KERNEL DENSITY ESTIMATION OF ARRIVALS IN GI[X]/M/C/N QUEUES

F. R. B. Cruz, M. A. C. Santos
Departamento de Estatística, Universidade Federal de Minas Gerais,
31270-901 – Belo Horizonte – MG, Brazil
fcruz@est.ufmg.br, msantos@est.ufmg.br

F. L. P. Oliveira
Departamento de Estatística, Universidade Federal de Ouro Preto,
35400-000 – Ouro Preto – MG, Brazil
fernandoluiz@iceb.ufop.br

N. L. C. Brito
Departamento de Ciências Exatas – CCET, Universidade Estadual de Montes Claros,
39401-089 – Montes Claros – MG, Brazil
nilson.brito@unimontes.br

ABSTRACT

We focus here on the use of kernels for the arrival process estimation of $GI^{[X]}/M/c/N$ queues, which in Kendal notation stands for independent general (GI) distributed bulk inter-arrival times of size X, Markovian (M) service times, c identical servers working in parallel, and a maximum capacity of N users simultaneously allowed in the systems, including those under service. For the problem on hand, we have selected some kernel methods and implemented them in R language. Some simulation results are presented to attest for the quality of the inferences obtained and real data obtained from a call center were analyzed.

KEYWORDS. Kernel estimator. Performance evaluation. Finite queues.

Main area EST – Estatística. MP – Modelos Probabilísticos

1. Introduction

The performance evaluation of $GI^{[X]}/M/c/N$ queueing systems is the main interest here. In Kendall (1953) notation, GI represents a general independent inter-arrival time distribution, in groups of size X, M represents an exponentially (Markovian) distributed service time, c, the number of servers in parallel, and N, the maximum number of users in the systems including those under service. These queues are interesting in practical terms specially when one can control the servers but has no control over how the groups of customers are coming into the system and their sizes.

Parametric models are the usual way to describe the arrival processes. However many times data unfortunately do not fit well into parametric models or else, when they fit, they lead to mathematically or computationally intractable models. Additionally, exact results are only available for Markovian systems or some very simple general queues (see Gross et al., 2009). And such systems are sometimes a bit away from the needs for the real life problems.

Then it is proposed here a combination of queues and kernels as a non-parametric approach that has been quite successful in some other similar practical applications. Indeed, kernel-based methods have received increasing attention from researchers from many areas (Lima & Atuncar, 2011; Bareche & Aïssani, 2014) and provided a simple way of finding structure in data sets without imposing a specific parametric model (Wand & Jones, 1985; Gustafsson et al., 2009). This brings flexibility to handle real-life data sets.

There is an extensive literature discussing as separate concepts queues (Allen, 1990) and kernels (Silverman, 1986; Malec & Schienle, 2014) but there are not so many papers bringing together these two concepts. Gontijo et al. (2010) proposed kernel based algorithms to performance evaluation and dimensioning of $GI^{[X]}/M/c/N$ systems but they have not provided an in-depth analysis of the arrival process as done here.

Two main contributions are brought by this paper. Firstly, a comprehensive simulation study is provided to attest for the quality of the inferences obtained for the queueing system. Secondly, real data obtained from a call center is analyzed by mean of kernel methods.

The rest of this paper is organized as follows. In the next section a brief review of the literature on queues and kernel estimators are presented. Then computational results are presented for simulated data and for a real data set. Finally some remarks and topics for future research in the area are discussed.

2. Materials and Methods

2.1 Bulk Arrival Queues

Vijaya Laxmi & Gupta (2000) have solved the bulk-arrival general-service finite queueing system $GI^{[X]}/M/c/N$. The clients arrive in groups of size X, with $\mathrm{P}(X=i)=g_i$, for $i \geq 1$, $E(X)=\bar{g}$, and traffic intensity defined as $\rho=(\lambda\bar{g}/c\mu)$, in which λ is the arrival rate and μ is the service rate.

They have also shown that the vector of arbitrary time invariant probabilities **p** related to the number of users an outsider observer finds in the system has its components given by:

$$p_k = \begin{cases} \dfrac{\rho c}{\min\{k,c\}\bar{g}} \displaystyle\sum_{i=0}^{k-1} \pi_i \sum_{j=k-i}^{\infty} g_j, & 0 < k \le N, \\ 1 - \displaystyle\sum_{i=1}^{N} p_i, & k = 0, \end{cases} \tag{1}$$

in which $\pi_i = \lim_{n\to\infty} P(Y_n = i), i = 0,1,\ldots,$ are the components of the vector of pre-arrival probabilities $\boldsymbol{\pi}$ related to the number of users an arriving client finds at the system.

The values of π_i are determined as a function of the transition probabilities p_{jk} from the state j to k by means of the linear equation system:

$$\begin{bmatrix} (p_{0,0}-1) & p_{1,0} & \cdots & p_{N,0} \\ p_{0,1} & (p_{1,1}-1) & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ p_{0,N-1} & p_{1,N-1} & \cdots & p_{N,N-1} \\ 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \pi_0 \\ \pi_1 \\ \vdots \\ \pi_{N-1} \\ \pi_N \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}. \tag{2}$$

The transition probabilities p_{jk} are given by:

$$p_{jk} = \begin{cases} \displaystyle\sum_{i=\max\{1,k-j\}}^{N-j} \beta_{j+i-k} g_i + \beta_{N-k} \sum_{i=N-j+1}^{\infty} g_i, & c \le k \le N, \\ \displaystyle\sum_{i=\max\{1,k-j\}}^{N-j} V_{j+i,k} g_i + V_{N,k} \sum_{i=N-j+1}^{\infty} g_i, & 0 < k < c, \\ 1 - \displaystyle\sum_{r=1}^{N} p_{jr}, & k = 0, \end{cases} \tag{3}$$

in which

$$V_{j,k} = \begin{cases} 0, & j < k < c, \\ \displaystyle\int_0^{\infty} \binom{j}{k} e^{-k\mu z} \left(1 - e^{-\mu z}\right)^{j-k} dA(z), & k \le j \le c, \\ \displaystyle\int_0^{\infty}\int_0^{z} \binom{c}{k} e^{-k\mu z} \frac{(c\mu y)^{j-c}}{(j-c)!} c\mu \left(e^{-\mu y} - e^{-\mu z}\right)^{c-k} dy\, dA(z), & k < c < j, \end{cases} \tag{4}$$

$$\beta_r = \int_0^{\infty} \frac{e^{-c\mu z} (c\mu z)^r}{r!} dA(z), \qquad r \ge 0, \tag{5}$$

and the interarrival times are independently identically distributed random variables (remind that the arrivals are in blocks of X users) with cumulative distribution function $A(z)$, probability density function $a(z)$, $z>0$, and mean a.

Once the probabilities p_k are known, Eq.(1), all the main performance measures follow easily, such as the blocking probability p_{block}, the effective arrival rate θ, the mean queue size L_q, and the mean queue time W_q, given by Little Law:

$$\left.\begin{aligned} & p_{\text{block}} = p_N , \\ & \theta = \lambda(1 - p_{\text{block}}), \\ & L_q = \sum_{k=0}^{N} (i - c)\, p_k , \\ & W_q = \frac{L_q}{\bar{g}\theta}. \end{aligned}\right\} \tag{6}$$

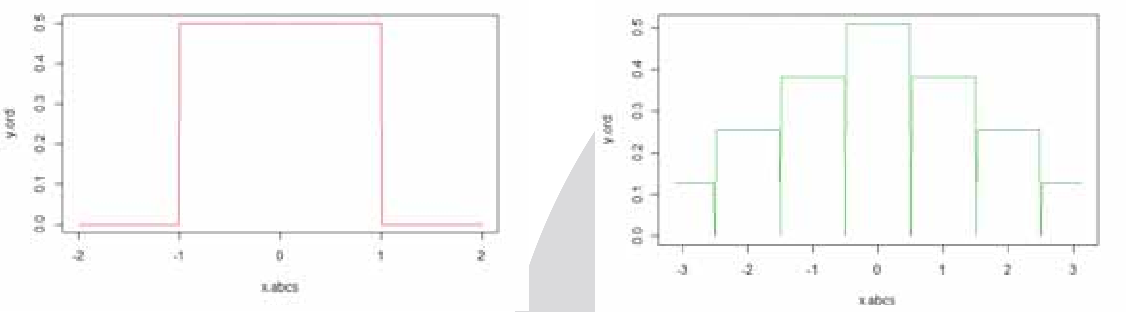
2.2. Kernel Estimators

Let be a sample X_1, X_2,..., X_n from an unknown interarrival time probability density funcion $a(z)$. A possible way to model the interarrival times is through a kernel estimator, which is an analytical tool to reveal the underlying structure of the samples.

The classical model is the Parzen-Rosemblatt estimator:

$$\hat{a}_n(x,h) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right), \tag{7}$$

in which $K(x)$ is a probability density function, symmetrical around 0, and h is a smoothing parameter called window (Wand & Jones, 1995). Figure 1-a presents the well-known kernel $I_{[-1;1]}$, which typically produces estimates as seen in Figure 1-b (that is, a histogram).



a) kernel b) kernel estimates

Figure 1: Kernel $I_{[-1;1]}$ and the underlying estimate

Various different kernels are considered here. Amongst the usual kernels there is the Gaussian (normal) kernel,

$$K_N(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2 / 2\sigma^2}, \tag{8}$$

which is recommended for estimating real functions.

There is also the Gamma kernel,

$$K_G\left(\frac{x}{h} + 1, h\right)(X_j) = \frac{X_j^{x/h} e^{-X_j/h}}{h^{(x/h)} \Gamma[(x/h) + 1]}, \tag{9}$$

proposed by Chen (2000), for non-negative random variables, which avoids allocating densities to negative values (notice that the inter-arrival times do not support negative values).

Finally, a more specialized estimator was tried, proposed by Zhang et al. (1999), which is an alternative to fix the problem of estimating a strictly positive function domain. In their method, data pass through a transformation $g(X_i)$ and the following estimator results:

$$\hat{a}_n(x,h) = \frac{1}{nh} \sum_{j=1}^{n} \left\{ K\left(\frac{x - X_j}{h}\right) + K\left(\frac{x + g(X_j)}{h}\right) \right\}, \tag{10}$$

in which $K(x)$ is some kernel (as one of those shown earlier). The Epanechnikov's kernel (which is based on the $I_{[-1;1]}$ kernel) will be used here:

$$K_E(x) = \frac{3}{4}(1 - x^2) I_{[-1,1]}. \tag{11}$$

3. Numerical Results

All algorithms were encoded in R (R Core Team, 2013). The code is available upon request for teaching and research purposes. The good performance of kernel methods depends on a proper choice of the window h. For the Gaussian and Gamma kernels, the least squares cross validation (LSCV) method is used, which is fast and easy to implement. In fact, the method includes the optimization (minimization) of a well-behaved function, as shown in Figure 2, for simulated data from a Weibull distribution (shape=2; scale=20), sample size n=1000, and a Gamma kernel estimator.

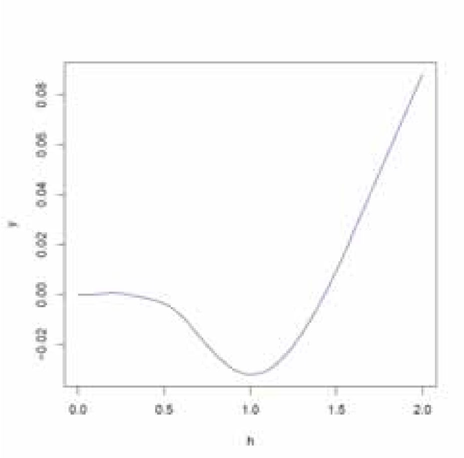
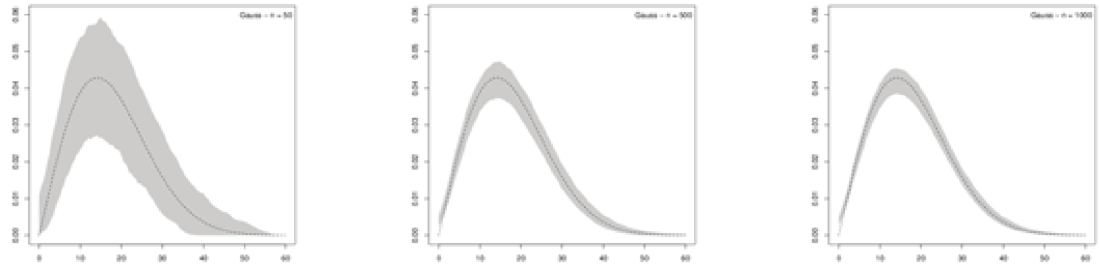


Figure 2: Squares as a function of the smoothing parameter *h*

For Zhang et al.'s kernel, the asymptotic mean integrated square error (AMISE) proposed by Chen (2000) is used combined with Chiu's modified plug-in method (Chiu, 1991; Wand & Jones, 1995).

3.1. Simulation Study

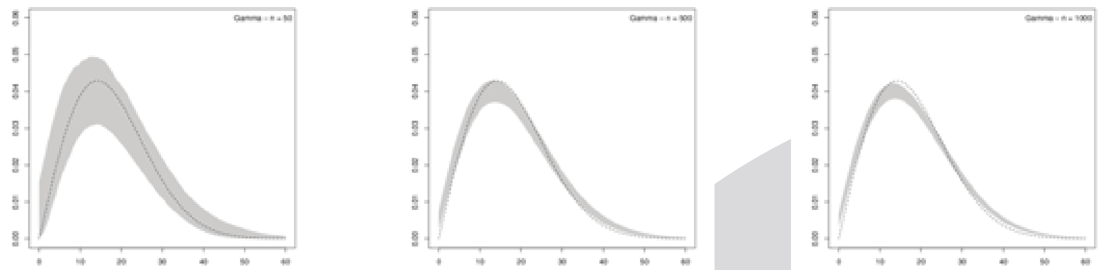
The results for Gaussian kernel, sample sizes n={50; 500; 1,000}, obtained from a Weibull distribution (shape=2; scale=20), can be seen in Figure 3, which shows the theoretical density and the bootstrapped 90% confidence interval based on 1,000 replications for each sample size. As the sample size increases, the variance reduces. For all cases, the confidence interval covers the theoretical density and there is no evidence of bias.



a) *n*=50 b) *n*=500 c) *n*=1,000

Figure 3: Theoretical density and 90% bootstrapped confidence intervals for Gaussian kernel

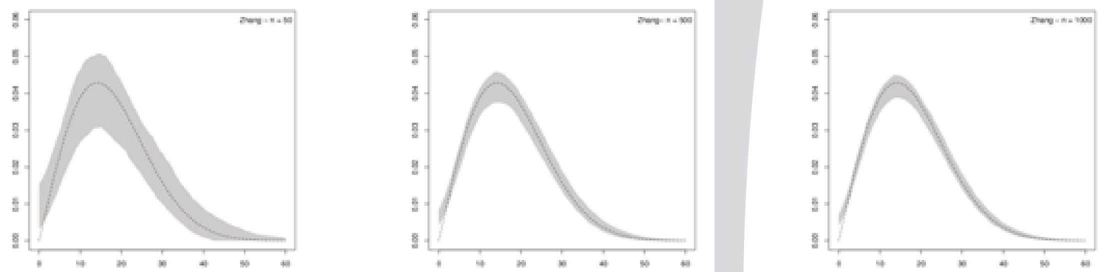
The results for Gamma kernel are seen in Figure 4, for the same Weibull distribution and samples sizes as earlier. The Gamma kernel is based on a non-symmetrical distribution and although we expected it to better model a non-negative random variable (*i.e.*, the interarrival times), what we see here is a large bias as the sample size increases. Definitely, this kernel should not be used to model interarrival times.



a) *n*=50 b) *n*=500 c) *n*=1,000

Figure 4: Theoretical density and 90% bootstrapped confidence intervals for Gamma kernel

Finally, Figure 5 shows the results for Zhang et al.'s kernel, for the same Weibull distribution and samples sizes as earlier. Zhang et al.'s kernel seems to be preferable because shows a similar performance as the Gaussian kernel in terms of low bias and low variability. Besides Zhang et al.'s kernel does not suffer from boundary problems and does not include probability densities in the negative values.



a) *n*=50 b) *n*=500 c) *n*=1,000

Figure 5: Theoretical density and 90% bootstrapped confidence intervals for Zhang et al.'s kernel

In summary, Zhang et al.'s method shows its superiority over the Gaussian and Gamma kernels in terms of mean square errors (MSE) as shown in Figure 6, with the mean square errors (MSE) obtained over 1,000 Monte Carlo simulations. For small sample sizes ($n \leq 400$), Gamma kernel tends to outperform Gaussian kernel. As seen in Figure 6, with sample sizes above 400, the advantage is in favor of the Gaussian kernel. However, it is noticeable that the MSE of Zhang et al.'s method is below both of them.

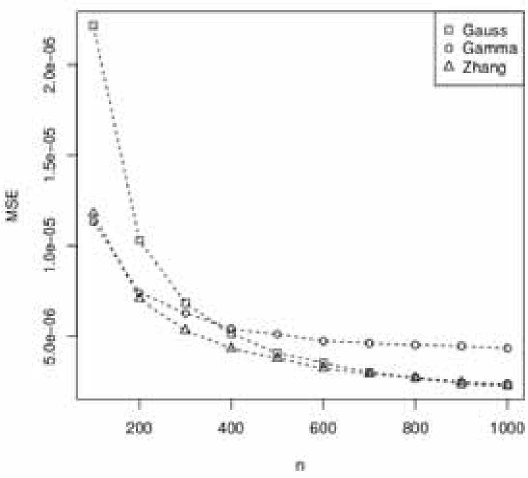


Figure 6: Mean square errors (MSE) as a function of the sample sizes *n* for different kernels

Finally, it is worthwhile mentioning that in other study (not shown) it was found that the use of Gaussian windows for non-negative and close to zero real-valued data tends to imply on probability mass losses in the negative region. Such losses can be associated to the fact that, when a Gaussian window is applied to the convolution integral which outputs the adjusted probability mass function, the negative values of the independent variable will necessarily be taken into account, due to the concentration of observed data values near zero.

3.2. Application to Real Data

Real data were obtained from a call center, from 8:00 am until 1:00 pm. Composed by 7,761 arrival times. Many ties were observed since the precision given by the data acquisition system is in seconds. Then all the calls arriving in the same second are regarded as part of a single arrival group. Table 1 shows some descriptive statistics for the data stratified per hour. It is clear that the arrival rates change significantly along the hours, which justifies data stratification per 1-hour time period. Table 2 shows the observed frequency of the group sizes X, which varies considerably along the time periods. It is possible to notice that the times between the arrivals hardly come from the same distribution.

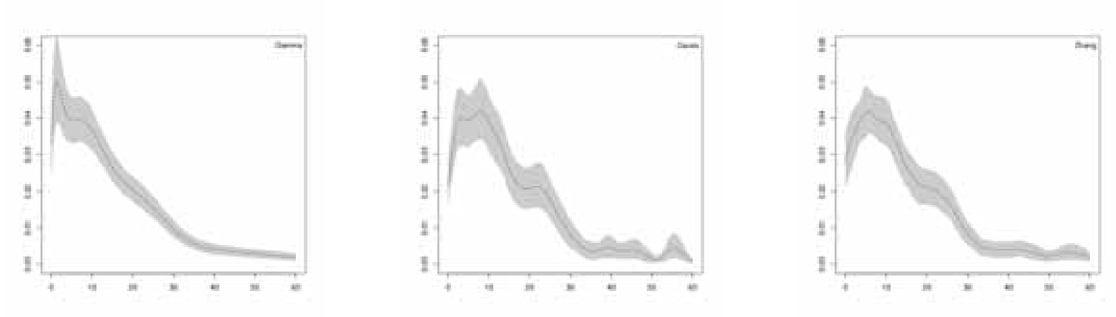
Table 1: Descriptive statistics for the real data analyzed

time period	sample size *n*	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
08 am -01 pm	3,318	1.00	2.00	3.00	5.425	6.00	129
08 am - 09 am	204	1.00	6.00	12.0	17.64	23.0	129
09 am - 10 am	942	1.00	1.00	3.00	3.820	5.00	36.0
10 am - 11 am	1,058	1.00	1.00	2.00	3.405	4.00	35.0
11 am - 12 pm	676	1.00	2.00	4.00	5.297	7.00	35.0
12 pm - 01 pm	438	1.00	2.00	6.00	8.263	12.0	63.0

Table 2: Observed frequency of group size X

time period	number of arrivals			
	1	2	3	4 and plus
08 am - 01 pm	0.915	0.085	0.000	0.000
08 am - 09 am	0.808	0.158	0.031	0.003
09 am - 10 am	0.794	0.172	0.030	0.004
10 am - 11 am	0.882	0.108	0.009	0.001
11 am - 12 pm	0.916	0.075	0.009	0.000

As an example, Figure 7 presents kernel estimations for the 8:00 am to 9:00 am period. Since the simulation studies shown earlier demonstrated that the Zhang et al.'s method has a superior performance in terms of low bias and low variability, this method should be preferred.



a) Gamma kernel b) Gaussian kernel c) Zhang et al.'s kernel

Figure 7: Estimated density and 90% bootstrapped confidence intervals

4. Conclusions and Final Remarks

It seems a promising idea to incorporate a step of statistical estimation by kernels into the analysis of queueing systems. An empirical analysis was presented for arrivals of a general bulk-arrival multi-server finite queue, more specifically a $GI^{[X]}/M/c/N$ queue, when the arrival process is estimated by kernels. From the simulation results presented, it is noticeable that the method may be quite effective, especially if some care is taken in regard to the kernel used. The kernel proposed by Zhang et al. (1999) proved to be the best one for the type of data found for this queue system, that is, for nonnegative random variables. Real data were analyzed and the results were also satisfactory in terms of processing times to compute the measures, which are directly obtained from the densities estimated from the kernels.

Future studies in this area should include the use of the estimated inter-arrival times to evaluate the performance measures given by using Eq. (1)-(6), and the implementation of an optimization steep to find the best configuration of the system, in terms of the number of servers c or the total capacity K to reach a certain performance level given, for instance, in terms of a low blocking probability, high throughput, low work-in-process, low cycle time (average time in the systems), and so on.

Acknowledgments

This research is partially supported by FAPEMIG (grant APQ-00613-12, CEX-PPM-00013-14) and CNPq (304671/2014-2).

References

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

Bareche, A. and Aïssani, D. (2014), Interest of boundary kernel density techniques in evaluating an approximation error of queueing systems characteristics, *International Journal of Mathematics and Mathematical Sciences*, 2014, 1-8.

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.

Gontijo, G. M., Atuncar, G. S., and Cruz, F. R. B. (2010), Performance evaluation of GIX/M/C/N systems through kernel estimation, *Proceedings of the XLII SBPO*, 1-11.

Gross, D., Shortle, J. F., Thompson, J. M., and Harris, C. M. (2009), *Fundamentals of Queueing Theory*, Wiley-Interscience, New York, NY,4 ed.

Gustafsson, J., Hagmann, M., Nielsen, J.P., and Scaillet, O. (2009), Local transformation kernel density estimation of loss distributions, *Journal of Business and Economic Statistics,* 27(2), 161-175.

Kendall, D. G. (1953), Stochastic processes occurring in the theory of queues and their analysis by the method of imbedded Markov chains, *Annals of Mathematical Statistics*, 24, 338-354.

Lima, M. S. and Atuncar, G. S. (2011), A Bayesian method to estimate the optimal bandwidth for multivariate kernel estimator, *Journal of Nonparametric Statistics*, 23(1), 137-148.

Malec, P. & Schienle, M. (2014), Nonparametric kernel density estimation near the boundary, *Computational Statistics & Data Analysis*,. 72, 57-76.

R Core Team. (2013), *R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing*, Vienna, Austria. URL http://www.R-project.org

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Monographs on Statistics and Applied Probability, Chapman and Hall, London.

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

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

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