The Staffing Requirements with Time-Varying Demand and Customer Abandonment in Call Centers

Feng CHENG \(^1\) and Jiazhen HUO \(^1\)

\(^1\) School of Economics & Management, Tongji University, Shanghai, 200092, China

Abstract: This paper considers staffing requirements problem in service systems especial call centers with random cyclic demands and customers' abandonment behavior by using stationary models. The queueing model can be denoted as, where arrival rates are sinusoidal, staffing requirements are time-varying and customers may abandon service if the offered waits exceed the customers' patience time. We use "stationary independent period by period" (SIPP) approach to setting staffing requirements during each planning period such as hours, half-hours, etc. The service levels are measured by customer abandonment probability and delay probability of customers. In order to find the minimum staffing needed in each period to achieve the service levels, we first obtain the minimum staffing by SIPP which satisfy the target delay probability and abandon probability respectively, then set the maximal of the two values as our objective staffing. Numerical analysis shows that in this way staffing requirements are reliable.

Key Words: Call centers; Time-varying demand; Customer abandonment; SIPP; Numerical analysis

1. Introduction

A call center consists of equipment and people capable of delivering products or services to customers by telephone and other communication tools. It is a bridge between industries and customers. See more information about call centers can reference Gans et al.[1]. While call center managers' objective is to reduce operation costs with keeping acceptable service levels. Because operators' (agents) costs are dominant in operation costs, call center managers should decide available number of operators to satisfy service levels. It is common to use Erlang delay models, which capture the uncertainty associated with arrivals and service times. These models usually assume that the key parameters such as arrival rates and service rates are constant. So a series of stationary queueing models which are often denoted as M/\(M/s\) models are constructed. Generally, in practice, customer demands for service are not only uncertainty captured by stochastic-process variability, but also uncertainty about the model parameters, usually vary with time. Figure 1 shows a telephone call arrival curve derived from actual data provided by a mobile communication company in China. We can see that from 08:00 to 22:00 customer demands are high, reach peak at 12:00 and 20:00. But from 02:00 to 06:00 customer demands are relatively low. And an arriving customer may abandon the service if his waiting time exceeds his patient time, otherwise he will get service. Then accordance with the former, we can denote them as \(M_t/M/s_t + M\) queuing model. Although the stationary queueing models may not suitable, we can reference this approach and make some necessary changes if necessary.

Corresponding Author: Feng Cheng
School of Economics & Management, Tongji University, Shanghai, 200092, China
cf06401005@163.com
(Received October 20, 2012)
(Revised December 20, 2012)
(Accepted January 30, 2013)

Fig. 1 A mobile communication company in China: income calls by half hour

Setting staffing requirement is one in a hierarchy of decisions that must be made in the design and management of a service system. Usually the daily staffing requirements should decide the number of agents needed to work during each staffing interval (planning period) over the day. The daily staffing might be an 8-hour day, a full 24-hour day and other continually available (24/7) call centers, or something in between. The staffing interval might be short (15-30 minutes), or medium (1 hour), or long (2 hours or more than that). Usually in call centers the staffing intervals are half an hour or one hour. Call center as a service system is to pursue service level with minimal cost. So call center managers/ market should balance between costs and service levels. There may be two situations: (1) the arrival rate may be lower than expected, resulting in operators being idle and hence, unnecessary costs, or (2) the arrival rate may be higher than expected, resulting long waiting times and hence, inability to meet the call center service-level goals. We
usually characterize the service level by (1) the probability of no delay, (2) the average waiting time, (3) the average number of customers abandon the queue and (4) the average number of throughput customers. As in Mandelbaum and Zelty[2] said the most popular measure of operational (positive) performance is the fraction of served customers that have been waiting less than some given time. However, if forced into choosing a single number as a proxy for operational performance, he recommends the probability to abandon. In this paper we consider delay probability as well as abandonment probability as main target service levels in our model. Because in call centers the operators’ (agents) costs are dominant in operation costs, call center managers should decide available number of operators to satisfy service levels.

A significant feature of our model is customer abandonment \( P(Ab) \). First, in Brown et al. [3] it is said that abandonment is important to be included in call-center models, because it is realistic: customers abandon if they must wait too long before starting service. Whitt[4] shows that abandonment feature plays an important role in analysis of model-parameter uncertainty. Without abandonment (and with unlimited waiting space), the queueing model will be unstable for a certain range of parameter values, in particular, when the input rate exceeds the maximum possible output rate, for some model parameters a proper steady-state distribution will not exist, and the long-run average cost is infinite. Assuming that instability can occur with positive probability, the overall average cost is necessary infinite as well. However, with abandonment, a proper steady-state distribution always should be finite. The abandonment ensures that we can average with respect to the probability distribution on the model parameters. More generally, abandonment tends to make the performance more robust in the model elements, as well as more realistic. For additional discussion about customer abandonment in queues, see Garnett et al.[5], Zohar et al.[6], Zelty and Mandelbaum[7].

Because customer abandonment is very important and modern call centers usually have management software to control the abandonment probability. It is common that place bound of abandonment probability to 5%. When arrival rate is constant, the queue model is \( M/M/s + M \), Palm[8] give the useful formulae for the steady-state distribution and some performance measures. So even the arrival rate is time-varying, it may be possible to use stationary models to determine staffing requirements. That is, chop time into segments and use a stationary model in each segment. Green, Kolesar, and Soares[9] refer to this as the stationary independent period-by-period approach(SIPP). In fact the SIPP method is common in industry, Green and Kolesar[10], Kwan et al.[11] study the conditions when such stationary models of nonstationary environment is reasonable. And Sze[12] and Thompson[13] recognize the potential problem of using a stationary model for a system with time-varying arrivals. In order to study the relationship between stationary and nonstationary models, Green and Kolesar[14][15][16] have series researches about it. Green, Kolesar, and Soares[9] identify how the system parameters affected SIPP reliability and determine specific situations in which SIPP is safe to use and those in which it is not. But they do not consider the customer abandonment.

In this paper we consider whether the abandonment behavior can influence the accuracy of SIPP if it is not safe in the \( M/M/s \) model. Numerical analysis shows abandonment probabilities exceed aimed value (usually 5%) if we only set delay probability as target service level. So in order to find the minimum staffing needed in the period to achieve the two targets, we first obtain the minimum staffing by SIPP which satisfy the target delay probability and abandon probability respectively, then set the maximal of the two values as our objective staffing

\[
\begin{align*}
\lambda(t) &= \lambda + A \sin(2\pi/24) \\
A &= A > 0 \text{ is the amplitude} \\
\mu &= \text{the service rate} \\
s_r &= \text{the number of servers on duty at time} \\
\theta &= \text{the abandonment rate} \\
PP &= \text{the planning period}
\end{align*}
\]

2. Model and Methodology

We study queueing system \( M/M/s + M \) with arrival rate \( \lambda(t) \), and the arrival rate at time \( t \) given by

\[
\lambda(t) = \lambda + A \sin(2\pi/24)
\]

In the following we give notations and their meanings,

\[
\begin{align*}
\lambda_t &= \text{arrival rate at time } t \\
A &= \text{the average arrival rate over the period} \\
A &= A > 0 \text{ is the amplitude} \\
\mu &= \text{the service rate} \\
s_r &= \text{the number of servers on duty at time} \\
\theta &= \text{the abandonment rate} \\
PP &= \text{the planning period}
\end{align*}
\]

We set the period of the sine function at 24 hours because of the many practical applications in which a daily cycle is evident.

From Palm[8] we know that for the \( M/M/n + M \) queue model (arrival rate is \( \lambda \), service rate is \( \mu \), abandonment rate is \( \theta \), the number of agents is constant as \( n \)), the delay probability and abandonment probability are in the following forms,

\[
\begin{align*}
P(W > 0) &= \frac{A(\frac{\mu}{\lambda} \cdot \frac{1}{\nu})}{1 + (A(\frac{\mu}{\lambda} \cdot \frac{1}{\nu}) - 1) \cdot E_{1,n}} \\
P(\text{Ab}|W > 0) &= \frac{1}{\rho(\frac{\mu}{\lambda} \cdot \frac{1}{\nu}) + 1 - \frac{1}{\rho}}
\end{align*}
\]

where

\[
\begin{align*}
A(x,y) &= \frac{x e^y}{y^x} \\
\gamma(x,y) &= \int_0^\gamma x^{s-1} e^{-t} ds, x > 0, y \geq 0 \\
E_{1,n} &= \frac{T_{\mu} \mu^t}{n!} \\
\rho &= \frac{\lambda}{n \mu}
\end{align*}
\]

Because in our model the arrive rate is not constant but time-varying, we should chop the 24hours into 24/PP periods, and average the arrival rate in each period, that is arrival rates are,

\[
\lambda_m = \int_{mPP}^{(m+1)PP} \lambda(t) dt, m = 0, 1, 2, \ldots, 24/PP
\]

thus in these periods we consider the arrival rates are constant.
We use $\Lambda_{\text{in}}$ instead of $\lambda$ in the above formulae in each period. Then analyze sequence for each scenario as the Green et al. [8], but there are something different, that is, we choose both delay and abandonment probability as our targets to set staffing requirements.

The analytic sequence for each scenario is as follow:

(a) Fix the scenario’s exogenous parameters: $A$, the mean arrival rate $\mu$, the service rate; and $RA = A/\lambda$, the relative amplitude. $\theta$, the abandonment rate; $PP$, the length of the planning period; and set the target probability of delay $r$, usually is 20%.

(b) Divide the cycle into nonoverlapping intervals of length $PP$, starting at $t = 0$. For each planning interval computes the average arrival rate by integrating Equation over the planning interval, i.e. formula (4). Then use this average arrival rate, the service rate, and iterative version of the Erlang-A equation to find minimal staffing needed in the interval to achieve the target delay probability. This produces a vector of staffing levels $\{s(t), t = 1, ..., 24/PP\}$.

(c) Use staffing levels $\{s(t), t = 1, ..., 24/PP\}$ to compute the abandonment probabilities in each interval by Erlang-A formula.

(d) For each planning interval computes the average arrival rate by integrating Equation over the planning interval. Then use this average arrival rate, the service rate, and iterative version of the Erlang-A equation to find minimal staffing needed in the interval to achieve the target abandonment probability (we set as 5%). This produces a vector of staffing levels $\{\bar{s}(t), t = 1, ..., 24/PP\}$.

(e) Use staffing levels $\{\bar{s}(t), t = 1, ..., 24/PP\}$ to compute the delay probabilities in each interval by Erlang-A formula.

(f) The staffing requirement in each interval is the maximal number of $s(t)$ and $\bar{s}(t)$.

We can also check the instantaneous delay probabilities if we only set delay probabilities as our target service level. Let $p_n(t)$ be the periodic steady-state probability that $n$ customers are in the system at time $t$, then the set of differential equations characterize the systems is as follow,

$$p_n'(t) = -\lambda(t)p_n(t) + \mu p_{n+1}(t)$$

if $0 \leq n < s(t)$ then

$$p_n'(t) = \lambda(t)p_{n-1}(t) + (n + 1)\mu p_{n+1}(t) - (\lambda(t) + n\mu)p_n(t)$$

if $n \geq s(t)$ then

$$p_n'(t) = \lambda(t)p_{n-1}(t) + [s(t)\mu + (n - s(t))\theta]p_{n+1}(t) - (\lambda(t) + s(t)\mu)p_n(t)$$

(5)

Let $p_n(t)$ be the instantaneous delay probabilities that a customer arriving at time $t$ is delayed. This is also the probabilities that all servers are busy at epoch $t$ and is given by,

$$p_D(t) = 1 - \sum_{n=0}^{s(t)-1} p_n(t)$$

(6)

After step (b), we get staffing levels $\{s(t), t = 1, ..., 24/PP\}$, thus we can calculate $P_D(t)$ and observe the instantaneous delay probabilities.

In the following we will fix all parameters except $n$, we consider staring at a low $\theta = 2$, doubling up to 64($\theta = 2, 4, 8, 16, 32, 64$), and test how they influence the staffing levels.

3. Numerical Analysis

3.1 An Example and its interpretation

Considering an incoming call center that is open 24 hours a day. The planning period is one-hour and set a service target of 20% probability of delay. Other parameters are as follow: $\lambda = 256, \mu = 16, \theta = 4, RA = 1, PP = 1, r = 0.2$. The standard SIPP method for this system suggests staffing levels (show in figure 2, the green line) that satisfy the target delay probability in each period.

However, by formulae (5) and (6) the peak instantaneous probability of delay is over 0.44, and the service target is exceeded in 16 of the 48 half-hour planning periods of the day, see figure 3. Thus, the actual performance of this system when staffing according to SIPP will be worse than desired.

In another way, after decide the staffing level in each period, we calculate the abandonment probability as show in figure...
4. The standard SIPP method for this system suggests staffing levels that result in an actual 24-hour average probability of abandonment is 0.062. However, the most peak abandonment probability over the planning period is 0.221. From 14:00 to 22:00 the probabilities exceed the target. There is one interpretation for this. Arrival rates decrease to the lowest from 14:00 to 18:00, then increase from the lowest from 18:00 to 22:00. That means during these times amount of customer are few, so one or two customers abandon will result in high abandonment probability. From figure 4 we can see that it is worse than we desired if we consider delay probability only.

So we also need to consider abandonment probability. We set a service target of 5% probability of abandonment. The standard SIPP method for this system suggests staffing levels (show in figure 2, the red line) that satisfy the target abandonment probability in each period. The standard SIPP method for this system suggests staffing levels that result in an actual 24-hour average probability of delay is 0.227. However, the most peak abandonment probability over the planning period is 0.477. From 0:00 to 11:00 the probabilities exceed the target. Arrival rates are relative higher during these intervals, and reach peak at 06:00, although the service level can satisfy the target of abandonment probability, many of them should wait in the queue. From figure 5 we can see that it is worse than we desired if we consider abandonment probability only.

If we consider both delay probability and abandonment probability as our service levels, how can we decide the staffing levels in each time period. An approach is to calculate staffing levels respectively in the above and make the maximal one as our final staffing level. Thus both targets can be satisfied. See figure 6.

For we set the abandonment probability as the main target, and the staffing levels can satisfy average delay probabilities, even though the instantaneous probability of delay is not satisfied as we expected, we think this is not important.

3.2 Other Tests

In practice there are many conditions accordance to different θ s, such as in banks companies θ may be 2, but in hospital emergency departments θ may be 60.

In this part we considered abandonment rates starting at a low of θ = 2, that is with the average expecting time as long as 30 minutes, doubling up to 64(θ = 2, 4, 8, 16, 32, 64), and other parameters are fixed as in 3.1. We can see how θ influence the staffing levels.

If we set delay probability as our target service level and get staffing levels, then abandonment probabilities in each interval are show in figure 7. It is obviously that abandonment probabilities increase as abandonment rates which stand for customers' impatient in systems. Large abandonment rates correspond to low impatient, while small ones to high impatient.

If we set abandonment probability as our target service level and get staffing levels, then delay probabilities in each interval are show in figure 8. It is obviously that delay probabilities increase as abandonment rates decrease. Low abandonment rates stand for high patience, thus more and more customer join the queue which leads to more customer delay in the systems.

Finally we give the staffing level satisfies each target service level in table 1 and we choose the large one in black bold as our desire staffing level which is the maximal one. From it we can see that if θ ≤ 4 most of staffing requirement in delay probability column can be chosen as our desired staffing, while
in $\theta \geq 8$ situations are inverse.

4. Conclusion and Future Research

In this paper we consider the staffing requirements in $M_r/M/1 + M$ queue models which used in several different service systems. We chop the 24 hour of a day into 24 time intervals (that means the planning period is an hour) and use the SIPP approach to deal with time-varying demand. We choose delay probability and abandonment probability as our target service levels and calculate staffing levels in each period respectively if consider only one target. As in the above sections show, it is not suitable just consider one target. Because if we choose one of the targets as the main target, the other one will not satisfy which will influence the accuracy of staffing deeply. The best way is to consider both target and set the maximal staffing number as our desired one. And numerical results show this dealing approach is very suitable. Because different abandonment rates will influence staffing levels we also consider the effect of abandon rate on staffing levels. For small abandonment rates, the delay probability will domin the abandonment probability, otherwise, the abandonment probability will domain. Although the findings described in this paper are based on sinusoidal arrival rates, we believe it is very useful in many applications. The shortage is we just give numerical analysis but some theoretical explanations, this is the next step of our research. However our findings suggest the following guidelines for 24-hour service system systems:

1) If SIPP average approach fails in $M_r/M/1$ models with delay probability as its target service level, it also fails in $M_r/M/1 + M$ model, but we can consider abandonment probability as its target service level, even consider both.

| $\theta = 2$ | $\theta = 4$ | $\theta = 8$ |
|--------------|--------------|--------------|
| Staffing levels: $P_D \leq 0.05$ | Staffing levels: $P_{A, B} \leq 0.05$ | Staffing levels: $P_D \leq 0.05$ | Staffing levels: $P_{A, B} \leq 0.05$ | Staffing levels: $P_D \leq 0.05$ | Staffing levels: $P_{A, B} \leq 0.05$ |
| 0.2 | 0.2 | 0.2 |
| 23 | 28 | 32 |
| 23 | 28 | 32 |
| 23 | 28 | 32 |
| 23 | 28 | 32 |
| 23 | 28 | 32 |
| 23 | 28 | 32 |
| 23 | 28 | 32 |
| 23 | 28 | 32 |
| 23 | 28 | 32 |
| 23 | 28 | 32 |
| 23 | 28 | 32 |
| 23 | 28 | 32 |
| 23 | 28 | 32 |
| 23 | 28 | 32 |
| 23 | 28 | 32 |
| 23 | 28 | 32 |
| 23 | 28 | 32 |
| 23 | 28 | 32 |
| 23 | 28 | 32 |
| 23 | 28 | 32 |

Table 1 The final staffing levels
(2) Abandonment rates influence staffing level in \( M_t/M_t/s\) + \( M \) models deeply, it is show that if \( \theta \) is relatively small, we can mainly consider delay probability as target service level, otherwise consider abandonment probability.

We do not consider customers’ behavior of retrials, but it is very important consider them in some settings, such as Abdalla and Boucherie[17], Aguri et al.[18].

References
[1] Gans, N. Koole, G. Mandelbaum, A. The telephone call centers: Tutorial, review and research prospects, Manufacturing and Service Operations Management, 5, 2003, pp.79-141.
[2] Mandelbaum A and Zeltyn S. Service Engineering in action: the Palm/Erlang-A queue, with applications to call centers.
[3] Brown, L. D., Gans, N., Mandelbaum, A., Sakov A., Shen H., Zeltyn, S., Zhao, L. Statistical analysis of a telephone call center: A queueing -science perspective, J. Amer. Statist. Assoc, 100(1), 2005, pp.36-50.
[4] Whitt, W. Staffing call center with uncertain arrival rate and absenteeism, Production and operations management, 15(1), 2006, pp.88-102.
[5] Garnett, O., Mandelbaum, A., Reiman, M. Designing a call center with impatient customers, Manufacturing and Service Operations Management, 4(3), 2002, pp.208-227.
[6] Zohar, E., Mandelbaum, A. Adaptive Behavior of impatient customers in Tele-Queues: Theory and Empirical support, Management Science, 48(4), 2002, pp.566-583.
[7] Zeltyn, S., Mandelbaum, A. Call centers with impatient customers: Many-server asymptotics of the \( M_t/M_t/n \) +G queue, Queueing Systems, 51, 2005, pp.361-402.
[8] Palm C. Research on telephone traffic carried by full availability groups, Tele, 1, pp.107.
[9] Green, L. V., Kolesar, P. J., Soares, J. Improving the SIPP approach for staffing service systems that have cyclic demands, Operations Research, 49(4), 2001, pp.549-564.
[10] Green, L. V., Kolesar, P. J. Testing the validity of a queueing model of police patrol, Management Science, 37, 1989, pp.127-148.
[11] Kwan, S. K., Davis, M., Greenwood, A. G. A simulation model for determining variable worker requirements in a service operation with time-dependent customer demand, Queueing Systems, 3, 1988, pp.265-276.
[12] Sze, D. Y. A queueing model for telephone operator staffing, Operations Research, 32, 1984, pp.229-249.
[13] Thompson, G. M. Accounting for the multi-period impact of service when determining employee requirements for labor scheduling, J. Oper. Management, 11, 1993, pp.269-287.
[14] Green, L. V., Kolesar, P. J. The pointwise stationary approximation for queues with nonstationary arrivals, Management Science, 37, 1991, pp.84-97
[15] Green, L. V., Kolesar, P. J. On the accuracy of the simple peak hour approximation for Markovian queues, Management Science, 41, 1995, pp.1353-1370.
[16] Green, L. V., Kolesar, P. J. The lagged PSA for estimating peak congestion in Markovian queues with periodic arrival rates, Management Science, 43, 1997, pp.80-87.
[17] Abdalla, N., Boucherie, N. R. Blocking probabilities in mobile communications networks with time-varying rates and redialing subscribers, Annals of Operations Research, 112(1), 2002, pp.15-22.
[18] Aguir, M. S., Karaesmen, F., Aksin, O. Z, Chauvet, F. The impact of retrials on call center performance, OR Spectrum, 26(3), 2004, pp.353-376.