Contents lists available at ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

Optimizing daily agent scheduling in a multiskill call center

Athanassios N. Avramidis [a], Wyean Chan [c], Michel Gendreau [b,*], Pierre L'Ecuyer [c], Ornella Pisacane [d]

[a] *School of Mathematics, University of Southampton Highfield, Southampton SO17 1BJ, United Kingdom*
[b] *Département d'informatique et de recherche opérationnelle and CIRRELT Université de Montréal, C.P. 6128, Succ. Centre-Ville Montréal, Québec, Canada H3C 3J7*
[c] *Département d'informatique et de recherche opérationnelle, CIRRELT and GERAD Université de Montréal, C.P. 6128, Succ. Centre-Ville Montréal, Québec, Canada H3C 3J7*
[d] *Dipartimento di Elettronica Informatica e Sistemistica Università della Calabria, Via P. Bucci, 41C Arcavacata di Rende (CS), Italy*

ARTICLE INFO

Article history:
Received 17 November 2007
Accepted 14 January 2009
Available online 6 February 2009

Keywords:
Call center
Stochastic optimization
Staffing
Scheduling
Service level
Cutting plane method

ABSTRACT

We examine and compare simulation-based algorithms for solving the agent scheduling problem in a multiskill call center. This problem consists in minimizing the total costs of agents under constraints on the expected service level per call type, per period, and aggregated. We propose a solution approach that combines simulation with integer or linear programming, with cut generation. In our numerical experiments with realistic problem instances, this approach performs better than all other methods proposed previously for this problem. We also show that the two-step approach, which is the standard method for solving this problem, sometimes yield solutions that are highly suboptimal and inferior to those obtained by our proposed method.

1. Introduction

The telephone call center industry employs millions of people around the world and is fast growing. In the United States, for example, customer service representatives held 2.1 million jobs in 2004, and employment in this job category is expected to increase faster than average at least through 2014 (Bureau of Labor Statistics, 2007). A few percent saving in workforce salaries easily means several million dollars.

Call centers often handle several types of calls distinguished by the required skills for delivering service. Training all agents to handle all call types is not cost-effective. Each agent has a selected number of skills and the agents are distinguished by the set of call types they can handle (also called their *skill set*). When such skill constraints exist, we speak of a *multiskill* call center. *Skill-based routing* (SBR), or simply *routing*, refers to the rules that control the call-to-agent and agent-to-call assignments. Most modern call centers perform skill-based routing (Koole and Mandelbaum, 2002; Gans et al., 2003).

In a typical call center, inbound calls arrive at random according to some complicated stochastic processes, call durations are also random, waiting calls may abandon after a random patience time, some agents may fail to show up to work for any reason, and so on. Based on forecasts of call volumes, call center managers must decide (among other things) how many agents of each type (i.e., skill set) to have in the center at each time of the day, must construct working schedules for the available agents, and must decide on the call routing rules. These decisions are made under a high level of uncertainty. The goal is typically to provide the required quality of service at minimal cost.

The most common measure of quality of service is the *service level* (SL), defined as the long-term fraction of calls whose time in queue is no larger than a given threshold. Frequently, multiple measures of SL are of interest: for a given time period of the day, for a given call type, for a given combination of call type and period, aggregated over the whole day and all call types, and so on. For certain call centers that provide public services, SL constraints are imposed by external authorities, and violations may result in stiff penalties (CRTC (2000)).

In this paper, we assume that we have a detailed stochastic model of the dynamics of the call center for one day of operation. This model specifies the stochastic processes for the call arrivals (these processes are usually non-stationary and doubly stochastic), the distributions of service times and patience times for calls, the call routing rules, the periods of unavailability of agents between calls (e.g., to fill out forms, or to go to the restroom, etc.), and so forth. We formulate a stochastic optimization problem where the objective is to minimize the total cost of agents, under various SL constraints. This could be used in long-term planning, to decide how many agents to hire and for what skills to train them, or for short-term planning, to decide which agents to call for work on a given day and what would be their work schedule. The problem is difficult because for any given fixed staffing of agents (the staffing determines how many agents of each type are available in each time period), no reliable formulas or quick numerical algorithms are available to estimate the SL; it can be estimated accurately only

[*] Corresponding author.
E-mail address: michel.gendreau@cirrelt.ca (M. Gendreau).

0377-2217/$ - see front matter © 2009 Published by Elsevier B.V.
doi:10.1016/j.ejor.2009.01.042

by long (stochastic) simulations. Scheduling problems are in general NP-hard, even in deterministic settings where each solution can be evaluated quickly and exactly. When this evaluation requires costly and noisy simulations, as is the case here, solving the problem exactly is even more difficult and we must settle with methods that are partly heuristic.

Staffing in the *single-skill* case (i.e., single call type and single agent type) has received much attention in the call center literature. Typically, the workload varies considerably during the day (Gans et al., 2003; Avramidis et al., 2004; Brown et al., 2005), and the planned staffing can change only at a few discrete points in time (e.g., at the half hours). It is common to divide the day into several periods during which the staffing is held constant and the arrival rate does not vary much. If the system can be assumed to reach steady-state quickly (relative to the length of the periods), then steady-state queueing models are likely to provide a reasonably good staffing recommendation for each period. For instance, in the presence of abandonments, one can use an Erlang-A formula to determine the minimal number of agents for the required SL in each period (Gans et al., 2003). When that number is large, it is often approximated by the *square root safety staffing formula*, based on the Halfin–Whitt heavy-traffic regime, and which says roughly that the capacity of the system should be equal to the workload plus some safety staffing which is proportional to the square root of the workload (Halfin and Whitt, 1981; Gans et al., 2003). This commonly used heuristic, known as the stationary independent period by period (SIPP) approach, often fails to meet target SL because it neglects the non-stationarity (Green et al., 2003). Non-stationary versions of these approximations have also been developed, still for the single-skill case (Jennings et al., 1996; Green et al., 2003).

Scheduling problems are often solved in two separate steps (Mehrotra, 1997): After an appropriate staffing has been determined for each period in the first step, a minimum-cost set of shifts that covers this staffing requirement can be computed in the second step by solving a linear integer program. However, the constraints on admissible working shifts often force the second step solution to overstaff in some of the periods. This drawback of the *two-step approach* has been pointed out by several authors, who also proposed alternatives (Keith, 1979; Thompson, 1997; Henderson and Mason, 1998; Ingolfsson and Cabral, 2003; Atlason et al., 2004). For example, the SL constraint is often only for the time-aggregated (average) SL over the entire day; in that case, one may often obtain a lower-cost scheduling solution by reducing the minimal staffing in one period and increasing it in another period. Atlason et al. (2004) developed a *simulation-based* methodology to optimize agents' scheduling in the presence of uncertainty and general SL constraints, based on simulation and cutting-plane ideas. Linear inequalities (cuts) are added to an integer program until its optimal solution satisfies the required SL constraints. The SL and the cuts are estimated by simulation.

In the *multiskill case*, the staffing and scheduling problems are more challenging, because the workload can be covered by several possible combinations of skill sets, and the routing rules also have a strong impact on the performance. Staffing a single period in steady-state is already difficult; the Erlang formulas and their approximations (for the SL) no longer apply. Simulation seems to be the only reliable tool to estimate the SL (Cez̧ik and L'Ecuyer, 2008) adapt the simulation-based methodology of Atlason et al. (2004) to the *optimal staffing* of a multiskill call center for a *single period*. They point out difficulties that arise with this methodology and develop heuristics to handle them. Avramidis et al. (2009) solve the same problem by using neighborhood search methods combined with an analytical approximation of SLs, with local improvement via simulation at the end. Pot et al. (2008) impose a constraint only on the aggregate SL (across all call types); they solve Lagrangean relaxations using search methods and analytical approximations.

Some authors have studied the special case where there are only two call types, and some have developed queueing approximations for the case of two call types, via Markov chains and under simplifying assumptions; see Stolletz and Helber (2004) for example. But here we are thinking of 20 to 50 call types or more, which is common in modern call centers, and for which computation via these types of Markov chain models is clearly impractical.

For the *multiskill scheduling problem*, Bhulai et al. (2008) propose a two-step approach in which the first step determines a staffing of each agent type for each period, and the second step computes a schedule by solving an IP in which this staffing is the right-hand side of key constraints. A key feature of the IP model is that the staff-coverage constraints allow *downgrading* an agent into any alternative agent type with smaller skill set, temporarily and separately for each period. Bhulai et al. (2008) recognize that their two-step approach is generally suboptimal and they illustrate this by examples.

In this paper, we propose a simulation-based algorithm for solving the multiskill scheduling problem, and compare it to the approach of Bhulai et al. (2008). This algorithm extends the method of (Cez̧ik and L'Ecuyer, 2008), which solves a single period staffing problem. In contrast to the two-step approach, our method optimizes the staffing and the scheduling simultaneously. Our numerical experiments show that our algorithm provides approximate solutions to large-scale realistic problem instances in reasonable time (a few hours). These solutions are typically better, sometimes by a large margin (depending on the problem), than the best solutions from the two-step approach. We are aware of no competitive faster method.

The remainder of this paper is organized as follows: in Section 2, we formally define the problem at hand and provide a mathematical programming formulation. The new algorithm is described in 3. We report computational results on several test instances in Section 4. The conclusion follows. A preliminary version of this paper was presented at the 2007 Industrial Simulation Conference (Avramidis et al., 2007a).

2. Model formulation

We now provide definitions of the multiskill staffing and scheduling problems. We assume that we have a stochastic model of the call center, under which the mathematical expectations used below are well defined, and that we can simulate the dynamics of the center under this model. Our problem formulations here do not depend on the details of this model.

There are K call types, labeled from 1 to K, and I agent types, labeled from 1 to I. Agent type i has the skill set $S_i \subseteq \{1, \ldots, K\}$. The day is divided into P periods of given length, labeled from 1 to P. The *staffing vector* is $\mathbf{y} = (y_{1,1}, \ldots, y_{1,P}, \ldots, y_{I,1}, \ldots, y_{I,P})^{\mathsf{t}}$ where $y_{i,p}$ is the number of agents of type i available in period p. Given $\mathbf{y}$, the *service level* (SL) in period p for type-k calls is defined as

$$g_{k,p}(\mathbf{y}) = \mathbb{E}[C_{g,k,p}]/\mathbb{E}[C_{k,p} + A_{k,p}],$$

where $\mathbb{E}$ denotes the mathematical expectation, $C_{k,p}$ is the number of type-k calls that arrive in period p and eventually get served, $C_{g,k,p}$ is the number of those calls that get served after waiting at most $\tau_{k,p}$ (a constant called the *acceptable waiting time*), and $A_{k,p}$ is the number of those calls that abandon after waiting at least $\tau_{k,p}$. Aggregate SLs, per call type, per period, and globally, are defined analogously. Given acceptable waiting times τ_p, τ_k, and τ, the aggregate SLs are denoted by $g_p(\mathbf{y})$, $g_k(\mathbf{y})$ and $g(\mathbf{y})$ for period p, call type k, and overall, respectively.

A *shift* is a time pattern that specifies the periods in which an agent is available to handle calls. In practice, it is characterized by its *start period* (the period in which the agent starts working), *break periods* (the periods when the agent stops working), and