Research Article

Simple and Yet Efficient Estimators for Markovian Multiserver Queues

Emilio Suyama, Roberto C. Quinino ⓘD, and Frederico R. B. Cruz ⓘD

Departamento de Estatística, Universidade Federal de Minas Gerais, 31270-901 Belo Horizonte, MG, Brazil

Correspondence should be addressed to Frederico R. B. Cruz; fcruz@est.ufmg.br

Estimators for the parameters of the Markovian multiserver queues are presented, from samples that are the number of clients in the system at arbitrary points and their sojourn times. As estimation in queues is a recognizably difficult inferential problem, this study focuses on the estimators for the arrival rate, the service rate, and the ratio of these two rates, which is known as the traffic intensity. Simulations are performed to verify the quality of the estimations for sample sizes up to 400. This research also relates notable new insights, for example, that the maximum likelihood estimator for the traffic intensity is equivalent to its moment estimator. Some limitations of the results are presented along with a detailed numerical example and topics for future developments in this research area.

1. Introduction

One of the major research interests in queuing theory is to study the performance characteristics of the queues as a function of their parameters, namely, the arrival rate λ, the service rate μ, or else the traffic intensity ρ. The mathematical problem dealt with here is how these parameters can be estimated using some statistical inferential method. This article focuses on the $M/M/s$ queuing system, which in Kendal notation represents the Markovian number of arrivals, exponential (Markovian) service times, s identical servers working in parallel, and an infinite maximum capacity of users simultaneously allowed in the system, which is one of the classical basic queuing models [1]. From the above parameters and reminding that for $M/M/s$ queues the traffic intensity is defined as the ratio $\rho = \lambda/s\mu$, important performance measures can be derived, such as the empty system probability P_0, the expected number of customers in the system L, the expected number of customers in the queue L_q, the expected time in the system W, and the expected time in the queue W_q.

One of the classical basic queuing models [1], $M/M/s$ queues have many applications and consequently great necessity of inference. In fact, many real-life systems are approximately described by queuing models, including computer and telecommunication networks [2–4], manufacturing and service systems [5–7], and heathcare systems [8–10].

Classical inference results for queues (based on the maximum likelihood) are unknown before the results reported in parallel by Clark [11], with the development of the maximum likelihood estimates for λ and μ for Markovian single-served queues, that is, $M/M/1$ queues, in Kendal notation, and by Benes [12], for Markovian infinite queues, that is, $M/M/\infty$ queues, in Kendal notation. More recently, Schruben and Kulkarni [13] proved that popular performance measures unfortunately do not have expected values and standard errors, Basawa and Prabhu [14] discovered some asymptotic properties of the estimators for the performance evaluation of $M/M/1$ queues, Zheng and Seila [15] showed that under a certain upper limit for the traffic intensity ($\rho < 1$), the estimators do exist and have the necessary properties, and Almeida et al. [16] proposed bias correction methods for classical estimators for $M/M/1$ queues.

However, probably because of the inherent computational difficulties, Bayesian inference has not gained enough interest until the pioneering results reported by Armero and Bayarri [17–19], which were followed by the improvements from Armero and Conesa [20–23], McGrath et al. [24], McGrath and Singpurwalla [25], Choudhury and Borthakur[26], and, more recently, Almeida and Cruz [27], Cruz et al. [28],

Quinino and Cruz [29], Cruz et al.[30], and Choudhury and Basak, [31], among many others.

Almost all the literature that addresses estimation issues describes methods that require the continuous observation of the process over a fixed interval of time. Among the exceptions is the article by Basawa et al. [32], which considers the estimation for $M/M/1$ queues from the sojourn time data, and the articles by Choudhury and Borthakur [26] and Chowdhury and Mukherjee [33], in which the number of customers present in the $M/M/1$ queue at successive departure epochs is used to estimate the traffic intensity. In Ross et al. [34], a method that requires the numbers in the queue at successive time points, which are not necessarily equally spaced, has been developed. The specific results presented by Ross et al. [34] apply to $M/M/s$ queues with large number of server s (in particular when $s > 40$, a limit that generates the smallest estimation errors).

Thus, this study aims to derive a maximum likelihood estimator for ρ in $M/M/s$ queues, from samples of the number of clients in the system at arbitrary points, which has been shown to be conveniently equivalent to the moment estimator. Additionally, with the use the additional information from samples of the sojourn times, a maximum likelihood estimator for the arrival and the service rates is derived. To the best of the authors' knowledge, these have not been done before.

Summarizing, the main contributions of this paper are as follows. The traffic intensity moment estimator is proved to be equal to the maximum likelihood estimator. Also, it is shown that if additional information of the sojourn times is used then estimators for the arrival and service rates can be deduced. Finally, it is demonstrated by extensive computational experiments that the method of moments is competitive to estimate arrival and service rates in $M/M/s$ queues.

The rest of this paper is organized as follows. In the next section, the probabilistic model of the $M/M/s$ queue is detailed, the respective maximum likelihood estimator for ρ is deduced, demonstrating its equivalence to the moment estimator, and the estimators for λ and μ are derived. In the following section, results from Monte Carlo simulations are presented to attest the efficiency of the estimators, along with an illustrative numerical example. In the last section, the paper is concluded, with discussions, final remarks, and the presentation of some topics for future research in this area.

2. Materials and Methods

2.1. Probabilistic Model.
In $M/M/s$ queues, the arrivals of the customers are assumed to follow a Poisson process with rate λ, which means that the times between successive arrivals are independent exponentially distributed random variables with mean $1/\lambda$. Upon arrival, a customer goes immediately into service, if any of the servers are free, or else joins the queue if all s servers are busy. After service, a customer leaves the system and if there are any customers waiting, the next one in line enters the service. The successive service times

are assumed to be independent exponential random variables with mean $1/\mu$.

Denoting by $X(t)$ the number in the system at time t, then $\{X(t), \ t \geq 0\}$ is a birth-and-death process with $\mu_n = n\mu$, if $1 \leq n \leq s$; $\mu_n = s\mu$, if $n > s$; and $\lambda_n = \lambda$, if $n \geq 0$. Defining $\rho = \lambda/s\mu$ as the traffic intensity (also known as the server utilization), it is required that $\rho < 1$, for the queue to be stable, which is a common assumption in queuing theory [1]. The traffic intensity ρ may also be seen as the average proportion of time that each of the servers is occupied (assuming that the jobs finding more than one vacant server choose their servers randomly). If the traffic intensity is such that $\rho < 1$, then the system is under equilibrium and it has a stationary distribution. From the theory developed for birth-and-death processes, the stationary distribution of the number of customers N in the system at the departure epochs is given by (for instance, see Gross et al. [1])

$$P_n \equiv P(N = n) = \begin{cases} \dfrac{(s\rho)^n}{n!} P_0, & 0 < n \leq s, \\ \dfrac{s^s \rho^n}{s!} P_0, & n > s, \end{cases} \quad (1)$$

in which $P_0 \equiv P(N = 0)$ is given from the usual boundary condition that the probabilities must sum to 1; that is,

$$P_0 = \left(\sum_{j=0}^{s-1} \frac{(s\rho)^j}{j!} + \frac{(s\rho)^s}{s!} \frac{1}{(1-\rho)} \right)^{-1}. \quad (2)$$

2.2. Maximum Likelihood Estimator for the Traffic Intensity.
In this section, the traffic intensity ρ is the focus; that is, the ratio $\rho = \lambda/(s\mu)$. To generate data, it is necessary to observe the system at arbitrary epochs. Such a scheme ensures that the data-generating process is consistent with the distribution given by (1). Assume that the number of customers in the queue is x_i. If $\mathbf{x} = \{x_1, x_2, ..., x_\ell\}$ constitutes our sample of size ℓ, then the corresponding likelihood function is

$$L(\mathbf{x} \mid \rho) = \prod_{i=1}^{\ell} \left[\frac{(s\rho)^{x_i}}{x_i!} P_0 I_{\{0 \leq x_i \leq s\}} + \frac{s^s \rho^{x_i}}{s!} P_0 I_{\{x_i > s\}} \right], \quad (3)$$

in which $I_{\{\bullet\}}$ is the indicator function. It is pertinent to note here that this data-generating scheme must ensure the independence of the sample observations, provided that they are sufficiently spaced in time.

The maximum likelihood estimator for the traffic intensity ρ is the value that maximizes the likelihood given by (3) for $0 < \rho < 1$. Observe that (3) may be written as $L(\mathbf{x} \mid \rho) = \alpha \prod_{i=1}^{\ell} \rho^{x_i} P_0$, in which α is a constant, and that the value that maximizes (3) also maximizes

$$U(\mathbf{x} \mid \rho) = \prod_{i=1}^{\ell} \rho^{x_i} P_0 = P_0^n \rho^{\sum_{j=1}^{\ell} x_j}. \quad (4)$$
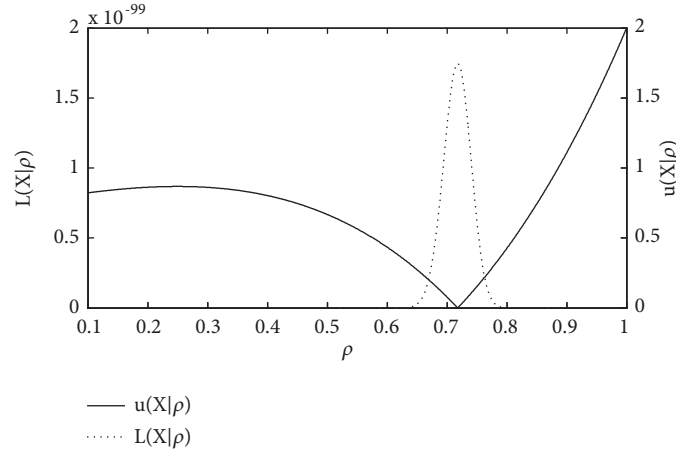
FIGURE 1: Behavior of the likelihood function, $L(\mathbf{x} \mid \rho)$, and the derivative of its natural logarithm with respect to ρ, $u(\mathbf{x} \mid \rho)$, for $s = 4$, and sample value $\overline{x} = 4.03$.

The value that maximizes (4) is the maximum likelihood estimator, $\widehat{\rho}_{\mathrm{MLE}}$. Applying the natural logarithm to (4), we obtain

$$\ln\left[U\left(\mathbf{x} \mid \rho\right)\right] = n\ln\left[\sum_{j=0}^{s-1} \frac{(s\rho)^j}{j!} + \frac{s^s \rho^s}{s!\left(1-\rho\right)}\right] + \sum_{i=0}^{\ell} x_i \ln\rho.$$

(5)

Making the derivative of (5) equal to zero will produce

$$u\left(\mathbf{x} \mid \rho\right)$$

$$= -nP_0\left[\sum_{j=0}^{s-1} \frac{s^j \rho^{j-1}}{(j-1)!} + \frac{s^s}{s!}\frac{s\left(1-\rho\right)\rho^{s-1} + \rho^s}{\left(1-\rho\right)^2}\right]$$

$$+ \frac{\sum_{i=0}^{\ell} x_i}{\rho} = 0.$$

(6)

The value that solves (6) is the maximum likelihood estimator, $\widehat{\rho}_{\mathrm{MLE}}$. However, because (6) is not algebraically simple, it is necessary to implement a numerical search to find $\widehat{\rho}_{\mathrm{MLE}}$, as we will show shortly. For our convenience, the software selected was MATLAB® [35], but many others could be used as well.

To better understand the behavior of (6), it is interesting to closely analyze some special cases of s. Table 1 illustrates some of these special cases for $s = 1, 2, 3,$ and 4, in which $\overline{x} = \sum_{i=1}^{\ell} x_i/\ell$. Observe that each polynomial is of a degree equal to the number of servers, s. Figure 1 illustrates the likelihood function, $L(\mathbf{x} \mid \rho)$, and the derivative of its natural logarithm with respect to ρ, $u(\mathbf{x} \mid \rho)$, for $s = 4$ and sample value $\overline{x} = 4.03$. The maximum likelihood estimated by the software is $\widehat{\rho}_{\mathrm{MLE}} = 0.7180$. It is noticeable that the functions were well behaved and seem to be unimodal. Other examples for different values of s were tested, and the results (not shown) were similar.

2.3. Traffic Intensity Moment Estimator versus the Maximum Likelihood Estimator.

The moment estimator [36] for ρ can be obtained by setting the analytical expression of the expected number of clients in the queuing system, L, equal to the average of the sample of the number of clients at the arbitrary epochs, $\overline{x}$. In accordance with Gross et al. [1], the expected number of clients in an $M/M/s$ queuing system is given by

$$L = \frac{\rho\left(\rho s\right)^s}{s!\left(1-\rho\right)^2}P_0 + s\rho.$$

(7)

To obtain the moment estimator, $\widehat{\rho}_{\mathrm{ME}}$, we set (7) equal to the average of the sample, $\overline{x}$. In other words, the moment estimator is the value $\widehat{\rho}_{\mathrm{ME}}$ that satisfies

$$L = \frac{\rho\left(\rho s\right)^s}{s!\left(1-\rho\right)^2}P_0 + s\rho = \overline{x},$$

(8)

which should be solved preferably by numerical methods since finding an analytical solution does not seem to be trivial.

After some computational experiments (not shown), the authors observed that the results from the maximum likelihood and moment estimators were actually equal, which led to the conjecture that, in $M/M/s$ queuing systems, the maximum likelihood estimator and the moment estimator for the traffic intensity ρ are equivalent. Indeed, it is easy to verify that both estimators are equal by showing that an algebraic manipulation of (6) will produce (7):

$$P_0\left[\rho s\left(\sum_{j=0}^{s-1} \frac{s^{j-1}\rho^{j-1}}{(j-1)!} + \frac{s^{s-1}\rho^{s-1}}{(s-1)!\left(1-\rho\right)}\right)\right.$$

$$\left.+ \frac{\rho\left(s\rho\right)^s}{s!\left(1-\rho\right)^2}\right] = \overline{x} \Longrightarrow$$

TABLE 1: Special cases for (6).

s	Eq. (6)
1	$\rho(\overline{x}+1)-\overline{x}=0$
2	$\rho^2\overline{x}+2\rho-\overline{x}=0$
3	$\rho^3(3\overline{x}-3)+\rho^2(6+\overline{x})+\rho(6+2\overline{x})-\overline{x}=0$
4	$\rho^4\left(\dfrac{\overline{x}}{2}-1\right)+\rho^3\left(\dfrac{\overline{x}}{4}+\dfrac{3}{4}\right)+\rho^2\left(-\dfrac{3\overline{x}}{16}+\dfrac{3}{2}\right)+\rho\left(\dfrac{3}{4}-\dfrac{3\overline{x}}{8}\right)-\dfrac{3\overline{x}}{16}=0$

$$P_0\left[\rho s\left(\sum_{y=0}^{s-2}\frac{s^y\rho^y}{y!}+\frac{s^{s-1}\rho^{s-1}}{(s-1)!(1-\rho)}\right)+\frac{\rho(s\rho)^s}{s!(1-\rho)^2}\right]$$

$$=\overline{x}\Longrightarrow$$

$$P_0\left[\rho s\left(\sum_{y=0}^{s-2}\frac{s^y\rho^y}{y!}+\frac{s^{s-1}\rho^{s-1}}{(s-1)!}\left(1+\frac{\rho}{1-\rho}\right)\right)\right.$$

$$\left.+\frac{\rho(s\rho)^s}{s!(1-\rho)^2}\right]=\overline{x}\Longrightarrow$$

$$P_0\left[\rho s\left(\sum_{y=0}^{s-1}\frac{s^y\rho^y}{y!}+\frac{s^{s-1}\rho^{s-1}}{(s-1)!}\frac{\rho}{1-\rho}\right)+\frac{\rho(s\rho)^s}{s!(1-\rho)^2}\right]$$

$$=\overline{x}\Longrightarrow$$

$$P_0\left[\rho s\left(\sum_{y=0}^{s-1}\frac{(s\rho)^y}{y!}+\frac{(s\rho)^s}{s!(1-\rho)}\right)+\frac{\rho(s\rho)^s}{s!(1-\rho)^2}\right]=\overline{x}\Longrightarrow$$

$$L=s\rho+\frac{\rho(s\rho)^s}{s!(1-\rho)^2}P_0=\overline{x}.$$

$$(9)$$

Thus, the moment estimator is in fact equal to the maximum likelihood estimator, which is very convenient because, in general, obtaining the moment estimator is simpler than obtaining the likelihood estimator and this facilitates the explanation of the estimation process. In addition, in this particular case, the moment estimator for the traffic intensity inherits the good asymptotic properties shared by all maximum likelihood estimators.

2.4. Arrival and Service Rate Estimators.

If the sample $\mathbf{x}=\{x_1,x_2,\ldots,x_\ell\}$ related to the number of clients in the system at arbitrary epochs is collected and also the sojourn times of the clients, it is possible to estimate the arrival rate λ and the service rate μ as well. Define the sojourn time of a client in the system, which includes both the waiting time in line and the service time. In accordance with Gross et al. [1], the expected sojourn time W may be expressed as

$$W=\frac{(\rho s)^s}{s!s\mu(1-\rho)^2}P_0+\frac{1}{\mu}.\qquad(10)$$

With a sample of the sojourn times in the system, it is possible to estimate W as $\overline{x}_W$. Also with the estimate $\widehat{\rho}$, which was obtained in the previous section, it is possible to estimate μ as

$$\widehat{\mu}=\frac{(\widehat{\rho}s)^s\widehat{P}_0+s!s(1-\widehat{\rho})^2}{\overline{x}_Ws!s(1-\widehat{\rho})^2}.\qquad(11)$$

Notice that in (11) $\widehat{P}_0$ may be calculated from (2) using the estimated traffic intensity, $\widehat{\rho}$. Then, from the traffic intensity, $\rho=\lambda/(s\mu)$, it is possible to estimate λ as

$$\widehat{\lambda}=s\widehat{\mu}\widehat{\rho}.\qquad(12)$$

3. Performance Analysis of the Estimation Methods

3.1. Simulations. In this section, the performances of the estimation methods described earlier are evaluated. The algorithms were implemented in MATLAB [35], and the codes are available directly from the authors upon request for research and teaching purposes. A Monte Carlo simulation study was conducted as follows to evaluate the methodology described earlier in this paper.

Random samples of the number of customers in the arbitrary epochs were obtained from simulations of an $M/M/s$ queuing system, with sample sizes $\ell=\{50,200,400\}$, traffic intensities $\rho=\{0.1,0.5,0.9\}$, and number of servers $s=\{3,9,27\}$. Without loss of generality, the arrival rate was fixed to $\lambda=10$ for all experiments, where the service rate is given by $\mu=\lambda/(s\rho)$. With these values, it is possible to evaluate a broad range of different situations that are likely to occur in practice.

Table 2 shows the average values (means) and the standard deviations (SDs) of estimates for ρ, μ, and λ, from (8), (11), and (12), respectively, obtained from a Monte Carlo simulation procedure, repeated 1,000 times. From Table 2, it is observed that, with the increase of the sample size, the difference between the estimates and the true values reduces and the standard deviation decreases.

3.2. Numerical Example. Now an application based on data collected in a large supermarket network in a region of interest is discussed for a better understanding of the presented methodology. The cashier's desks in this network are separated into normal desks and fast desks. The focus of our example is the latter, which serve exclusively consumers with up to 15 items. The primary interest of those responsible

TABLE 2: Averages (means) and standard deviations (SDs) of estimates for ρ, λ, and μ.

| | | | | Sampling distribution | | | | | |
| ℓ | ρ | s | μ | ρ estimated | | μ estimated | | λ estimated | |
				Mean	SD	Mean	SD	Mean	SD
		3	33.3333	0.1006	0.0253	36.0326	4.7752	10.4347	1.4326
	0.1	9	11.1111	0.0998	0.0150	11.9889	1.6397	10.3399	1.4758
		27	3.7037	0.0996	0.0090	4.0059	0.5537	10.3394	1.4949
		3	6.6667	0.4979	0.0452	7.1426	1.0592	10.2850	1.5888
50	0.5	9	2.2222	0.4992	0.0320	2.4031	0.3244	10.3540	1.4599
		27	0.7407	0.4990	0.0186	0.8015	0.1112	10.3596	1.5006
		3	3.7037	0.8744	0.0138	3.3016	0.5743	9.1311	1.5507
	0.9	9	1.2346	0.8643	0.0137	1.0493	0.1328	8.7706	1.0757
		27	0.4115	0.8661	0.0126	0.4256	0.0630	9.7771	1.5296
		3	33.3333	0.1002	0.0130	35.9311	2.3430	10.3781	0.7029
	0.1	9	11.1111	0.0998	0.0073	11.9786	0.8625	10.3374	0.7762
		27	3.7037	0.1001	0.0042	3.9896	0.2456	10.3651	0.6631
		3	6.6667	0.4990	0.0228	7.1038	0.4836	10.2856	0.7254
200	0.5	9	2.2222	0.4992	0.0167	2.3884	0.1575	10.3282	0.7088
		27	0.7407	0.4994	0.0093	0.7973	0.0516	10.3430	0.6965
		3	3.7037	0.8755	0.0067	3.2548	0.2719	9.0676	0.7341
	0.9	9	1.2346	0.8649	0.0068	1.0335	0.0616	8.6964	0.4987
		27	0.4115	0.8668	0.0064	0.4234	0.0310	9.7694	0.7543
		3	33.3333	0.1001	0.0090	35.9298	1.6050	10.3695	0.4815
	0.1	9	11.1111	0.1001	0.0051	11.9613	0.5730	10.3618	0.5157
		27	3.7037	0.1002	0.0029	3.9871	0.1878	10.3697	0.5071
		3	6.6667	0.4986	0.0158	7.0955	0.3618	10.2760	0.5427
400	0.5	9	2.2222	0.5001	0.0120	2.3877	0.1108	10.3486	0.4985
		27	0.7407	0.4991	0.0064	0.7985	0.0377	10.3397	0.5085
		3	3.7037	0.8758	0.0047	3.2426	0.1902	9.0476	0.5136
	0.9	9	1.2346	0.8652	0.0045	1.0294	0.0423	8.6808	0.3424
		27	0.4115	0.8671	0.0042	0.4229	0.0209	9.7613	0.5083

(Header also notes $\lambda=10$ in the top-left of the table.)

for the supermarket is to evaluate the time between 15:00 and 22:00 on Saturdays, mainly because there is a large volume of buyers at that time, and the volume has increased because the supermarket chain has an advertising agency conducting regular campaigns. The fear is that users will give up shopping or not return to this supermarket because of dissatisfaction with the sojourn time. The past experiences of those responsible for supermarkets indicate that the traffic intensity ρ should not exceed 85% to minimize the risk of compromising current and future sales. Based on several previous studies, it is reasonable to assume that consumers arrive during a given period of time approximately according to a Poisson process and that the service times are exponentially distributed. The objective is to evaluate the traffic intensity, arrival rates, and the service rates for an $M/M/6$ queuing system, when the current number of servers (that is, 6) is insufficient.

In this scenario, 200 random observations of the number of customers in the system were collected, 28 at each of 8 consecutive Saturdays, at times sufficiently spaced and previously defined by the person responsible for data collection.

The observed values (O) and frequencies (F) are presented in Table 3. For example, from the 200 observations, at 2 times, only 1 customer was found in the system and at 6 times, 2 customers were found.

Along with the observation of the number of clients in the system, the sojourn times for 200 randomly selected customers were also collected (which was performed simultaneously with the observation of the number of customers in the system). Figure 2 depicts the histogram for the observed results.

Using the collected data, the estimates from (8), (11), and (12) are $\hat{\rho} = 0.9013$, $\hat{\mu} = 2.0046$ clients/min, and $\hat{\lambda} = 10.8247$ clients/min, respectively. The data suggests that $s = 6$ servers are insufficient to handle the clients and to keep the traffic intensity below 85%. Using the previous estimates for μ and λ, we can verify that increasing the number of servers to $s = 7$ would reduce the traffic intensity to $\hat{\rho} = \hat{\lambda}/(s\hat{\mu}) = 10.8247/(7 \times 2.0046) = 0.7714$, which would result in an acceptable performance for the queuing system (that is, below 85%).

TABLE 3: Observed values and frequencies.

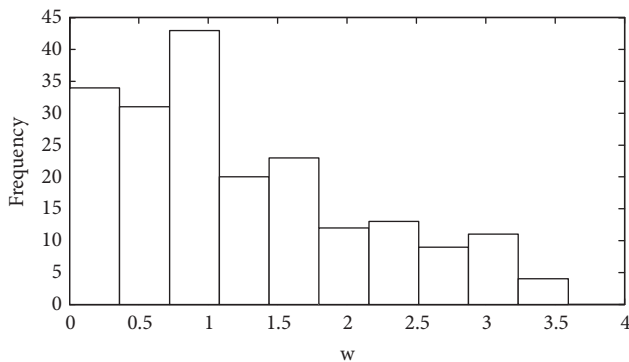| O	F	O	F	O	F	O	F
1	2	7	14	13	4	19	4
2	6	8	14	14	11	20	6
3	14	9	11	15	8	21	4
4	23	10	8	16	7	22	8
5	18	11	10	17	3	23	2
6	12	12	8	18	2	25	1

FIGURE 2: Histogram of the observed sojourn times for 200 clients.

4. Conclusions and Final Remarks

The estimation of the traffic intensity ρ, the arrival rate λ, and the service rate μ, for Markovian multiservice queues, known in Kendall notation as $M/M/s$ queues, is a challenging mathematical problem that was solved here by statistical inferential methods. For the $M/M/s$ queues, even without knowledge of the arrival rate and queue length of service, it is possible to make inferences about the traffic intensity by collecting a random sample of the number of customers in the system at the arbitrary epochs. If, in addition to the number of customers, the sojourn times in the queue are also collected, then it is possible to accurately estimate the arrival and service rates by means of simple expressions, as demonstrated here. In fact, a comprehensive set of Monte Carlo simulations were performed to show that estimates for ρ, μ, and λ may be obtained that are very close to the true values, especially when the sample sizes are close to 400.

Finally, it is worthwhile to mention that future studies in this area could include extensions to more general single-server queuing systems, such as $M/G/1$ and $G/M/1$, and general multiserver queues, such as $M/G/s$ and $G/M/s$. These are only a few of the many topics for future research in this area.

Data Availability

The data used to support the findings of this study are included within the article.

Disclosure

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

References

[1] D. Gross, J. F. Shortle, J. M. Thompson, and C. M. Harris, *Fundamentals of Queueing Theory*, Wiley-Interscience, New York, NY, 4th edition, 2008.

[2] L. Kleinrock, *Queueing Systems*, vol. I, Theory, John Wiley & Sons, New York, NY, 1975.

[3] L. Lakatos, L. Szeidl, and M. Telek, *Introduction to Queueing Systems with Telecommunication Applications*, Springer Science & Business Media, New York, NY, 2013.

[4] F. R. Cruz, M. A. Santos, F. L. Oliveira, and R. C. Quinino, "Estimation in a general bulk-arrival Markovian multi-server finite queue," *Operational Research*, pp. 1–17, 2018.

[5] H. T. Papadopolous, C. Heavey, and J. Browne, *Queueing Theory in Manufacturing Systems Analysis and Design*, Springer Science & Business Media, 1993.

[6] M. K. Govil and M. C. Fu, "Queueing theory in manufacturing: A survey," *Journal of Manufacturing Systems*, vol. 18, no. 3, pp. 214–240, 1999.

[7] G. Koole and A. Mandelbaum, "Queueing models of call centers: an introduction," *Annals of Operations Research*, vol. 113, pp. 41–59, 2002.

[8] E. Almehdawe, B. Jewkes, and Q.-M. He, "A Markovian queueing model for ambulance offload delays," *European Journal of Operational Research*, vol. 226, no. 3, pp. 602–614, 2013.

[9] S. P. J. van Brummelen, W. L. de Kort, and N. M. van Dijk, "Waiting time computation for blood collection sites," *Operations Research for Health Care*, vol. 7, pp. 70–80, 2015.

[10] E. Almehdawe, B. Jewkes, and Q.-M. He, "Analysis and optimization of an ambulance offload delay and allocation problem," *OMEGA - The International Journal of Management Science*, vol. 65, pp. 148–158, 2016.

[11] A. B. Clarke, "Maximum likelihood estimates in a simple queue," *Annals of Mathematical Statistics*, vol. 28, pp. 1036–1040, 1957.

[12] V. E. Beneš, "A Sufficient Set of Statistics for a Simple Telephone Exchange Model," *Bell System Technical Journal*, vol. 36, no. 4, pp. 939–964, 1957.

[13] L. Schruben and R. Kulkarni, "Some consequences of estimating parameters for the M/M/1 queue," *Operations Research Letters*, vol. 1, no. 2, pp. 75–78, 1982.

[14] I. V. Basawa and N. U. Prabhu, "Large sample inference from single server queues," *Queueing Systems*, vol. 3, no. 4, pp. 289–304, 1988.

[15] S. Zheng and A. F. Seila, "Some well-behaved estimators for the M/M/1 queue," *Operations Research Letters*, vol. 26, no. 5, pp. 231–235, 2000.

[16] M. A. C. Almeida, F. R. B. Cruz, F. L. P. Oliveira, and G. de Souza, "Bias correction for estimation of performance measures of a Markovian queue," *Operational Research*, Article ID 10.1007/s12351-017-0351-4, pp. 1–20, 2017.

[17] C. Armero and M. J. Bayarri, "Bayesian prediction in M/M/1 queues," *Queueing Systems*, vol. 15, no. 1-4, pp. 401–417, 1994.

[18] C. Armero and M. J. Bayarri, "A Bayesian analysis of a queueing system with unlimited service," *Journal of Statistical Planning and Inference*, vol. 58, no. 2, pp. 241–261, 1997.

[19] C. Armero and M. J. Bayarri, "Dealing with uncertainties in queues and networks of queues: A Bayesian approach," in *Multivariate, Design and Sampling*, S. Ghosh, Ed., pp. 579–608, Marcel Dekker, Inc, New York, 1999.

[20] C. Armero and D. Conesa, "Inference and prediction in bulk arrival queues and queues with service in stages," *Applied Stochastic Models in Business and Industry*, vol. 14, no. 1, pp. 35–46, 1998.

[21] C. Armero and D. Conesa, "Prediction in Markovian bulk arrival queues," *Queueing Systems*, vol. 34, no. 1-4, pp. 327–350, 2000.

[22] C. Armero and D. Conesa, "Statistical performance of a multiclass bulk production queueing system," *European Journal of Operational Research*, vol. 158, no. 3, pp. 649–661, 2004.

[23] C. Armero and D. Conesa, "Bayesian hierarchical models in manufacturing bulk service queues," *Journal of Statistical Planning and Inference*, vol. 136, no. 2, pp. 335–354, 2006.

[24] M. F. Mcgrath, D. Gross, and N. D. Singpurwalla, "A subjective Bayesian approach to the theory of queues I - Modeling," *Queueing Systems*, vol. 1, no. 4, pp. 317–333, 1987.

[25] M. F. McGrath and N. D. Singpurwalla, "A subjective Bayesian approach to the theory of queues II - Inference and information in M/M/1 queues," *Queueing Systems*, vol. 1, no. 4, pp. 335–353, 1987.

[26] A. Choudhury and A. C. Borthakur, "Bayesian inference and prediction in the single server Markovian queue," *Metrika*, vol. 67, no. 3, pp. 371–383, 2008.

[27] M. A. C. Almeida and F. R. B. Cruz, "A note on Bayesian estimation of traffic intensity in single server Markovian queues," *Communications in Statistics - Simulation Computation*, vol. 47, no. 9, pp. 2577–2586, 2017.

[28] F. R. Cruz, R. C. Quinino, and L. L. Ho, "Bayesian estimation of traffic intensity based on queue length in a multi-server M/M/s queue," *Communications in Statistics - Simulation and Computation*, vol. 46, no. 9, pp. 7319–7331, 2017.

[29] R. C. Quinino and F. R. B. Cruz, "Bayesian sample sizes in an M/M/1 queueing systems," *The International Journal of Advanced Manufacturing Technology*, vol. 88, no. 1-4, pp. 995–1002, 2017.

[30] F. R. B. Cruz, M. A. C. Almeida, DM. F. S. V. Angelo, and T. van Woensel, "Traffic Intensity Estimation in Finite Markovian Queueing Systems," *Mathematical Problems in Engineering*, vol. 2018, Article ID 3018758, 15 pages, 2018.

[31] A. Choudhury and A. Basak, "Statistical inference on traffic intensity in an M/M/1 queueing system," *International Journal of Management Science and Engineering Management*, vol. 13, no. 4, pp. 274–279, 2018.

[32] I. V. Basawa, U. N. Bhat, and R. Lund, "Maximum likelihood estimation for single server queues from waiting time data," *Queueing Systems*, vol. 24, no. 1-4, pp. 155–167, 1997.

[33] S. Chowdhury and S. P. Mukherjee, "Estimation of Traffic Intensity Based on Queue Length in a Single M/M/1 Queue," *Communications in Statistics—Theory and Methods*, vol. 42, no. 13, pp. 2376–2390, 2013.

[34] J. V. Ross, T. Taimre, and P. K. Pollett, "Estimation for queues from queue length data," *Queueing Systems*, vol. 55, no. 2, pp. 131–138, 2007.

[35] MATLAB® Release R2008a. (2008). The MathWorks™, Inc., Natick, Massachusetts, USA.

[36] C. H. Kapadia and R. L. Thomasson, "On estimating the parameter of a truncated geometric distribution by the method of moments," *Annals of the Institute of Statistical Mathematics*, vol. 27, no. 1, pp. 269–272, 1975.