**American Journal of Operations Management and Information Systems**
2020; 5(4): 74-83
http://www.sciencepublishinggroup.com/j/ajomis
doi: 10.11648/j.ajomis.20200504.11
ISSN: 2578-8302 (Print); ISSN: 2578-8310 (Online)

# Nonstationarity and Abandonment in Markovian Queues with Application to Call Centers

**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. Nonstationarity and Abandonment in Markovian Queues with Application to Call Centers. *American Journal of Operations Management and Information Systems.* Vol. 5, No. 4, 2020, pp. 74-83. doi: 10.11648/j.ajomis.20200504.11

**Received**: October 3, 2020; **Accepted**: October 19, 2020; **Published**: October 26, 2020

**Abstract:** Both large and small firms maintain call centers to establish contact between themselves and their customers. A call center is staffed by Customer Service Representatives (CSRs). In addition to CSRs, a call center needs computers and telecommunication equipment such as an Automatic Call Distributor (ACD). If calls arrive according to a Poisson arrival process and the service times have an exponential distribution, the call center can be modeled as an M/M/s queue where s is the number of CSRs. Typical calculations include finding the number of CSRs required and finding the number of trunks lines required. However, if call center models ignore the abandonment behavior of customers, they distort information that is relevant to management. Typically, ignoring abandonment would lead to overstaffing, as fewer CSRs are actually needed, because some of the callers abandon the system. Also, nonstationarity of arrivals is highly prevalent in call centers. Green, Kolesar and Whitt plot hourly arrival rates for a financial services call center. There is significant variation in arrivals by time of day. In this paper, we model call centers as multiserver Markovian queues with both nonstationarity and abandonment. Nonstationarity is modeled by having a Poisson arrival stream with time dependent mean, which varies according to a sinusoid. Abandonment is defined in terms of an exponential patience random variable, which is the amount of time the caller would be on hold before abandoning the call. We numerically study the performance of this nonstationary M(t)/M/s+M queue with abandonment and compare its performance measures with those of the stationary M/M/s+M queue. We find that approximating a nonstationary system by a stationary system, even at low levels of nonstationarity, can lead to significant errors. Similar results in a system without abandonment, have been obtained by Green, Kolesar and Svoronos. Additionally, we find that abandonment dampens the effect of nonstationarity. Since abandonment and nonstationarity are both present together in real call centers, a real call center with a high level of abandonment behaves closer to an ideal stationary system.

**Keywords:** Queueing, Nonstationarity, Abandonment, Markovian Queues, Call Centers

## 1. Introduction

Both large and small firms maintain call centers to establish contact between themselves and their customers. In recent years, especially in the past decade, the number of firms maintaining call centers has grown dramatically. The firm could be an online retailer offering sales and returns through the call center or a logistics company offering tracking services for its parcels. Alternatively, it could be a financial services firm offering advice on financial products or else a software firm providing sales and after-sales support.

A call center is staffed by Customer Service Representatives (CSRs), a large pool of personnel, who together may handle thousands of calls per day. In addition to CSRs, a call center needs computers and telecommunication equipment such as an Automatic Call Distributor (ACD), that manages the queue of customers and connects customers to CSRs. Call centers may be either generic or involve skill-based routing. In generic call centers any CSR can handle any call. In skill-based routing different CSRs possess different skills and handle subsets of customers. For example different CSRs may be fluent in different languages and handle calls accordingly. Alternatively, for a financial services firm, some CSRs may be trained to sell health insurance while some others sell car insurance. Therefore incoming calls to the call center have to be routed to the CSRs keeping in view their differing skills.

Typically, in practice, see Cleveland [1] and Reynolds [2], call centers are modeled as queueing systems in which it is assumed that there is one generic set of CSRs who handle all calls, i.e. skill-based routing is ignored. Also it is assumed that once customers join the queue, they do not abandon the queue. So customers wait for service as long as is required. In addition it is assumed that incoming traffic at the call center does not vary with time of day, i.e. the arrival process is stationary.

With these assumptions, if calls arrive according to a Poisson arrival process and the service times have an exponential distribution, the call center can be modeled as an M/M/s queue where s is the number of CSRs. Typical calculations include finding the number of CSRs required and finding the number of trunks lines required, see Reynolds [2]. A trunk line connects to the ACD and the number of customers who succeed in getting through at a time when all CSRs are busy depends on the number of trunk lines available. For finding the number of CSRs, a target probability of delay in receiving service is chosen. Given this target probability of delay in the M/M/s model, the value of s is chosen. For finding the number of trunk lines, the M/M/s/s loss system (Erlang B) is used, i.e. there are s servers and s places in queue, so no call can wait. A target probability of blockage is chosen and s is found from the formula for the Erlang B queueing model, where s is the number of trunks to be made available.

In this paper, we model call centers as multiserver Markovian queues with both nonstationarity and abandonment. We first discuss abandonment.

### *1.1. Abandonment in Call Centers*

If call center models ignore abandonment, they distort information that is relevant to management. Typically, ignoring abandonment would lead to overstaffing, as fewer CSRs are actually needed, because some of the callers abandon the system. Consider the following numerical example, which has been taken from Garnett, Mandelbaum and Reiman [3]. There are 50 CSRs, 48 calls per minute, the average service time is 1 minute and the average patience is 2 minutes.

***Table 1.*** *Comparing results for models with and without abandonment.*

| | M/M/s | M/M/s+M |
|---|---|---|
| Fraction Abandoning | -- | 3.1% |
| Average speed of answer | 20.8 sec | 3.6 sec |
| Waiting time's 90th percentile | 58.1 sec | 12.5 sec |
| CSR Utilization | 96% | 93% |

It can be shown that the performance measures of the system with 54 CSRs and no abandonment would be very similar to the performance measures of the system with 50 CSRs and a 3% abandonment rate. Therefore if we do not include abandonment in our queueing model we would need to use 4 extra CSRs. While if we include abandonment we realize we do not need these extra CSRs and save 8% (because of the 4 CSRs) on personnel costs. It is usually estimated that personnel costs easily account for more than 50% of total cost in call centers. Therefore the above saving is significant and including abandonment in the model improves performance.

We next briefly discuss capacity selection in the absence and presence of abandonment. This is the square root staffing rule. Here capacity refers to the numbers of CSR's to have at any point in time.

Consider the M/M/s queue. Let λ be the arrival rate in number of callers per hour and μ the service rate. We have that 1/μ is the mean call duration in hours. Define R = λ/μ. Then R is the offered load or the number of hours of calling time that needs to be serviced per hour. This demand is met using s servers. Define ρ = λ/sμ, to be the utilization.

The square root staffing rule in the absence of abandonment is based on the following heavy traffic limit theorem in Halfin and Whitt [4].

Let $W_{s\rho}$ be a random variable with the steady state waiting time distribution for each positive integer s and each ρ, 0< ρ<1.

Theorem (Halfin and Whitt 1981). The limit $P\left(W_{s\rho} > 0\right) \to \alpha$ as $s \to \infty$, where 0< α<1, holds if and only if

$$(1-\rho)\sqrt{s} \to \beta \text{ as } s \to \infty$$

where $0 < \beta < \infty$.

Here $\alpha = \left[1 + \frac{\beta\varphi(\beta)}{\phi(\beta)}\right]^{-1}$, where $\phi(.)$ is the standard normal density function and $\varphi(.)$ is the standard normal cumulative distribution.

Whitt [5] showed that the condition

$$(1-\rho)\sqrt{s} = \beta$$

is equivalent to

$$s = R + \beta\sqrt{R}$$

if $\beta$ is small compared to $\sqrt{s}$ .

The square root staffing rule then is,

$$s = R + \beta\sqrt{R} ,$$

where $\beta$ is a constant that depends on the service level. The first term in the above equation is the offered load and the second term is the excess capacity needed to meet service level requirements. This grows less than proportionately with R.

In practice, the number of CSRs required, can be determined as follows. Determine a target P (Wait >0) = α based on service level considerations. Choose β from the above equation in the Theorem, relating α and β. Find the capacity or the number of CSRs as, $s = R + \beta\sqrt{R}$ .

There is also a square root staffing rule in the presence of abandonment, which has been derived by Garnett, Mandelbaum and Reiman [3]. See also Daskin [6]. We first need to define abandonment. Abandonment is defined in terms of a patience random variable, which is the amount of time the caller would be on hold before abandoning the call. We assume this patience random variable has an exponential distribution with mean 1/θ. For call centers, Garnett, Mandelbaum and Reiman [3] define a mean patience of 10 minutes as very patient, 1 minute as moderately patient and 6 seconds as very impatient.

The number of CSRs required in the presence of abandonment, can be determined in a similar way to what has been described above for the case of no abandonment. First, we determine a target P (Wait >0) = α based on external service level considerations. Garnett, Mandelbaum and Reiman [3], have developed an approximation for the probability of waiting in the presence of abandonment. This is,

$$\alpha = \left[ 1 + \frac{\sqrt{\theta/\mu}.h\left(\beta/\sqrt{\theta/\mu}\right)}{h(-\beta)} \right]^{-1}$$

where $h(x) = \dfrac{\phi(x)}{1-\varphi(x)}$.

Given α, we determine β from the above equation. Then, we find s as

$$s = R + \beta\sqrt{R}$$

For a survey of research in the application area of call centers, refer Gans, Koole and Mandelbaum [7].

Motivated by applications to service systems, Yunan and Whitt [8] develop simple engineering approximation formulas for the steady-state performance of heavily loaded G/ GI/ n+ GI multiserver queues, which can have non-Poisson and nonrenewal arrivals and non-exponential service-time and patience-time distributions. Good performance across a wide range of parameters is obtained by making heuristic refinements, the main one being truncation of the queue length and waiting time approximations to nonnegative values. Simulation experiments show that the proposed approximations are effective for large-scale queuing systems for a significant range of the traffic intensity ρ and the abandonment rate θ.

Batt and Terwiesch [9] study queue abandonment from a hospital emergency department. The authors show that abandonment is influenced by the queue length and the observable queue flows during the waiting exposure, even after controlling for wait time. They also show that patients are sensitive to being "jumped" in the line and that patients respond differently to people more sick and less sick moving through the system. This customer response to visual queue elements is not currently accounted for in most queuing models. Additionally, to the extent the visual queue information is misleading or does not lead to the desired behavior, managers have an opportunity to intervene by altering what information is available to waiting customers.

### *1.2. Nonstationarity of Arrivals in Call Centers*

Nonstationarity of arrivals is highly prevalent in call centers. Green, Kolesar and Whitt [10] plot hourly arrival rates for a financial services call center. There is significant variation in arrivals by time of day. Reynolds [2], 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'.

Reynolds [2] illustrates her method for time of day forecasting using 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 2.*** *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 |

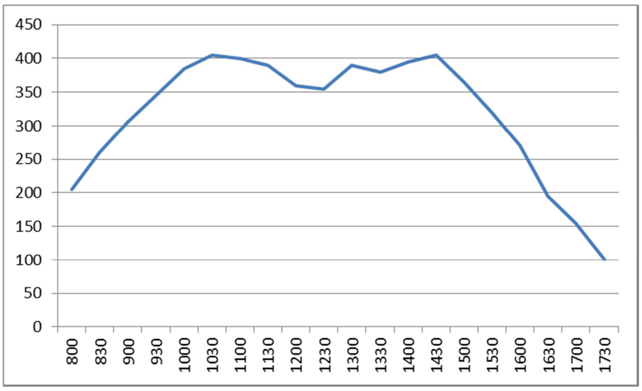

***Figure 1.*** *Average half-hourly call volume for Monday.*

From Figure 1, we see 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.

Typically, nonstationarity is accounted for by dividing the day into many smaller time intervals and assuming a stationary model applies to each of the smaller time intervals. Still, it is important to know that if a nonstationary arrival process is approximated by a stationary model, how accurate will that approximation be. Green, Kolesar and Svoronos [11], numerically study the behavior of a nonstationary multiserver Markovian queue with sinusoidal Poisson input. The authors find that if a queueing system has a periodic arrival process with a relative amplitude (amplitude normalized by the average arrival rate) of only 10%, a stationary model is not likely to do well. The authors further find in their numerical work that at a relative amplitude of only 33%, the actual expected delay is more than twice the stationary expected delay, while at a relative amplitude of 100% the actual expected delay is ten times the stationary expected delay.

Theoretical papers which compare stationary systems and nonstationary systems include Ross [12], Rolski [13] and Svoronos and Green [14]. Ross [12] put forward two conjectures. Firstly, that in a single server infinite capacity queueing system, the more nonstationary the arrival process, the greater the average delay. Secondly, in a finite capacity system, the proportion of lost customers is greater when arrivals are nonstationary than when they are stationary. Rolski [13] proved the second Ross conjecture for pure loss systems with exponential service times and one or two servers. Svoronos and Green [14] showed that for single server loss systems with exponential service times and periodic Poisson input, the proportion of losses in convex increasing in the amplitude.

Aksin and Harker [15] study the problem of determining capacity for a call center, where capacity consists of multiple types of resources that are required simultaneously to provide service. In addition, the system is characterized by the presence of a common resource that is shared across multiple types of customers. In the case of call centers, the capacity is the optimal number of servers that need to be allocated to different call types. Heuristics are proposed for the problem. The authors show that for systems experiencing heavy loads and serving a diverse set of customers, the proposed heuristics outperform current methods that ignore the role of a shared resource.

Tirdad et. al [16] consider a nonstationary M(t)/M/c/c queue with periodic arrival rates and two levels of the number of servers. The authors define a cost function which needs to be minimized and find the cost using transient solutions of the M(t)/M/c/c queue. The results of the model are used to build a Markov Decision Process (MDP) and applied in the area of healthcare.

Cho and Ko [17] investigate the stabilization of the mean virtual response time in a single-server processor sharing (PS) queueing system with a time-varying arrival rate and a service rate control. The authors propose and compare a modified square-root (SR) control and a difference-matching (DM) control to stabilize the mean virtual response time. Extensive simulation studies with various settings of arrival processes and service times show that the DM control outperforms the SR control for heavy-traffic conditions, and that the SR control performs better for light-traffic conditions.

The natural queueing models for many operations research applications have time-varying arrival rates. In addition, the natural models often are not Markov stochastic processes, so that they are not amenable to exact mathematical analysis. Whitt and You [18] propose a time-varying robust queueing algorithm to approximate the time-varying distribution of the workload (virtual waiting time) in a non-Markovian single-server queue with a time-varying arrival-rate function. They apply simulation and asymptotic methods to examine the performance of periodic robust queueing. Simulation examples show that the mean and the full distribution (specified by the quantiles) of the periodic steady-state workload are remarkably well approximated.

For a detailed review and analysis of queueing models in the presence of nonstationarity, refer Green, Kolesar and Whitt [10].

We numerically study nonstationarity and abandonment in multiserver Markovian queues. Nonstationarity is modeled by having a Poisson arrival stream with time dependent mean, which varies according to a sinusoid. Abandonment is defined in terms of an exponential patience random variable, which is the amount of time the caller would be on hold before abandoning the call. We obtained our numerical results by solving a system of differential equations, which represent the birth death equations for the M(t)/M/s+M system. We numerically study the performance of this nonstationary queue with abandonment and compare its performance measures with those of the stationary M/M/s+M queue.

We find that approximating a nonstationary system by a stationary system, even at low levels of nonstationarity can lead to significant errors. Similar results in a system without abandonment, have been obtained by Green, Kolesar and Svoronos [11]. Additionally, we find that abandonment dampens the effect of nonstationarity. At a high level of abandonment, a nonstationary system behaves closer to a stationary one. Since abandonment and nonstationarity are both present together in real call centers, a real call center with a high level of abandonment behaves closer to an ideal stationary system.

The paper is organized as follows. In Section 2 we describe our Methodology. We next discuss our numerical results. In Section 3, we study the effects of nonstationarity, while in Section 4 we analyze the effects of abandonment. Conclusions are presented in Section 5.

## 2. Methodology

We obtained our numerical results by solving the following set of differential equations, which represent the birth death equations for a M(t)/M/s+M system.

$$
\begin{aligned}
p_0'(t) &= -\lambda(t)\,p_0(t)+\mu p_1(t)\\
p_n'(t) &= -\big(\lambda(t)+n\mu\big)p_n(t)+\lambda(t)\,p_{n-1}(t)\\
&\quad +(n+1)\mu p_{n+1}(t),\ \ 1\le n<s\\
p_n'(t) &= -\big(\lambda(t)+s\mu+(n-s)\theta\big)p_n(t)+\lambda(t)\,p_{n-1}(t)\\
&\quad +\big(s\mu+(n+1-s)\theta\big)p_{n+1}(t),\ \ \mathrm{s}\le n
\end{aligned}
$$

Here $p_n(t)$ is the probability of n customers in the system at time t, $\lambda(t)$ is the arrival rate at time t, μ is the service rate, θ is the abandonment rate and s is the number of servers.

Thus the above is a system of linear first order differential equations, which needs to be solved numerically for the unknown functions $p_n(t)$.

Similar to Green, Kolesar and Svoronos [11], the arrival rate $\lambda(t)$ is given by,

$$\lambda(t) \;=\; \bar{\lambda} \;+\; A\cos(2\pi t/24)$$

where $\bar{\lambda}$ is the daily average arrival rate and A is the amplitude of the sinusoidal process. Without loss of generality, the period is 24 hours. We need $\bar{\lambda}$ - $A \ge 0$, so that $\lambda(t) \ge 0$, for all t. Here A is a measure of the nonstationarity in the process. As in Green, Kolesar and Svoronos [11], we consider Relative Amplitude (RA) = $A/\bar{\lambda}$, as a measure of the degree of nonstationarity in the process. We have that RA varies between 0 and 1.

We carried out the numerical solution of the system of ordinary differential equations using MATLAB. We used the numerical solver ode45 which combines a fourth order and a fifth order Runge-Kutta method. This solver is described in Hunt et al [19]. The Appendix contains the equations that are used for a fourth order Runge-Kutta method for solving a system of 2 differential equations, see Mathews and Fink [20]. The solver ode45 varies the step size at each step in order to achieve the desired accuracy. It is suitable for a wide variety of initial value problems. For numerical work, we divided the cycle of 24 hours into 144, 10 minute intervals. So the function $p_n(t)$ is approximated by a discrete function at 145 time points, from t = 0 to t = 144 (i.e. 24 hours). For each *n*, we requested the solver ode45 to provide solution values at these specific time points.

### 2.1. Initialization of the System of Differential Equations and Performance Measures

The system of differential equations needs to be initialized with a probability vector $p_n(0)$, n=0,...,N , where N is the maximum number of equations solved. Define $\pi_n$ as the probability of having n customers in the system for a steady state M/M/s+M queue. Then $\pi_n$ is given as below, as defined in Garnett, Mandelbaum, Reiman [3].

$$\pi_k = \begin{cases} \dfrac{(\lambda/\mu)^k}{k!}\pi_0 & 0\le k\le s\\[2ex] \displaystyle\prod_{j=s+1}^{k}\left(\frac{\lambda}{s\mu+(j-s)\theta}\right)\frac{(\lambda/\mu)^s}{s!}\pi_0 & s<k\le N\end{cases}$$

where

$$\pi_0 = \left[\sum_{k=0}^{s}\frac{(\lambda/\mu)^k}{k!}+\sum_{k=s+1}^{N}\prod_{j=s+1}^{k}\left(\frac{\lambda}{s\mu+(j-s)\theta}\right)\frac{(\lambda/\mu)^s}{s!}\right]^{-1}$$

We have that $p_n(0)$ is taken equal to $\pi_n$, with $\lambda=\lambda(0) \;=\; \bar{\lambda} \;+\; A$.

We next describe the two performance measures for the system. Let $p_{ni}$ be the probability of *n* customers in the system at the start of interval *i*, *i* = 1,.., 144. Let $\lambda_i$ be the arrival rate at the start of interval *i*, so that $\bar{\lambda} \;=\sum_{i=1}^{144}\lambda_i/144$. Then, similar to Green, Kolesar and Svoronos [11], our two performance measures are the daily (customer average) probability of delay, $p_d$, and the daily average queue length, $L_q$. These are defined below.

$$p_d = \sum_{i=1}^{144}\frac{\lambda_i\left(1-\sum_{n=0}^{s-1}p_{ni}\right)}{144\bar{\lambda}}$$

$$L_q = \sum_{i=1}^{144}\sum_{n=s}^{N}(n-s)p_{ni}/144$$

### 2.2. Results for the Stationary M/M/s+M Queue for Comparison

For comparison purposes, we need to compare our performance measures for the nonstationary queue with abandonment, with those of the stationary M/M/s+M queue. However, analytical results for the performance measures for the stationary multiserver queue with abandonment are almost entirely intractable, to be of any use. See for example, Abou-El-Ata and Hariri [21].

Fortunately, Garnett, Mandelbaum and Reiman [3], have developed heavy traffic approximations for the performance measures of the M/M/s+M queue. The authors have found these approximations to be excellent for practical use, even in the case of medium sized call centers with 20 or more CSRs and moderate traffic intensities in the range of 0.6 or higher.

These approximations are given as follows. Let,

$$\phi(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2},\quad \Phi(x)=\int_{-\infty}^{x}\phi(y)dy$$

as well as the hazard rate by

$$h(x)=\frac{\phi(x)}{1-\Phi(x)}$$

Let

$$w(x,y)=\left[1+\frac{h(-xy)}{yh(x)}\right]^{-1}$$

Define

$$\rho=\frac{\bar{\lambda}}{s\mu}$$

and $\beta=\sqrt{s}\ (1-\rho)$

Then in the stationary model,

$$p_d = \text{P(Wait} > 0) \approx w\left(-\beta,\sqrt{\frac{\mu}{\theta}}\right)$$

and the expected number waiting in queue

$$L_q \approx \left[1-\frac{h\left(\beta\sqrt{\mu/\theta}\right)}{h\left(\beta\sqrt{\mu/\theta}+\sqrt{\theta/(s\mu)}\right)}\right] w\left(-\beta,\sqrt{\frac{\mu}{\theta}}\right)\frac{\bar{\lambda}}{\theta}$$

We used the above approximations to determine performance measures for the stationary system for comparison, keeping in mind the ranges of parameters in which these approximations are applicable.

We next discuss our numerical results.

## 3. Effects of Nonstationarity

We measure nonstationarity by the Relative Amplitude (RA) and our performance measures of interest are Expected Queue Length and Probability of Delay. For our numerical experiments, we fix s = 20, μ = 60, θ = 12 and we vary ρ and RA.

We find that as we vary RA from 0.1 to 0.9, the expected queue length is a multiple times higher. From Table 3 below we see that the multiple is near 73 for ρ = 0.6 and near 5 for ρ = 0.95. Thus as nonstationarity in the system increases (from RA = 0.1 to RA = 0.9), expected queue length increases significantly.

***Table 3.*** *Ratio of expected queue lengths at RA = 0.1 and RA = 0.9 for different values of ρ.*

| Relative Amplitude Rho | 0.1 | 0.9 | Ratio |
|---|---|---|---|
| 0.6 | 0.04 | 3.06 | 73.33 |
| 0.7 | 0.22 | 7.90 | 35.68 |
| 0.8 | 0.84 | 13.47 | 15.98 |
| 0.9 | 2.53 | 18.06 | 7.13 |
| 0.95 | 4.07 | 19.92 | 4.89 |

We can also compare the expected queue length for the nonstationary system to the expected queue length for the stationary system. At ρ = 0.9, the expected queue length for the stationary system is 1.99. For ρ = 0.9 and RA = 0.3, the expected queue length for the nonstationary system is 5.52 or 2.7 times higher. At RA = 0.9, the expected queue length for the nonstationary system increases to 18.06 or 9 times higher.

Below, we plot the expected queue length versus relative amplitude for different values of ρ. The expected queue length increases as relative amplitude increases.

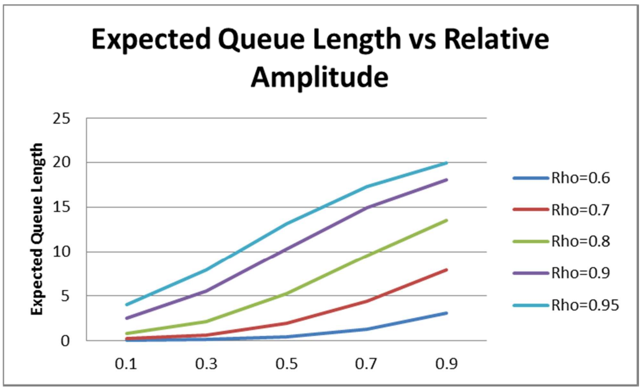

***Figure 2.*** *Expected queue length versus relative amplitude.*

Below we also plot the probability of delay versus relative amplitude. As we would expect the probability of delay also increases with relative amplitude.

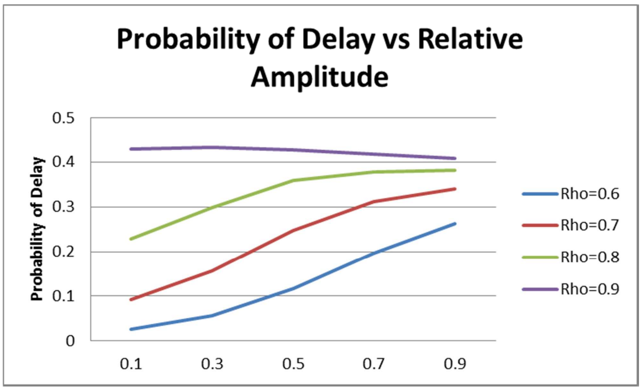

***Figure 3.*** *Probability of delay versus relative amplitude.*

Similar to Green, Kolesar and Svoronos [11], we define an Error Measure of Expected Queue Length as

$$\text{Error Measure} = \frac{\text{Actual Expected Queue Length} - \text{Stationary Expected Queue Length}}{\text{Actual Expected Queue Length}}$$

From our numerical results we find that for a RA of 0.3 or higher, the Error Measure is 50% or higher for all values of ρ. In some cases (RA = 0.9, ρ = 0.6, 0.7, 0.8), the Error Measure is even higher than 95%. Below, we show the graph of Error Measure in expected queue length versus relative amplitude.

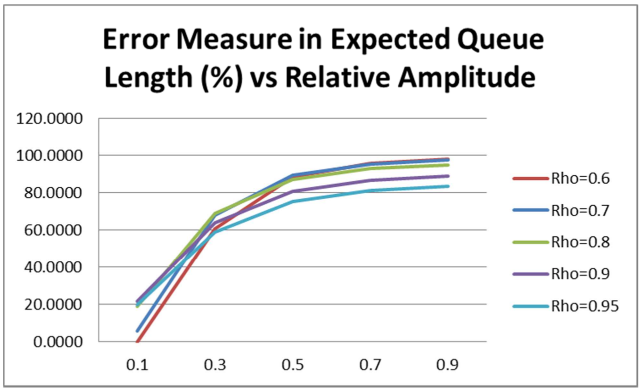

***Figure 4.*** *Error Measure in expected queue length versus relative amplitude.*

In Table 4 below, we consider the stationary expected queue length as a percentage of the nonstationary expected queue length. For ρ = 0.6 and RA = 0.1, the two queue lengths are the same. However, for ρ = 0.6, by the time we increase RA to 0.9 and introduce significant nonstationarity, the stationary expected queue length is only about 2% of the nonstationary expected queue length. Or the nonstationary expected queue length is about 50 times higher. Similarly for RA = 0.9 and ρ = 0.95, the nonstationary expected queue is about 6 times higher than the stationary expected queue length.

***Table 4.*** *The stationary expected queue length as a percentage of the nonstationary expected queue length.*

| Rho Relative Amplitude | 0.6 | 0.95 |
|---|---|---|
| 0.1 | 100% | 80.15% |
| 0.3 | 39.37% | 41.3% |
| 0.5 | 11.71% | 24.92% |
| 0.7 | 4.17% | 18.91% |
| 0.9 | 1.83% | 16.38% |

Thus approximating a nonstationary system by a stationary system, even at low levels of Relative Amplitude can lead to significant errors. Similar results have been obtained by Green, Kolesar and Svoronos [9], in a system without abandonment. It is therefore important to explicitly model the nonstationary behavior of arrivals in a call center and ignoring this behavior can lead to erroneous decision making.

In the next section, we discuss the effects of abandonment.

# 4. Effects of Abandonment

For studying the effects of abandonment, we fix s = 20, μ = 60 and ρ = $\bar{\lambda}$ /sμ = 0.9. We vary RA from 0.1 to 0.9 and θ from 12 to 60. That is, the mean value of the patience random variable, which is exponentially distributed, varies from 5 minutes to 1 minute.

The graph of expected queue length versus θ, for different values of RA is shown below.

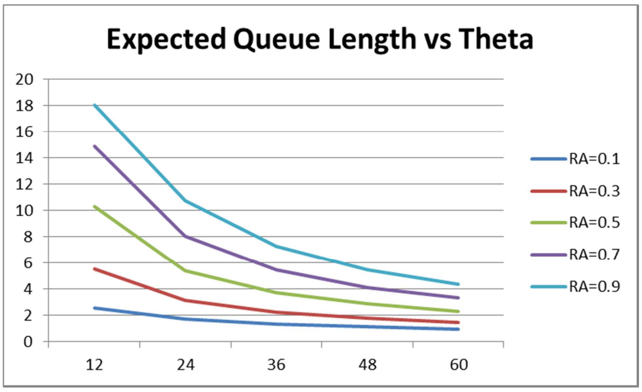

***Figure 5.*** *Expected queue length versus θ for different values of RA.*

From the graph we see that for a fixed value of θ, as RA increases, expected queue length increases. Secondly, for a fixed value of RA, as θ increases from 12 to 60, the expected queue length decreases. Once RA gets fixed, the level of nonstationarity gets fixed. For the same level of nonstationarity, as abandonment increases expected queue length decreases.

Table 5 below shows the ratio of expected queue length for RA = 0.1 and RA = 0.9, for different values of θ. We see that the ratio is around 7 for θ = 12 and decreases to about 4.5 for θ = 60.

Thus abandonment dampens the effect of nonstationarity. At a high level of abandonment, a nonstationary system behaves closer to a stationary one. Since abandonment and nonstationarity are both present together in real systems, a real system with a high level of abandonment behaves closer to an ideal stationary system.

***Table 5.*** *Ratio of expected queue lengths at RA = 0.1 and RA = 0.9 for different values of θ.*

| Relative Amplitude θ | 0.1 | 0.9 | Ratio |
|---|---|---|---|
| 12 | 2.53 | 18.06 | 7.13 |
| 24 | 1.72 | 10.77 | 6.24 |
| 36 | 1.35 | 7.25 | 5.37 |
| 48 | 1.12 | 5.47 | 4.87 |
| 60 | 0.96 | 4.4 | 4.54 |

Below, we show a plot of the Error Measure in expected queue length, as θ varies from 12 to 60 for different values of RA.

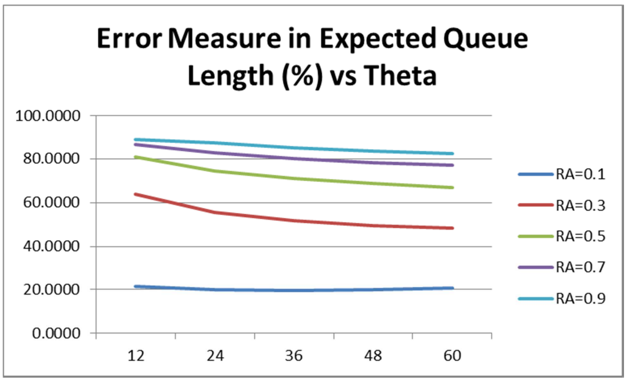

***Figure 6.*** *Error Measure in expected queue length (%) versus θ.*

The Table 6 below shows the values of the Error Measure in expected queue length at RA = 0.1 and RA = 0.9 for different values of θ. We find that the Error Measure of the stationary approximation is primarily affected by nonstationarity or relative amplitude. Abandonment has very little effect on the Error Measure. The Error Measure remains almost flat as θ varies from 12 to 60. For a RA of 0.1, the Error Measure remains almost constant with θ at about 20%. While for a RA of 0.9, the Error Measure remains almost constant at about 85%.

***Table 6.*** *Error measure in expected queue length at RA = 0.1 and RA = 0.9 for different values of θ.*

| Relative Amplitude Θ | 0.1 | 0.9 |
|---|---|---|
| 12 | 21.50% | 88.98% |
| 24 | 19.81% | 87.15% |
| 36 | 19.76% | 85.05% |
| 48 | 20.11% | 83.6% |
| 60 | 20.6% | 82.52% |

For our next set of numerical experiments related to abandonment, we fix s = 20, μ = 60 and RA = 0.8. We vary ρ from 0.6 to 0.95 and θ from 12 to 60.

Figure 7 below shows the graph of expected queue length versus θ, as θ varies from 12 to 60 for different values of ρ. For a fixed value of θ as ρ increases, expected queue length increases. For a fixed value of ρ, as θ or the rate of abandonment increases, expected queue length decreases.

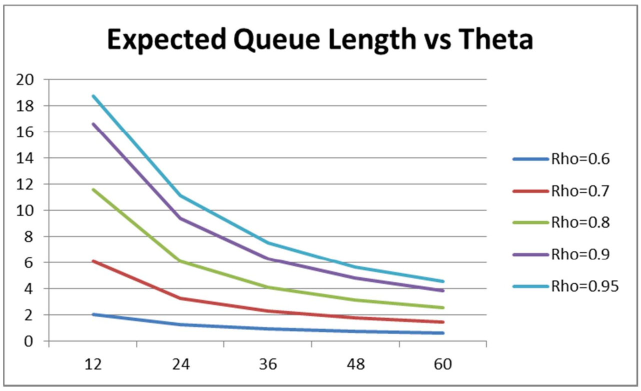

***Figure 7.*** *Expected queue length versus θ for different values of ρ.*

Table 7 below shows the ratio of expected queue lengths at ρ = 0.6 and ρ = 0.95 for different values of θ. Firstly, as we would expect the effect of utilization is significant. At ρ = 0.95 the expected queue lengths are roughly 8 times higher than at ρ = 0.6. Secondly, we see that abandonment does not dampen the effect of high utilization, to the same extent that it dampened the effect of nonstationarity. As θ increases from 12 to 60, the ratio of expected queue lengths only decreases from about 9 to about 7.

***Table 7.*** *Ratio of expected queue lengths at ρ = 0.6 and ρ = 0.95 for different values of θ.*

| P θ | 0.6 | 0.95 | Ratio |
|---|---|---|---|
| 12 | 2.05 | 18.74 | 9.13 |
| 24 | 1.27 | 11.17 | 8.78 |
| 36 | 0.94 | 7.52 | 7.94 |
| 48 | 0.76 | 5.67 | 7.42 |
| 60 | 0.64 | 4.56 | 7.06 |

In Figure 8 below, we show the show the graph of the Error Measure in expected queue length versus θ for different values of ρ. The Error Measure remains in a tight range of 75%-97%. It is highest at 97.2% for θ = 12 and ρ = 0.6. It is lowest at 75.5% for θ = 60 and ρ = 0.95. It hardly varies with θ or ρ. What drives its value is the Relative Amplitude, which is a constant 0.8 in this case. From our previous numerical work, we would expect the Error Measure to be in the range of 80%-90% in this case, given a RA of 0.8. This turns out to be the case.

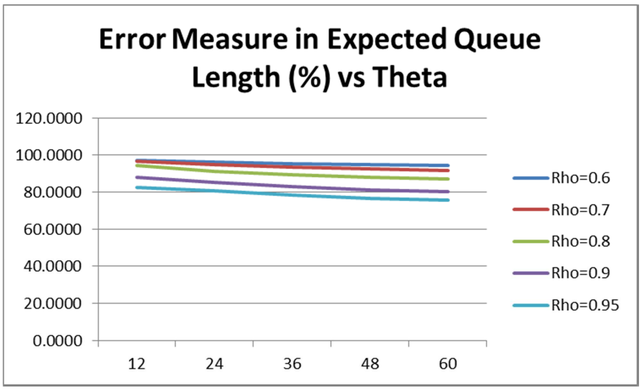

***Figure 8.*** *Error Measure in Expected Queue Length(%) versus θ for different values of ρ.*

We next present our Conclusions.

## 5. Conclusions

We numerically study nonstationarity and abandonment in multiserver Markovian queues. We obtain our numerical results by solving a system of differential equations, which represent the birth death equations for the M(t)/M/s+M system. We numerically study the performance of this nonstationary queue with abandonment and compare its performance measures with those of the stationary M/M/s+M queue. We find that approximating a nonstationary system by a stationary system, even at low levels of nonstationarity can lead to significant errors. Similar results in a system without abandonment, have been obtained by Green, Kolesar and Svoronos [11]. It is therefore important to explicitly model the nonstationary behavior of arrivals in a call center and ignoring this behavior can lead to erroneous decision making.

Additionally, we find that abandonment dampens the effect of nonstationarity. At a high level of abandonment, a nonstationary system behaves closer to a stationary one. Since abandonment and nonstationarity are both present together in real call centers, a real call center with a high level of abandonment behaves closer to an ideal stationary system. However, we numerically find that abandonment does not dampen the effect of high utilization, to the same extent that it dampened the effect of nonstationarity.

Future work to study nonstationary Markovian queues with abandonment can consider the Pointwise Stationary Approximation (PSA) of Green and Kolesar [22]. Consider a performance measure such as the expected queue length in a nonstationary system with abandonment. The PSA approximates this as follows. First, it uses the expected queue length of the stationary system with the arrival rate being the time-dependent arrival rate of the nonstationary system. It then finds the time average expected queue length. Green and Kolesar [22] found that the PSA is easy to compute and provides good estimates of key performance measures of a nonstationary Markovian system without abandonment. It would be interesting to see if those findings are applicable to a nonstationary system with abandonment.

# Appendix

Runge-Kutta Method of Order 4 for the solution of a System of 2 Differential Equations

Consider the system of differential equations

$$x'(t) = f(t, x(t), y(t))$$
$$y'(t) = g(t, x(t), y(t))$$

with

$$\mathrm{x}(t_0) = x_0 \text{ and } \mathrm{y}(t_0) = y_0$$

A solution to the above system, is a pair of differentiable functions *x(t)* and *y(t)*, which satisfy the above equations, refer Mathews and Fink [20].

A numerical solution to the above system, over the interval a≤t≤b is desired. The interval is divided into M subintervals of width *h = (b-a)/M* and the mesh points are $t_{k+1} = t_k + h$, k = 0,..., M-1.

The Runge-Kutta method of Order 4 is as follows.

$$x_{k+1} = x_k + \frac{h}{6}(f_1 + 2f_2 + 2f_3 + f_4)$$

$$y_{k+1} = y_k + \frac{h}{6}(g_1 + 2g_2 + 2g_3 + g_4)$$

where

$$f_1 = f(t_k, x_k, y_k)$$
$$\mathrm{g}_1 = g(t_k, x_k, y_k)$$
$$f_2 = f\left(t_k + \frac{h}{2}, x_k + \frac{h}{2} f_1, y_k + \frac{h}{2} g_1\right)$$
$$g_2 = g\left(t_k + \frac{h}{2}, x_k + \frac{h}{2} f_1, y_k + \frac{h}{2} g_1\right)$$
$$f_3 = f\left(t_k + \frac{h}{2}, x_k + \frac{h}{2} f_2, y_k + \frac{h}{2} g_2\right)$$
$$g_3 = g\left(t_k + \frac{h}{2}, x_k + \frac{h}{2} f_2, y_k + \frac{h}{2} g_2\right)$$
$$f_4 = f(t_k + h, x_k + hf_3, y_k + hg_3)$$
$$g_4 = g(t_k + h, x_k + hf_3, y_k + hg_3)$$

The function *x(t)* is numerically approximated by the set of points $x_k$, *k = 1,..., M*. Similarly, the function *y(t)* is numerically approximated by the set of points $y_k$, *k = 1,..., M*.

# References

[1] Cleveland, B. 2009. Call center management on fast forward. Colorado Springs, Colorado: ICMI.

[2] Reynolds, P. 2003. Call center staffing: The complete practical guide to workforce management. Nashville (Tennessee): Call Center School Press.

[3] Garnett, O, Mandelbaum, A, Reiman, MI. 2002. Designing a call center with impatient customers. Manufacturing and Service Operations Management. 4 (3): 208-227.

[4] Halfin, S, Whitt, W. 1981. Heavy-traffic limits for queues with many exponential servers. Operations Research. 29: 567-588.

[5] Whitt, W. 1992. Understanding the efficiency of multi-server service systems. Management Science. 38 (5): 708-723.

[6] Daskin, MS. 2010. Service science. New Jersey: John Wiley.

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

[8] Yunan, L, Whitt, W. 2016. Approximations for heavily loaded G/GI/n+GI queues. Naval Research Logistics. 63 (3): 187-217.

[9] Batt, R. J, Terwiesch, C. 2015. Waiting patiently: An empirical study of queue abandonment in an emergency department. Management Science. 61 (1): 39-59.

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

[11] Green, L, Kolesar, P, Svoronos, A. 1991. Some effects of nonstationarity on multiserver Markovian queueing systems. Operations Research. 39 (3): 502-511.

[12] Ross, SM. 1978. Average delay in queues with nonstationary Poisson arrivals. Journal of Applied Probability. 15: 602-609.

[13] Rolski, T. 1984. Comparison theorems for queues with dependent interarrival times. In: Baccelli F, Fayolle G, editors. Modeling and performance evaluation methodology. p. 42-67.

[14] Svoronos, A, Green, L. 1988. A convexity result for single server exponential loss systems with nonstationary arrivals. Journal of Applied Probability. 25: 224-227.

[15] Aksin, ZO, Harker, PT. 2003. Capacity sizing in the presence of a common shared resource: Dimensioning an inbound call center. European Journal of Operational Research. 147 (3): 464-483.

[16] Tirdad, A, Grassman, WK, Tavakoli, J. 2016. Optimal policies of M(t)/M/c/c queues with two different levels of servers. European Journal of Operational Research. 249 (3): 1124.

[17] Cho, Y, Ko, YM. 2020. Stabilizing the virtual response time in single-server processor sharing queues with slowly time-varying arrival rates. Annals of Operations Research. 293 (1): 27-55.

[18] Whitt, W, You, W. 2019. Time-varying robust queueing. Operations Research. 67 (6): 1766-1782.

[19] Hunt, BR, Lipsman, RL, Osborn, JE, Rosenberg, JM. 2012. Differential equations with MATLAB. 3rd ed. New York: John Wiley.

[20] Mathews, JH, Fink, KD. 2004. Numerical methods using MATLAB. 4th ed. New Delhi: Pearson Education.

[21] Abou-El-Ata, MO, Hariri, AMA. 1992. The M/M/c/N queue with balking and reneging. Computers and Operations Research. 19 (8): 713-716.

[22] Green, L, Kolesar, P. 1991. The pointwise stationary approximation for queues with nonstationary arrivals. Management Science. 37 (1): 84-97.