ESTIMATION IN *GI*[*X*]/*M*/*C*/*N* QUEUES AND THEIR DIMENSIONING

F. R. B. Cruz
Departamento de Estatística, Universidade Federal de Minas Gerais,
31270-901 – Belo Horizonte – MG, Brazil
fcruz@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

ABSTRACT

Queues with general (*GI*) inter-arrival times in batches of random sizes (*X*), Markovian service times (*M*), multi-servers (*c*), and finite buffer spaces (*N*) are an appropriate model in many situations of practical interest. The focus of this paper is on the estimation of parameters of such queues and their dimensioning. For adjusting the arrival process, a kernel-based method is used, along with numerical integrations to approximate the performance measures. Computational simulations are performed to attest the quality of the estimations. Interesting new insights are raised and limitations of the results are discussed.

KEYWORDS. Estimation. Dimensioning. General finite queues. Performance evaluation.

Paper topics. EST – Statistics. MP – Probabilistic Models

1. Introduction

Finite queues are prevalent in many real-life situations, in particular the $GI[X]/M/c/N$ queues, which, in Kendall notation [Kendall 1953], stands for independent general (GI) distributed inter-arrival times of bulk arrivals 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. Studied before [Vijaya Laxmi & Gupta 2000], such a finite queue is of practical interest especially in situations in which there are limitations in the buffer spaces and some control on the servers but no control over how groups and how many users are coming into the system. An interesting application of such queues was presented to model a cellular telephone center in which the inter-arrival times were truncated to the next integer second because of limitations of the data acquisition system, consequently producing ties in the arrival times, which may be seen as group arrivals [Gontijo et al. 2011]. Moreover, the group inter-arrival times were far from any known distribution, as seen in Figure 1, along with a kernel approximation detailed in a previous paper [Cruz et al. 2015].

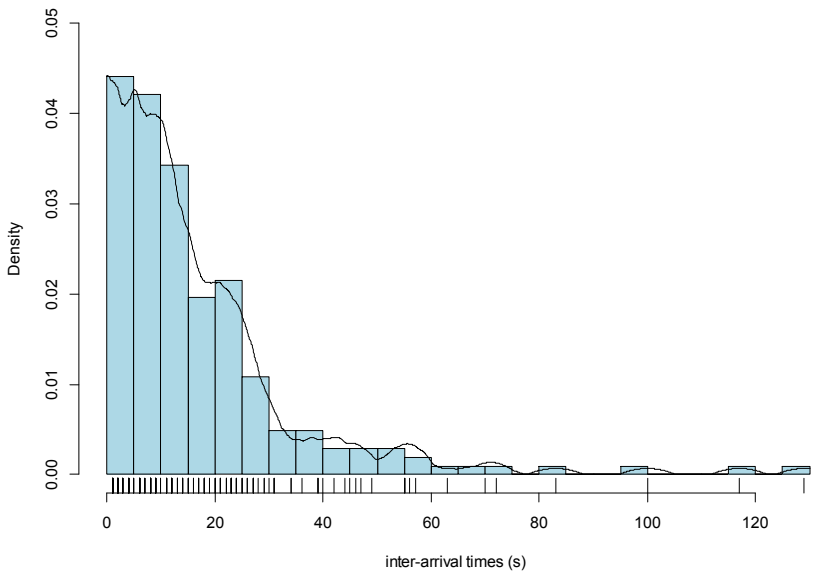


Figure 1: Inter-arrival times in a cellular telephone central from 8am to 9am and kernel estimate

The interest usually lies in the optimization of a queueing system, which in the present case consists in the dimensioning of the number of servers (c), buffer sizes (N), and service rates (μ), to achieve some level of quality evaluated in terms of some performance measure of the queueing system. However, an important and difficult first step before actually computing the performance measures that will guide the optimization process is the estimation of some parameters, such as the arrival process (and its rate λ) and the service rate μ, which is delicate because it involves data collecting and processing.

A straightforward way to address arrival process modeling is through parametric models. However, since data will not fit well into parametric models, in this paper a kernel-based method is applied, which is a non-parametric approach that has been successful in such cases [Cruz et al. 2015]. Indeed, methods based on kernels have received increasing attention from researchers from many areas [Lima & Atuncar 2011; Bareche & Aïssani 2014] mainly because they provide a simple way of finding structure in data sets without imposing a specific parametric model [Wand & Jones 1985; Gustafsson et al., 2009]. To model the service process, a well-known classical approach is used in this paper, the method of moments [Pearson 1934].

Thus, the objective of this paper is to present an estimation method for some performance measures for *GI*[*X*]/*M*/*c*/*N* queues, to investigate their behavior and dimensioning under finite sample settings. The rest of this paper is organized as follows. In the next section, we present a brief review of the literature on queues and kernel estimators. Then, some results from computational experiments are presented and discussed. Final remarks and topics for future research in the area close this paper.

2. Materials and Methods

2.1 Markovian and Bulk Arrival Queues

A first attempt to model a particular practical situation involving finite queues could be performed with one of the simplest queueing models with parallel channels and truncation, that is, an *M*/*M*/*c*/*N* queue, that is, Markovian arrivals and service times, *c* parallel servers, and a total capacity of *N* users, including those in service. From the literature [Gross et al. 2009], it follows that the steady-state, system-size probabilities are given by

$$p_n = \begin{cases} \dfrac{\lambda^n}{n!\,\mu^n}\,p_0, & 1 \le n < c, \\ \dfrac{\lambda^n}{c^{n-c}c!\,\mu^n}\,p_0, & c \le n \le N, \end{cases} \tag{1}$$

in which p_0 is given as follows,

$$p_0 = \left(\sum_{n=0}^{c-1} \frac{\lambda^n}{n!\,\mu^n} + \sum_{n=c}^{N} \frac{\lambda^n}{c^{n-c}c!\,\mu^n} \right)^{-1}. \tag{2}$$

From the probabilities given by Eq. (1) and (2), important performance measures for the queue can be drawn, such as the blocking probability (p_{block}), that is, the probability that a user finds the systems full, one of the most important performance measures, the throughput (θ), the expected queue length (L_q), the expected system size (L), and the expected values for waiting times in the queue (W_q) and in the system (W), readily obtained from Little's formula [Little 1961], as given by:

$$\begin{cases} p_{\text{block}} = p_N, \\ \theta = \lambda\left(1 - p_{\text{block}}\right) \\ L_q = \displaystyle\sum_{n=c+1}^{N} (i-c)p_n, \\ L = \displaystyle\sum_{n=1}^{N} np_n, \\ W_q = \dfrac{L_q}{\theta}, \\ W = \dfrac{L}{\theta}. \end{cases} \tag{3}$$

Although a pure Markovian, finite-queue system could be helpful, the bulk-arrival, general-service, finite queueing system, *GI*[*X*]/*M*/*c*/*N*, might be a more appropriate modeling tool. Such a queue is of practical interest, especially in situations in which we can control the servers but we have no control over how many groups of customers are coming into the system (see an application to modeling a cellular telephone center [Gontijo et al. 2011]).

These *GI*[*X*]/*M*/*c*/*N* systems have been solved [Vijaya Laxmi & Gupta 2000], in which the clients arrive in groups of size *X*, with $P(X=i) = g_i$, for $i \ge 1$, $E(X) = \bar{g}$, and traffic intensity defined as $\rho = (\lambda\bar{g}/c\mu)$, in which λ is the arrival rate (or the inverse of the mean inter-arrival time), and μ is the service rate (or the inverse of the mean service time). The vector of arbitrary time

invariant probabilities **p**, related to the number of users that an outside observer finds in the systems, has its components given by:

$$p_n = \begin{cases} \dfrac{\rho c}{\min\{n,c\}\overline{g}} \displaystyle\sum_{i=0}^{n-1} \pi_i \sum_{j=n-i}^{\infty} g_j, & 0 < n \le N, \\ 1 - \displaystyle\sum_{i=1}^{N} p_i, & n = 0, \end{cases} \tag{4}$$

in which $\pi_i = \lim_{n\to\infty} P(Y_n=i)$, $i = 0, 1, \ldots$, are the components of the vector of the pre-arrival probabilities $\boldsymbol{\pi}$ related to the number of users an arriving client finds at the system for partial rejection (that is, in which an arriving batch of clients is partially lost if it does not fit completely into the buffer space; for total rejection, see details elsewhere [Vijaya Laxmi & Gupta 2000]).

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:

$$\begin{bmatrix} (p_{0,0}-1) & p_{1,0} & \cdots & p_{N,0} \\ p_{0,1} & (p_{1,1}-1) & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ p_{0,N-1} & p_{1,N-1} & \cdots & p_{N,N-1} \\ 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \pi_0 \\ \pi_1 \\ \vdots \\ \pi_{N-1} \\ \pi_N \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}. \tag{5}$$

Transition probabilities p_{jk} are given by:

$$p_{jk} = \begin{cases} \displaystyle\sum_{i=\max\{1,k-j\}}^{N-j} \beta_{j+i-k} g_i + \beta_{N-k} \sum_{i=N-j+1}^{\infty} g_i, & c \le k \le N, \\ \displaystyle\sum_{i=\max\{1,k-j\}}^{N-j} V_{j+i,k} g_i + V_{N,k} \sum_{i=N-j+1}^{\infty} g_i, & 0 < k < c, \\ 1 - \displaystyle\sum_{r=1}^{N} p_{jr}, & k = 0, \end{cases} \tag{6}$$

in which

$$V_{j,k} = \begin{cases} 0, & j < k < c, \\ \displaystyle\int_0^{\infty} \binom{j}{k} e^{-k\mu z} \left(1 - e^{-\mu z}\right)^{j-k} dA(z), & k \le j \le c, \\ \displaystyle\int_0^{\infty}\int_0^{z} \binom{c}{k} e^{-k\mu z} \frac{(c\mu y)^{j-c}}{(j-c)!} c\mu \left(e^{-\mu y} - e^{-\mu z}\right)^{c-k} dy dA(z), & k < c < j, \end{cases} \tag{7}$$

and

$$\beta_r = \int_0^{\infty} \frac{e^{-c\mu z} (c\mu z)^r}{r!} dA(z), \qquad r \ge 0, \tag{8}$$

in which the inter-arrival times are independent, identically distributed random variables with cumulative distribution $A(z)$.

Similarly to $M/M/c/N$ queues, once the probabilities p_n are known, Eq.(4), performance measures follow easily from Eq. (3), the only difference being that:

$$\begin{cases} p_{\text{block}} = \displaystyle\sum_{i=0}^{N} \pi_i \sum_{j=N-i}^{\infty} (1/\overline{g}) \sum_{k=j+1}^{\infty} g_k, \\ \theta = \lambda \overline{g} \left(1 - p_{\text{block}}\right). \end{cases} \tag{9}$$

2.2 Kernel Estimators

Let us suppose that we have a sample X_1, X_2, ..., X_n, but we do not know from which cumulative distribution $A(z)$, mentioned in the integrals in Eq.(7) and (8), such data come. A possible way to model the inter-arrival times is through a kernel estimator, which is an analytical tool to reveal the underlying structure of a sample. The classical model is the Parzen-Rosemblatt estimator

$$\hat{A}_n(x,h) = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right), \tag{6}$$

in which $K(y)$ is a symmetrical density function and h is a smoothing parameter also called the window [Wand & Jones, 1995].

We consider a more specialized estimator [Zhang et al. 1999], which is an alternative to overcome the problem of estimating a strictly positive function domain:

$$\hat{A}_n(x,h) = \frac{1}{nh}\sum_{j=1}^{n}\left\{K\left(\frac{x - X_j}{h}\right) + K\left(\frac{x + g(X_j)}{h}\right)\right\}, \tag{9}$$

in which $K(y)$ is Epanechnikov's kernel:

$$K_E(x) = \frac{3}{4}(1 - x^2)I_{[-1,1]}. \tag{10}$$

The satisfactory performance of kernel methods depends on the proper choice of the window h. Following previous results [Gontijo et al. 2011], we consider the asymptotic mean integrated square error (AMISE) and to estimate h, a constant c is used [Chiu 1991] to determine a cutoff value of λ in $|\varphi(\lambda)|^2 < c/n$, where $\varphi(\lambda)$ is the estimated characteristic function, and n is the sample size. The constant $A > 1/3$ is also needed to generate a set of pseudo data, $g(X_i)$. Both must be chosen accordingly [Zhang et al. 1999]. From previous studies, A=2/3 and c=3 was used.

3. Numerical Results and Discussion

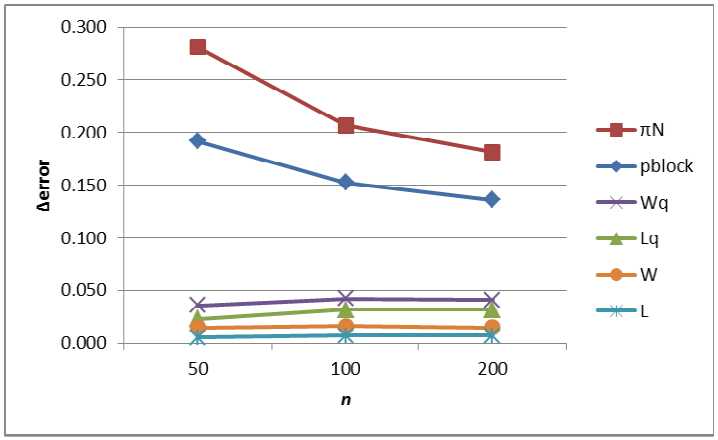
3.1 Simulation Study

All algorithms were encoded for an R platform [R Core Team 2015] and are available upon request for teaching and research purposes. To study the finite sample behavior of the estimates in more detail, Monte Carlo simulations were performed for the traffic intensities ρ = {0.1, 0.2, 0.4} and samples sizes n = {50, 100, 200}. The number of Monte Carlo replications for each combination was 100. For each scenario, the average estimates (avrg) and standard error of the means (se) are reported along with the relative errors, $\Delta\varepsilon = \left(\theta - \hat{\theta}\right)/\theta$, in comparison with the true parameter (θ), computed here numerically from the cumulative distribution $A(z)$, which in this paper is the exponential distribution, in contrast to the estimates ($\hat{\theta}$) calculated from finite size samples. The processing times are low usually taking few minutes to run, in a CORETM i7 laptop running Windows® 7. The results are given in Table 1 and are summarized in Figure 2 for the blocking probabilities for an arbitrary customer and for the first customer (p_{block} and π_N, respectively), the expected queue length (L_q), the expected waiting time in the queue (W_q), the expected system size (L), and the expected waiting time in the system (W).

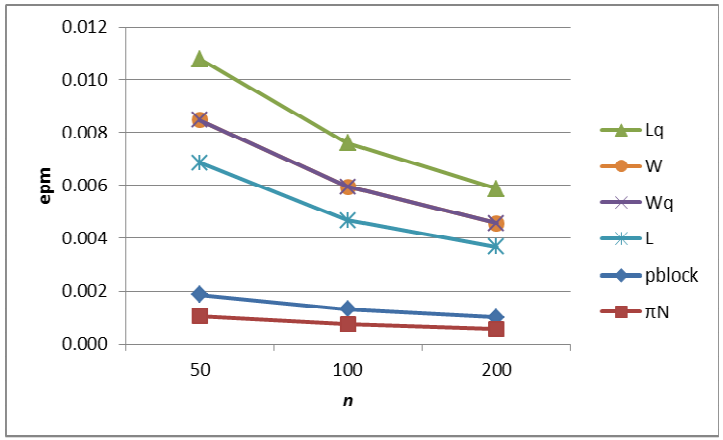
Figure 2 shows the averages of the estimates presented in Table 1, from which it is easy to observe some interesting patterns. The standard error of the means (Figure 2-a and 2-c) is less than the relative error of the estimates (Figure 2-b and 2-d), which indicates that the Monte Carlo procedure provides reliable conclusions. The relative errors decrease when the sample sizes increase, as expected. If estimates with less than 15% relative error are required, the sample size should be approximately 200 (Figure 2-b). The blocking probabilities (p_{block} and π_N) are difficult to estimate when the system has low traffic intensity, that is, $\rho \approx 0$ (see Figure 2-d).

Table 1: Average estimates (avrg) and the respective standard error of the mean (se) and relative errors (Δε)

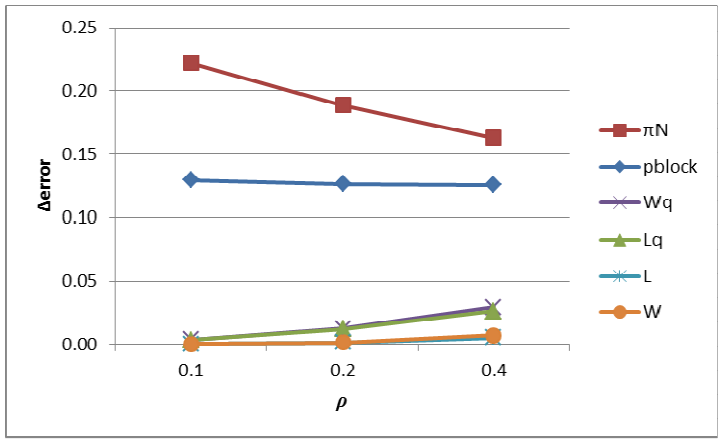
ρ	θ	n=50			n=100			n=200		
		avrg	se	Δε	avrg	se	Δε	avrg	se	Δε
0.1	p_{block}	0.000204	0.000023	0.000033	0.000198	0.000016	0.000027	0.000194	0.000012	0.000023
	π_N	0.000039	0.000006	0.000010	0.000037	0.000004	0.000008	0.000035	0.000003	0.000006
	L_q	0.096541	0.002064	-0.000458	0.097750	0.001536	0.000750	0.098026	0.001268	0.001026
	W_q	0.080471	0.001723	-0.000376	0.081476	0.001281	0.000629	0.081705	0.001058	0.000858
	L	1.296296	0.002038	-0.000498	1.297512	0.001518	0.000718	1.297793	0.001254	0.000999
	W	1.080471	0.001723	-0.000376	1.081476	0.001281	0.000629	1.081705	0.001058	0.000858
0.2	p_{block}	0.001611	0.000162	0.000278	0.001508	0.000093	0.000175	0.001502	0.000076	0.000169
	π_N	0.000444	0.000057	0.000109	0.000398	0.000031	0.000064	0.000395	0.000025	0.000061
	L_q	0.177734	0.004480	0.001512	0.178985	0.003009	0.002763	0.179579	0.002465	0.003357
	W_q	0.148408	0.003773	0.001360	0.149403	0.002527	0.002355	0.149890	0.002070	0.002842
	L	1.375801	0.004297	0.001178	1.377175	0.002900	0.002553	1.377777	0.002376	0.003155
	W	1.148408	0.003773	0.001360	1.149402	0.002527	0.002355	1.149890	0.002070	0.002842
0.4	p_{block}	0.014444	0.001065	0.002354	0.013791	0.000660	0.001701	0.013523	0.000503	0.001433
	π_N	0.005422	0.000484	0.001126	0.005063	0.000291	0.000767	0.004943	0.000223	0.000647
	L_q	0.397541	0.009503	0.011008	0.399087	0.006447	0.012555	0.398512	0.005006	0.011980
	W_q	0.337018	0.008489	0.010966	0.337594	0.005703	0.011542	0.336865	0.004417	0.010813
	L	1.580207	0.008283	0.008183	1.582538	0.005675	0.010514	1.582285	0.004416	0.010261
	W	1.337018	0.008489	0.010966	1.337594	0.005703	0.011542	1.336865	0.004417	0.010813



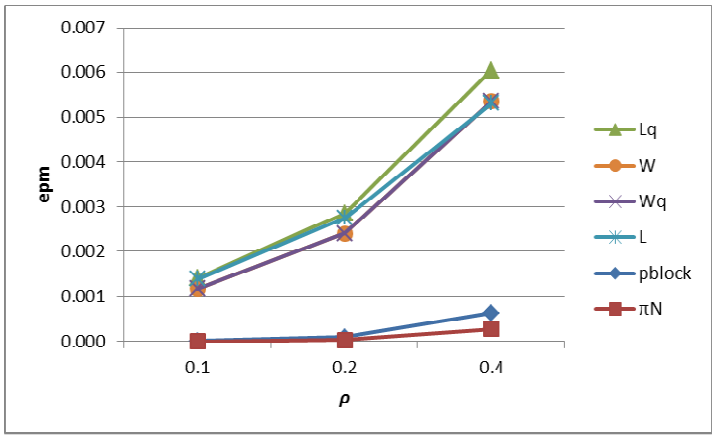
a) standard error of the means versus sample sizes



b) relative errors versus sample sizes



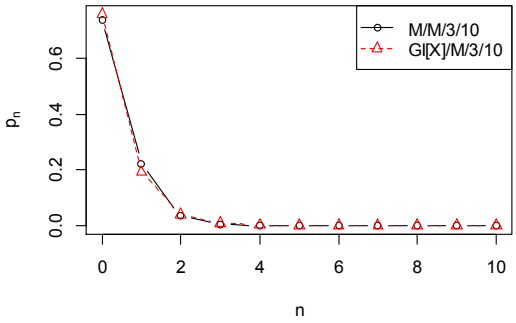
c) standard error of the means versus traffic intensity



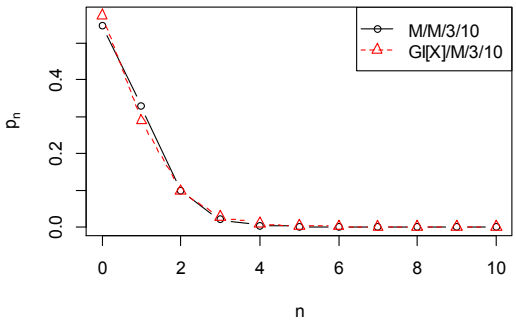
d) relative errors versus traffic intensities

Figure 2: Finite-sample behavior of the performance measure estimates

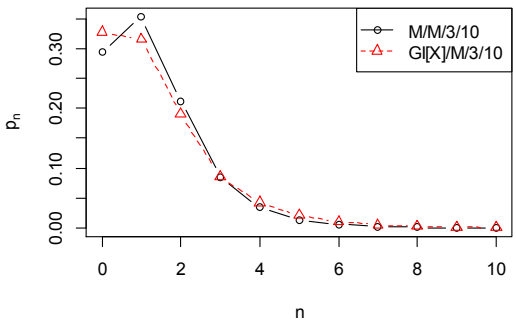
From Figure 3, it is possible to observe that if one considers an $M/M/3/10$ queue, with $\lambda = 1/17.64$, as a modeling tool for the data presented in Figure 1 [Cruz et al. 2015], the blocking probabilities (and consequently, the other performance measures) may be satisfactory if the system is under light traffic ($\rho \leq 0.2$), but a significant difference is noticed otherwise. A more precise estimate should be expected from a $GI[X]/M/3/10$ queue, with arrivals modeled by kernels (same λ as earlier) and groups sizes with $g_1 = 0.808$, $g_2 = 0.158$, $g_3 = 0.031$, $g_4 = 0.003$, and $g_i = 0$, otherwise [Cruz et al. 2015]. Indeed, when the system is under traffic $\rho = 0.4$, the Markovian model clearly overestimates the steady-state probabilities for 1 and 2 customers (although both models show similar probabilities for the remaining values of n, as seen in Figure 3-c).



a) for $\rho = 0.1$



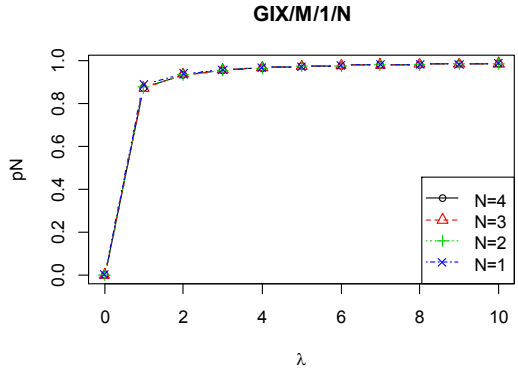
b) for $\rho = 0.2$



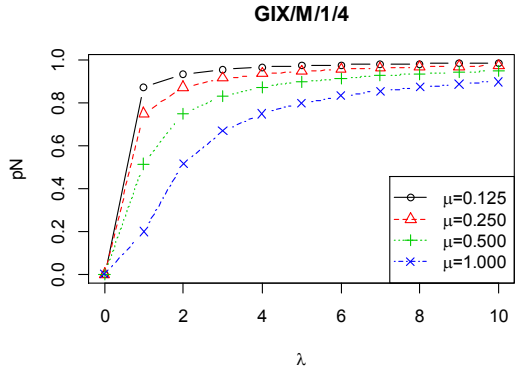
c) for $\rho = 0.4$

Figure 3: Steady-state, system-size probabilities

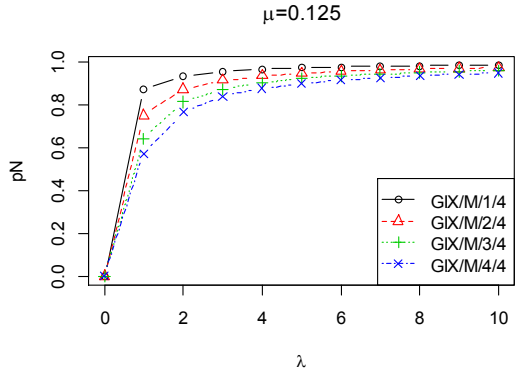
How a *GI*[*X*]/*M*/*c*/*N* queueing model would behave under an optimization framework was also investigated. Figure 4 shows the effects on the blocking probability as a function of the arrival rate caused by changes in the maximum capacity (*N*), service rate (μ), and number of servers (*c*). The service rate plays a key role and is the most significant factor in the reduction of the blocking probability. However, because the service rate may not be a variable under control, a better way to improve he system in terms of a low blocking probability is to increase the number of servers. The number of buffer spaces also improves the system performance but not as effectively as servers.



a) effect of the total system size (*N*)



b) effect of the service rate (μ)



c) effect of the number of servers (*c*)

Figure 4: Blocking probability pN behavior in GI[X]/M/c/N queues as a function of the arrival rate

4. Conclusions and Final Remarks

It seems to be a promising idea to incorporate statistical estimation into the context of analysis and optimization of queueing systems. In this paper, an empirical analysis is performed of a *GI*[*X*]/*M*/*c*/*N* queue. The inter-arrival distribution may be efficiently and accurately estimated by a kernel density-based method as it has been shown. The results can be used to evaluate the performance measures of the underlying queueing system. From the simulation results presented, it is possible to conclude that the method is effective, especially if some care is taken in regard to the kernel used and if there are enough data available (at least approximately 200). A specialized kernel which overcomes the problem of estimating strictly positive probability density functions [Zhang et al. 1999] proved to be a good alternative for the type of data found in such a queue system, that is, the inter-arrival times are nonnegative random variables. Competing queueing models were analyzed showing better results in favor of *GI*[*X*]/*M*/*c*/*N* queues, under block arrivals and high traffic intensities ($\rho \approx 0.4$), in terms of low processing times and accuracy. However, pure Markovian queues, *M*/*M*/*c*/*N*, seem to be applicable under low traffic intensities.

Future studies in this area should include the development of some algorithmic approach to optimize *GI*[*X*]/*M*/*c*/*N* queueing systems, preferably focusing on their arrival rates or number of servers, as variables under control.

Acknowledgments

This research is partially supported by CNPq (grant 303388/2010-2), FAPEMIG (grant CEX-PPM-00013-14), CAPES, and Universidade Federal de Ouro Preto (grant Edital PROPP 09/2016).

References

Bareche, A. & 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(Article ID 871357):1–8.

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

Cruz, F. R. B.; Santos, M. A. C.; Oliveira, F. L. P. & Brito, N. L. C. (2015). Kernel density estimation of arrivals in *GI*[*X*]/*M*/*c*/*N* queues. *In Anais do XLVII SBPO*, 1-9, Porto de Galinhas. SOBRAPO.

Gontijo, G. M.; Atuncar, G. S.; Cruz, F. R. B. e Kerbache, L. (2011). Performance Evaluation and Dimensioning of *GIX*/*M*/*c*/*N* Systems through Kernel Estimation. *Mathematical Problems in Engineering*, 2011(Article ID 348262):1–20.

Gross, D.; Shortle, J. F.; Thompson, J. M. e Harris, C. M. (2009). Fundamentals of queueing theory. Wiley-Interscience, New York, 4th ed.

Gustafsson, J.; Hagmann, M.; Nielsen, J.P. e Scaillet, O. (2009). Local transformation kernel density estimation of loss distributions. *Journal of Business & 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. & Atuncar, G. S. (2011). A Bayesian method to estimate the optimal bandwidth for multivariate kernel estimator. *Journal of Nonparametric Statistics Statistics*, 23(1):137–148.

Little, J. D. C. (1961). A proof for the queuing formula: $L=\lambda W$. *Operations Research*, 9(3):383–387.

Pearson, K. (1934). Tables of the incomplete beta function. London: University College, Biometriks Office.

R Core Team. (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.r-project.org/

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

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.