Chapter 19
Flexible Staffing for Call Centers with Non-stationary Arrival Rates

Alex Roubos, Sandjai Bhulai, and Ger Koole

Abstract We consider a multi-period staffing problem of a single-skill call center. The call center is modeled as a multi-server queue in which the staffing levels can be changed only at specific moments in time. The objective is to set the staffing levels such that a service level constraint is met in the presence of time-varying arrival rates. We develop a Markov decision model to obtain time-dependent staffing levels for both the case where the arrival rate function is known as well as unknown. The characteristics of the optimal policies associated to the two cases are illustrated through a numerical study based on real-life data. We show that the optimal policies provide a good balance between staffing costs and the penalty probability for not meeting the service level.

Key words: Call centers, Markov decision processes, Staffing, Time-varying arrival rates

19.1 Introduction

Call centers have become the central focus of many companies, as these centers stay in direct contact with the firm's customers and form an integral part of their customer relationship management. Running a successful call center operation means managing by the numbers. One of the most important numbers in call centers is the

A. Roubos (✉)
CCmath, H.J.E. Wenckebachweg 48, 1096 AN Amsterdam, The Netherlands
e-mail: alex@ccmath.com

S. Bhulai • G. Koole
Faculty of Sciences, Vrije Universiteit Amsterdam, De Boelelaan 1081a,
1081 HV Amsterdam, The Netherlands
e-mail: s.bhulai@vu.nl; ger.koole@vu.nl

R.J. Boucherie, N.M. van Dijk (eds.), *Markov Decision Processes in Practice*,
International Series in Operations Research & Management Science 248,
DOI 10.1007/978-3-319-47766-4_19

number of agents serving incoming calls at each moment of time. Since more than two-thirds of the operating costs can be attributed to personnel, getting the right number of agents in place is critical in terms of both the offered service and the operating costs. This agent staffing problem is a complex problem in which many issues have to be taken into account, e.g., demand forecasting, variability in the call arrival patterns, quality of service, and flexibility of the workforce. We refer the reader to the comprehensive surveys in [1, 12].

In this chapter, we consider the staffing problem in a single-skill call center for a given working day. The inherent randomness in the call center, due to variability in the duration of the calls and fluctuations in the call arrival rates, makes the staff problem complex. The randomness is the root cause of deviations of the performance measures from the predicted values at the moment of planning, see, e.g., [4, 16, 17, 19, 24, 31]. Traditionally, most call center literature assume known and constant mean arrival rates, mainly for the purpose of tractability. However, in addition to the usual uncertainty that is intrinsic to stochastic modeling, real call center data shows that there is also uncertainty in the process parameters. Since most performance indicators are sensitive to fluctuations in the parameters [18], both types of uncertainty should be accounted for in any staffing algorithm.

A substantial body of the literature has focused on the probability distribution of the arrival rates from a statistical perspective, see, e.g., [2, 4, 8, 9, 22, 26, 27, 29, 30]. These papers mostly deal with modeling the time-varying arrival process such that the essential features of call center arrivals are captured, e.g., a variance larger than the mean for the number of arrivals, a time-varying arrival intensity, and nonzero correlation between arrival counts in different periods.

Staffing in the presence of time-varying arrival processes was analyzed first by using the pointwise stationary approximation (PSA), see, e.g., [13, 15, 16], in which it is assumed that the arrival rates are known, deterministic, and non-stationary. However, the PSA does not explicitly consider non-stationary behavior that may be induced by abrupt changes in the arrival rate, and it appears to perform less well in these cases. Further numerical methods have been studied by Yoo [32], Ingolfsson et al. [18] and Feldman et al. [11]. The first two are based on methods that solve the Chapman-Kolmogorov forward equations by using small, discrete intervals to approximate the continuously varying parameter. The latter is based on an iterative simulation-based staffing method to achieve time-stable performance.

The case of unknown non-stationary arrival rates has been studied by Jongbloed and Koole [20], Steckley et al. [28], Harrison and Zeevi [17], Whitt [31], Robbins [24] and Liao et al. [21]. The first paper mainly focuses on the characterization of the uncertainty by providing bounds on the number of agents needed. The second paper studies the impact of different performance measures under this uncertainty. The next two papers focus on fluid approximations to determine the number of agents. The next paper uses simulation to derive the number of agents, whereas the last paper uses a robust programming formulation. Characteristic to these papers is that they study the staffing problem under uncertainty using a fixed staffing approach.

In contrast to these papers, in our problem there exists some flexibility to change the number of agents at fixed moments of the working day and thus creates more flexibility.

We model the arrival process as a non-stationary stochastic process with uncertain rates. Moreover, as is common in call centers, the call center operates under a service level that constrains the waiting time for incoming calls. The distinguishing feature of our model is that the staffing levels can only be changed at specific moments of the day, but still have to respect the service level constraint. We assume that there is a fixed number of employees with a permanent contract and a number of flexible agents that can be changed throughout the day on specific moments. The costs of using an agent differs between the fixed and flexible agents. The objective is to find the optimal staffing level that minimizes the total call center operating cost while meeting the service level constraint. We develop a Markov decision model that determines the optimal agent staffing policies in case the arrival rate function is both known and unknown. We conduct a numerical study in order to illustrate the main characteristics of the optimal solutions corresponding to these approaches. In the numerical illustration, we use real call center data and show how the optimal policy balances the staffing costs and the penalty probability for not reaching the service level. Furthermore, we show how the number of periods in which the staffing level can be changed affects the staffing costs.

The paper that is closest to our model is [22]. In this paper intraday updates of the call arrival rate are also allowed. The updates are based on the cumulative number of actual arrivals and the cumulative number of expected arrivals. The ratio between these numbers is used to adjust the forecast of the next intervals. Based on the update, the new staffing levels in each period are updated using the stationary independent period by period (SIPP) approach [14]. Moreover, the performance measure used in the paper is the expected service level. In our work, we do not only look at the expected service level, but at the whole distribution of the service level, and incur a penalty when the service level at the end of the planning period is not above a certain target. Hence, we need to address the effect of a change in the number of agents on the service level at the next decision epoch. Clearly, the SIPP approach is not sufficient to address this issue. Hence, we use a dynamic programming approach to assess this impact. It is this aspect of the problem that distinguishes our model from [22], but also from flexible staffing models in service facilities other than call centers, e.g., [5–7, 10, 23].

The remainder of the chapter is structured as follows. In Sect. 19.2, we describe the call center model under consideration and formulate the associated staffing problem. In Sect. 19.3 we formulate our staffing algorithm for the case of both known and unknown arrival rate functions. In Sect. 19.4, we conduct a numerical study to evaluate the alternative formulations. We illustrate the impact of the number of moments at which the staffing level can be changed, and thus the benefits of flexibility in the call center. The chapter ends in Sect. 19.5 with concluding remarks and highlights some future research.

19.2 Problem Formulation

Consider a call center to which customers arrive according to a non-homogeneous Poisson process with parameter λ_t for $t \geq 0$. We assume that the call center has s_t fixed permanent agents and f_t flexible agents at each time t and only N workplaces available so that $s_t + f_t \leq N$ for all $t \geq 0$. If upon arrival of a new customer at time t no agent out of the $s_t + f_t$ agents is available, then the customer joins an infinite buffer. In the other case, the customer is directly taken into service by an idle agent and has an exponentially distributed service duration with parameter μ_t. Queued customers are served in a first-come first-served order.

The objective of the call center manager is to meet a service level requirement by varying the number of flexible agents over the day. More precisely, divide the length of the day into m smaller intervals, each of length θ. We assume that the arrival rate function λ_t is constant over each interval and unknown. Hence, we also take s_t to be constant over each interval. Let SL_i represent the realized service level over interval $i = 1, \ldots, m$, given by the fraction of customers that has waited less than the acceptable waiting time τ upon starting service within that interval. The requirement of the call center is that SL, the service level over the whole day, is at least α, where SL is computed by the average of the SL_i's weighted by the arrival counts in each interval. The decision variable of the call center manager to achieve this requirement is the variable f_t that can only be changed at epochs determined at the start of certain intervals, namely at the start of interval $t \in \mathcal{T} = \{1, \kappa + 1, 2\kappa + 1, \ldots, m - \kappa + 1\}$, where κ is a divisor of m. Hence, the variable f_t is fixed for a longer period than λ_t and s_t, and needs to take into account the variability inherent to these variables. This is especially challenging since the arrival rate function is not known.

The problem as described above is common in call centers. It is not realistic to assume that the number of flexible agents are changed continuously over time. The assumption that λ_t is constant over small time intervals is also not unrealistic, since this is usually the result of data estimation procedures that reliably approximate the true arrival rate function when the interval length is small. We assume that the permanent agents have a cost c_1 per unit of time for each agent, and that the flexible agents cost c_2 per unit of time for each agent, with $c_2 > c_1$. Note that for any given staffing policy, one cannot guarantee that $\mathrm{SL} \geq \alpha$ is always met at the end of the day, due to randomness. When the service level at the end of the day is not met, we impose that the call center manager incurs a penalty P. We model this service level constraint as a soft constraint. With these additional cost definitions the problem under study becomes

$$\min \sum_{i=1}^{m} (c_1 s_i \theta + c_2 f_i \theta) + P \mathbb{1}_{\{\mathrm{SL} < \alpha\}}$$

subject to

$$\begin{aligned} &f_t = f_{t+1} = \cdots = f_{t+\kappa-1}, && \forall t \in \mathcal{T}, \\ &s_t + f_t \leq N, && t = 1,\ldots,m, \\ &f_t \in \mathbb{N}_0, && t = 1,\ldots,m. \end{aligned}$$

19.3 Solution Approach

In order to solve the call center staffing problem, we cast the problem as a finite-horizon Markov decision problem on epochs $\mathcal{T}$. However, several simplifying approximations are required for purposes of implementation. We refer to Appendix for an exact formulation that solves the problem theoretically.

Let $\mathcal{X}$ denote the state space, where at epoch $t \in \mathcal{T}$ the state $x_t \in \mathcal{X}$ denotes the service level realized up to epoch t, i.e., $x_t = \sum_{i=1}^{t-1} \tilde{\lambda}_i \mathrm{SL}_i / \sum_{i=1}^{t-1} \tilde{\lambda}_i$ for $t \in \mathcal{T}$, where $\tilde{\lambda}_i$ is the value of λ_i derived from the observed arrival counts. Normally, the state space would be modeled by $[0,1]$, however, we discretize the state space to $\mathcal{X} = \{0, 1/\omega, 2/\omega, \ldots, 1\}$, where the parameter ω controls how well the continuous state space is approximated. The realized service level at each epoch is rounded down to the nearest value in the new state space.

Let the action space be denoted by $\mathcal{A}_t = \{0,\ldots,N - \bar{s}_t\}$, where $\bar{s}_t = \max\{s_t, s_{t+1}, \ldots, s_{t+\kappa-1}\}$. Action $a_t \in \mathcal{A}_t$ means that the call center manager schedules $a_t = a_{t+1} = \cdots = a_{t+\kappa-1}$ flexible agents at epoch t after observing x_t.

Note that the definition of the state space is such that the Markov property does not hold. Therefore, it is impossible to give exact transition probabilities. Given that the service level x_t is known, we simulate the system to obtain the service level $x_{t+\kappa}$, given that $s_t + a_t, \ldots, s_{t+\kappa-1} + a_{t+\kappa-1}$ agents are available. We assume that at the beginning of each interval t, the system with arrival rate λ_t, service rate μ_t and $s_t + a_t$ agents has reached stationarity, which is not an unrealistic assumption when changes in the dynamics are not too severe. Starting from a steady-state situation, we apply simulations for the duration of an interval to obtain the service level distribution. Then, by convoluting this distribution over the κ intervals, we derive the distribution for the next epoch. This approach has the advantage that we can simulate the transition probabilities up front for each combination of x_t and a_t. Hence, we can store a table with combinations of x_t and a_t that give $p_t(x_t, a_t, x_{t+\kappa})$, i.e., the probability of moving from state x_t to $x_{t+\kappa}$ when action a_t is chosen.

Finally, the direct costs are given by

$$c_t(x_t, a_t) = \sum_{i=t}^{t+\kappa-1} \left(c_1 s_i \theta + c_2 a_i \theta + P \mathbb{1}_{\{i=m\}} \mathbb{1}_{\{\mathrm{SL} < \alpha\}}\right).$$

The first and second terms are related to the staffing of permanent and flexible agents. The last term corresponds to the penalty P that is incurred if the service level at the end of the day is not met.

The tuple $(\mathcal{X},\mathcal{A},p,c)$ completely describes the Markov decision process for this problem.

Note that in the problem above, the values of s_t are given. However, in practice, the values for s_t would be obtained by having an estimate of the values of λ_t for the specific day. This would typically be done in light of long-term personnel planning by using the Erlang C formula. The decision variable a_t can then be seen as short-term planning that adjusts for deviations on this estimate. It is worthwhile to mention that the use of the Erlang C formula for deriving values for s_t is not optimal in general (see Sect. 19.4 for some examples), but provides a good starting point for the staffing problem at hand.

In the description of our algorithm, we mention that at epoch t we define the state x_t as the realized service level up to epoch t. This service level is easy to compute, since the arrival counts in intervals $i < t$ are known. At epoch t we also need the values λ_i for $i \geq t$ to determine the optimal actions. However, these values are unknown and need to be obtained via an estimation procedure. Note that this estimation procedure can be different than the procedure used to determine the values of s_t, since the realized values up to epoch t can be used as well and provide better information on the future values of λ_t. Examples of estimation procedures can be found in, e.g., [2, 4, 27].

19.4 Numerical Experiments

In this section we show the characteristics of the optimal policies by means of numerical experiments. The parameters of the experiments are based on real-life data, or otherwise chosen to represent parameters that can be found in practice. We start with an example that demonstrates the benefits of the flexibility in staffing. This example assumes a known and constant arrival rate.

Remark 19.1. In order to evaluate the optimal policies, we apply independent simulations. We mainly focus on two performance measures: the total staffing costs and the probability that a penalty is incurred if at the end of the day the service level is lower than the target. Because the penalty probability is extremely small, many simulations are necessary to obtain an accurate estimate, see, e.g., the topic of rare-event simulation in [3]. For instance, for a probability $p = 0.01$ and a 95% confidence interval with half-width equal to $0.1p$, the number of simulations should be at least $n = 40{,}000$. We perform $n = 1{,}000{,}000$ simulations, which can be calculated within a few seconds. We present the half-width of the 95% confidence intervals between parentheses, but omit confidence intervals for values that have a negligible half-width.

19.4.1 Constant Arrival Rate

Consider a call center with no flexibility, i.e., only a fixed number of agents are scheduled for the whole time horizon. The arrival rate is $\lambda = 3$ per minute and the service rate is $\mu = 0.2$ per minute. The acceptable waiting time is $\tau = 1/3$ min (20 s). Based on these parameters, the Erlang C formula tells us that we need $s = 19$ agents in order to meet the 80% service level target. Each agent costs $c_1 = 1$ unit per minute. Suppose the call center operates for a time horizon of $T = 720$ minutes (12 h). The call center starts empty, and waiting customers at the end of the time horizon are ignored.

With these parameters the costs for staffing are $C = 1 \times 19 \times 720 = 13{,}680$. However, despite the fact that the expected service level is above 80% as predicted by the Erlang C formula, simulations show that the realized service level is in many cases below 80%. With probability 0.34 the service level falls below the required target and hence a penalty is incurred. This phenomenon that the service level is not always reached in a finite-time horizon is discussed in [25]. One way to deal with this problem, without flexibility, is to try a higher staffing level. With $s = 20$ the costs for staffing are increased to $C = 14{,}400$. Furthermore, the probability that a penalty is incurred is reduced to only 0.03.

We now allow flexible agents. We have a base staffing level of $s = 19$ for the whole time horizon. Each 30 min there is the opportunity to add additional flexible agents (i.e., $m = 24$, $\theta = 30$ and $\kappa = 1$), against a cost of $c_2 = 1.2$ units per minute per agent. When choosing to do so, the extra agents are immediately available. We take $N = 30$, which is sufficiently large for this example. Furthermore, we incur a sufficiently high penalty of $P = 10^6$ for failing to meet the target service level at the end of the day. The state space is discretized according to $\omega = 400$. The transition probabilities $p_t(x_t, a_t, x_{t+\kappa})$ are determined from 10,000 sample path simulations for each action a_t. By application of our method we find the optimal policy. Evaluation by means of simulations shows that the costs for staffing are $C = 14{,}192$ and that the probability of a penalty is 0.0018 (8.4×10^{-5}). This is a considerable improvement in both staffing costs and penalty probability compared to staffing 20 agents with no flexibility.

The optimal policy is displayed in the left plot in Fig. 19.1. This figure should be interpreted as follows. For a decision moment at time t, and given a realized service level up to t, the number of flexible agents staffed is given by the figure. The optimal policy suggests that in some cases 30 agents should be scheduled (a base staffing level of 19 plus 11 flexible agents). With that many agents the probability to reach a service level of 1 in a 30-minute interval is already 0.996. The large white area below the curve denotes pairs of time epochs and service levels that always result in an expected service level lower than the target. Hence, no flexible agents are staffed.

The right plot in Fig. 19.1 shows boxplots of the service level at a decision epoch. The small circles are outliers, where an outlier is a data value more extreme than 1.5 times the interquartile range from the box. This figure shows that the average service

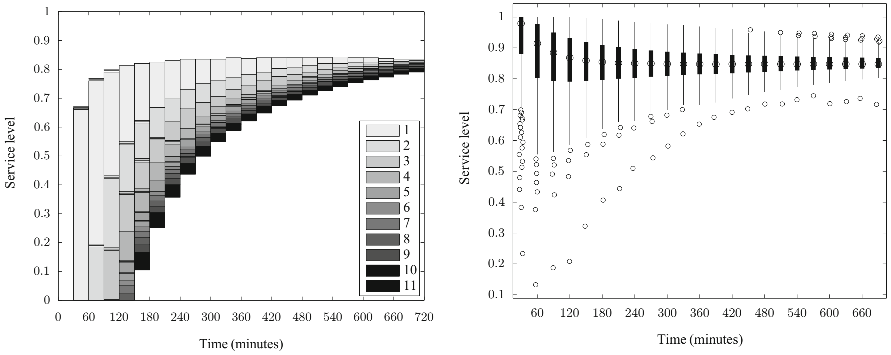

Fig. 19.1: *Left plot*: the optimal policy for the example with a constant arrival rate. *Right plot*: boxplots of the service level at a decision epoch

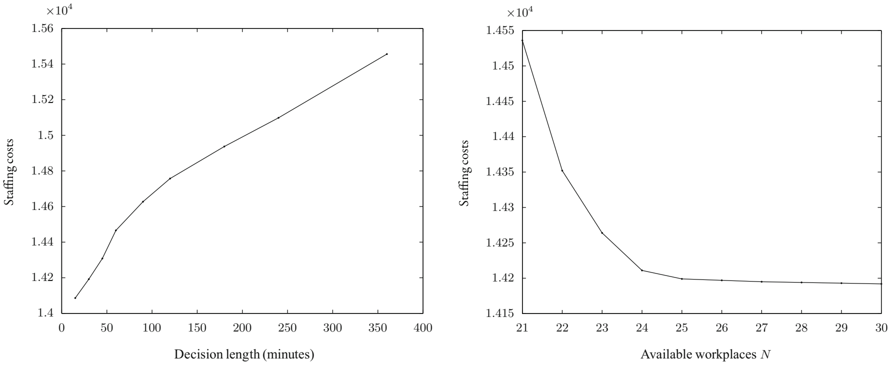

Fig. 19.2: Value of flexibility

level nicely converges to approximately 0.85. Moreover, the variability greatly diminishes over time, and outliers are becoming more sparse.

19.4.1.1 Value of Flexibility

It is clear that staffing only a few flexible agents at the right moments keeps both the staffing costs and the penalty probability low. In the previous example we could vary the number of staffed flexible agents each 30 min. Allowing flexibility on this time scale might not be possible for all call centers. Also, the optimal policy was not restricted by the number of available workplaces. Therefore, we are interested in the effect of different levels of flexibility on the performance measures.

Figure 19.2 shows how the staffing costs depend on the frequency with which decisions can be made and on the number of available workplaces. The staffing costs in the left plot increase up to a maximum of $C = 15{,}840$, which is attained at staffing

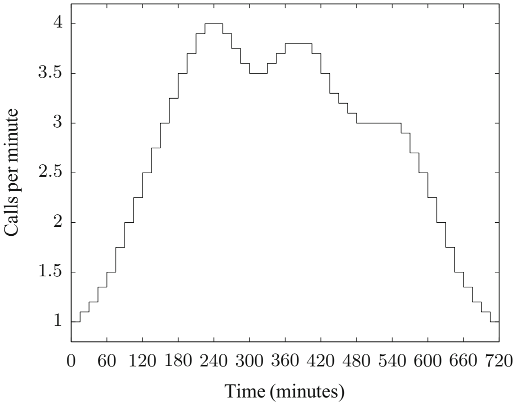

Fig. 19.3: The time-dependent arrival rate

22 agents for the whole day of 720 min. This plot shows that large improvements can be obtained if the call center can react on a small time scale. But there is also a significant gain if the call center can only adjust the staffing level once a day, after 360 min, since this reduces the staffing costs to $C = 15{,}456$. The right plot shows that most of the improvement comes from the first few flexible agents. The plot starts from a minimum of 21 workplaces, because with only 20 available workplaces there will always be a considerable penalty probability of 3%. With 21 workplaces, the staffing costs will be relatively high compared to a larger number of available workplaces, since the flexible agents are almost always used.

19.4.2 Time-Dependent Arrival Rate

We now consider a call center with a time-dependent arrival rate. We still assume that the arrival rate is known. In Fig. 19.3 the typical pattern of arrivals over the day is depicted. Here we model the arrival rate as a piecewise constant function, where each interval equals 15 min. All other parameters related to the model remain the same. Based on the stationary Erlang C formula, we find the base staffing level in each interval such that the target will be met. These staffing levels have the same shape as the arrival rate. Performance assessment concludes that with no flexibility the staffing costs are $C = 12{,}855$ and that the probability of failing to meet the target service level at the end of the day is 0.18. When staffing one agent more in each interval, the penalty probability is reduced to 0.01, but the staffing costs are then $C = 13{,}575$.

With the opportunity to add flexible agents we can improve this situation. We assume that decisions about flexible agents can only be made each consecutive 30 min, and that we have a limited number of available workplaces of $N = 30$. This implies, e.g., that we can only choose up to 5 flexible agents in the time periods $[210,240)$

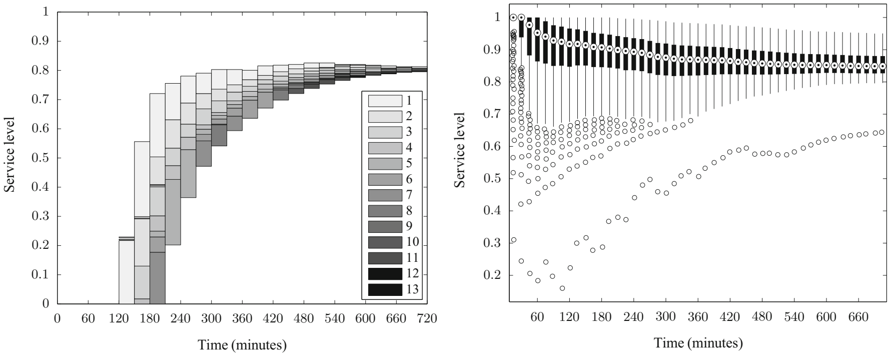

Fig. 19.4: *Left plot*: the optimal policy for the example with a time-dependent arrival rate. *Right plot*: boxplots of the service level at a staffing period

and [240, 270), since there are already 25 permanent agents scheduled at [225, 255). All other parameters related to our method and performance assessment remain the same. When we apply our method we find the optimal policy as shown in Fig. 19.4. The corresponding costs for staffing are $C = 13{,}101$ and the probability of a penalty is 0.0062 (1.5×10^{-4}). Again, this is a considerable improvement. It is astonishing to notice that this performance can be achieved by requiring on average 6.8 flexible agents at the right moments, which corresponds to only 3.4 agent hours.

The optimal policy reveals a very interesting characteristic. Until 120 min, no flexible agents are needed at all. This is, of course, due to the low arrival rates at the beginning of the day, which means that the realized service level up to 120 min is not that important. Consequently, this provides an excellent opportunity to better estimate the arrival rate in the remainder of the day, in case the arrival rate is unknown.

The right plot in Fig. 19.4 shows boxplots of the service level at the beginning of each 15 min staffing period. Most notably from this figure is that the whiskers extending from the bottom of the boxes are becoming shorter. There are hardly any realized service levels below the 80% service level target at the end of the day, which demonstrates that our method works well for this example.

19.4.2.1 Optimal Permanent Agents

Although the scope of this chapter is on flexible agents, we do make a short remark about the choice of the permanent agents. The flexibility of adding agents at the decision epochs can, and should, be taken into account when making a long-term planning of the permanent agents. Our method can also be used to do this. Consider for example the following heuristic approach. Start with the vector s such that $s_i = \min\{k \in \mathbb{N} \mid k > \lambda_i/\mu_i\}$ for $i = 1, \ldots, m$. That is, the number of agents in each interval is higher than the offered load, but as small as possible. Apply our

Table 19.1: The number of permanent agents

Base	8, 9, 9, 10, 11, 12, 14, 15, 17, 18, 19, 21, 22, 23, 24, 25,
	25, 24, 23, 23, 22, 22, 23, 23, 24, 24, 24, 23, 22, 21, 20, 20,
	19, 19, 19, 19, 19, 19, 18, 17, 15, 14, 12, 11, 10, 9, 9, 8
Optimal	7, 7, 8, 9, 10, 11, 13, 15, 16, 17, 19, 21, 23, 24, 25, 25,
	25, 25, 24, 23, 22, 22, 23, 23, 23, 23, 24, 23, 22, 21, 20, 19,
	18, 18, 18, 18, 19, 18, 16, 15, 13, 11, 9, 8, 7, 6, 6, 6

method to find the policy $\pi(s)$ and the corresponding costs $C^{\pi(s)}$. The next step consists of adding an additional permanent agent to exactly one interval, namely the interval that will result in the largest decline in costs. Let $s+e_i$ denote the vector with an additional agent at interval i, and let $j = \arg\min_{i=1,\ldots,m} C^{\pi(s+e_i)}$. Then, if $C^{\pi(s+e_j)} < C^{\pi(s)}$, update s to $s+e_j$. Continue this iteration until no improvement can be found anymore.

We use this heuristic approach on the previous example with the time-dependent arrival rate. All parameters remain the same, with the exception of P. We increased the penalty to $P = 10^{12}$ in order to keep the penalty probability low. We find the optimal policy similar to the one in Fig. 19.4. However, due to an overall decrease in the number of permanent agents, the staffing costs turn out to be much lower. They are $C = 12{,}935$, and the penalty probability is 0.0099 (1.9×10^{-4}). More flexible agents are used now, in order to reach the target service level. On average 18.5 flexible agents are needed for specific 30-minute intervals per day.

In Table 19.1 the number of permanent agents are given, for each interval of 15 min. This table compares the heuristic optimal staffing levels with the base staffing levels, where in each interval the target service level will be met according to the Erlang C formula. The optimal levels are for the most part lower, indicating that the availability of flexible agents is better utilized when necessary. Staffing is higher in five intervals with a high arrival rate. This ensures a higher expected service level in these intervals, and possibly compensating for other intervals of lesser importance. It is interesting to note that the last couple of intervals are really understaffed. This is due to the fact that the average service level can only be changed very limitedly near the end of the day.

19.4.3 Unknown Arrival Rate

In most practical situations the real arrival rate λ_t is not known. What is available is a best estimate $\hat{\lambda}_t$ that is estimated or forecast from historical data. It goes without saying that if this estimate is accurate ($\hat{\lambda}_t$ is close to λ_t) our method works well, in the sense that the service level requirement will always be achieved, because it reduces to the case of a known arrival rate. What we are interested in is the performance in case the arrival rate estimate is inaccurate.

As more information becomes available over the course of the day, our algorithm updates the arrival rate estimate. In practice this can be done quite accurately, since a large database with historical arrival rates are available, and sophisticated updating procedures can be used (see, e.g., [27]). However, what we will show is that even with no knowledge of previous arrival rates, and therefore using a very basic updating method, our algorithm works just as well. The updating method we consider is the historical proportion method [27], which works as follows. At decision epoch $t \in \mathcal{T}$ calculate the ratio R between the realized and estimated arrival rate up to t, i.e., $R = \sum_{i=1}^{t-1} \tilde{\lambda}_i / \sum_{i=1}^{t-1} \hat{\lambda}_i$. Then, update the estimate for the remainder of the day: $\hat{\lambda}_i' = R\hat{\lambda}_i$, $i = t, \ldots, m$. This new estimate, together with the realized arrival rate, is then used to give an updated optimal policy.

As a result of this updating procedure, we need to evaluate the optimal policy multiple times per day. This computation takes roughly 1 min to carry out, namely we update 24 times a day for a calculation that runs for approximately a few seconds. Hence, an accurate evaluation by means of extensive simulations becomes hardly doable if $n = 1{,}000{,}000$ (see also Remark 19.1). Therefore, we have to settle for less accuracy in our simulations with $n = 1{,}000$. Also, the state space is now discretized according to $\omega = 200$. As in the examples before, we take $N = 30$, a penalty of $P = 10^6$ and allow flexible agents each consecutive 30 min.

In our experiments, we consider several cases with respect to the pattern of the arrival rate. In the first example the estimated arrival rate is the real arrival rate multiplied by a constant scalar, $\hat{\lambda}_t = \lambda_t \cdot \beta$. In the second and third examples we correctly estimate the arrival rate pattern, but we make a fixed under- or overestimation, $\hat{\lambda}_t = \lambda_t + \beta$, with $\beta = -0.5$ and $\beta = 0.5$. Finally, the fourth and fifth examples are examples with a wrongly estimated pattern, $\hat{\lambda}_t = \lambda_t \cdot \beta_t$, with $\beta_t = 1 - 0.005t$ and $\beta_t = 1 + 0.005t$. That is, the estimate becomes increasingly more wrong. In all examples, the true arrival rate is the one shown in Fig. 19.3.

For a fair comparison between the performance of the different examples we use the same number of permanent agents in each interval across the examples, which is the number determined by the Erlang C formula using λ_t, $\mu = 0.2$, $\tau = 1/3$ and $\alpha = 0.8$ in each interval (i.e., the base staffing levels). A reason against using the Erlang C formula with $\hat{\lambda}_t$ is that in the overestimated situations the server costs would be high and the penalty probabilities low, even without using flexible agents. Moreover, from the previous examples we have seen that the base staffing levels do require flexible agents in order to balance the server costs and the penalty probability.

The results of the experiments are shown in Table 19.2. The results for the first example are independent of β, because the β disappears in the updated estimate after the first epoch. As the day progresses, the estimate for the remainder of the day naturally becomes more accurate. Hence, this example can be seen in light of the previous example with a known and time-dependent arrival rate, though with more uncertainty. The results are also very similar. The underestimation of the arrival rate in the second example actually becomes an overestimation, because of the updating

Table 19.2: Results of experiments with an unknown and time-varying arrival rate

Example	Service level	Server costs	Penalty probability
1	0.853 (0.002)	13,100 (24)	0.024 (0.009)
2	0.874 (0.002)	13,903 (31)	0.018 (0.008)
3	0.855 (0.002)	13,140 (30)	0.007 (0.005)
4	0.851 (0.002)	13,067 (28)	0.031 (0.011)
5	0.865 (0.002)	13,245 (25)	0.004 (0.004)

method. Therefore, more flexible agents are used resulting in higher server costs, a higher service level and a decrease in penalty probability. The third example is exactly the opposite, in the sense that for most intervals the arrival rate will be underestimated. However, an overestimation will still happen in the last ten intervals. This example shows that our method works by adapting only when the service level is too low. That the penalty probability is low, is due to the overestimation in the intervals at the end. Examples four and five show results as could be expected for under- and overestimated arrival rates. That the penalty probability is not equal to zero is again due to the approximate transition probabilities.

19.5 Conclusion and Discussion

In this chapter we have shown that significant improvements can be obtained by introducing flexible agents. The improvements are expressed in the form of lower staffing costs or a lower probability of failing to meet the service level target at the end of the day, compared to the traditional approach that does not exploit this flexibility. Numerical experiments showed that our approach works remarkably well, even in the case of an unknown and time-varying arrival rate, with a forecast that is not necessarily accurate.

We model the call center as a Markov decision process in a non-traditional manner where our state variable denotes the service level as opposed to the number of customers in the system. The transition probabilities are, due to the complexities of calculating them exactly, obtained via simulations. This allows us to look further than (non-homogeneous) Poisson arrivals and exponential service times. As more information becomes available over the course of the day, we make use of a better estimated arrival rate to update the optimal policy. In the same way, we can also update the service time distribution. The case of agent absenteeism (e.g., a permanent agent is scheduled to work, but did not show up) is easily handled by decreasing the number of permanent agents s_t. The absent agent will be taken care of by a flexible agent, if that turns out to be necessary.

Our approach is highly relevant to call center practice. Uncertainty in call arrivals demands flexibility from a call center to guarantee good performance without incurring excessive staffing costs. In practice, many call centers indeed have this flexibility. Flexibility in the workforce is achieved by, e.g., managers that help answering telephone calls during busy periods, or due to people that are flexible in their working hours, and can be requested to work on an ad-hoc basis with flexible contracts, such as students and agents that work from home. Additional flexibility can be obtained at the moment a shift of an agent ends and that agent can be requested to work overtime. This is practically relevant, since we observe that the demand for flexible agents increases at the end of the day, see Fig. 19.4. The algorithm in this chapter exploits this flexibility in call centers in an easily implementable fashion, and therefore has the potential to be integrated in workforce management software of call centers.

Appendix: Exact Solution

In this section, we formulate a discrete-time Markov decision problem for our original continuous-time problem. We only discretize time into small intervals, but make no other approximation. Hence, the formulation is nearly exact for small time intervals, and thus computes nearly optimal policies for the original problem. We denote the length of a time interval by $1/\eta$, thus every $1/\eta$ time units the system is observed.

In order to model the transitions of the system after each observation, we need a large state space that contains all information to calculate the next state. Hence, define the state space $\mathcal{X}$ to consist of tuples $(n, s_c, s_d, m, z, w_1, \dots, w_n)$. In this tuple, $n \in \mathbb{N}_0$ denotes the number of customers in the system at the time of an observation. The realized waiting times of each of the n customers at the moment of observation is given by $w_1, \dots, w_n \in \mathbb{R}_0^+$. We will adopt the convention that customers in service have a waiting time of 0. Further, let $s_c \in \mathbb{N}_0$ denote the number of servers currently in use, and $s_d \in \mathbb{N}_0$ the number of servers that is desired to have. The service level can be computed by the ratio of $z \in \mathbb{N}_0$, the number of customers served within τ time units, and $m \in \mathbb{N}_0$, the number of customers served. These variables are sufficient to model the state transitions in a Markovian way. Hence, the dynamic programming backward recursion formula becomes

$$
\begin{aligned}
\eta V_{k+1}(n, s_c, s_d, m, z, w_1, \dots, w_n) &= c_1 s_k + c_2(s_c - s_k) \\
&+ \lambda \mathbb{1}_{\{s_c<n\}} H_k(n+1, s_c, s_d, m, z, w_1, \dots, w_{s_c}, w_{s_c+1} + 1/\eta, \dots, w_n + 1/\eta, 0) \\
&+ \lambda \mathbb{1}_{\{s_c=n\}} H_k(n+1, s_c, s_d, m, z, w_1, \dots, w_n, 0) \\
&+ \lambda \mathbb{1}_{\{s_c>n\}} H_k(n+1, s_c, s_d, m+1, z+1, w_1, \dots, w_n, 0) \\
&+ \mu s_c \mathbb{1}_{\{s_c=s_d\}} \mathbb{1}_{\{s_c<n\}} \big[H_k(n-1, s_c, s_d, m+1, z + \mathbb{1}_{\{w_{s_c+1}+1/\eta<\tau\}}, \\
&\quad w_2, \dots, w_{s_c}, 0, w_{s_c+2} + 1/\eta, \dots, w_n + 1/\eta)\big] \\
&+ \mu s_c \mathbb{1}_{\{s_c>s_d\}} \mathbb{1}_{\{s_c<n\}} \big[H_k(n-1, s_c - 1, s_d, m, z, \\
&\quad w_2, \dots, w_{s_c}, w_{s_c+1} + 1/\eta, \dots, w_n + 1/\eta)\big]
\end{aligned}
$$

$$
\begin{aligned}
&+\mu n \mathbb{1}_{\{s_c=s_d\}}\mathbb{1}_{\{s_c\geq n\}}\left[H_k(n-1,s_c,s_d,m,z,w_2,\ldots,w_n)\right]\\
&+\mu n \mathbb{1}_{\{s_c>s_d\}}\mathbb{1}_{\{s_c\geq n\}}\left[H_k(n-1,s_c-1,s_d,m,z,w_2,\ldots,w_n)\right]\\
&+\left(\eta-\lambda-\min\{n,s_c\}\mu\right)\left[\mathbb{1}_{\{s_c\geq n\}}H_k(n,s_c,s_d,m,z,w_1,\ldots,w_n)\right.\\
&\left.+\,\mathbb{1}_{\{s_c<n\}}H_k(n,s_c,s_d,m,z,w_1,\ldots,w_{s_c},w_{s_c+1}+1/\eta,\ldots,w_n+1/\eta)\right].
\end{aligned}
$$

The index k counts the number of intervals to go until the end of the complete period, the last interval. The first two terms describe the cost of using s_k permanent and $s_c - s_k$ flexible agents. If upon arrival, the number of servers currently in use is less than n, then there are $n - s_c$ customers in the queue. Hence, these customers add $1/\eta$ time units to their waiting time (term 3). If $s_c = n$, then everyone is in service, and the arriving customer has to wait (term 4). If $s_c > n$, then there are idle servers. Hence, an arriving customer is served immediately and satisfies the service level directly as well (term 5). The next two terms, terms 6 and 7, model the case where a customer leaves the system when there are customers waiting in the queue. The first case is where the number of servers currently in use is equal to the desired number. Hence, $1/\eta$ is added to the waiting times and s_c remains unchanged. When the customer is taken into service, then the service level is also adjusted. The second case is when s_c is higher than s_d, then additionally s_c is decreased by one and no customer is taken into service (thus, the service level is not updated either). Terms 8 and 9 model a similar situation, however, in this case there are a sufficient number of servers available so that no customer is waiting. Hence, the waiting times are not adjusted and neither is the service level. The final terms deal with the similar cases in which no event occurs within the interval. Hence, only the waiting times are updated when $s_c < n$.

In the dynamic programming backward recursion formula, we have adopted the notation H for the action operator. This action operator is equal to V for all intervals k that are not a decision epoch, i.e., $k \notin \mathcal{T}$. However, for all $k \in \mathcal{T}$, we have that

$$
\begin{aligned}
H_k(n,s_c,s_d,m,z,w_1,\ldots,w_n) = \min_{l\in\mathbb{N}_0}\Big\{&\{V_k(n,s_c,l,m,z,w_1,\ldots,w_n)\,|\,l<s_c\}\\
&\cup\{V_k(n,l,l,m,z,w_1,\ldots,w_n)\,|\,l\geq s_c\}\Big\}.
\end{aligned}
$$

The first set in the minimization models the case in which the number of servers is decreased, hence s_d is adjusted. The second set models the case in which the number of servers is increased. Since this happens immediately, both s_c and s_d are set to the desired level.

Finally, we finish the model by describing what happens at the last interval. In this case, we can evaluate the realized service level and compare it to α. If the service level is not met, then a penalty of P is incurred, and otherwise no additional cost is incurred. This is given by the following equation.

$$
\eta V_0(n,s_c,s_d,m,z,w_1,\ldots,w_n) = c_1 s_k + c_2(s_c - s_k) + P\mathbb{1}_{\{z/m<\alpha\}}.
$$

References

1. O.Z. Akşin, M. Armony, V. Mehrotra, The modern call center: a multi-disciplinary perspective on operations management research. Prod. Oper. Manag. **16**(6), 665–688 (2007)
2. S. Aldor-Noiman, P.D. Feigin, A. Mandelbaum, Workload forecasting for a call center: methodology and a case study. Ann. Appl. Stat. **3**(4), 1403–1447 (2009)
3. S. Asmussen, P.W. Glynn, *Stochastic Simulation: Algorithms and Analysis* (Springer, New York, 2007)
4. A.N. Avramidis, A. Deslauriers, P. L'Ecuyer, Modeling daily arrivals to a telephone call center. Manag. Sci. **50**(7), 896–908 (2004)
5. J. Bard, H. Purnomo, Short-term nurse scheduling in response to daily fluctuations in supply and demand. Health Care Manag. Sci. **8**(4), 315–324 (2005)
6. R. Batta, O. Berman, Q. Wang, Balancing staffing and switching costs in a service center with flexible servers. Eur. J. Oper. Res. **177**(2), 924–938 (2007)
7. O. Berman, R.C. Larson, A queueing control model for retail services having back room operations and cross-trained workers. Comput. Oper. Res. **31**(2), 201–222 (2004)
8. L. Brown, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Multi-factor Poisson and gamma-Poisson models for call center arrival times. Working Paper, 2004
9. L.D. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao, Statistical analysis of a telephone call center: a queueing-science perspective. J. Am. Stat. Assoc. **100**(469), 36–50 (2005)
10. F.F. Easton, J.C. Goodale, Schedule recovery: unplanned absences in service operations. Decis. Sci. **36**(3), 459–488 (2005)
11. Z. Feldman, A. Mandelbaum, W.A. Massey, W. Whitt, Staffing of time-varying queues to achieve time-stable performance. Manag. Sci. **54**(2), 324–338 (2008)
12. N. Gans, G.M. Koole, A. Mandelbaum, Telephone call centers: tutorial, review, and research prospects. Manuf. Serv. Oper. Manag. **5**(2), 79–141 (2003)
13. L.V. Green, P.J. Kolesar, The pointwise stationary approximation for queues with nonstationary arrivals. Manag. Sci. **37**(1), 84–97 (1991)
14. L.V. Green, P.J. Kolesar, J. Soares, Improving the SIPP approach for staffing service systems that have cyclic demands. Oper. Res. **49**(4), 549–564 (2001)
15. L.V. Green, P.J. Kolesar, J. Soares, An improved heuristic for staffing telephone call centers with limited operating hours. Prod. Oper. Manag. **12**(1), 46–61 (2003)
16. L.V. Green, P.J. Kolesar, W. Whitt, Coping with time-varying demand when setting staffing requirements for a service system. Prod. Oper. Manag. **16**(1), 13–39 (2007)
17. J. Harrison, A. Zeevi, A method for staffing large call centers based on stochastic fluid models. Manuf. Serv. Oper. Manag. **7**(1), 20–36 (2005)

18. A. Ingolfsson, E. Akhmetshina, S. Budge, Y. Li, X. Wu, A survey and experimental comparison of service-level-approximation methods for nonstationary $M(t)/M/s(t)$ queueing systems with exhaustive discipline. INFORMS J. Comput. **19**(2), 201–214 (2007)
19. T. Jiménez, G.M. Koole, Scaling and comparison of fluid limits of queues applied to call centers with time-varying parameters. OR Spectr. **26**(3), 413–422 (2004)
20. G. Jongbloed, G.M. Koole, Managing uncertainty in call centers using Poisson mixtures. Appl. Stoch. Model. Bus. Ind. **17**(4), 307–318 (2001)
21. S. Liao, G.M. Koole, C. van Delft, O. Jouini, Staffing a call center with uncertain non-stationary arrival rate and flexibility. OR Spectr. **34**, 1–31 (2012)
22. V. Mehrotra, O. Ozlük, R. Saltzman, Intelligent procedures for intra-day updating of call center agent schedules. Prod. Oper. Manag. **19**(3), 353–367 (2010)
23. E.J. Pinker, R.C. Larson, Optimizing the use of contingent labor when demand is uncertain. Eur. J. Oper. Res. **144**(1), 39–55 (2003)
24. T. Robbins, Managing service capacity under uncertainty. Ph.D. Thesis, Penn State University, 2007
25. A. Roubos, G.M. Koole, R. Stolletz, Service level variability of inbound call centers. Manuf. Serv. Oper. Manag. **14**(3), 402–413 (2012)
26. H. Shen, J.Z. Huang, Forecasting time series of inhomogeneous Poisson processes with application to call center workforce management. Ann. Appl. Stat. **2**(2), 601–623 (2008)
27. H. Shen, J.Z. Huang, Interday forecasting and intraday updating of call center arrivals. Manuf. Serv. Oper. Manag. **10**(3), 391–410 (2008)
28. S.G. Steckley, S.G. Henderson, V. Mehrotra, Service system planning in the presence of a random arrival rate. Working Paper, 2004
29. J.W. Taylor, A comparison of univariate time series methods for forecasting intraday arrivals at call a center. Manag. Sci. **54**(2), 253–265 (2008)
30. J. Weinberg, L.D. Brown, J.R. Stroud, Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. J. Am. Stat. Assoc. **102**(480), 1186–1199 (2007)
31. W. Whitt, Staffing a call center with uncertain arrival rate and absenteeism. Prod. Oper. Manag. **15**(1), 88–102 (2006)
32. J. Yoo, Queueing models for staffing service operations. Ph.D. Thesis, University of Maryland, 1996