimum (see the discussion in Section 4.2.2). Therefore, we refer to the best solution found as the *estimated optimum*.

3. Problem statement and notations

We focus on a single-stage multiserver $M_t/G/s_t+G$ queue, as depicted in Fig. 1. The current time is represented by t and ranges between 0 and time horizon T (i.e., the opening hours of the service system). Customer arrivals have a time-varying arrival rate λ_t (in our numerical experiments, we assume Poisson arrivals; given that the system is evaluated by simulation, any other type of arrival distribution could have been used). The service process is generally distributed with per server service rate μ (which equals the inverse of the expected service time per unit); the abandonment process is generally distributed with rate θ (the inverse of the expected time-to-abandon).

The main objective is to estimate an optimal shift schedule, such that the target customer service is achieved at minimum cost. The cost is measured in worker hours. In line with the related literature [4,9,18,20], customer service is measured by the *virtual waiting time* W_t at given time instants t, i.e., the waiting time that an infinitely patient (fictive) customer encounters upon arrival at time t [12,23,24]. More formally, let $\mathbf{t}_p = \{0, \Delta_p, 2\Delta_p, \ldots, T-\Delta_p\}$ represent the set of time instants at which performance is evaluated (the notations are illustrated in Fig. 2). We then require the following hard constraint to be met:

$$\Pr(W_t > \tau) \le \alpha \quad \text{for all } t \in \mathbf{t}_p, \tag{1}$$

with τ the maximum allowed waiting time, and α the target probability of excessive waiting. The validity of this constraint is checked by simulation. Note that for $\tau = 0$, Expression (1) corresponds to the delay probability.

Capacity changes can only take place at specific points in time, i.e., at the start of a *staffing interval*. Staffing intervals have length Δ_s. The set of staffing interval indices is $\mathbf{I}_s = \{1, \ldots, I_s\}$ with $I_s \equiv T/\Delta_s$ (see Fig. 2). $\mathbf{t}_s = \{0, \Delta_s, 2\Delta_s, \ldots, T-\Delta_s\}$ contains the staffing interval start times, for all $i \in \mathbf{I}_s$ (with Δ_s an integer multiple of Δ_p, which implies that $\mathbf{t}_s \subseteq \mathbf{t}_p$).

Let vector $\mathbf{s} = \{s_1, \ldots, s_{I_s}\}$ represent the staffing vector, containing the number of workers in each staffing interval. Assume that W different pre-defined shift types exist. For any staffing vector $\mathbf{s}$, the minimum-cost shift solution can be determined by solving the

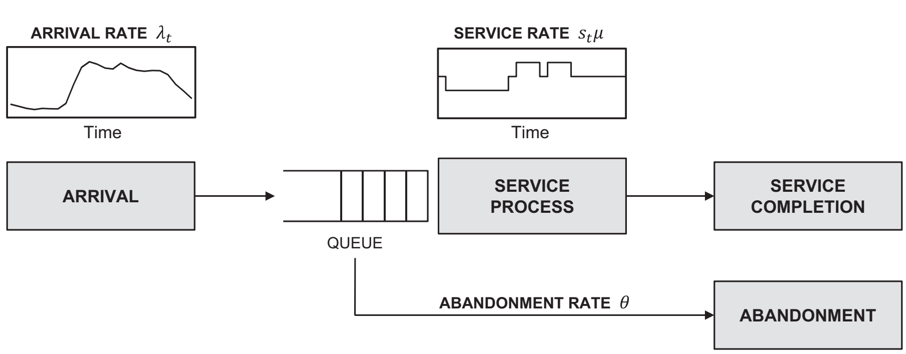

Fig. 1. Schematic representation of a single-stage queueing system with time-varying demand.