Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Performance Evaluation and Dimensioning of $GI^X/M/C/N$ Systems Through Kernel Estimation

G. M. Gontijo, G. S. Atuncar,
F. R. B. Cruz, and L. Kerbache

Relatório Técnico
RTP-04/2010

Relatório Técnico
Série Pesquisa

Performance evaluation and dimensioning of $GI^X/M/c/N$ systems through kernel estimation

G. M. Gontijo[*], G. S. Atuncar[*], F. R. B. Cruz[*], and L. Kerbache[†]

[*]*Universidade Federal de Minas Gerais, 31.270-901 – Belo Horizonte – MG, Brazil*

[†]*HEC School of Management, Paris, France*

gmgontijo@ufmg.br, gregorio@est.ufmg.br, fcruz@est.ufmg.br, kerbache@hec.fr

Abstract

We extend the analysis of queueing systems for real-life situations where the arrival pattern of customers is unknown. In real systems, we must understand how the choice of a method of estimation influences the configuration of the system. Using kernel smoothing, we evaluate algorithms to estimate performance measures of a $GI^X/M/c/N$ system, including the invariant probability distribution of the number of customers in the system, the blocking probability, the average queue size, and the average client queue time. We successfully apply the method to the calls from a calling center to plan and improve the performance of these important queueing systems.

Keywords: Statistics, kernel estimator, performance evaluation.

1. Introduction

There is a large practical interest in investigating the behavior of general-arrival queueing systems, namely those of $GI^X/M/c/N$ type, because when managing real queueing systems the behavior of the arrival process generally is not known *a priori*. In Kendall (1951) notation, in these queueing systems the inter-arrival times are independent and do not follow any specific distribution (*GI*). The service times follow a Markov process (exponential time). We have c identical servers working in parallel and a maximum capacity of N users that are simultaneously allowed in the system, including those in service. Finally, X is a random variable representing the size of group (bulk) arrivals.

Such queueing systems could be used in situations where we have relative control over how the servers work, but we do not know beforehand how the customers arrive at the system.

Naturally, the mathematical model depends on the type of queueing system considered, and there are several methods to obtain such models. The most widely used methods are those that attempt to explain the density functions of inter-arrival and service times by means of parametric statistical models. Nevertheless, real data rarely fit well into parametric models; instead they often produce intractable models. Exact results for performance evaluation of Markovian and some simple general queueing systems are know (e.g., Gross & Harris, 1985), but such systems are rarely found in real life. Kalashnikov (1994) has warned that "many parts of the theory of queues were developed as a 'pure science' with no practical application". Bareche & Aïssani (2008) assert that real systems are "generally very complicated, so their analysis cannot lead to analytical results or it leads to complicated results which are not useful in practice".

An alternative approach is using nonparametric methods to study queueing systems. Nonparametric methods that use kernel smoothing have received much attention lately (Lima & Atuncar, 2010). Kernel estimators provide a simple way of finding structure in data sets without imposing a specific parametric model (Wand & Jones, 1995), which gives us flexibility to handle virtually any data set. There is extensive literature discussing queueing systems (Allen, 1990) and kernel smoothing (Wand & Jones, 1995) as separate concepts but virtually no study has brought the two concepts together.

The contribution of this paper is twofold. First, we develop algorithms to calculate performance measures of queueing systems where the density of the inter-arrival time is determined by kernel estimators. Second, we evaluate the performance of these algorithms as a function of the kernel estimator, the smoothing window, the intensity rate, and the system size. We also present a study case in a call center that illustrates the usefulness of the method.

First we present a literature review about queueing systems and kernel estimators as well as the fundamental concepts required to understand the proposed queueing system model. We also discuss the use of kernel estimators, the chosen models, and the issue of

selecting the smoothing parameter. Then, we describe how we estimated system performance and present the comparative results of simulations with the different methods discussed. We then apply our methods to a call center case and end with our main conclusions and some ideas for future work in the area.

2. Literature Review and Fundamental Concepts

2.1 Previous Works

There are many situations in real life where queues occur and queueing models may be helpful. Recently, queueing models have been used successfully in manufacturing processes (Andriansyah et al., 2010; Dimitriou & Langaris, 2010; Smith et al., 2010); transportation (Cruz et al. 2010b); airports, ports, and product distribution systems (vanWoensel et al., 2008); computer and telecommunication systems (Tang et al., 2010; Cruz et al., 2010a); call center modeling (Jouini et al., 2010); and the analysis of health systems (Osorio & Bierlaire, 2008). Queues may cause the quality of the services or the prices of the goods to rise or fall, depending on their efficiency (vanWoensel & Cruz, 2009), which may be estimated by means of the mathematical tools developed in queueing theory.

As we mentioned earlier, there is not much literature that incorporates both general arrival queueing systems and kernel smoothing. Takács (1962) analyzed a closed solution for various systems that have non-specific distributions, including some multiserver queues such as *GI/M/c* and *M/G/c*. Hokstad (1975) established some closed form results to the *GI/M/c/N* system. Chaudhry & Templeton (1983) analyzed various types of queue with bulk arrivals. Vijaya Laxmi & Gupta (2000) defined the linear equations needed to solve the $GI^X/M/c/N$ system. Zhao (2004) proposed a closed form solution for the $GI^X/M/c$ system, and Bareche & Aïsani (2008) proposed a method to evaluate the proximity of *GI/M/1* and *M/M/1* systems when the density of the inter-arrival time is estimated by kernel estimators.

Concerning kernel smoothing, Wand & Jones (1995) introduced general fundamental concepts. Regarding the issues related to the asymmetry of the random variable under analysis (e.g., non-negativity), Zhang et al. (1999) proposed a boundary corrected kernel estimator based on pseudodata generation, transformation, and reflection around the *Y*-

axis. Chen (2000) proposed the use of a gamma kernel to avoid boundary problems present in certain situations. Scaillet (2004) studied the application of other asymmetric kernels. Bouezmarni & Scaillet (2005) were concerned about the consistency of these asymmetrical estimators. For recent developments in the area of kernel smoothing and a thorough literature review, see Atuncar et al. (2008) and Lima & Atuncar (2010).

2.2 The $GI^X/M/c/N$ Model with Partial Blocks

Vijaya Laxmi & Gupta (2000) described a generalization of the system *GI/M/c* when customers arrive in groups of size X with $P(X = i) = g_i$ ($i \geq 1$) and mean $E(X) = \bar{g}$. Let B_n be the number of clients who were served between the arrival of the nth customer and its successor. Therefore, the number of clients the nth customer finds in the system at the arrival, Y_n, would depend on X_n and B_n, such that $Y_{n+1} = [Y_n + X_n - B_n]^+$. Because Y_{n+1} depends only on Y_n, B_n, and X_n but not on Y_{n-1}, Y_{n-2}, and so on, the stochastic process $\{Y_n\}$ is a first order Markov Chain.

The $GI^X/M/c/$N is a finite capacity system such that a customer that arrives to the saturated system is refused with a probability that we will call P_{BL}, or partial blocks. This refers to the case in which an arrival group of a size greater than the remaining spots in the system is partially denied according to the number of the remaining vacancies until the system is complete.

Vijaya Laxmi & Gupta (2000) report that when the traffic intensity rate $\rho = (\lambda\bar{g}/c\mu)$ is smaller than 1, this Markov Chain has an invariant probability distribution of $\pi_k = \lim_{n\to\infty} P(Y_n = k), k = 0,1,2,\ldots$ associated with the number of clients an arbitrary customer finds in the system at arrival. The π_k 's are often called pre-arrival probabilities.

Pre-arrival probabilities can be determined by the following system of linear equations:

$$\pi_k = \sum_{j=0}^{N} p_{jk}\pi_j ,$$

for k = 0, 1, 2, ..., N-1, and

$$\sum_{j=0}^{N} \pi_j = 1,$$

which gives

$$\begin{bmatrix} (p_{0,0}-1) & p_{1,0} & \cdots & p_{N,0} \\ p_{0,1} & (p_{1,1}-1) & & \vdots \\ \vdots & & \ddots & \vdots \\ p_{0,N-1} & \cdots & \cdots & (p_{N,N-1}-1) \\ 1 & \cdots & \cdots & 1 \end{bmatrix} \begin{bmatrix} \pi_0 \\ \pi_1 \\ \vdots \\ \vdots \\ \pi_N \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix},$$

where p_{jk} are named transition probabilities such that

$$\begin{aligned} p_{jk} &= P\{Y_{n+1} = k \mid Y_n = j\} = \sum_{i=1}^{\infty} P\{X_n = i, Y_{n+1} = k \mid Y_n = j\} \\ &= \sum_{i=1}^{\infty} g_i P\{Y_{n+1} = k \mid Y_n = j, X_n = i\} = \sum_{i=1}^{\infty} g_i P(B_n = j-k+i) \\ &= \sum_{i=1}^{\infty} g_i \int_0^{\infty} \pi_{j+i,k}(z) d\tau(z). \end{aligned}$$

where $i \geq k-j$, $j \leq N$, $k \leq N$, $\pi_{j+i,k}(z)$ is the serving probability of ($j - k + i$) clients under the assumption of the inter-arrival time $\tau_n = z$, and $\tau(z)$ is the inter-arrival time distribution. We shall analyze how $\pi_{j+i,k}(z)$ behaves.

When $j+i \geq c$ and $k \geq c$, there will be more clients than the servers can handle in the entire interval. Because the service process is Markovian, we can treat the server group as a single unit that serves customers at a rate $c\mu z$ and a Poisson distributed transition probability:

$$\pi_{j+i,k}(z) = e^{-c\mu z} \frac{(c\mu z)^{j+i-k}}{(j+i-k)!}.$$

When $j+i<c$, all clients within the system are being serviced and only *k* customers will remain in the system to time *z*. Knowing that the probability of a service time greater than *z* is $e^{-\mu z}$, we can describe this transition probability as a Binomial distribution:

$$\pi_{j+i,k}(z)=\binom{j+i}{k}e^{-k\mu z}\left(1-e^{-\mu z}\right)^{j+i-k}.$$

When $j+i\geq c$ and $k<c$, there will be $(j+i-c)$ customers waiting and c customers being served at the beginning of the interval, but $(c-k)$ spots at the end. Let y be the interval ending immediately before $(j+i-c+1)$ clients are served. If each service time is exponentially distributed with a rate $c\mu$, then y is gamma distributed with a shape $(j+i-c+1)$ and rate a $c\mu$. The other c customers will be served in a time $(z-y)$ and only k will remain. The transition probability of this subinterval will follow a Binomial distribution with success probability $e^{-\mu(z-y)}$. The values $\pi_{j+i,k}(z)$ can be obtained by the convolution of these two variables:

$$\begin{aligned}
\pi_{j+i,k}(z) &= \int_0^z e^{-c\mu y}\frac{(c\mu)^{j+i-c+1}y^{j+i-c}}{(j+i-c)!}\binom{c}{k}e^{-k\mu(z-y)}(1-e^{-\mu(z-y)})^{c-k}\,dy \\
&= \binom{c}{k}e^{-k\mu z}\int_0^z \frac{(c\mu y)^{j+i-c}}{(j+i-c)!}e^{k\mu y}e^{-c\mu y}(1-e^{\mu y}e^{-\mu z})^{c-k}\,c\mu dy \\
&= \binom{c}{k}e^{-k\mu z}\int_0^z \frac{(c\mu y)^{j+i-c}}{(j+i-c)!}\left[e^{-\mu y}(1-e^{\mu y}e^{-\mu z})\right]^{c-k}c\mu dy \\
&= \binom{c}{k}e^{-k\mu z}\left[\int_0^z \frac{(c\mu y)^{j+i-c}}{(j+i-c)!}\left(e^{-\mu y}-e^{-\mu z}\right)^{c-k}c\mu dy\right].
\end{aligned}$$

Therefore, we can get the transitions probabilities given by Vijaya Laxmi & Gupta (2000) when we take each transition probability $\pi_{j+i,k}(z)$ as $\pi_{N,k}(z)$ when $i>N-j$:

$$p_{jk}=\begin{cases}\sum_{i=\max\{1,k-j\}}^{N-j}\beta_{j+i-k}g_i+\beta_{N-k}\sum_{k=N-j+1}^{\infty}g_i\,, & k\geq c,\\ \sum_{i=\max\{1,k-j\}}^{N-j}V_{j+i,k}g_i+V_{N,k}\sum_{k=N-j+1}^{\infty}g_i\,, & 0<k<c,\\ 1-\sum_{r=1}^{N}p_{jr}\,, & k=0.\end{cases}$$

They defined the integration of $\pi_{j+i,k}(z)$ by $V_{j+i,k}$ and β_{j+i-k} as

$$V_{j,k}=\begin{cases}0\,, & j<k<c,\\ \int_0^{\infty}\binom{j}{k}e^{-k\mu z}\left(1-e^{-\mu z}\right)^{j-k}d\tau(z)\,, & k\leq j\leq c\,,\\ \int_0^{\infty}\int_0^{z}\binom{c}{k}e^{-k\mu z}\frac{(c\mu y)^{j-c}}{(j-c)!}c\mu\left(e^{-\mu y}-e^{-\mu z}\right)^{c-k}dyd\tau(z)\,, & k<c<j,\end{cases}$$

$$\beta_r=\int_0^{\infty}e^{-c\mu z}\frac{(c\mu z)^r}{r!}d\tau(z),\quad r\geq 0.$$

There is a relationship between the vector of pre-arrival probabilities $\boldsymbol{\pi}$ and the vector of arbitrary time probabilities **P**, related to the number of people that an outside observer finds in the system. Viajaya Laxmi & Gupta (2000) established a method that proved the relationship:

$$P_k=\begin{cases}\frac{\rho c}{\min\{k,c\}\bar{g}}\sum_{i=0}^{k-1}\pi_i\sum_{j=k-i}^{\infty}g_j, & 0<k\leq N,\\ 1-\sum_{i=1}^{N}P_i, & k=0.\end{cases}$$

Performance measures (below) are used, including the average queue length L_q, blocking probability of an arbitrary customer P_{BL}, and average waiting time in the queue W_q, to analyze the efficiency of queueing systems:

$$L_q = \sum_{i=0}^{N} (i-c)P_i,$$
$$P_{BL} = \sum_{i=0}^{N} \pi_i \sum_{j=N-i}^{\infty} \bar{g}^{-1} \sum_{k=j+1}^{\infty} g_j,$$
$$W_q = L_q / [\bar{g}\lambda(1-P_{BL})].$$

2.3 Kernel Estimators

Suppose that we have a sample of the inter-arrival times, X_1, ..., X_n, with an unknown density $\tau(t)$. The kernel estimator is an analytical tool that provides an effective way of revealing the structure behind such a sample.

2.3.1 Gamma Kernel Estimator

Recently, Chen (2000) suggested an asymmetric kernel with naturally varying shape, as a way to avoid allocating weight for negative values. The gamma kernel estimators are always non-negative, free of boundary bias, and achieve the optimal rate of convergence for the mean square error (MSE) in the non-negative kernel estimator class. Bouezmarni & Scaillet (2005) showed that this estimator is consistent and able to avoid boundary bias. Be $K_G(p, q)$ the gamma density function with shape p and rate q. The gamma kernel considered is

$$K_G\left(\frac{t}{b}+1, b\right)(X_j) = \frac{X_j^{t/b} e^{-X_j/b}}{b^{(t/b)+1}\Gamma\left[(t/b)+1\right]},$$

where b is a smoothing parameter that satisfies the condition $b \to 0$, $nb \to \infty$ as $n \to \infty$. The gamma kernel estimator is

$$\hat{\tau}(t;b) = n^{-1} \sum_{j=1}^{n} K_G\left(\frac{t}{b}+1, b\right)(X_j).$$

The smoothing parameter b is critical for the overall performance of the kernel estimator considered. A small b leads to a relatively bumpy density while a large one results in a smooth density. There are several methods to determine the best fit, from a minimization

of the mean integrated squared error (MISE) of $\widehat{\tau}(t;b)$ to the asymptotic behavior of the MISE (AMISE).

a. Least Squares Cross Validation (LSCV) Method

The least squares cross validation (LSCV) method starts from the MISE expansion

$$\mathrm{MISE}\{\widehat{\tau}(x;b)\} = E\int \widehat{\tau}(x;b)^2\,dx - 2E\int \widehat{\tau}(x;b)\tau(x)dx + E\int \tau(x)^2\,dx.$$

The minimization of the first term is equivalent to the minimization of

$$\mathrm{MISE}\{\widehat{\tau}(x;h)\} - E\int \tau(x)^2\,dx = E[\int \widehat{\tau}(x;h)^2\,dx - 2\int \widehat{\tau}(x;h)\tau(x)dx].$$

The right-hand side is unknown because it depends on τ. However, an unbiased estimator for this quantity is

$$\mathrm{LSCV}(h) = \int \widehat{\tau}(x;h)^2\,dx - 2n^{-1}\sum_{i=1}^{n}\widehat{\tau}_{-i}(X_i;h),$$

where $\widehat{\tau}_{-i} = (X_i;h)$ is the density estimate based on the sample with X_i deleted; this is often called the "leave-one-out" density estimator. A disadvantage of this method is that it suffers from high variation.

b. Asymptotic Behavior of the MISE (AMISE) Method

An alternative parameter selector is to consider the asymptotic behavior of the MISE of the gamma kernel estimator. Chen (2000) uses some aspects of the gamma distribution and a Taylor expansion to determine the MISE as follows:

$$\mathrm{MISE}(\hat{\tau}) = b^2\int_0^\infty \left\{x\tau'(x) + \frac{1}{2}x\tau''(x)\right\}^2 dx + (2n\sqrt{b\pi})^{-1}\int_0^\infty x^{-1/2}\tau(x)dx + O(n^{-1}b^{-1/2} + b^2).$$

The asymptotic MISE disregards the last term; therefore, the optimal *b* that minimizes the leading terms above is

$$b_{\text{AMISE}} = \left[\frac{\left(2\sqrt{\pi}\right)^{-1} \int_0^{\infty} x^{-1/2} \tau(x) dx}{4n \int_0^{\infty} \left(x\tau'(x) + 2^{-1} x\tau''(x)\right)^2 dx} \right]^{2/5},$$

where the functions τ, τ' e τ'' are unknown. These quantities are obtained from the fitted gamma density with parameters adjusted from the sample. This solution still requires further study, but our paper shows promising results.

2.3.2. Zhang et al. (1999) Estimator

Zhang et al. (1999) submitted a model that works particularly well when $\tau(0) > 0$ and combines pseudodata creation, its transformation, and its reflection around the *Y*-axis in the following three steps:

Step 1: Transform the original data X_1 , ..., X_n to $g(X_1)$, ..., $g(X_n)$, while keeping the original data, where g is a nonnegative, continuous, and monotonically increasing function from $[0,\infty)$ to $[0,\infty)$.

Based on extensive simulations, the transformation that best suits a broad variety of densities is

$$g(x) = x + dx^2 + Adx^3,$$

where $A > 1/3$ and $d=f'(0)/f(0)$.

Step 2: Reflect the pseudodata, $g(X_1)$, ..., $g(X_n)$, around the origin.

Step 3: Based on the enlarged data sample, $-g(X_1)$, ..., $-g(X_n)$, X_1, ..., X_n, define the new estimator as

$$\hat{\tau}_n(x,h) = \frac{1}{nh} \sum_{j=1}^{n} \left\{ K\left(\frac{x - X_j}{h} \right) + K\left(\frac{x + g\left(X_j\right)}{h} \right) \right\}, \qquad x \geq 0,$$

where h is a smoothing parameter and K is a symmetric probability function with support [-1, 1] like the Epanechnikov kernel:

$$K(t) = \frac{3}{4}(1 - t^2) I_{[-1,1]}.$$

Notice that the transformation g defined above is not available in practice because d is unknown. A good estimator can be obtained when d is written as $(d/dx)\log f(x)|_{x=0}$,

$$d_n = \frac{\log f_n(h) - \log f_n(0)}{h},$$

where

$$f_n(h) = f_n^*(h) + \frac{1}{n^2}, \qquad f_n(0) = \max\left(f_n^*(0), \frac{1}{n^2} \right),$$

$$f_n^*(h) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h} \right), \qquad f_n^*(0) = \frac{1}{nh_0} \sum_{j=1}^{n} K_0\left(\frac{X_j}{h_0} \right),$$

and K_0 is a so-called endpoint kernel, satisfying

$$\int_{-1}^{0} K_0(t)dt = 1, \ \int_{-1}^{0} tK_0(t)dt = 0, \int_{-1}^{0} t^2 K_0(t)dt \neq 0,$$

and

$$h_0 = \left\{ \frac{\left[\int_{-1}^{1} K(t)dt \right]^2 \int_{-1}^{0} K_0(t)^2 dt}{\left[\int_{-1}^{0} t^2 K_0(t)dt \right]^2 \int_{-1}^{1} K(t)^2 dt} \right\} h.$$

Zhang et al. (1999) proved that for $t \geq h$, the effect of reflected pseudodata is insignificant and the estimator can be reduced to the Parzen-Rosenblatt estimator:

$$\hat{\tau}_n(t,h) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{t - X_j}{h} \right).$$

They also stated that

$$\int_0^{\infty} \hat{\tau}_n(t)dt = 1 + \frac{1}{n} \sum_{i=1}^{n} \int_{-X_i/h}^{-g(X_i)/h} K(z)dz.$$

Thus, $\hat{\tau}_n(t)$ only integrates to 1 when $d_n = 0$, so $g_n(X_i) = X_i$, or when $X_i = 0$ for all X_i's, because $g_n(0) = 0$. However, when $n \to \infty$, both limits of the second term will eventually converge to 0 and $\hat{\tau}_n(t)$ will integrate to 1 asymptotically.

Zhang & Karunamuni (1998) used the endpoint kernel

$$K_0(t) = 12(1+t)\left(\frac{1}{2}+t\right)I_{[-1,0]},$$

and showed that this kernel minimizes the MSE when estimating $\tau(0)$. Therefore, $h_0 = 2h$ is approximately the optimal smoothing parameter for estimating $\tau(0)$ except when $\tau(0) = 0$.

Chiu (1991) described a parameter selecting method that considered the optimal h that minimizes the asymptotic MISE when K is a symmetric probability function with up to the fourth moment being finite

$$h_{AMISE} = \left[\frac{\int K(x)^2\,dx}{n\left(\int x^2 K(x)dx\right)^2 \int \tau''(x)^2\,dx}\right]^{1/5},$$

where the function $\int \tau''(x)^2\,dx$ is unknown.

Chiu's (1991) "plug in" method consists of estimating this quantity through the characteristic function of the sample

$$\hat{\varphi}(\lambda) = n^{-1}\sum_{j=1}^{n} e^{i\lambda X_j},$$

and calculating the optimal h with the previous formula. The characteristic function of τ is

$$\varphi(\lambda) = \int e^{i\lambda x}\tau(x)dx.$$

By the inversion formula, we have

$$\tau(x) = (2\pi)^{-1}\int e^{i\lambda x}\varphi(\lambda)d\lambda,$$

thus,

$$\tau''(x) = (2\pi)^{-1}\int \lambda^2 e^{i\lambda x}\varphi(\lambda)d\lambda,,$$

and

$$\int \tau''(x)^2\,dx = \int [(2\pi)^{-1}\int \lambda^2 e^{i\lambda x}\varphi(\lambda)d\lambda]^2\,dx = \int \lambda^4[(2\pi)^{-1}\int e^{i\lambda x}\varphi(\lambda)d\lambda]^2\,dx = \int \lambda^4[\tau(x)]^2\,dx.$$

Using Parseval's identity, we can show that

$$\int \lambda^4[\tau(x)]^2\,dx = (2\pi)^{-1}\int \lambda^4\left|\varphi(\lambda)\right|^2 d\lambda.$$

Chiu (1991) introduced a cutoff value Λ for λ, such that $\left|\hat{\varphi}(\lambda)\right|^2 < c/n$. Extensive computational experiments from Bessegato et al. (2002) show that $c = 3$ is the value that minimizes the estimator variance. The final "plug in" estimator is then

$$\hat{h}_{\text{AMISE}} = \left[\frac{\int K(x)^2\,dx}{n\left(\int x^2K(x)dx\right)^2\pi^{-1}\int_0^{\Lambda}\lambda^4\left[\left|\hat{\varphi}(\lambda)\right|^2 - n^{-1}\right]d\lambda}\right]^{1/5}.$$

3. Experimental Results

This section presents some results of simulations for $GI^X/M/c/N$ systems with partial blocks, where the inter-arrival time is estimated through the following kernel methods:

- Gamma kernel estimator with LSCV method;
- Gamma kennel estimator with optimal b_{AMISE}.
- Zhang et al. (1999) estimator with Chiu's (1991) "plug in" method.

To evaluate the performance of the estimators above, we will compare the mean square error (MSE) of each estimated pre-arrival probability in the following way:

Step 1: Generate a sample of size n of general inter-arrival distribution τ;

Step 2: Calculate the mean service rate $\mu = \bar{g}\,(\rho c\mathrm{E}(\tau))^{-1}$;

Step 3: Estimate the optimal smoothing parameter h or b;

Step 4: Use the kernel density method to estimate the theoretical density function $\tau(x)$;

Step 5: Find each estimated transition probability;

Step 6: Solve the linear system:

$$\begin{bmatrix} (\hat{p}_{0,0}-1) & \hat{p}_{1,0} & \cdots & \hat{p}_{N,0} \\ \hat{p}_{0,1} & (\hat{p}_{1,1}-1) & & \vdots \\ \vdots & & \ddots & \vdots \\ \hat{p}_{0,N-1} & \cdots & \cdots & (\hat{p}_{N,N-1}-1) \\ 1 & \cdots & \cdots & 1 \end{bmatrix} \begin{bmatrix} \hat{\pi}_0 \\ \hat{\pi}_1 \\ \vdots \\ \vdots \\ \hat{\pi}_N \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

The algorithm above was coded in R 2.8.0 (or earlier versions; see R Development Core Team, 2010). The code is available upon request from the authors or directly from the web[1] for educational and research purposes.

The experiments are based on random samples of inter-arrival times of size n = 100. Slightly larger and smaller samples were also tested, but the results (not shown) are similar. The theoretical inter-arrival distributions considered in this experiment were

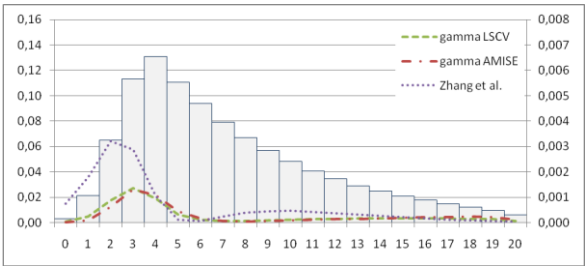
- Weibull distribution with shape = 2 and rate = 20,
- Gamma distribution with shape = 10 and rate = 2, and
- Gamma mixture distribution of 0.45×gamma(5; 2) + 0.55×gamma (30; 1).

The Weibull density has $\tau'(0) \neq 0$, the pure gamma has $\tau'(0) = 0$, and the gamma mixture is bimodal. We choose Weibull and gamma distributions because of their well-known flexibility for modeling real databases. For simplicity, the group size X is constant and equal to 1. Two different numbers of servers were considered, c = 5 and 10, with maximum capacities of N = 20 and 25, respectively (which result in buffers of fixed sizes equal to 10) (Figs. 1, 2).

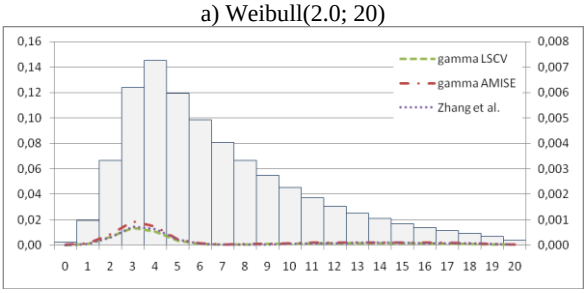
In general, the errors decreased as the number of servers c increased, and they were dependent on the theoretical distribution considered. We obtained the largest errors for the last distribution (gamma mixture). The LSCV method is better than b_{AMISE} on the first and last distributions. Zhang et al.'s (1999) estimator has the worst performance on the first and a competitive performance on the second distribution (gamma distribution). All the estimators performed well with the second distribution. The last distribution (gamma mixture) has a particular behavior. Although the errors are large for most of the

[1] URL: http://www.est.ufmg.br/ftp/fcruz/kernel/

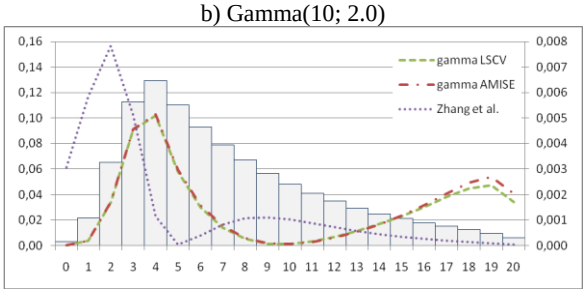
estimates, the error is small for the blocking probability (i.e., $\hat{\pi}_{20}$, for the system with c = 5, Figure 1, and $\hat{\pi}_{25}$, for c = 10, Figure 2). The blocking probability is an important performance measure because it indicates the fraction of costumers lost by the system.



a) Weibull(2.0; 20)

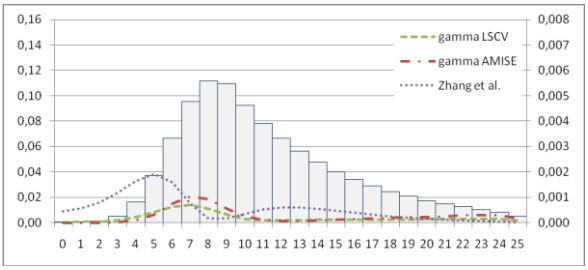


b) Gamma(10; 2.0)

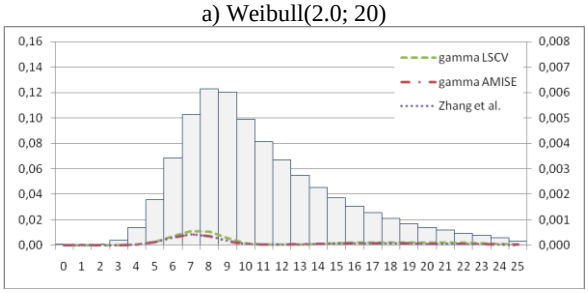


c) 0.45Gamma(5.0; 2.0) + 0.55Gamma(30; 1.0)

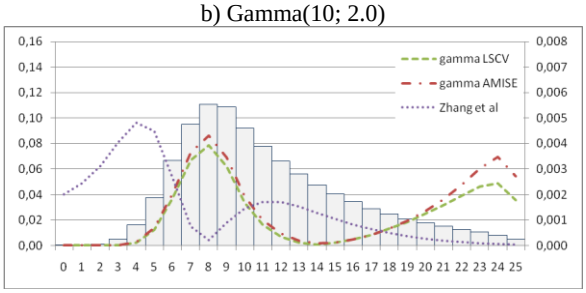
Figure 1: πn and EQM($\hat{\pi}_n$) for c = 5



a) Weibull(2.0; 20)



b) Gamma(10; 2.0)



c) 0.45Gamma(5.0; 2.0) + 0.55Gamma(30; 1.0)

Figure 2: πn and EQM($\hat{\pi}_n$) for c = 10

4. Application to a Real Call Center Data Set

We analyzed a real database of 7,761 phone calls made on April 7, 2006, from 8:00 am until 1:00 pm to a call center (data available from the authors upon request or directly from the web[1]). We worked on a set of inter-arrival and service times to find the required minimum size of the server facility taking into account that the arrival rates are not homogeneous. In fact, larger inter-arrival times were detected toward the beginning of the day rather than throughout the day. This will lead to the adjustment of a different system for each hour of service that had an approximately homogeneous arrival rate. Using this data, we will model a $GI^X/M/c/N$ system with a kernel inter-arrival density estimate for each hour of service.

The data set is presented in seconds, which is the precision given by the data acquisition system. This constraint will lead to ties in the arrival times (i.e., more than one call can arrive in the same second). Therefore, we treated same-time calls as part of a single arrival group and fit a discrete probability distribution. Table 1 shows the observed frequency of group sizes X and the fitted distribution D by hour.

Table 1: Observed frequency of group sizes *X* and fitted distributions *D*

Time:	8am – 9am	9am – 10am	10am – 11am	11am – 12am	12am – 1pm
1 arrival	0,915	0,808	0,794	0,882	0,916
2 arrivals	0,085	0,158	0,172	0,108	0,075
3 arrivals	0,000	0,031	0,030	0,009	0,009
4 arrivals	0,000	0,003	0,004	0,001	0,000
$D \sim$	Poisson(0.082)	Geom.(0.813)	Geom.(0.804)	Geom.(0.885)	Geom.(0.914)

We used the gamma kennel-estimator with optimal b_{AMISE} as the kernel density estimation method. This method gave estimated densities bounded at $\tau = 0$ and illustrates the information we miss when data is rounded. We also considered the gamma kernel estimator with LSCV method, but it did not behave as well with discrete data. Table 2 shows the smoothing parameter calculated, and Figure 3 shows the estimated densities.

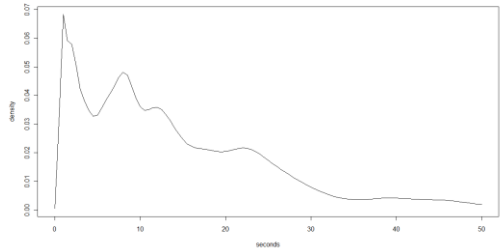
Table 2: AMISE smoothing parameter for gamma kernel

Time:	8am – 9am	9am – 10am	10am – 11am	11am – 12am	12am – 1pm
b_{AMISE}	0,1509	0,0505	0,0476	0,0625	0,0932

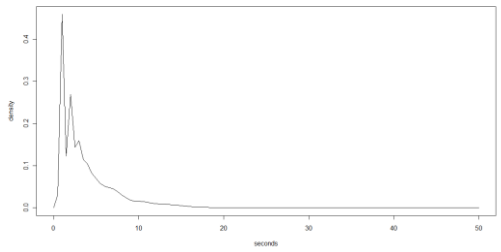
Service time distribution fits an exponential distribution with parameter $\lambda = 0.003339$ and standard error of 3.8×10^{-5}. The behavior of the system then becomes restricted to the choice of the number of servers c and capacity of the system N. In Table 3, the minimum number of servers required is set to maintain the stability of the system at each hour considered.

Table 3: Minimum number of servers

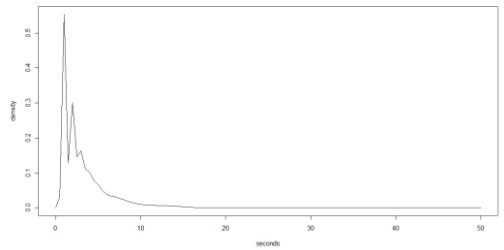
Time:	8am – 9am	9am – 10am	10am – 11am	11am – 12am	12am – 1pm
min{c}	19	97	110	64	40



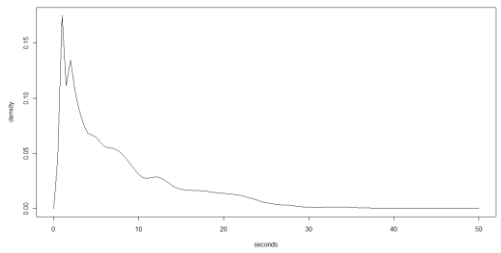
a) 8:00am-9:00am



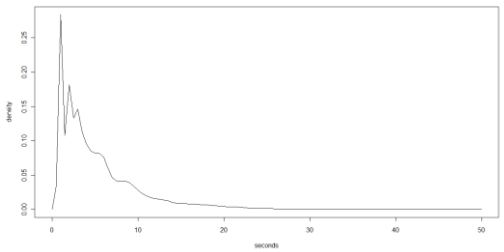
b) 9:00am-10:00am



c) 10:00am-11:00am



d) 11:00am-12:00am



e) 12:00am-1:00pm

Figure 3: Estimated inter-arrival times by hour

To optimize system performance, it is necessary to establish some criterion function for the adopted design. If we only consider the effective arrival rate ($\lambda_{EF} = \lambda[1\text{-}P_{BL}]$), we can describe this performance measure according to the number of servers and the maximum queue size. Figure 4 shows this relationship for 8:00 to 9:00am. We saw similar results (not shown) for different periods. A good criterion would be the $C = (C_1/\lambda_{EF} + C_2c + C_3N)$, where C_1, C_2 , and C_3 are costs related to each parameter and defined by the system environment. As an example, if we have $C_1 = 500$, $C_2 = 1$ and $C_3 = 1$, the state with

minimal criterion would be $c = 19$ and $N = 20$. If cost C_1 is raised to 600, the best configuration would be $c = 21$ and $N = 21$.

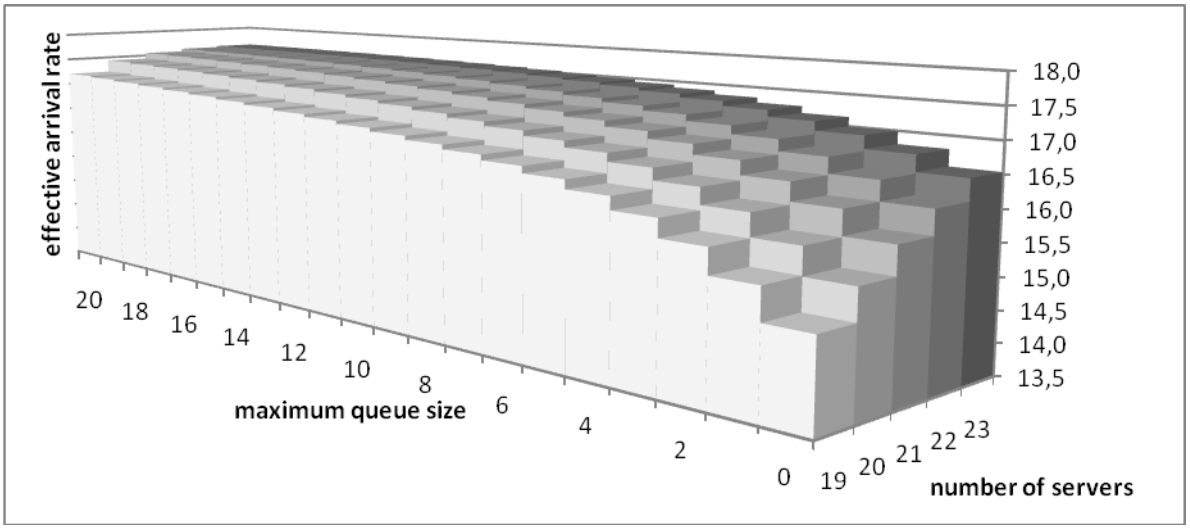


Figure 4: Effective arrival rate as a function of N and c

Figure 5 shows the invariant distributions estimated for other periods with different queue and server facility sizes. The method seems helpful for adjusting the maximum size *N*, and it is clear that the period plays a key role.

5. Concluding Remarks

We studied the adequacy of the kernel estimator methods for calculating the invariant probability distribution and performance measures of queueing systems that have general inter-arrival distribution times with bulk arrivals. Simulations showed that when $\tau'(0) \neq 0$, the gamma kernel method had the best performance. This suggests the Zhang et al. (1999) method does not work well when $\tau(0) > 0$. At the same time, its behavior with the bimodal density showed a very low EQM for probabilities near the maximum state. This implies a good estimation of the blocking probability and other performance measures.

The method used to select the smoothing parameter for the gamma kernel estimator had no effect. A better selector would have the function of its AMISE optimal parameter estimated like the "plug in" method for symmetric kernels. The combined use of smoothing parameter selection and Bayesian techniques is promising (see Lima & Atuncar, 2010).

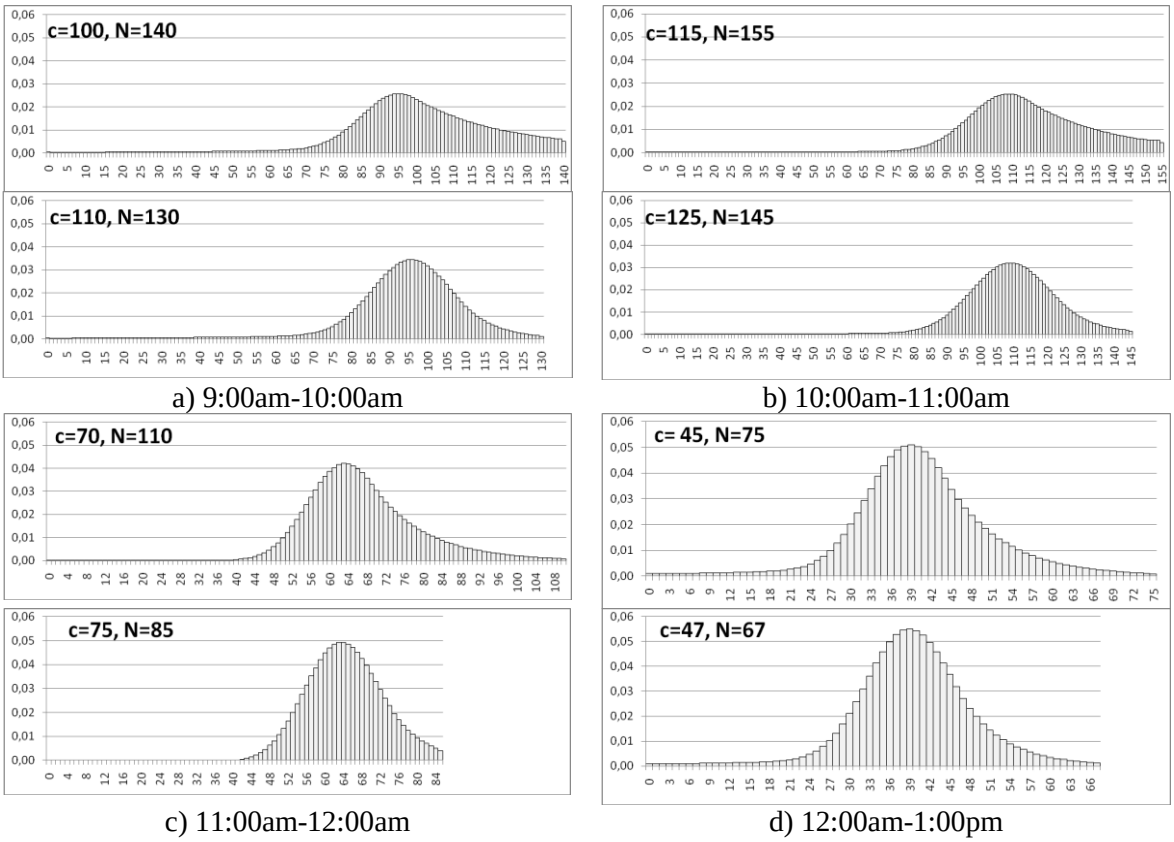


Figure 5: Estimated invariant distributions

Future research may take other directions—for instance, determining how the estimate of the pre-arrival invariant distribution moves away from its real value. One approach is from the variance of the bias and variance of each probability estimated. For example, the variance of an estimate of transition probability p_{ij} could depend on the variance of term $\hat{\beta}_0$, so we would need to find

$$\text{var}(\hat{\beta}_0) = \text{var}(\int_0^{\infty} e^{-c\mu z} \hat{\tau}(z,h)dz).$$

This research can be applied to several areas of practical interest, including health and industry. For example, it could be critical in a medical emergency room that must optimize resource allocation.

Acknowledgments

This research is supported by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico; grants 201046/1994-6, 301809/1996-8, 307702/2004-9, 472066/2004-8,

304944/2007-6, 561259/2008-9, 553019/2009-0, 550207/2010-4, 501532/2010-2, 303388/2010-2), by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior; grant BEX-0522/07-4), by FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais, grants CEX-289/98, CEX-855/98, TEC-875/07, CEX-PPM-00401/08, and CEX-PPM-00390-10), and by PRPq-UFMG (Pró-Reitoria de Pesquisa da Universidade Federal de Minas Gerais).

Role of the Funding Source

The Brazilian government funding agencies mentioned above had no role in the study.

References

Allen, A. O. (1990). Probability Statistics and Queuing Theory with Computer Science Applications, Academic Press, Sand Diego, CA, USA.

Andriansyah, R.; van Woensel, T.; Cruz, and F. R. B. & Duczmal, L. (2010) Performance optimization of open zero-buffer multi-server queueing networks, Computers & Operations Research 37(8):1472-1487.

Atuncar, G. S.; Dorea, C. C. Y. & Gonçalves, C. R. (2008). Strong consistency of kernel density estimates for Markov chains failure rates, Statistical Inference for Stochastic Processes 11(1):1-10.

Bareche, A. & Aïssani, D. (2008). Kernel density in the study of strong stability of the *M*/*M*/1 queuing system, Operation Research Letters 36, 535-538.

Bessegato, L. F.; Atuncar, G. S. and Duczmal, L. H. (2006). Routines in R for Kernel Smoothing Techniques (in Portuguese), Technical Report, Departamento de Estatística – ICEx – UFMG, Belo Horizonte, Brazil (available at url: http://www.est.ufmg.br/portal/arquivos/rts/rtp0601.pdf).

Bouezmarni, T. & Scaillet, O. (2005). Consistency of asymmetric kernel density estimators and smoothed histograms with application to income data, Econometric Theory 21, 390-412.

Chaudhry, M. L. & Templeton, J. G. C. (1983). First Course in Bulk Queues. John Wiley & Sons, Inc., Somerset, NJ, USA.

Chen, S. X. (2000). Probability density function estimation using gamma kernels, Annals of the Institute of Statistical Mathematics 52, 471-480.

Chiu, S. T. (1991). Bandwidth selection for kernel density estimation, Annals of Statistics 33, 1883-1905.

Cruz, F. R. B.; Oliveira, P. C. & Duczmal, L. (2010a). State-dependent stochastic mobility model in mobile communication networks, Simulation Modelling Practice and Theory 18(3):348-365.

Cruz, F. R. B.; van Woensel, T.; Smith, J. M. & Lieckens, K. (2010b). On the system optimum of traffic assignment in *M*/*G*/*c*/*c* state-dependent queueing networks, European Journal of Operational Research 201(1):183-193.

Dimitriou, I. & Langaris, C. (2010). A repairable queueing model with two-phase service, start-up times and retrial customers, Computers & Operations Research 37(7):1181-1190.

Gross, D., Harris, C. M., (1985). Fundamentals of Queueing Theory. Wiley, New York. NY, USA.

Hokstad, P. (1975). *G*/*M*/*m* queue with finite waiting room, Journal of Applied Probability. 12, 779–792.

Jouini, O.; Pot, A.; Koole, G. & Dallery, Y. (2010). Online scheduling policies for multiclass call centers with impatient customers, European Journal of Operational Research 207(1):258-268.

Kalashnikov, V. V. (1994). Mathematical Methods in Queuing Theory, Springer, New York, NY, USA.

Kendall, D. G. (1951). Some problems in the theory of queues, Journal of the Royal Statistical Society Serie B 13:151-173.

Lima, M. S. & Atuncar, G. S. (2010). A Bayesian method to estimate the optimal bandwidth for multivariate kernel estimator, Journal of Nonparametric Statistics (in press), doi: 10.1080/10485252.2010.485200.

Osorio, C. & Bierlaire, M. (2009). An analytic finite capacity queueing network model capturing the propagation of congestion and blocking, European Journal of Operational Research, 196(3):996-1007.

R Development Core Team (2010). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Scaillet, O. (2004). Density estimation using inverse and reciprocal inverse Gaussian kernels, Journal of Nonparametric Statistics 16:217–226.

Smith, J. M.; Cruz, F. R. B. & van Woensel, T. (2010). Topological network design of general, finite, multi-server queueing networks, European Journal of Operational Research 201(2):427-441.

Takács, L. (1962). Introduction to the Theory of the Queues, Oxford University Press, Oxford, UK.

Tang, L.; Xi, H-s.; Zhu, J. & Yin, B-q. (2010). Modeling and optimization of *M*/*G*/1-type queueing networks: An efficient sensitivity analysis approach, Mathematical Problems in Engineering (Article ID 130319) 20 pages.

VanWoensel, T. & Cruz, F. (2009). A stochastic approach to traffic congestion costs, Computers & Operations Research 36 (6):1731-1739.

VanWoensel, T.; Kerbache, L.; Peremans, H. & Vandaele, N. (2008). Vehicle routing with dynamic travel times: A queueing approach, European Journal of Operational Research 186 (3) 990–1007.

Vijaya Laxmi, P. & Gupta, U.C. (2000). Analysis of finite-buffer multi-server queues with group arrivals: GI^X/M/c/N, Queueing Systems 36:125-140.

Wand, M. P. & Jones, M. C. (1995). Kernel Smoothing. Chapman and Hall/CRC; Boca Raton, FL. USA.

Zhang, S. & Karunamuni, R. J. (1998). On kernel density estimation near endpoints, Journal of Statistical Planning and Inference 70:301-316.

Zhang, S., Karunamuni, R. J. & Jones, M. C. (1999). An improved estimator of density function at the boundary, Journal of the American Statistical Association 94:1231-1241.

Zhao, Y. (1994). Analysis of the $GI^X/M/c$ model, Queueing Systems 15:347-364.