Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: www.tandfonline.com/journals/tstf20

Bayesian inference and prediction in multi-server Markovian queueing system with reverse balking: A simulation-based approach

Asmita Tamuli & Dhruba Das

To cite this article: Asmita Tamuli & Dhruba Das (2025) Bayesian inference and prediction in multi-server Markovian queueing system with reverse balking: A simulation-based approach, Statistical Theory and Related Fields, 9:4, 367-383, DOI: 10.1080/24754269.2025.2555032

To link to this article: https://doi.org/10.1080/24754269.2025.2555032

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 25 Sep 2025.

Submit your article to this journal

Article views: 146

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tstf20

STATISTICAL THEORY AND RELATED FIELDS
2025, VOL. 9, NO. 4, 367–383
https://doi.org/10.1080/24754269.2025.2555032

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Bayesian inference and prediction in multi-server Markovian queueing system with reverse balking: A simulation-based approach

Asmita Tamuli and Dhruba Das

Department of Statistics, Dibrugarh University, Dibrugarh, India

ABSTRACT

This article presents an in-depth exploration of classical and Bayesian inference methods to estimate the traffic intensity parameter, providing a comprehensive comparison of these two statistical paradigms in a novel multiserver Markovian queueing model (M/M/s) incorporating the phenomenon of reverse balking. The classical inference relies on maximum likelihood (ML) estimation, while the Bayesian approach leverages prior distributions and posterior analysis to enhance estimates. The results indicate that Bayesian methods offer better flexibility and precision compared to traditional ML estimates. Additionally, the predictive probabilities for the number of customers in the system are calculated for different hyper-parameter values of the prior through extensive simulation techniques. The results provide valuable insights for optimizing queue management and improving service efficiency in systems where reverse balking occurs. Moreover, a real-life example is presented to demonstrate the practical implementation of the proposed methodology. This work not only advances the theoretical understanding of queueing dynamics, but also offers practical implications for industries relying on efficient service mechanisms.

ARTICLE HISTORY
Received 28 December 2024
Accepted 22 August 2025

KEYWORDS
M/M/s queueing system; reverse balking; maximum likelihood estimation; Bayesian estimation; prediction

1. Introduction

Queueing systems have been a significant area of study in operations research for decades due to their extensive applications in telecommunications, manufacturing, transportation, and the service industries. Multi-server Markovian queueing systems, in particular, have received considerable amounts of attention for their ability to model and analyze systems in which several servers manage incoming units with stochastic service times. The traditional framework of these systems has been extensively explored, providing vital insights into performance measures such as traffic intensity, expected system size, and queue lengths. However, real-world problems often involve various customer behaviours that classical models cannot fully capture. One such behaviour is reverse balking, in which customers are more likely to join a system as it increases, in contrast to the traditional balking phenomenon in which longer queues discourage new arrivals. The concept of reverse balking was pioneered by Jain et al. (2014). Reverse balking is common in investment

CONTACT Dhruba Das dhrubadas16@gmail.com Department of Statistics, Dibrugarh University, Dibrugarh-786004, Assam, India

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

firms such as stock markets, crowdfunding platforms, and renowned financial advisors, where high participation attracts more customers. Although the concept of reverse balking has received less attention, it is observed in various real-world scenarios where congestion attracts more users. Some of the works considering the concept of reverse balking can be found in Jyothsna et al. (2022); Kumar and Som (2020); Saikia and Choudhury (2021); Som and Kumar (2018); Tamuli et al. (2024, 2025), and others cited therein. Som and Kumar (2018) studied a finite capacity Markovian queueing system with two heterogeneous servers, reverse balking, and reneging, obtained stationary system size probabilities using an iterative method and conducted a sensitivity analysis. Jyothsna et al. (2022) analyzed a steady-state finite buffer M/M/1 feedback queue with reverse balking, reverse reneging, and multiple working vacations, calculated steady-state system length distributions and optimized costs using ACO. Tamuli et al. (2025) investigated an M/M/2/K heterogeneous queueing system with reverse balking and reneging till the end of service and derived steady-state probabilities. They also performed the sensitivity analysis and cost optimization of the queueing model.

The effective management of a queueing system necessitates the control of various performance measures to design an efficient and reliable system. Among these measures, traffic intensity, denoted as ρ, stands out as the most crucial performance measure. Traffic intensity is defined as the ratio of the arrival rate of customers to the service rate of the system. This measure encapsulates the overall load on the queueing system and significantly influences other performance measures. The statistical inference of traffic intensity serves as a cornerstone for assessing system stability and making informed decisions for the efficient management of queueing systems. The estimation of traffic intensity in a multi-server Markovian queueing system considering reverse balking presents novel challenges and opportunities. The primary objective is to develop a robust framework for classical and Bayesian inference of the proposed queueing system. Clarke (1957) made a pioneering contribution in the domain of statistical inference of the queueing parameters by estimating the queueing parameters λ, μ, and ρ using the classical approach. Since then a surge of interest has been witnessed in the statistical inference of queueing parameters and performance measures which can be found in the works of Acharya and Singh (2019); Basawa and Prabhu (1981); Bingham and Pitts (1999); Clarke (1957); Cruz et al. (2018); Goyal and Harris (1972); Singh et al. (2024) and others cited therein. However, Bayesian inference provides a powerful approach incorporating prior knowledge and for dealing with uncertainty. Muddapur (1972) pioneered the Bayesian approach to inference in queueing systems by extending Clarke's methodology to get Bayesian estimates of λ, μ, and ρ. Singh et al. (2021) investigated statistical methods to estimate the traffic intensity viz., maximum likelihood and Bayesian estimators, by observing the number of customers present in the system at successive departure epochs, and provided computational results from Monte Carlo simulations to establish the efficiency and effectiveness of the proposed approaches. The Bayesian approach to inference in queueing systems has gained significant attention in recent times. Basak and Choudhury (2021) derived Bayesian estimators for traffic intensity in a single-server queuing model with exponentially distributed inter-arrival and service times, comparing their performance with classical maximum likelihood estimators and selecting priors based on a model comparison criterion using Bayes factor. Singh et al. (2023) obtained Bayes estimators for traffic intensity ρ under the squared error loss function in an M/D/1 queueing system, assuming three forms of prior information (incomplete gamma, left-truncated beta, and improper Jeffreys priors), proposed the Bayes factor as a model comparison criterion, and conducted Monte

Carlo simulations to validate the proposed algorithms, with a real case study illustrating method applicability. Numerous other studies on Bayesian inference of queueing systems can be found in works by Bura and Sharma (2023); Deepthi and Jose (2020); Singh et al. (2024), and among others.

This article developed a multi-server Markovian queueing model considering reverse balking. To date, no research endeavours have estimated performance measures for queueing systems incorporating reverse balking. In the case of reverse balking, the probability of joining the system (b_m) is an increasing function of the system size (m) which makes the computational analysis more complex. In this study, we propose the probability of joining the system as $b_m = \frac{m+1}{m+2}$ to effectively capture the essence of reverse balking. This formulation ensures that when the system is idle ($m = 0$), the probability of joining the system is unbiased, reflecting a natural scenario where customers have an equal chance of either entering or balking. This choice is crucial as it maintains a realistic and mathematically tractable framework for analyzing reverse balking, which has not been extensively studied in classical and Bayesian estimation contexts. The motivation behind this study lies in addressing the complexities introduced by state-dependent joining probabilities in queueing systems, which significantly impact performance evaluation and decision-making. By incorporating both classical and Bayesian estimation techniques, we provide a comprehensive statistical framework for estimating key parameters, enhancing model applicability in real-world scenarios such as customer flow management in service industries. Given the critical importance of traffic intensity as a performance measure, the study focuses on estimating traffic intensity using both classical and Bayesian approaches. Specifically, the Bayesian estimation is performed using the Markov Chain Monte Carlo (MCMC) method implemented in R software. By precisely estimating traffic intensity, businesses can make informed decisions about system capacity, and process improvements, ensuring that they can meet customer demand without overburdening their resources. This leads to more efficient operations, cost savings, and improved customer experiences, particularly in environments with fluctuating demand and complex customer behaviours such as reverse balking.

2. Mathematical modelling

This article proposed a novel M/M/s queueing model considering reverse balking. Customers arrive at the queueing system following a Poisson process with arrival rate λ and an arriving customer joins the system with probability $b_m = \frac{m+1}{m+2}$ such that b_m is a monotonically increasing function of the number of customers (denoted as m) in the system. When the system is void, the probability of an arriving customer joining the system is assumed to be unbiased, i.e. $\frac{1}{2}$. It suggests that in the absence of prior system congestion information, an arriving customer is equally likely to join or not join the system. Customers are served according to the first-come, first-serve basis (FCFS) principle by s homogeneous servers, each having exponential service times. If there are r ($1 \leq r \leq s$) customers in the system, then only r servers are busy, and the interval between two consecutive service completions is exponentially distributed with a mean $\frac{1}{r\mu}$. If there are r ($r \geq s$) customers in the system, then all s servers are busy, and the interval between two consecutive service completions is exponentially distributed with a mean $\frac{1}{s\mu}$. Therefore, the service rate is

$$\begin{cases} r\mu, & 1 \leq r < s, \\ s\mu, & r \geq s. \end{cases}$$

The traffic intensity ρ is the average proportion of time where each server is busy. Queueing theory often operates under the assumption of equilibrium. When $\rho<1$, then the system is in equilibrium and has a stationary distribution. The following are the differential difference equations of the proposed queueing system obtained using the Birth and Death equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}P_0(t) = \mu P_1(t) - \frac{\lambda}{2}P_0(t); \tag{1}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}P_m(t) = \frac{m}{m+1}\lambda P_{m-1}(t) + (m+1)\mu P_{m+1}(t) - \frac{m+1}{m+2}\lambda P_m(t) - m\mu P_m(t), \quad 0 < m < s; \tag{2}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}P_m(t) = \frac{m}{m+1}\lambda P_{m-1}(t) + s\mu P_{m+1}(t) - \frac{m+1}{m+2}\lambda P_m(t) - s\mu P_m(t), \quad m \geq s. \tag{3}$$

Assuming that steady-state exists, then $\lim_{m\to\infty} P_m(t) = P_m$, and $\lim_{m\to\infty} P'_m(t) = 0$. Therefore, the steady-state equations of the queueing model under investigation are

$$0 = \mu P_1 - \frac{\lambda}{2}P_0; \tag{4}$$

$$0 = \frac{m}{m+1}\lambda P_{m-1} + (m+1)\mu P_{m+1} - \frac{m+1}{m+2}\lambda P_m - m\mu P_m, \quad 0 < m < s; \tag{5}$$

$$0 = \frac{m}{m+1}\lambda P_{m-1} + s\mu P_{m+1} - \frac{m+1}{m+2}\lambda P_m - s\mu P_m, \quad m \geq s. \tag{6}$$

Solving Equations (4)–(6), the stationary distribution of the number of customers (M) in the system at departure epochs is given by

$$P_m = \begin{cases} \dfrac{(s\rho)^m}{(m+1)!}P_0, & 0 \leq m < s, \\ \dfrac{s^s\rho^m}{s!(m+1)}P_0, & s \leq m, \end{cases} \tag{7}$$

where $\rho = \frac{\lambda}{s\mu}$ (Shortle et al., 2018).

Using the normality condition, $P_0 = P(M = 0)$ can be obtained as

$$\begin{aligned} P_0 &= \left(\sum_{m=0}^{s-1}\frac{(s\rho)^m}{(m+1)!} + \sum_{m=s}^{\infty}\frac{s^s\rho^m}{s!(m+1)}\right)^{-1} \\ &= \frac{\mathrm{e}^{\rho s}\Gamma(s+1, s\rho) - s\Gamma(s)}{\rho s^2\Gamma(s)} + \frac{s^s\rho^s}{s!}\Phi_{HL}(\rho, 1, s+1), \end{aligned} \tag{8}$$

where $\Phi_{HL}(\cdot)$ is a Hurwitz-LerchPhi function (McPhedran et al., 2007).

3. Performance measures and sensitivity analysis

In this section, we present the key performance measures of the system, including expected system size and average reverse balking, which help evaluate the efficiency of the queueing model. Performance measures are crucial for understanding system behaviour, optimizing resource allocation, and improving service quality. Additionally, sensitivity analysis is conducted to examine how variations in system parameters, such as arrival and service rates, affect these measures, providing insights into system performance.

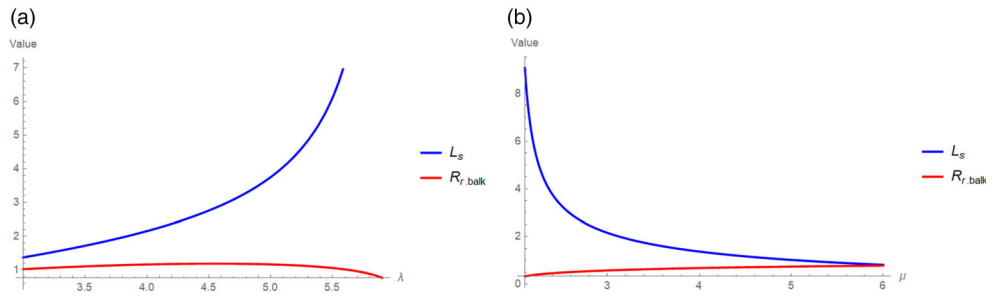

Figure 1. Impact of system parameter's on performance measures. (a) Variations in Expected system size w.r.t. λ (where $s = 4$ and $\mu = 6$). (b) Variations in Expected system size w.r.t. μ (where $s = 4$ and $\lambda = 2$).

3.1. Performance measures

Expected System Size L_s is

$$L_s = \sum_{m=1}^{\infty} mP_m = \sum_{m=1}^{s-1} \left(m \frac{(sp)^m}{(m+1)!} \right) + \sum_{m=s}^{\infty} \left(m \frac{s^s \rho^m}{s!(m+1)} \right). \tag{9}$$

Expected Reverse Balking Rate $R_{r.\text{balk}}$ is

$$\begin{aligned} R_{r.\text{balk}} &= \sum_{m=0}^{\infty} \left(1 - \frac{m+1}{m+2} \right) \lambda P_m \\ &= \sum_{m=0}^{s-1} \left(1 - \frac{m+1}{m+2} \right) \lambda \frac{(sp)^m}{(m+1)!} + \sum_{m=s}^{\infty} \left(1 - \frac{m+1}{m+2} \right) \lambda \frac{s^s \rho^m}{s!(m+1)}. \end{aligned} \tag{10}$$

3.2. Sensitivity analysis

The variations in performance measures for different values of arrival rate λ are illustrated in Figure 1(a). As the arrival rate λ increases, the expected system size also increases, indicating a higher number of customers in the system. Conversely, the expected reverse balking decreases, as customers are less likely to hesitate balk when the system is more occupied. This outcome aligns with the model's expectation, where a larger system size encourages customers to join rather than balk.

Further, the variations in the performance measures for various values of service rate μ are shown in Figure 1(b). As service rate μ increases, the expected system size decreases due to faster service completion. Consequently, the expected reverse balking increases, as a larger system size fosters trust, making customers less likely to balk. These findings are consistent with the model's expectation.

4. Classical estimation

In this section, the maximum likelihood estimator of traffic intensity ρ for the proposed model is obtained. In formulating the likelihood function, this article considers an ensemble of n independent and identically distributed instances of the proposed queueing model. To generate the data, it is necessary to observe the system at departure epochs. It is assumed

that each departing customer leaves behind a certain number of customers, denoted as x_i at ith observation. Therefore, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ constitutes our sample of size n and each x_i $(i = 1, 2, \ldots)$ follows the derived distribution (7). Then the corresponding likelihood function is obtained as

$$\begin{aligned} L(\mathbf{x}|\rho) &= \prod_{i=1}^{n} \left[\frac{(s\rho)^{x_i}}{(x_i+1)!} P_0 I_{\{0 \le x_i < s\}} + \frac{s^s \rho^{x_i}}{s!(x_i+1)} P_0 I_{\{x_i \ge s\}} \right] \\ &= \frac{(s\rho)^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n}(x_i+1)!} P_0 I_{\{0 \le x_i < s\}} + \frac{s^s \rho^{\sum_{i=1}^{n} x_i}}{s!\left(\sum_{i=1}^{n} x_i + n\right)} P_0 I_{\{x_i \ge s\}}, \end{aligned} \tag{11}$$

where $P_0 = \frac{e^{\rho s}\Gamma(s+1, s\rho) - s\Gamma(s)}{\rho s^2 \Gamma(s)} + \frac{s^s \rho^s}{s!}\Phi_{HL}(\rho, 1, s+1)$ and $I(\cdot)$ is the indicator function.

The data-generating process described above ensures the independence of sample observations, provided they are sufficiently spaced. This is due to the ergodic property of the Markov chain. Succinctly, if X_m denotes the number of customers in the system at the departure epoch of the mth customer, and $P_{ij}^k = P(X_{m+k} = j \,|\, X_m = j)$, then the ergodicity property asserts that the limiting probabilities $v_j = \lim_{k\to\infty} P_{ij}^k$, for $j = 1, 2, \ldots$, exist and are independent of the initial state i, provided the system is in a stationary state (Choudhury & Borthakur, 2008).

Then taking logarithm of Equation (11), the log-likelihood function for ρ is obtained as

$$\begin{aligned} \log L(\mathbf{x} \,|\, \rho) = &\left[\sum_{i=1}^{n} x_i \log s + \sum_{i=1}^{n} x_i \log \rho - \log\left\{ \prod_{i=1}^{n} (x_i + 1) \right\} + \log P_0 \right] I_{\{0 \le x_i\}} \\ &+ \left[s \log s + \sum_{i=1}^{n} x_i \log \rho - \log(s!) - \log\left(\sum_{i=1}^{n} x_i + \log n \right) + \log P_0 \right] I_{\{x_i \ge s\}}. \end{aligned} \tag{12}$$

By differentiating Equation (12) with respect to ρ and equating it to zero, (i.e., $\frac{\partial}{\partial \rho} \log L(\mathbf{x} \,|\, \rho = 0)$), and then solving it, the maximum likelihood estimator of ρ can be obtained. However, a direct solution of the aforementioned equation is not mathematically sound. Therefore, this article employed a numerical approximation using GenSA package in R software (version 4.3.3). The algorithm for generating samples using MH sampler is as follows.

5. Bayesian estimation

In this section, the Bayesian estimator of ρ for the proposed queueing model is presented. Here, ρ is assumed to be a random variable, and its randomness is quantified through a prior distribution, denoted as $\pi(\rho)$. This prior information is updated in light of the observed data and is represented by the posterior distribution, which is given by

$$\Pi(\rho \,|\, \mathbf{x}) = \frac{L(\mathbf{x} \,|\, \rho)\pi(\rho)}{\int_0^1 L(\mathbf{x} \,|\, \rho)\pi(\rho)\, \mathrm{d}\rho} \propto L(\mathbf{x} \,|\, \rho)\pi(\rho),$$

and any inference about the traffic intensity ρ can be drawn based on its posterior distribution derived from the observed data.

Algorithm 1: Algorithm to generate sample observations from P_m

Step 1: Set an initial state x_0 at $t = 0$.
Step 2: Consider the proposal distribution as geometric distribution, $G(p), 0 \leq p \leq 1$.
Step 3: Generate a candidate point y from a proposal distribution $G(\cdot)$ which depends on the current state x_t.
Step 4: Compute the acceptance probability $\alpha(x_t, y)$:

$$\alpha(x_t, y) = \left(1, \frac{\Psi(y)G(x_t \mid y)}{\Psi(x_t)G(y \mid x_t)}\right).$$

Here, $\Psi(\cdot)$ is the target distribution denoted as $P_m$in Equation (7).
Step 5: Generate a uniform random number $U \sim \text{Uniform}(0, 1)$.
Step 6: If $u \leq \alpha(x_t, y)$, acceptthe candidate and set $x_{t+1} = y$; otherwise, reject thecandidate and set $x_{t+1} = x_t$.
Step 7: Increment t by 1. Repeat Steps 2to 4. Obtain the sample distributed as $\Psi(\cdot)$.

Since $0 < \rho < 1$, a natural prior for ρ is

$$\pi(\rho) = \frac{1}{\beta(a, b)}\rho^{a-1}(1-\rho)^{b-1}, \quad a, b > 0. \tag{13}$$

The beta distribution is chosen due to its flexibility, which enables it to accommodate a wide range of distributional shapes and functions as a natural distribution for Equation (7). The determination of parameters a and b for the prior distribution can be approached indirectly. This can be achieved through various techniques such as utilizing percentiles (e.g. 5%, 50%, and 95%) or by considering the average and standard deviation of the prior distribution of traffic intensity (ρ) (Cruz et al., 2017). To account for different prior beliefs regarding the traffic intensity ρ, various Beta prior distributions are considered. Beta(1, 1) was utilized as non-informative prior to indicate an absence of prior information about the traffic intensity (ρ). Subsequently, Beta(5, 5) was adopted to represent a prior belief that ρ is likely concentrated around 0.5. Additionally, scenarios where prior beliefs leaned towards either $\rho > 0.5$ or $\rho < 0.5$ were considered, employing Beta distributions Beta(5, 2) and Beta(2, 5) respectively. If the prior belief suggests that the traffic intensity ρ is more likely to be greater than 0.5, then the Beta distribution Beta(5, 2) is a suitable choice. In the case of Beta(5, 2), the distribution is skewed towards the right. This means that the probability mass is concentrated towards higher values of ρ, making it more probable that $\rho > 0.5$.

Figure 2 visually presents the shapes of the Beta distributions utilized in this study, offering insight into their respective probability density functions. This graphical representation helps in understanding how the choice of prior distributions influences the Bayesian estimation process. This article systematically examines various combinations of sample sizes (n), and traffic intensities (ρ), considering beta prior distribution for different hyper-parameter values to comprehensively evaluate the performance of the Bayesian approach in modelling customers at departure epochs.

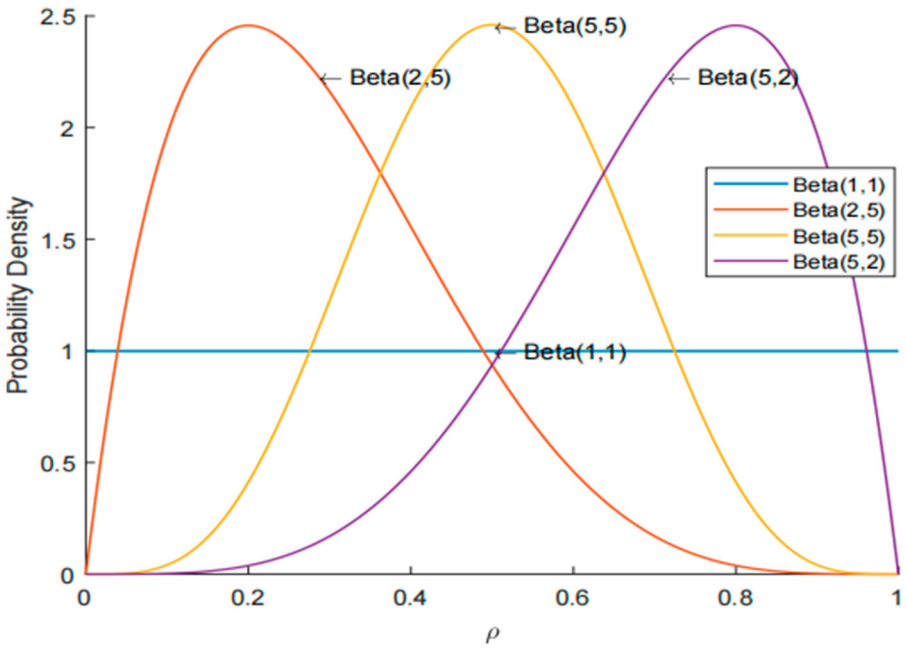

Figure 2. Different Beta prior distributions under consideration.

The posterior distribution of ρ given the data corresponding to the prior distribution in Equation (13) is given by

$$\begin{aligned}
&\Pi(\rho \mid \mathbf{x}) \\
&\propto \frac{1}{\beta(a,b)}\rho^{a-1}(1-\rho)^{b-1}\left[\frac{(s\rho)^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}(x_i+1)!}P_0 I_{\{0\le x_i<s\}} + \frac{s^s\rho^{\sum_{i=1}^{n}x_i}}{s!(\sum_{i=1}^{n}x_i+n)}P_0 I_{\{x_i\ge s\}}\right] \\
&\propto \frac{s^{\sum_{i=1}^{n}x_i}}{\beta(a,b)\prod_{i=1}^{n}(x_i+1)!}\rho^{\sum_{i=1}^{n}x_i+a-1}(1-\rho)^{b-1}P_0 I_{\{0\le x_i<s\}} \\
&\quad + \frac{s^s}{\beta(a,b)s!(\sum_{i=1}^{n}x_i+n)}\rho^{\sum_{i=1}^{n}x_i+a-1}(1-\rho)^{b-1}P_0 I_{\{x_i\ge s\}}. \qquad (14)
\end{aligned}$$

Since it is difficult to obtain the posterior distribution of $\Pi(\rho|\text{data})$ given in Equation (14), this article employed the Monte Carlo Markov Chain (MCMC) method. The MCMC method offers a solution for problems where direct computation of the Bayesian estimate is not feasible. The implementation of MCMC method was carried out in R software. The algorithm for Bayesian estimation is as follows.

5.1. Predictive distribution

After observing $\mathbf{x}$, one can compute the stationary predictive distribution of the number of customers (M) in the system at the departure epochs, $P_{m,i}^{\text{pred}} = P(M = m \mid \mathbf{x})$. For obtaining

Algorithm 2: Algorithm to generate sample observations from $\Pi(\rho|\mathbf{x})$

Step 1: Set ρ_0 at $t = 0$.
Step 2: Generate a random sample of *n*observations from the pmf inEquation (7).
Step 3: Consider the posterior distributionas the target distribution given inEquation (14).
Step 4: Consider the proposal distributionas beta distribution given inEquation (13),$B(a, b), a, b > 0$.
Step 5: Generate a candidate point y from aproposal distribution $B(\cdot)$ which depends on the currentstate ρ_t.
Step 6: Compute the acceptance probability $\alpha(\rho_t, y)$:

$$\alpha(\rho_t, y) = \left(1, \frac{\Pi(y)B(\rho_t \mid y)}{\Pi(\rho_t)B(y \mid \rho_t)}\right).$$

Here, $\Pi(\cdot)$ is the posterior distribution given inEquation (14), which isconsidered as target distribution.
Step 7: Generate a uniform random number $U \sim \text{Uniform}(0, 1)$.
Step 8: If $u \leq \alpha(\rho_t, y)$,accept the candidate and set $\rho_{t+1} = y$; otherwise,reject the candidate and set $\rho_{t+1} = \rho_t$.
Step 9: Increment t by 1. Repeat Steps 2to 4. Obtain the sample distributed as $\Pi(\cdot)$.
Note: The above steps are repeated for $k = 10000$ iterations to obtain a Markov chain whose stationarydistribution approximates the posterior$\Pi(\rho \mid \mathbf{x})$.

$P_{m,i}^{\text{pred}}$, the initial step involves computing $P_{m,i}^{\text{pred}}, i = 1, 2, \ldots, k$ and is defined as

$$P_{m,i}^{\text{pred}} = \begin{cases} \dfrac{(s\rho_i^*)^m}{(m+1)!}P_0, & 0 \leq m < s, \\ \dfrac{s^s(\rho_i^*)^m}{s!(m+1)}P_0, & s \leq m, \end{cases} \tag{15}$$

where $P_{0,i}$ is given by

$$P_{0,i} = \left(\sum_{m=0}^{s-1} \frac{(s\rho_i^*)^m}{(m+1)!} + \sum_{m=s}^{\infty} \frac{s^s(\rho_i^*)^m}{s!(m+1)}\right)^{-1}, \quad i = 1, 2, \ldots, k. \tag{16}$$

The predictive probability is approximated by the average,

$$P_m^{\text{pred}} \cong \frac{1}{k}\sum_{i=1}^{k} P_{m,i}^{\text{pred}}. \tag{17}$$

6. Computational results

To evaluate the methodology discussed in the preceding sections, we computed the ML and Bayes estimates of ρ using samples derived from the described procedure for sample sizes, $n = 25, 50, 100, 200$. Under the Bayesian approach, the estimates were derived under different hyperparameter settings (a, b) of the prior distribution. Table 1 summarizes the results,

Table 1. ML and Bayes estimates of traffic intensity ρ.

(a, b) ρ	n	ML estimates		Bayes estimates	
		Estimate	RMSE	Estimate	RMSE
(1, 1) 0.20	25	0.197442	0.063655	0.219623	0.055428
	50	0.196347	0.044164	0.214076	0.039204
	100	0.198326	0.032473	0.212445	0.028222
	200	0.199105	0.023173	0.208881	0.021107
(2, 5) 0.31	25	0.295701	0.073999	0.305889	0.022492
	50	0.296986	0.050853	0.305260	0.016158
	100	0.296219	0.035400	0.304720	0.011466
	200	0.296473	0.025237	0.305152	0.007919
(5, 5) 0.59	25	0.571122	0.110916	0.584596	0.013246
	50	0.580815	0.081607	0.584914	0.009896
	100	0.585455	0.061028	0.585688	0.007188
	200	0.296473	0.025237	0.585673	0.004833
(5, 2) 0.91	25	0.942583	0.138363	0.899796	0.011003
	50	0.942025	0.088487	0.901801	0.007763
	100	0.921724	0.041340	0.906526	0.005732
	200	0.918899	0.012700	0.906866	0.003946

including the ML and Bayesian estimates of ρ along with their corresponding Root Mean Squared Errors (RMSE). The replication process has been carried out 1000 times across four distinct sample sizes.

Table 1 demonstrates that as the sample size increases, the RMSEs of both the Maximum Likelihood (ML) and Bayes estimates decrease. Consequently, the estimates gradually converge toward the true parameter values as the sample size increases. This convergence aligns with theoretical expectations, indicating the asymptotic consistency of the ML and Bayes estimates for larger sample sizes. It is observed that the computation time is higher for the Bayesian estimation approach compared to the maximum likelihood (ML) estimation method for all the parameter values. However, it is clearly evident from Figure 3 that the Bayes estimates consistently exhibit superior performance with lower RMSEs compared to the ML estimates.

Using Equation (17), the posterior predictive probabilities of the number of customers at the departure epoch are computed with a simulated sample of size 100. Table 2 presented the posterior predictive probabilities for customer counts ranging from 0 to 5. Furthermore, Figure 4 illustrates a plot of these probabilities for 0 to 10 customers, providing a visual representation of the predictive distribution. It is clearly evident from Table 2 and Figure 4 that the predictive probability of the system being idle decreases as the number of servers increases from $s = 4$ to $s = 7$.

7. Special case

The methodology outlined herein caters to an M/M/s queueing system with reverse balking where multiple servers exist. While such a model finds broad applicability in various practical settings, it's imperative to consider the special case. Therefore, following are the special cases of the proposed queueing system.

Case I: Single Server ($s = 1$)

When there is only one server, transform the system into an M/M/1 queueing system with reverse balking. For this scenario, the probability of the number of customers (M) in

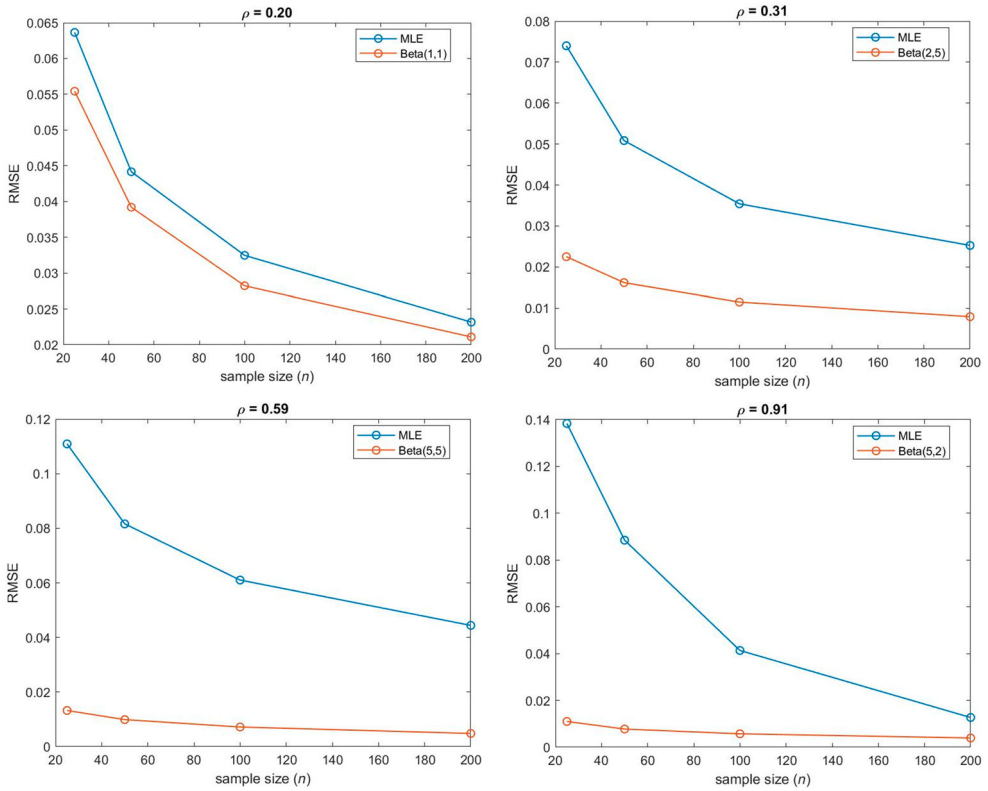

Figure 3. RMSE versus *n*.

the system at the departure epoch is given by

$$P_m = \begin{cases} \dfrac{\rho^m}{m+1} P_0, & m = 0, 1, 2, \ldots, \\ 0, & \text{Otherwise}, \end{cases} \tag{18}$$

where P_0 is obtained from utilizing the normality condition and is given by

$$P_0 = P\{M = 0\} = \left(1 + \sum_{j=1}^{\infty} \frac{\rho^j}{j+1}\right)^{-1}. \tag{19}$$

Assume that each departing customer leaves behind a certain number of customers, denoted as x_i. Taking a random sample of size n, $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$, then the resulting likelihood function is given by

$$L(\mathbf{x} \mid \rho) = \prod_{i=1}^{n} \left\{(x_i + 1)^{-1} \rho^{x_i} P_0\right\}. \tag{20}$$

Case II: Finite Capacity System (K)

In scenarios where physical space or other constraints limit the system's total capacity to K customers (including those in service and waiting), the queueing model is characterized as an M/M/s/K system. In this setting, if an arriving customer encounters the system at full

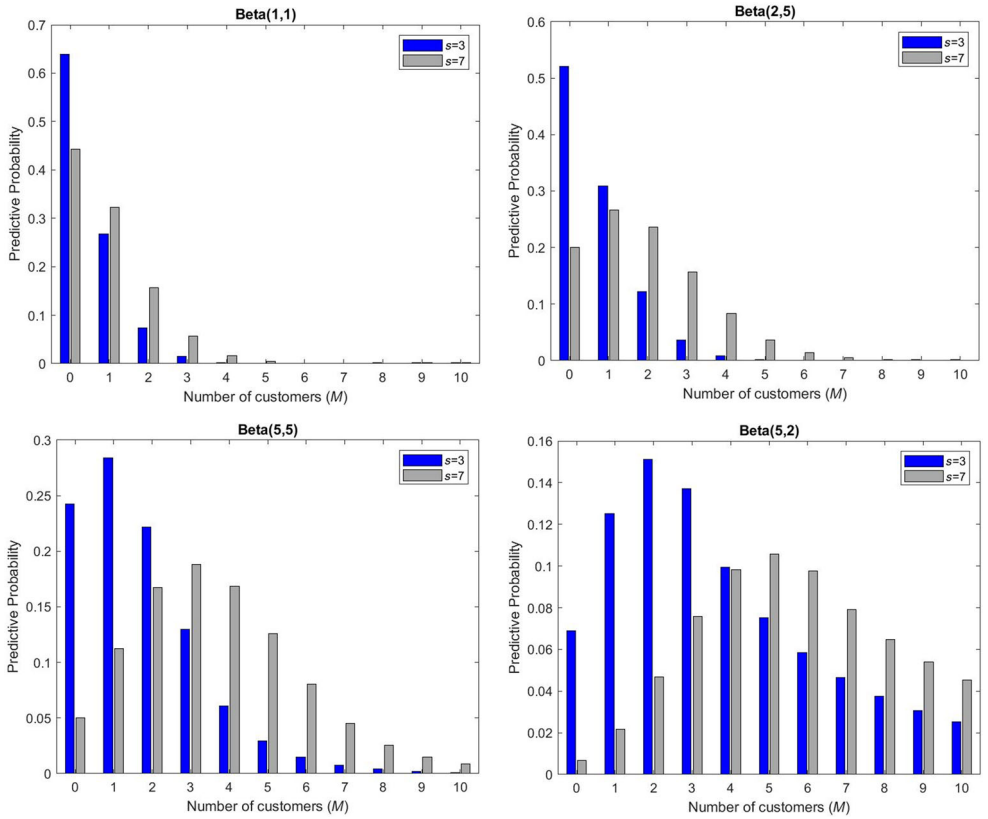

Figure 4. Posterior predictive probabilities for different values of *a* and *b*.

capacity, they are not allowed to enter and considered lost. This model is essential for analyzing systems with finite capacity, where the potential for customer loss due to space limitations must be accounted for. Therefore, the probability of the number of customers (M) present in the system at the departure epoch is given by

$$P_m = \begin{cases} \dfrac{(s\rho)^m}{(m+1)!}P_0, & 0 \le m < s, \\ \dfrac{s^s\rho^m}{s!(m+1)}P_0, & s \le m \le K, \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

Then, using the normality condition P_0 is obtained as

$$P_0 = \left(\sum_{m=0}^{s-1} \frac{(s\rho)^m}{(m+1)!} + \sum_{m=s}^{K} \frac{s^s\rho^m}{s!(m+1)} \right)^{-1}. \tag{22}$$

Similarly, assume that each departing customer leaves behind a certain number of customers, denoted as x_i. Taking a random sample of size n, $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$, then the likelihood

Table 2. Posterior predictive distribution of number of customers (M).

(a, b)	Customers (M)	$s = 4$	$s = 7$
(1, 1)	0	0.639657	0.442256
	1	0.267224	0.322335
	2	0.074424	0.156621
	3	0.015546	0.057076
	4	0.002598	0.016640
	5	0.000452	0.004043
(2, 5)	0	0.521165	0.200101
	1	0.309023	0.266024
	2	0.122156	0.235778
	3	0.036216	0.156728
	4	0.008590	0.083345
	5	0.002122	0.036934
(5, 5)	0	0.242520	0.050031
	1	0.284075	0.112169
	2	0.221834	0.167654
	3	0.129922	0.187940
	4	0.060873	0.168544
	5	0.029710	0.125958
(5, 2)	0	0.068969	0.006733
	1	0.125091	0.021773
	2	0.151254	0.046935
	3	0.137167	0.075882
	4	0.099514	0.098147
	5	0.075205	0.105786

function is given by

$$L(\mathbf{x} \mid \rho) = \frac{(s\rho)^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n}(x_i+1)!} P_0 \cdot I_{\{0 \le x_i < s\}} + \frac{s^s \rho^{\sum_{i=1}^{n} x_i}}{s!\left(\sum_{i=1}^{n} x_i + n\right)} P_0 \cdot I_{\{s \le x_i \le K\}}, \tag{23}$$

where $I(\cdot)$ is the indicator function.

Thus, the methodology elucidated in this article provided for an M/M/s queueing system with reverse balking can be extended for this special cases. Notably, the algorithms presented in this article would require necessary modifications to accommodate the consideration of one server into the model.

8. Real-life application

To demonstrate the methodology discussed in this article, we applied it to a real-life scenario in the Radiology department of a nursing home located in Dibrugarh, Assam, India. In this case study, we focus on the arrival of patients for ultrasound in the Radiology department, where patients are served according to first come first serve basis. The ultrasound unit is equipped with four separate cabins with four different ultrasound technologists, (i.e. $s = 4$). The phenomenon of reverse balking is inherent in such systems, as patients are more likely to visit nursing homes where a larger number of patients are already present, under the assumption that the diagnostic services provided are of high quality and that the attending physicians are experienced and trustworthy. This behaviour stems from the perception that a busy department reflects reliable care and accurate diagnostics. Estimating the traffic intensity (ρ) allows the nursing home management to assess system load, ensures effective utilization of ultrasound rooms, and reduces patient waiting time, thereby supporting better

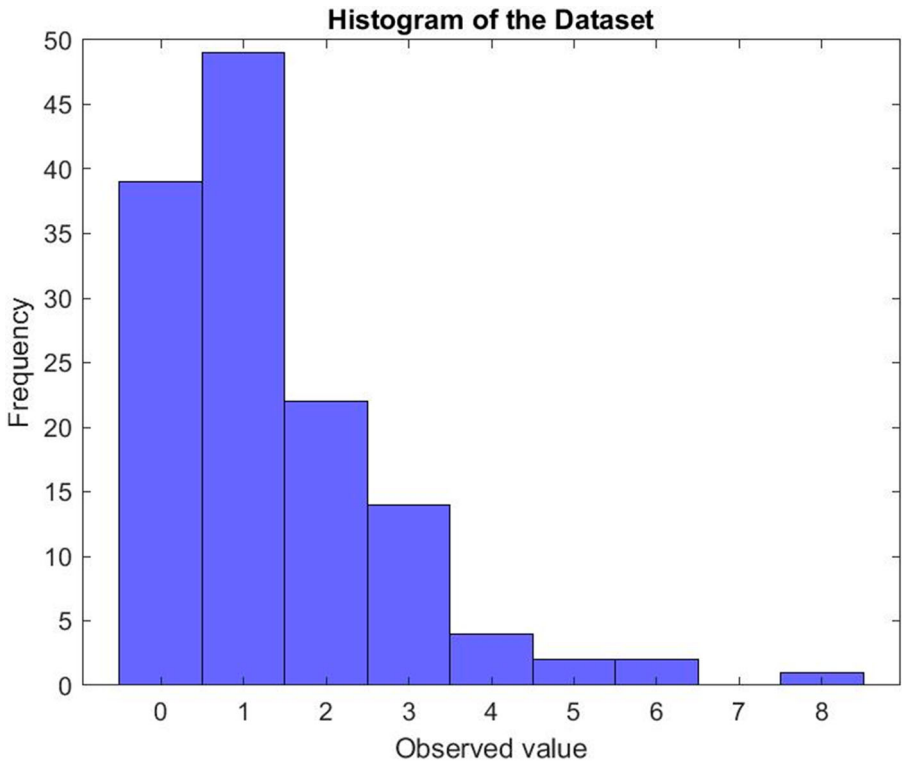

Figure 5. Histogram of the collected dataset.

scheduling, resource allocation, and service delivery. Therefore, the primary objective is to estimate the traffic intensity (ρ).

The ultrasound unit operates from Monday to Saturday, between 7:00 AM and 6:00 PM. We collected data on the number of patients present in the system at various time points during the operational hours, resulting in a total of 133 observations. The collected data are provided as follows and also presented visually in Figure 5.

```
Dataset:[6, 2, 0, 1, 1, 4, 1, 3, 4, 1, 3, 3, 6, 0, 2, 2, 0,
1, 0, 3, 0, 0, 1, 3, 1, 1, 4, 0, 0, 1, 0, 0, 2, 0, 2, 3, 0,
0, 0, 2, 3, 1, 1, 3, 2, 0, 0, 2, 2, 1, 1, 1, 1, 5, 1, 0, 1,
2, 0, 1, 0, 1, 2, 1, 2, 5, 1, 2, 3, 2, 1, 2, 2, 1, 1, 1, 2,
0, 1, 0, 1, 1, 1, 1, 0, 0, 3, 0, 8, 0, 1, 3, 0, 3, 1, 1, 1,
1, 1, 1, 1, 1, 0, 0, 2, 1, 0, 1, 1, 0, 1, 0, 1, 0, 3, 1, 2,
1, 0, 2, 2, 1, 1, 3, 0, 0, 0, 1, 0, 4, 2, 0, 0]
```

To highlight the advantages of our proposed queueing model, we compared it with the classical M/M/s queueing model. The MLE and χ^2 goodness-of-fit test results are depicted in Table 3. It is observed that the proposed queueing model with reverse balking provides a better fit to the data, as it exhibits lower AIC and BIC values compared to the classical M/M/s model. Therefore, the findings suggested that the proposed queueing model more accurately represents the real-life scenario compared to the classical model. Since the nursing home

Table 3. Result of goodness-of-fit test.

	Proposed Queueing Model	M/M/s Queueing Model
MLE	0.496963	0.333718
χ^2 statistic	4.285300	9.946100
p-value	0.746400	0.191600
AIC	415.984900	419.602700
BIC	418.875300	422.493100

administration lacks prior knowledge about traffic intensity, a non-informative Beta prior $B(1, 1)$ is chosen. Based on the observed data, the Bayesian estimate of traffic intensity is found to be 0.5006332.

9. Conclusion

This article develops and analyzes a multi-server Markovian queueing model considering reverse balking. Additionally, the classical and Bayesian estimation of traffic intensity ρ is presented through a comprehensive simulation-based approach. The findings emphasize the significance of reverse balking in the design and management of queueing systems by estimating the parameter of the queueing system, particularly when customer behaviour contrasts with traditional balking patterns. Using the Markov chain Monte Carlo (MCMC) method, both maximum likelihood (ML) and Bayesian estimates are derived. The study reveals that as sample size increases, both classical and Bayesian estimates converge to the true value. Moreover, Bayesian estimation outperforms classical methods in terms of RMSE. Additionally, predictive probabilities are obtained enhancing the understanding of the system's dynamics. The results show that the posterior predictive probability of the system being idle decreases as the number of servers increases. A real-life application of the proposed model is presented to demonstrate the application of the methodology discussed in this study.

Future research could extend this work by considering a heterogeneous multi-server Markovian queueing model, incorporating various customer behaviours, such as reneging, feedback mechanisms, retrials, and other realistic dynamics.

Acknowledgments

The authors would like to thank the anonymous reviewers for his/her detailed, careful, and exhaustive comments. These have led to a very substantial improvement in the paper. Furthermore, the first author expresses gratitude to the Department of Science and Technology (DST), Government of India, for supporting the research through an INSPIRE fellowship.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Data and code availability statement

The data used to support the findings of this study and the accompanying R codes will be made available from the corresponding author upon request.

ORCID

Asmita Tamuli https://orcid.org/0009-0006-4215-8633
Dhruba Das http://orcid.org/0000-0001-8546-0193

References

Acharya, S. K., & Singh, S. K. (2019). Asymptotic properties of maximum likelihood estimators from single server queues: A martingale approach. *Communications in Statistics–Theory and Methods*, *48*(14), 3549–3557. https://doi.org/10.1080/03610926.2018.1477958

Basak, A., & Choudhury, A. (2021). Bayesian inference and prediction in single server M/M/1 queuing model based on queue length. *Communications in Statistics–Simulation and Computation*, *50*(6), 1576–1588. https://doi.org/10.1080/03610918.2019.1586924

Basawa, I., & Prabhu, N. (1981). Estimation in single server queues. *Naval Research Logistics Quarterly*, *28*(3), 475–487. https://doi.org/10.1002/nav.v28:3

Bingham, N. H., & Pitts, S. M. (1999). Non-parametric estimation for the M/G/∞ queue. *Annals of the Institute of Statistical Mathematics*, *51*(1), 71–97. https://doi.org/10.1023/A:1003831118254

Bura, G. S., & Sharma, H. (2023). Maximum likelihood and Bayesian estimation on M/M/1 queueing model with balking. *Communications in Statistics–Theory and Methods*, *53*(14), 5117–5145. https://doi.org/10.1080/03610926.2023.2208695

Choudhury, A., & Borthakur, A. C. (2008). Bayesian inference and prediction in the single server Markovian queue. *Metrika*, *67*(3), 371–383. https://doi.org/10.1007/s00184-007-0138-3

Clarke, A. B. (1957). Maximum likelihood estimates in a simple queue. *The Annals of Mathematical Statistics*, *28*(4), 1036–1040. https://doi.org/10.1214/aoms/1177706808

Cruz, F. R., Almeida, M. A., D'Angelo, M. F., & van Woensel, T. (2018). Traffic intensity estimation in finite Markovian queueing systems. *Mathematical Problems in Engineering*, *2018*(2), 1–15. https://doi.org/10.1155/2018/3018758

Cruz, F. R., Quinino, R. D. C., & Ho, L. L. (2017). Bayesian estimation of traffic intensity based on queue length in a multi-server M/M/s queue. *Communications in Statistics–Simulation and Computation*, *46*(9), 7319–7331. https://doi.org/10.1080/03610918.2016.1236953

Deepthi, V., & Jose, J. K. (2020). Bayesian estimation of M/E_k/1 queueing model using bivariate prior. *American Journal of Mathematical and Management Sciences*, *40*(1), 88–105. https://doi.org/10.1080/01966324.2020.1835589

Goyal, T. L., & Harris, C. M. (1972). Maximum-likelihood estimates for queues with state-dependent service. *Sankhya: The Indian Journal of Statistics, Series A*, *34*(1), 65–80.

Jain, N., Kumar, R., & Som, B. K. (2014). An M/M/1/N queuing system with reverse balking. *American Journal of Operational Research*, *4*(2), 17–20.

Jyothsna, K., Laxmi, P. V., & Kumar, P. V. (2022). Optimization of a feedback working vacation queue with reverse balking and reverse reneging. *Reliability: Theory & Applications*, *17*(1), 154–163.

Kumar, R., & Som, B. K. (2020). A multi-server queue with reverse balking and impatient customers. *Pakistan Journal of Statistics*, *36*(2), 91–101.

McPhedran, R. C., Botten, L. C., Nicorovici, N. A. P., & John Zucker, I. (2007). Symmetrization of the Hurwitz zeta function and Dirichlet *L* functions. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *463*(2077), 281–301. https://doi.org/10.1098/rspa.2006.1762

Muddapur, M. (1972). Bayesian estimates of parameters in some queueing models. *Annals of the Institute of Statistical Mathematics*, *24*(1), 327–331. https://doi.org/10.1007/BF02479762

Saikia, G., & Choudhury, A. (2021). A single server Markovian queuing system with limited buffer and reverse balking. *Independent Journal of Management & Production*, *12*(7), 1774–1784. https://doi.org/10.14807/ijmp.v12i7.1471

Shortle, J. F., Thompson, J. M., Gross, D., & Harris, C. M. (2018). *Fundamentals of queueing theory*. (Vol. 399). John Wiley & Sons.

Singh, S. K., Acharya, S. K., Cruz, F. R., & Quinino, R. C. (2021). Estimation of traffic intensity from queue length data in a deterministic single server queueing system. *Journal of Computational and Applied Mathematics*, *398*,113693. https://doi.org/10.1016/j.cam.2021.113693

Singh, S. K., Acharya, S. K., Cruz, F. R., & Quinino, R. C. (2023). Bayesian inference and prediction in an M/D/1 queueing system. *Communications in Statistics–Theory and Methods*, *52*(24), 8844–8864. https://doi.org/10.1080/03610926.2022.2076120

Singh, S. K., Cruz, F. R., Gomes, E. S., & Banik, A. D. (2024). Classical and Bayesian estimations of performance measures in a single server Markovian queueing system based on arrivals during service times. *Communications in Statistics–Theory and Methods*, *53*(10), 3517–3546. https://doi.org/10.1080/03610926.2022.2155789

Som, B. K., & Kumar, R. (2018). A heterogeneous queuing system with reverse balking and reneging. *Journal of Industrial and Production Engineering*, *35*(1), 1–5. https://doi.org/10.1080/21681015.2017.1297739

Tamuli, A., Das, D., & Choudhury, A. (2024). Optimizing the performance of multiserver heterogeneous queueing systems with dynamic customer behaviour. *Sankhya B*, *86*(2), 366–414. https://doi.org/10.1007/s13571-024-00340-0

Tamuli, A., Das, D., Choudhury, A., & Kushvaha, B. (2025). Optimal service design for heterogeneous queueing system with reverse balking and reneging. *Operational Research*, *25*(2), 1–29. https://doi.org/10.1007/s12351-025-00911-7