American Journal of Operations Management and Information Systems
2019; 4(3): 71-79
http://www.sciencepublishinggroup.com/j/ajomis
doi: 10.11648/j.ajomis.20190403.11
ISSN: 2578-8302 (Print); ISSN: 2578-8310 (Online)

SciencePG
Science Publishing Group

Determining Trunk Lines in Call Centers with Nonstationary Arrivals and Lognormal Service Times

Siddharth Mahajan

Production and Operations Management Area, Indian Institute of Management, Bangalore, India

Email address:
s_mahajan100@yahoo.co.in

To cite this article:
Siddharth Mahajan. Determining Trunk Lines in Call Centers with Nonstationary Arrivals and Lognormal Service Times. *American Journal of Operations Management and Information Systems*. Vol. 4, No. 3, 2019, pp. 71-79. doi: 10.11648/j.ajomis.20190403.11

Received: June 26, 2019; **Accepted**: July 18, 2019; **Published**: August 5, 2019

Abstract: Two important resources in a call center are the number of staff and the number of trunk lines required. In this paper, we focus on the decision of the number of trunk lines that a call center should have. The current practice is to use the Erlang B or the *M/M/s/0* queueing model which assumes Poisson arrivals, exponential service times, *s* servers and no places in queue, i.e. no customers can wait. In this paper, we improve on the state of practice in determining the required number of trunk lines, by including two realistic features present in call centers. The first realistic feature is to consider nonstationarity of arrivals. The second feature is to consider the lognormal service time distribution instead of the exponential distribution. There is extensive empirical evidence for both features. In order to carry out our computations we use the results of a paper by Massey and Whitt, Operations Research, 44(6), 1996. We have two main findings. Firstly, we find numerically that in our nonstationary Erlang loss model, $M_t/G/s/0$, an insensitivity result holds. The blocking probability of arrivals at the call center depends only on the mean of the lognormal service time distribution and not on its variance. Our second finding is that current practice is quite robust. In particular, we find the number of trunk lines required using a stationary Poisson approximation. This approximation assumes stationary Poisson arrivals with an appropriately chosen arrival rate and exponential service times. The approximation does quite well in predicting the number of trunk lines required.

Keywords: Queueing, OR in Service Industries, Call Centers, Nonstationary Arrivals, Lognormal Distribution

1. Introduction

Two important resources in a call center are the number of staff and the number of trunk lines required to connect calls to the call center. The first resource is certainly very important since it accounts for a major portion of the costs incurred in operating a call center. However the second resource, the number of trunk lines, has also got to be considered. The number of trunks in place determines the routing of calls to the staff and therefore has an important part to play in the realization of service level measures.

There are quite a few service level measures. One is ASA (Average Speed of Answer), i.e. how many seconds does a call have to wait on average before being answered. This includes those calls that do not have to wait at all. A second service level measure is, Delay of Delayed Calls. As mentioned previously, some calls do not have to wait at all. For those that have to wait, what is the average wait? A third measure is Service Level, what x% of calls are answered within a fixed time interval, y seconds. A common Service Level measure in call centers is to have 80% of calls answered within 20 seconds.

Usually, the cost of a trunk line is less than the cost of a single staff member or Customer Service Representative (CSR). If there are too few trunk lines, then of course calls get blocked from entering the call center. But if there are too many trunk lines relative to agents then that also can be a problem. Calls simply get put on hold and customer wait times increase significantly, resulting in bad service levels. Also many times the trunk lines are toll-free lines, so it is the call center instead of the customer, that is paying for the cost of waiting. This can also significantly increase costs for the call center.

Current practice in determining the number of trunk lines in a call center, is to first set a service level target. The service level target here is what percentage of calls can be allowed to be blocked. A typical service level target as mentioned in Reynolds [1] is to allow 2% blocked calls.

Then the Erlang B or the *M/M/s/0* queueing model is used. This model implies Poisson arrivals, exponential service times, *s* servers and no places in queue, i.e., no customers can wait. In this model, servers refer to trunk lines and do not have anything to do with CSRs (or Customer Service Representatives). A formula for the blocking probability in the Erlang B model is used, see for example, Gross et al. [2]. The blocking probability from this formula is equated to the required service level (e.g. 2% blocked calls). Given this, the value of *s*, or the number of trunk lines is found.

For a more extensive discussion of call center trunking requirements, refer Reynolds [1], which is a good reference on call center workforce management.

In this paper, the state of practice is improved, in determining the required number of trunk lines, by including two realistic features present in call centers. The first realistic feature is to consider nonstationarity of arrivals. It has been repeatedly found in call center data that the number of arrivals vary quite significantly over the period of time the call center operates. One such small data set from Reynolds [1], is considered in our analysis here. The second realistic feature is to consider the lognormal distribution as the distribution of service times in the call center, instead of the exponential distribution. Empirical analysis of service time data in call centers has shown that service times follow the lognormal distribution. This evidence is discussed below. So the usual assumption of exponential service times may not be a good one. We next discuss more on these two realistic features.

Nonstationarity of arrivals is highly prevalent in call centers. Green, Kolesar and Whitt [3] plot hourly arrival rates for a financial services call center. There is significant variation in arrivals by time of day. Reynolds [1], reiterates this by mentioning, `The most accurate approach for call center forecasting involves time series analysis, which takes into account both trend and seasonality. It is the approach used in most call centers and serves as the basis for most of the automated workforce management forecasting models'. Brown et al [4] perform extensive statistical analysis of call center data. Their data comprises a complete operational history of a small Israeli banking call center, call by call, over a full year. Their plot of arrivals in calls/hour by time of day shows clear nonstationarity of arrivals.

Next evidence for the lognormal service time distribution is discussed. Firstly, a service time refers to the time spent by a CSR in talking to the customer, the time spent `on hold' while the CSR is processing the customer's request and lastly the time spent after the customer hangs up but while the CSR is still doing work related to the customer's request. Gans et al [5] provide extensive evidence of the lognormal distribution as the distribution of service times in a call center. Confirmations of the lognormal fit are provided in Bolotin [6], Chlebus [7], and Mandelbaum et al [8]. The authors also mention that the lognormal distribution has been found in unpublished call center data of a Dutch bank.

We now quote from Brown et al [4]. `Looking at the figure we see that the distribution of service times is clearly not exponential, as is assumed by standard queueing theory. In fact, after separating the calls with very short service times, our analysis reveals a remarkable fit to the lognormal distribution'. For all these reasons, the lognormal distribution as the service time distribution has been chosen.

The queueing model we consider in the paper, is the nonstationary Erlang loss model, $M_t/G/s/0$, which has nonstationary Poisson arrivals, *s* servers (i.e. trunk lines) in parallel, no extra waiting spaces and i.i.d. service times which follow the lognormal distribution. Massey and Whitt [9] have analyzed this model for a general service time distribution. Their results have been used in this paper to do computations for determining the number of trunk lines required.

Massey and Whitt [9] in their paper, approximate a queueing model with a nonstationary arrival process with a queueing model with a stationary arrival process. The authors consider a fixed time interval and divide it into subintervals. They consider approximations for the blocking probabilities over subintervals by replacing the nonstationary arrival process over the subinterval by a stationary arrival process. The authors act as if the nonstationary M_t arrival process were a stationary *G* arrival process and then try to approximate the stochastic variability. The approximation they propose is based on a heavy traffic peakedness formula. We discuss more on the results from their paper, as we use them in our calculations, in Section 3.

As mentioned before, in this paper we find the number of trunk lines required in call centers using nonstationary arrival data and the lognormal service time distribution. For our analysis, we consider two different lognormal distributions. Both have the same mean, but the second distribution has a variance double that of the first.

We have two main findings. Firstly, we find numerically that in our nonstationary Erlang loss model, an insensitivity result holds. The blocking probability of arrivals at the call center depends only on the mean of the lognormal service time distribution and not on its variance. In particular, both the lognormal service time distributions predict the same requirement of trunk lines.

Our second finding is that current practice is quite robust. In particular, we find the number of trunk lines required using a stationary Poisson approximation. This approximation assumes stationary Poisson arrivals with an appropriately chosen arrival rate and exponential service times (the Erlang B formula). The approximation predicts the same requirement of trunk lines as the original model. That this happens is not too surprising, given a statement made in this regard in the paper by Davis, Massey and Whitt [10]. However, it is still worthwhile to go through the numerical analysis and verify that the original model and the Poisson approximation match quite closely.

Kim and Park [11] consider 'two-stage' call centers where some incoming calls are completed by first service while others require an additional second service. The authors develop an effective outsourcing strategy in 'two-stage' call centers. To this end, they model a 'two-stage' service system and propose several call routing structures. The structures are compared through numerical testing.

Kilincli and Zhang [12] consider staff scheduling in call centers with cross-trained workers. Call centers face demand variations over time across multiple service categories and typically employ a cross-trained workforce with flexible schedules to hedge against these fluctuations. In practice, it is often impossible to cross-train agents in each category, thus partial and limited cross-training are the norm. To solve the problem an integer program that addresses cross-training, shift schedule, days off and break assignments across multiple service categories is proposed. The model is hard to solve and a two-phase sequential approach is developed. Experimental results with data from a call center with nine categories clearly demonstrate the significance of cross-training. The authors find that partial limited cross-training, where 30% of staff is cross-trained with two skills or 10% of staff is cross-trained with three skills, could result in considerable cost saving. However, these savings could diminish quickly with the increase of efficiency loss in secondary skills.

Yu et al [13] consider delay announcements for call centers with hyperexponential patience modelling. Using real call center data, the patience distribution is modeled by the hyperexponential distribution. A state-dependent Markovian approximation is applied for computing abandonment. An empirical study shows good fit to real call center data. Simulation experiments indicate that the model and approximation are reasonable. The authors conclude that the approach could be applied to a voice response system of real call centers.

Li et al [14] state that efficient management of call centers needs accurate modeling of customer waiting behavior. This customer waiting behavior contains important information about customer patience (how long a customer is willing to wait) and service quality (how long a customer needs to wait to get served). The authors develop a two-way functional hazards model to study customer waiting behavior. The authors analyze data from a US Bank call center and provide insights about customer patience and service quality. The findings also provide inputs for call center agent staffing and scheduling.

Bimpikis and Markakis [15] motivated by applications such as call centers and healthcare, consider service systems that process two types of tasks that are unknown beforehand. There are also two kinds of servers with different skillsets. The service provider wants to maximize the systems long term throughput. Given this, what should be the resource allocation policy, i.e. how to assign tasks to severs over time? The authors show that the performance loss of the system due to uncertainty in task types can be significant. The authors also show that one optimal design could be a hierarchical structure. Tasks are initially routed to the least skilled servers and then progressively moved to more skilled servers.

The paper is organized as follows. In Section 2, the performance measure, i.e. call congestion is discussed. In Section 3, we discuss the results of Massey and Whitt [9] as used in the paper for our calculations. Section 4 describes the lognormal distribution while in Section 5 Boole's rule on numerical integration is presented. In Section 6, we present the data and the application of the method discussed in Section 3. Section 7 considers the stationary Poisson approximation and Section 8 presents Conclusions.

2. Performance Measure

The arrival process is a nonstationary Poisson process and is described by the deterministic arrival rate function, $\lambda(t)$, defined over the interval [0, T]. The average arrival rate over the time interval is, $\bar{\lambda} = \frac{\int_0^T \lambda(t)\,dt}{T}$. As in Massey and Whitt [9], we assume that the mean service time of a customer is 1. We also assume that T is not too small or too large. For a justification of this, please refer Massey and Whitt [9]. The authors recommend that T should be between 6 and 20. We take T=10. We assume that the average length of a call at the call center is 3 minutes and the planning interval under consideration is 30 minutes. Thus, if 3 minutes is taken as a service time of 1 time unit, we have T = 10 time units.

Let $Q(t)$ be the number of busy servers at time t. The blocking probability is,

$$\beta(t) = P(Q(t) = s) \tag{1}$$

which is the probability that all s servers (trunk lines) are busy at time t.

The performance measure that is considered is the call congestion, β_c. This is the ratio of the expected number of lost customers to the expected number of arrivals. As in Massey and Whitt [9], let $B(t)$ be the number of blocked calls in the interval [0, t] and $A(t)$ be the number of arrivals in the interval [0, t]. The ratio is,

$$\beta_c = \frac{EB(T)}{EA(T)} = \frac{\int_0^T \lambda(t)\beta(t)\,dt}{\int_0^T \lambda(t)\,dt} \tag{2}$$

We next discuss the results of Massey and Whitt [9].

3. Results of Massey and Whitt [9] as Used in the Paper

As mentioned before, a queueing model with a nonstationary Poisson arrival process is approximated by a queueing model with a general stationary arrival process. In particular, the distribution of $Q(t)$ in the nonstationary $M_t/G/s/0$ model over the interval (0, T] is approximated by the distribution of $Q(t)$ in the stationary $G/G/s/0$ model. Here, in our case, the distribution G is the lognormal distribution.

The first approach in Massey and Whitt [9] is to construct a stationary point process from the nonstationary Poisson arrival process. Let $N=\{N(t):t\geq 0\}$ be the stationary point process constructed, please see Massey and Whitt [9], for details. The authors characterize the variability of the stationary point process using *the index of dispersion for counts, I(t),* i.e.

$$I(t)=\frac{Var[N(t)]}{E[N(t)]}=\frac{Var[N(t)]}{\bar{\lambda}t} \tag{3}$$

The authors define $c^2=I(1)$ and find the value of c^2 for the special case of the linear arrival rate function, i.e. when $\lambda(t)=a+bt,\ 0\leq t\leq T$. For this special case,

$$c^2=1+\frac{b^2T^2}{6(2a+bT)} \tag{4}$$

We need this expression for c^2 in the approximation for the blocking probability that we use.

The second expression that we need for approximating the blocking probability is the expression for *peakedness*. According to Massey and Whitt [9], peakedness is defined as `the ratio of the variance to the mean of the steady-state number of busy servers in an associated infinite-server model with the same service time distribution and the same arrival process'.

The authors consider the limiting behaviour of the peakedness as the arrival rate grows. The heavy traffic peakedness is, (see Massey and Whitt [9])

$$z=\ 1+\left(c^2-1\right)\frac{1}{E[S]}\int_0^\infty\left[1-G(t)\right]^2dt \tag{5}$$

with E[S]=1 and as mentioned before, G(.) being the cdf of the lognormal distribution.

Let *B(s, a)* be the Erlang blocking formula for *s* servers and offered load *a*, extended to nonintegral *s*. Then, as the final result, we have the Hayward approximation. The call congestion is approximated by,

$$\beta_c\ \cong\ B\left(s/z,\bar{\lambda}/z\right) \tag{6}$$

Here we need to specify the expression for *B(s, a)*. For nonintegral *s*,

$$B(s,a)=\frac{a^se^{-a}}{\Gamma(s+1,a)} \tag{7}$$

where $\Gamma(s+1,a)$ is the incomplete gamma function,

$$\Gamma(s+1,a)=\int_a^\infty t^se^{-t}dt \tag{8}$$

By partial integration, it is possible to show that for integral values of *s*,

$$\Gamma(s+1,a)=s!e^{-a}\left(1+\frac{a}{1!}+...+\frac{a^s}{s!}\right) \tag{9}$$

So, in this case for integral *s*, the above formula becomes,

$$B(s,a)=\frac{\frac{a^s}{s!}}{1+\frac{a}{1!}+...+\frac{a^s}{s!}} \tag{10}$$

which is the standard Erlang loss formula for integral *s*, see Gross et al. [2].

We next discuss the lognormal distribution.

4. The Lognormal Distribution

For a discussion of the lognormal distribution, refer Wikipedia. The lognormal distribution is a continuous probability distribution. The lognormal random variable is such that its logarithm follows the normal distribution.

Define, $Y=\mu+\sigma Z$, where Z has the standard normal distribution. Then Y has a normal distribution with mean μ and standard deviation σ. We have,

$X=e^Y$ has the lognormal distribution.

Its support is the positive real line. The pdf is,

$$f(x)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}\ ,\ x>0 \tag{11}$$

$$\text{The mean is, }\exp(\mu+\sigma^2/2) \tag{12}$$

and the variance is,

$$[\exp(\sigma^2)-1]*\exp(2\mu+\sigma^2)\,. \tag{13}$$

$$\text{The cdf is }F(x)=\frac{1}{2}+\frac{1}{2}\text{erf}\Big[\frac{\ln x-\mu}{\sqrt{2}\sigma}\Big]. \tag{14}$$

In the above erf(.) or the error function is a special function of sigmoid shape.

We have,

$$\text{erf}(x)=\frac{1}{\sqrt{\pi}}\int_{-x}^{x}e^{-t^2}\,\text{dt}\ =\ \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,\text{d}t \tag{15}$$

For non-negative values of *x*, the error function has the following interpretation, as given in Wikipedia. For a random variable *X* that is normally distributed with mean 0 and variance ½, erf(x) is the probability of *X* lying in the range [*-x, x*].

Next numerical integration is discussed.

5. Numerical Integration: Boole's Rule

In the expression for the peakedness z, the following integral needs to be evaluated numerically,

$$\int_0^\infty \left[1-G(t)\right]^2 dt$$

where G(.) is the cdf of the lognormal distribution. This is done using Boole's rule, which is a quadrature formula. Before stating the rule, we state some definitions. For a discussion of Numerical Integration, refer Mathews and Fink [16].

Definition 1. Suppose that $a = x_0 < x_1 < \ldots < x_M = b$. A formula of the form

$$Q[f] = \sum_{k=0}^{M} w_k f\left(x_k\right)$$

with the property that

$$\int_a^b f\left(x\right) dx = Q[f] + E[f] \qquad (16)$$

is a numerical integration or quadrature formula. The term $E[f]$ is the truncation error for integration. The values $\left\{x_k\right\}_{k=0}^{M}$ are the quadrature nodes and $\left\{w_k\right\}_{k=0}^{M}$ are the weights.

Definition 2. The degree of precision of a quadrature formula is the positive integer n such that $E\left[P_i\right] = 0$ for all polynomials $P_i\left(x\right)$ of degree $i \le n$, but for which $E\left[P_{n+1}\right] \ne 0$ for some polynomial $P_{n+1}\left(x\right)$ of degree n+1.

Theorem. The general form of the truncation error term is, $E\left[f\right] = Kf^{(n+1)}\left(c\right)$, where K is a suitably chosen constant and n is the degree of precision.

Boole's rule, along with the Trapezoidal rule and the Simpson's rule, is an example of a closed Newton-Cotes Quadrature formula.

Boole's Rule. Assume that $x_k = x_0 + kh$ are equally spaced nodes and $f_k = f\left(x_k\right)$. Boole's rule is,

$$\int_{x_0}^{x_4} f\left(x\right)dx \approx \frac{2h}{45}\left(7f_0 + 32f_1 + 12f_2 + 32f_3 + 7f_4\right) \quad (17)$$

Theorem. Boole's Rule has degree of precision n=5. If $f \in C^6[a,b]$, then

$$\int_{x_0}^{x_4} f\left(x\right)dx = \frac{2h}{45}\left(7f_0 + 32f_1 + 12f_2 + 32f_3 + 7f_4\right) - \frac{8h^7}{945} f^{(6)}\left(c\right) \quad (18)$$

Next the data and the application of the method above is discussed.

6. Data and Application of the Method

Reynolds [1] has the following data (Chapter 3, page 35), which are samples from a call center that takes calls from 8:00 AM to 6:00 PM daily. The data represents half-hourly call volumes for the previous three Mondays.

Table 1. *Half-hourly call volumes for Monday.*

	June 5	June 12	June 19	Average
0800	205	200	210	205
0830	265	255	260	260
0900	300	310	305	305
0930	345	345	345	345
1000	380	385	390	385
1030	400	405	410	405
1100	395	400	405	400
1130	385	395	390	390
1200	355	360	365	360
1230	350	355	360	355
1300	385	390	395	390
1330	375	385	380	380
1400	395	395	395	395
1430	400	405	410	405
1500	360	365	370	365
1530	320	320	320	320
1600	270	265	275	270
1630	190	195	200	195
1700	160	155	150	155
1730	105	100	95	100
Total				6385

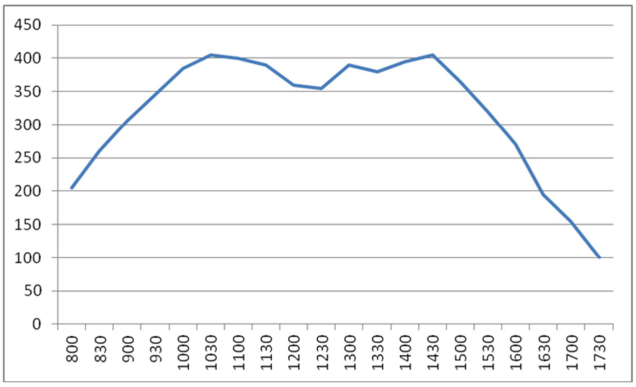

Figure 1. *Average half-hourly call volume for Monday.*

From Figure 1, it is seen that there is significant variation in call volumes by time of day. The maximum number of calls in a half-hour are 405 from 10:30-11:00, while the minimum calls during a half-hour are 100 from 5:30-6:00. This variation indicates a significant amount of nonstationarity in the arrival process.

It is assumed that the 10 hour interval is divided into 20 half-hour periods and that $\lambda\left(t\right)$ varies linearly during each of the half-hour periods. For each of the 20 periods, given $\lambda\left(0\right)$ and $\lambda\left(10\right)$ (since T=10), we calculate a, b, c^2 and $\overline{\lambda}$. These calculations are shown below.

Table 2. *Calculations of* a, b, c^2 and $\bar{\lambda}$ *for each of the 20 time intervals.*

30 min Interval	Avg. Call Volume	$\lambda(0)$	$\lambda(10)$	a	b	c^2	$\bar{\lambda}$
1	205	20.5	20.5	20.5	0	1	20.5
2	260	20.5	26	20.5	0.55	1.10842	23.25
3	305	26	30.5	26	0.45	1.05973	28.25
4	345	30.5	34.5	30.5	0.4	1.04103	32.5
5	385	34.5	38.5	34.5	0.4	1.03653	36.5
6	405	38.5	40.5	38.5	0.2	1.00844	39.5
7	400	40.5	40	40.5	-0.05	1.00052	40.25
8	390	40	39	40	-0.1	1.00211	39.5
9	360	39	36	39	-0.3	1.02	37.5
10	355	36	35.5	36	-0.05	1.00058	35.75
11	390	35.5	39	35.5	0.35	1.0274	37.25
12	380	39	38	39	-0.1	1.00216	38.5
13	395	38	39.5	38	0.15	1.00484	38.75
14	405	39.5	40.5	39.5	0.1	1.00208	40
15	365	40.5	36.5	40.5	-0.4	1.03463	38.5
16	320	36.5	32	36.5	-0.45	1.04927	34.25
17	270	32	27	32	-0.5	1.07062	29.5
18	195	27	19.5	27	-0.75	1.20161	23.25
19	155	19.5	15.5	19.5	-0.4	1.07619	17.5
20	100	15.5	10	15.5	-0.55	1.19771	12.75

As regards the service time distribution, it is lognormal with mean 1. For our numerical work, took two lognormal distributions were taken, one with variance 1 and one with variance 2. The first lognormal distribution has a SCV (squared coefficient of variation, variance divided by the square of the mean) of 1, the same as an exponential, and so that is an important benchmark. The second lognormal distribution has a SCV of 2, much higher than the exponential. Since performance tends to degrade in a system with higher variance, we decided to use this distribution. In particular, in our application, degraded performance would imply that a higher number of trunk lines are required to achieve the same low blocking probability. We wanted to investigate if that is the case.

For a lognormal distribution with mean 1 and variance 1, it follows from equations (12) and (13),

$$\exp(\mu+\sigma^2/2)=1$$

and

$$[\exp(\sigma^2)-1]*\exp(2\mu+\sigma^2)=1$$

or

$$[\exp(\sigma^2)-1]=1$$

$$\text{i.e } \sigma^2=\ln(2) \text{ and } \mu=-0.5\ln(2)$$

Substituting these values in the cdf of the lognormal, the cdf G(t) is found. Then the integral $\int_a^b (1-G(t))^2\,dt$, numerically evaluated using Boole's rule is as given below.

Table 3. *Numerical integration using Boole's rule for the lognormal distribution with variance 1.*

a	b	$\int_a^b (1-G(t))^2\,dt$
0	1	0.50843
1	2	0.04273
2	3	0.0049
3	4	0.00084
4	5	0.00018
5	6	5E-05
Overall Integral, $\int_0^6 (1-G(t))^2\,dt$		0.55713

Although in the equation for peakedness, z, the upper limit of the integral is infinity, for practical purposes, it suffices to evaluate the integral upto an upper limit of 6 to get sufficient numerical accuracy. The upper limit of 6 in this case corresponds to Mean + 5*Standard Deviation.

Similarly, for the case of the lognormal distribution with variance 2, we have, $\sigma^2=\ln(3)$ and $\mu=-0.5\ln(3)$.

These values are again substituted in the cdf of the lognormal. Then the integral $\int_a^b (1-G(t))^2\,dt$, numerically evaluated using Boole's rule is as given below.

Table 4. *Numerical integration using Boole's rule for the lognormal distribution with variance 2.*

a	b	$\int_a^b (1-G(t))^2\,dt$
0	1	0.40371
1	2	0.03884
2	3	0.00723
3	4	0.00195
4	5	0.00066
5	6	2.6E-04
6	7	1.1E-04
7	8	5.3E-05
Overall Integral, $\int_0^8 (1-G(t))^2\,dt$		0.45281

It is again found that an upper limit of 8 gives sufficient numerical accuracy and corresponds to roughly Mean + 5.6*Standard Deviation, similar to the previous case.

Now that c^2 and the two integrals have been evaluated, one each for each lognormal distribution, we can find the value of the peakedness z using equation (5) for each of the 20 time intervals and for each distribution.

Table 5. *Peakedness z for each of the 20 time intervals and for each distribution.*

30 minute interval	Peakedness, z	
	Variance =1	Variance = 2
1	1	1
2	1.060406	1.049095

30 minute interval	Peakedness, *z*	
	Variance =1	Variance = 2
3	1.03328	1.027048
4	1.022857	1.018577
5	1.020352	1.016541
6	1.004702	1.003821
7	1.000288	1.000234
8	1.001175	1.000955
9	1.011143	1.009056
10	1.000325	1.000264
11	1.015268	1.012409
12	1.001206	1.00098
13	1.002696	1.002191
14	1.001161	1.000943
15	1.019295	1.015682
16	1.02745	1.02231
17	1.039345	1.031978
18	1.112325	1.091292
19	1.042448	1.0345
20	1.110152	1.089526

We are now ready to do the final calculation of the number of trunk lines required for each time interval and for each of the two distributions. Based on Reynold's [1], we set a blocking probability of 0.02. That is, number of trunk lines to make the blocking probability just below 0.02 is chosen.

This is done as follows. We start with a low integral number of trunk lines, *s*, and keep increasing them till the blocking probability, $B\left(s/z,\bar{\lambda}/z\right)$ as given by equation (7), just falls below 0.02 for the first time. The value of *s* for which this happens is chosen as the number of trunk lines. This is done for each time interval and for each distribution. Results are shown below.

Table 6. *Number of trunk lines required for each of the 20 time intervals and for each distribution.*

30 minute interval	Number of trunk lines required, *s*	
	Variance =1	Variance = 2
1	29	29
2	32	32
3	38	38
4	42	42
5	47	47
6	50	50
7	50	50
8	50	50
9	48	48
10	46	46
11	47	47
12	49	49
13	49	49
14	50	50
15	49	49
16	44	44
17	39	39
18	33	32
19	26	26
20	20	20

For each distribution, the number of trunk lines required is chosen as the maximum over all the 20 time intervals.

Table 7. *Number of trunk lines required for each distribution.*

Number of trunk lines chosen, *s*	
Variance =1	Variance =2
50	50

Thus for both the lognormal distributions, 50 trunk lines are required to achieve a low 2% blocking probability.

If we look at Table 6 above, we find that the same number of trunk lines are required for each of the two service time distributions for each time interval. This is so barring one time interval. The two service time distributions have the same mean and differ significantly in their variance. For the standard *M/M/s/0* Erlang loss model, it is known that an insensitivity result holds. That is, the blocking probability depends only on the mean of the service time distribution and not on the second moment and higher moments.

This insensitivity result need not hold for the nonstationary Erlang loss model. Davis, Massey and Whitt [10] find that `the service time distribution beyond the mean can have a significant impact on the time-dependent blocking probability in the nonstationary Erlang loss model'.

The authors consider the nonstationary $M_t/PH/s/0$ model with a phase type (*PH*) service time distribution consisting of two phases. In the paper, the authors consider five different service time distributions, all of them being special two phase *PH* distributions. These are an Erlang (E_2), an exponential (*M*) and three hyperexponential (H_2) distributions. All the five distributions have mean 1. The SCVs (squared coefficient of variation) are 0.5 for E_2, 1 for *M* and 4 for the three H_2 distributions. The authors find that the peak time-dependent blocking probabilities for the different service time distributions can differ by as much as a factor of 3.5. So the service time distribution beyond the mean can affect the blocking probability in the nonstationary Erlang loss model.

However the above situation does not always have to hold for the nonstationary Erlang loss model. In our case, both the lognormal service time distributions have mean 1. The first distribution has variance 1, the second has variance 2. We numerically find that in our case, in the $M_t/G/s/0$ model, an insensitivity result holds. The blocking probability depends only on the mean of the lognormal service time distribution and not on its variance.

Next the stationary Poisson approximation is discussed.

7. Stationary Poisson Approximation

In this approximation, the nonstationary Poisson process is replaced with a stationary Poisson process having arrival rate $\bar{\lambda}$. Again the 10 hour time period is divided into 20 half hour intervals. The interval which has the highest average arrival rate is found. This is the arrival rate $\bar{\lambda}$ that we use in the Erlang B formula of the Erlang loss model, stated below. The service rate is 1. The Erlang B blocking formula is given below for the blocking probability $\hat{B}\left(s,\bar{\lambda}\right)$ in the *M/M/s/0* model, see Gross et al. [2].

$$\hat{B}\left(s,\bar{\lambda}\right)=\frac{\frac{\bar{\lambda}^{s}}{s!}}{\sum_{k=0}^{s}\left(\frac{\bar{\lambda}^{k}}{k!}\right)} \tag{19}$$

The calculation of the Erlang B formula can cause numerical problems, because of the large values of s!, as s is reasonably large. It is discussed in Gross et al [2] that $\hat{B}\left(s,\bar{\lambda}\right)$ can be computed using the following iterative method.

$$\hat{B}\left(s,\bar{\lambda}\right)=\frac{\bar{\lambda}\hat{B}\left(s-1,\bar{\lambda}\right)}{s+\bar{\lambda}\hat{B}\left(s-1,\bar{\lambda}\right)},\ s\geq 1 \tag{20}$$

with $\hat{B}\left(0,\bar{\lambda}\right)=1$.

The above iterative method is used to compute $\hat{B}\left(s,\bar{\lambda}\right)$. We substitute $\bar{\lambda}=40.25$, which is the highest average arrival rate for time interval 7 (see Table 2). The value of s such that $\hat{B}\left(s,\bar{\lambda}\right)$ goes below 0.02 for the first time is found.

We have the following results. For s=49, $\hat{B}\left(s,\bar{\lambda}\right)$= 0.0253 and for s=50, $\hat{B}\left(s,\bar{\lambda}\right)$= 0.0199.

Therefore using the stationary Poisson approximation, the number of trunk lines required to achieve a 2% blocking probability is 50.

We can compare this result with the result in Table 7. We thus find in our numerical work that the stationary Poisson approximation is robust and works quite well.

In this paper, we have considered finding the number of trunk lines required using a nonstationary Poisson process and the lognormal service time distribution. The nonstationary Poisson arrival process has been observed in many call centers and the lognormal service time distribution is also justified by empirical work. Existing practice for determining the number of trunk lines assumes a stationary Poisson arrival process and exponential service times and uses the Erlang loss model. Our numerical work finds that existing practice is quite robust and works well in determining the number of trunk lines required, despite the limiting assumptions in the analysis.

This finding of the robustness of existing practice, also finds support in a statement made in Davis, Massey and Whitt [10]. According to the authors, a common engineering approach is to use the stationary Erlang loss model with a constant arrival rate during the time interval over which the system is most heavily loaded. With this approach, the assumed arrival rate in the model is usually greater than the real arrival rate the majority of the time. According to the authors, this approach has been very successful in designing systems with a fixed number of servers that must be able to satisfy demand at any time.

Given the above, it is not too surprising that the stationary Poisson approximation works so well, in this case.

The computations discussed in this section and the previous section, were carried out in Excel and MATLAB.

We next present Conclusions.

8. Conclusions

Two important resources in a call center are the number of staff and the number of trunk lines required. In this paper, we focus on the decision of the number of trunk lines to have. The current practice is to use the Erlang B or the *M/M/s/0* queueing model which assumes Poisson arrivals, exponential service times, s servers and no places in queue, i.e. no customers can wait. In this paper, we improve on the state of practice in determining the required number of trunk lines, by including two realistic features present in call centers. There is extensive empirical evidence for both features as found in the papers by Gans et al [5] and Brown et al [4].

In order to carry out our computations we use the results of a paper by Massey and Whitt [9]. The authors approximate a queueing model with a nonstationary arrival process with a queueing model with a stationary arrival process. In particular, the distribution of the number of busy servers in the nonstationary $M_t/G/s/0$ model is approximated by the distribution of the number of busy servers in the stationary *G/G/s/0* model. Here, in our case, the distribution *G* is the lognormal distribution.

We have two main findings. Firstly, we find numerically that in our nonstationary Erlang loss model, $M_t/G/s/0$, an insensitivity result holds. The blocking probability of arrivals at the call center depends only on the mean of the lognormal service time distribution and not on its variance. In particular, both the lognormal service time distributions predict the same requirement of trunk lines. In the stationary *M/M/s/0* Erlang loss model, the insensitivity result theoretically holds, as is well known. Davis, Massey and Whitt [10] have shown that in the nonstationary Erlang loss model, the insensitivity result need not hold. We find numerically find that in our model, it holds.

Our second finding is that current practice is quite robust. In particular, we find the number of trunk lines required using a stationary Poisson approximation. This approximation assumes stationary Poisson arrivals with an appropriately chosen arrival rate and exponential service times. The approximation does quite well in predicting the number of trunk lines required. Based on a statement in Davis, Massey and Whitt [10], this finding is not too surprising. However, it is worthwhile to go through the numerical analysis and come to this conclusion.

Future work could consider determining call center staffing levels under these more realistic assumptions. We are presently working along those lines.

References

[1] P. Reynolds, Call Center Staffing: The Complete Practical Guide to Workforce Management, 2003, Call Center School Press, Nashville, Tennessee.

[2] D. Gross, J. F. Shortle, J. M. Thompson, and C. M. Harris, Fundamentals of Queueing Theory (4th ed.), 2008, John Wiley, New Delhi, India.

[3] Green, L., P. Kolesar and W. Whitt (2007). Coping with time-varying demand when setting staffing requirements for a service system. Production and Operations Management 16 (1), 13-39.

[4] Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn and L. Zhao (2005). Statistical analysis of a telephone call center: A Queueing-Science perspective. Journal of the American Statistical Association 100 (469), 36-50.

[5] Gans, N., G. Koole and A. Mandelbaum (2003). Telephone call centers: Tutorial, review and research prospects. Manufacturing and Service Operations Management 5 (2), 79-141.

[6] Bolotin, V. A. (1994). Telephone circuit holding time distributions. Proc. 14th International Teletraffic Conference, 125-134.

[7] Chlebus, E.(1997). Empirical validation of call holding time distributions in cellular communication systems. Proc. 15th International Teletraffic Conference, Elsevier, Amsterdam, 1179-1188.

[8] Mandelbaum, A., A. Sakov and S. Zeltyn (2001). Empirical analysis of a call center. Technical Report, Technion, Haifa, Israel.

[9] Massey, W. A. and W. Whitt (1996). Stationary-process approximations for the nonstationary Erlang loss model. Operations Research 44 (6), 976-983.

[10] Davis, J. L., W. A. Massey and W. Whitt (1995). Sensitivity to the service-time distribution in the nonstationary Erlang loss model. Management Science 41 (6), 1107-1116.

[11] Kim, J. W. and S. C. Park (2010). Outsourcing strategy in two-stage call centers. Computers & Operations Research 37 (4), 790-805.

[12] Klincli, T. G. and X. Zhang (2017). Mathematical models and solution approach for cross-training staff scheduling at call centers. Computers and Operations Research, 87, 258-269.

[13] Yu, M., J. Gong, J. Tang and F. Kong (2017). Delay announcements for call centers with hyperexponential patience modelling. Industrial Management and Data Systems, 117 (6), 1037-1057.

[14] Li, G., J. Z. Huang and H. Shen (2018). To wait or not to wait: Two-way functional hazards model for understanding waiting in call centers. Journal of the American Statistical Association, 113, 1503-1514.

[15] Bimpikis, K and G. M. Markakis (2019). Learning and hierarchies in service systems. Management Science, 65 (3), 1268-1285.

[16] J. H. Mathews and K. D. Fink (2004), Numerical Methods Using MATLAB (4th ed.), Pearson Education, India.