Pak. J. Statist.
2026 Vol. 42(1), 1-10

PARETO DISTRIBUTION BASED G/G/1 QUEUE MODEL TO COMPARE THE SERVICE PERFORMANCE OF BANKS

Muhammad Amin[1§], Ahad Yasin[1] and **Saddam Qureshi[2]**
[1] Department of Statistics, University of Sargodha
Sargodha, Pakistan
[2] Department of Statistics, Emerson University
Multan, Pakistan.
[§] Corresponding author Email: muhammad.amin@uos.edu.pk

ABSTRACT

This study introduces a G/G/1 queue model to evaluate and compare the service performance of two leading banks in Pakistan i.e., United Bank Limited (UBL) and Habib Bank Limited (HBL), focusing on customer waiting times and queue efficiency. By assuming Pareto distributions for both arrival and service times, the model captures the heavy-tailed nature of banking service processes. Data from 200 customers per bank were analyzed and with the support of the Cramér-Von Mises test, we found that the waiting and service times were well fitted to the Pareto distributions. To compare the service performance of both banks, we derived key performance metrics, including arrival and service rates, server utilization factor, average queue length, and waiting times. Results reveal that UBL outperforms HBL, exhibiting shorter waiting times (2.78 vs. 7.43 minutes), fewer customers in the queue (1.63 vs. 4.55), and a lower server utilization factor (0.75 vs. 0.89), indicating greater system stability and efficiency. These findings highlight the efficacy of queueing theory in optimizing banking operations and enhancing customer satisfaction. The study provides actionable information for bank management to reduce queue lengths and service times, with potential applications in other service-oriented sectors like healthcare and hospitality. The use of Pareto distributions ensures a robust, data-driven approach to modeling real-world banking dynamics.

KEYWORDS

G/G/1 Queue Model, Pareto Distribution, Service Performance, Waiting Time.

1. INTRODUCTION

In an increasingly competitive global market, delivering superior services is paramount for fostering customer satisfaction and sustaining organizational growth (Amin et al., 2016). Banks, as pivotal institutions in driving economic development, face intense pressure to optimize service delivery to meet rising customer expectations. Efficient queue management is critical, as prolonged waiting times can erode customer trust, discourage account openings, and tarnish a bank's reputation. Queueing theory provides a powerful analytical framework to address these challenges by modeling customer arrival and service processes, enabling banks to minimize delays and enhance operational efficiency. In Pakistan, where banking services are integral to economic activity, optimizing queue

© 2026 Pakistan Journal of Statistics

performance is particularly vital to support financial inclusion and customer retention. This study leverages queueing theory to compare the service performance of two prominent banks, United Bank Limited (UBL) and Habib Bank Limited (HBL), focusing on their ability to manage customer queues effectively. By identifying and addressing inefficiencies in service delivery, banks can improve customer experiences, strengthen competitive positioning, and contribute to broader economic goals.

The global banking industry has increasingly prioritized service quality to maintain a competitive edge, particularly in emerging economies like Pakistan, where economic growth hinges on efficient financial systems. Numerous studies have explored bank service quality using diverse methodologies to address customer expectations and operational challenges. For instance, Lewis (1991) utilized descriptive metrics to evaluate service quality across international banks, highlighting variations in customer perceptions. Arasli et al. (2005) employed factor and multivariate regression analyses to establish a strong link between service quality and customer satisfaction in Cyprus, emphasizing the role of responsiveness and reliability. Similarly, Oyetunji et al. (2009) applied gap analysis to compare service quality in Ghanaian banks, revealing significant shortcomings in meeting customer expectations. In India, Banerjee and Sah (2012) conducted a comparative study of public and private banks, focusing on customer perceptions of service dimensions such as tangibility and empathy. Ahmad et al. (2010) used linear regression to confirm a robust relationship between service quality and customer satisfaction in Pakistani banks, underscoring the need for consistent service delivery. Queueing models have also been instrumental in analyzing waiting times. Famule and Daisi (2010) applied the M/M/1 model in Nigeria to assess customer waiting times, while Sundari and Srinivasan (2011) used the M/M/C model in India to recommend increasing service channels to reduce queue lengths. These studies collectively highlight the critical need for tailored analytical approaches to address queue-related inefficiencies, which can lead to customer dissatisfaction, reduced loyalty, and operational bottlenecks in banking environments.

In the literature, Famule and Daisi (2010) and Sundari and Srinivasan (2011) applied the M/M/1 queue model for the evaluation of banks'service, where the probability of arrival and service times is exponential. However, it is not always true that the probability distributions of the arrival and service times are exponential. This study addresses this gap by adopting a G/G/1 queue model, where both arrival and service times follow Pareto distributions, to compare the service efficiency of UBL and HBL in Pakistan. By first identifying the appropriate distributions through empirical data analysis and then applying queueing theory, we aim to provide actionable insights for reducing queue lengths, enhancing service delivery, and improving customer experiences in the banking sector. This approach not only fills a methodological void but also offers a scalable framework for other service industries facing similar challenges.

2. METHODOLOGY

2.1 Queueing Theory and Kendall's Notation

Queueing systems are formalized using Kendall's notation, introduced by Kendall (1953), which provides a standardized framework to describe queue characteristics. The notation is expressed as A/B/C, where:

A: Represents the probability distribution of inter-arrival times (the time between consecutive customer arrivals).
B: Represents the probability distribution of service times (the time required to serve a customer).
C: Denotes the number of servers in the system.

This study focuses on the G/G/1 queue model, where both inter-arrival and service times follow general distributions, specifically, Pareto distributions and the system operates with a single server. The choice of the G/G/1 model is driven by empirical evidence that customer arrival and service times in the studied banks exhibit heavy-tailed characteristics, best modeled by Pareto distributions. Table 1 provides an overview of common queue models, their notations, and descriptions, adapted from Kendall (1953) and Cohen (1982).

Table 1
Notation and Description of Some Queue Models

Queue Model Notation	A	B	Description
M/M/1	Exponential	Exponential	Single server with exponential inter-arrival and service times.
M/G/1	Exponential	General	Single server with exponential inter-arrival times and general service times.
G/M/1	General	Exponential	Single server with general inter-arrival times and exponential service times.
G/G/1	General	General	Single server with general distributions for both inter-arrival and service times.
$E_k/G/1$	Erlang	General	Single server with Erlang-distributed inter-arrival times and general service times.

Consider $A_i\,(i = 1,2,\dots,n)$ is the inter-arrival time of customers entered into the system with an inter-arrival rate λ i.e. $\lambda = \frac{1}{E(A)}$, and S_i is the service time avail by the ith customer in the system with service rate μ i.e. $\mu = \frac{1}{E(S)}$. We also assume that Wrepresents the customer's time spent in the system, D is the departure time of the customer and I is the idle time of the customers.

For a single server queue model with general distribution of the arrival and service times i.e. Pareto distribution and is represented as G/G/1. In our study, we assume that A_i comes from the Pareto distribution with parameter α_1, also S_i assumed to be Pareto distributed with parameterα_2. Consider C_A and C_S are the coefficients of variation of inter-arrival and service times, respectively. Moreover, some basic formulas of the considered G/G/1 queue model are given by.

$$\lambda = \text{Arrival rate} = \frac{1}{E(T)} = \frac{\alpha_1 - 1}{\alpha_1}, \alpha_1 > 1 \quad (1)$$

$$\mu = \text{Service rate } = \frac{\alpha_2 - 1}{\alpha_2},\ \alpha_2 > 1 \quad (2)$$

$$C_a^2 = \frac{var(T)}{\big(E(T)\big)^2} = \lambda^2 var(T) = \frac{1}{\alpha_1(\alpha_1 - 2)}, \alpha_1 > 2 \quad (3)$$

$$C_s^2 = \frac{var(S)}{\big(E(S)\big)^2} = \mu^2 var(S) = \frac{1}{\alpha_2(\alpha_2 - 2)}, \alpha_2 > 2 \quad (4)$$

2.2 Performance Measures of G/G/1 Model

In this section, we proposed some performance measures of the G/G/1 queue model.

2.2.1 Probability of no Customer in the System

The probability of no customer in the system means that when the system is empty, then we will use the following expression.

$$p_0 = 1 - \rho = 1 - \frac{\lambda}{\mu}, \quad (5)$$

For our proposed G/G/1 using Eq. (1) and Eq. (2), Eq. (5) can be written as

$$p_0 = \frac{\alpha_2 - \alpha_1}{\alpha_1(\alpha_2 - 1)}, \text{with } \alpha_2 > \alpha_1 \text{ and } \alpha_2 > 1 \quad (6)$$

Similarly, the probability of exactly *n* customers in the system is defined by

$$p_n = \left(\frac{\lambda}{\mu}\right) p_0 \quad (7)$$

For the G/G/1 queue model, using Eq. (1), Eq. (2), Eq. (6) in Eq. (7), we have

$$p_n = \frac{\alpha_2(\alpha_1 - 1)(\alpha_2 - \alpha_1)}{[\alpha_1(\alpha_2 - 1)]^2}, \alpha_2 > \alpha_1 \text{ and } \alpha_1, \alpha_2 > 1 \quad (8)$$

2.2.2 Average Waiting Time

The average waiting time of G/G/1 queue model is approximated by the following equation

$$W_q = \frac{p_0}{(1-\rho)^2 \mu}\left[\frac{C_A^2 + C_S^2}{2}\right] \text{ or equivalently } W_q = \frac{C_A^2 + C_S^2}{2(\mu - \lambda)}, \quad (9)$$

where W_q represents the average waiting time of a customer to avail the service, the waiting time depends on the arrival and service rate. So, using Eq. (1), Eq. (2). Eq. (3) and Eq. (4) in Eq. (9), we have

$$W_q = \frac{\alpha_1(\alpha_1 - 2) + \alpha_2(\alpha_2 - 2)}{2(\alpha_1 - 2)(\alpha_2 - 2)(\alpha_2 - \alpha_1)}. \quad (10)$$

2.2.3 Average Number of Waiting Customers

The average of customers who are waiting to avail of the service for the G/G/1 queue model (James et al., 2009) is defined by

$$L_q = \lambda W_q, \tag{11}$$

where L_q represents the average number of waiting customers to utilize the service; the number of waiting customers depends on the service and arrival rate. The service time effects inversely while the arrival time has a direct effect on the waiting time. Using Eq. (1) and Eq. (10) in Eq. (11), we have

$$L_q = \frac{(\alpha_1 - 1)[\alpha_1(\alpha_1 - 2) + \alpha_2(\alpha_2 - 2)]}{2\alpha_1(\alpha_1 - 2)(\alpha_2 - 2)(\alpha_2 - \alpha_1)}. \tag{12}$$

2.2.4 Average Utilization Server

In queuing theory, it is used to optimize process performance of the queuing system. The average utilization server is the proportion of time of the server that is busy in performing a specific task to the total available time. It is denoted by ρ and using Eq. (1) and Eq. (2), we have

$$\rho = \frac{\lambda}{\mu} = \frac{\alpha_2(\alpha_1 - 1)}{\alpha_1(\alpha_2 - 1)}. \tag{13}$$

The utilization factor is also used to measure the stability of the queuing system i.e. for a stable system, if $\rho < 1$ (Ravindran, 2008). If $\rho > 1$, then the queuing system is unstable. It is also used to compare the service stability of two or more queuing systems, a queue system is said to be more stable if it has a smaller utilization factor as compared to the utilization of other systems.

2.2.5 Average Number of Customers in the System

The average number of customers in the system means the total number of users in the system. It is used to understand the behavior of the system. The general definition of G/G/1 queue model was given by James et al. (2016) with the average number of customers is computed as

$$L = \lambda W,$$

where L represents the number of customers in the system. For our considered model, it is defined as

$$L = \left(\frac{\alpha_1 - 1}{\alpha_1}\right)\left(\frac{(\alpha_2 - 1)\{\alpha_1(\alpha_1 - 2) + \alpha_2(\alpha_2 - 2)\} + 2\alpha_2(\alpha_1 - 2)(\alpha_2 - 2)(\alpha_2 - \alpha_1)}{2(\alpha_1 - 2)(\alpha_2 - 1)(\alpha_2 - 2)(\alpha_2 - \alpha_1)}\right). \tag{14}$$

2.2.6 Average Waiting Time in the System

The average waiting time in the system means the total time waiting in the system, which is denoted by W and is defined as

$$W = W_q + \frac{1}{\mu} \tag{15}$$

Using Eq. (2) and Eq. (10) in Eq. (15), we have

$$W = \frac{\alpha_1(\alpha_1 - 2) + \alpha_2(\alpha_2 - 2)}{2(\alpha_1 - 2)(\alpha_2 - 2)(\alpha_2 - \alpha_1)} + \frac{\alpha_2}{(\alpha_2 - 1)}$$

Or equivalently

$$W = \frac{(\alpha_2 - 1)\{\alpha_1(\alpha_1 - 2) + \alpha_2(\alpha_2 - 2)\} + 2\alpha_2(\alpha_1 - 2)(\alpha_2 - 2)(\alpha_2 - \alpha_1)}{2(\alpha_1 - 2)(\alpha_2 - 1)(\alpha_2 - 2)(\alpha_2 - \alpha_1)}. \quad (16)$$

3. APPLICATION

3.1 Data Collection

Data were collected from 200 customers at UBL and HBL branches in Pakistan during peak operational hours to ensure representative samples. Inter-arrival and service times were recorded meticulously using stopwatch measurements and verified through bank transaction logs. The whole process is shown in Figure 1. This figure highlights the system's structure, with annotations for customer arrival, queuing, service, and departure processes, provides a visual aid to the mathematical framework presented in Equations (1)–(16). Statistical tests, including goodness-of-fit analyses based on Cramar Von-Mises test, confirmed that both arrival and service times follow Pareto distributions, characterized by heavy-tailed behavior, thus validating the use of the G/G/1 model for this study.

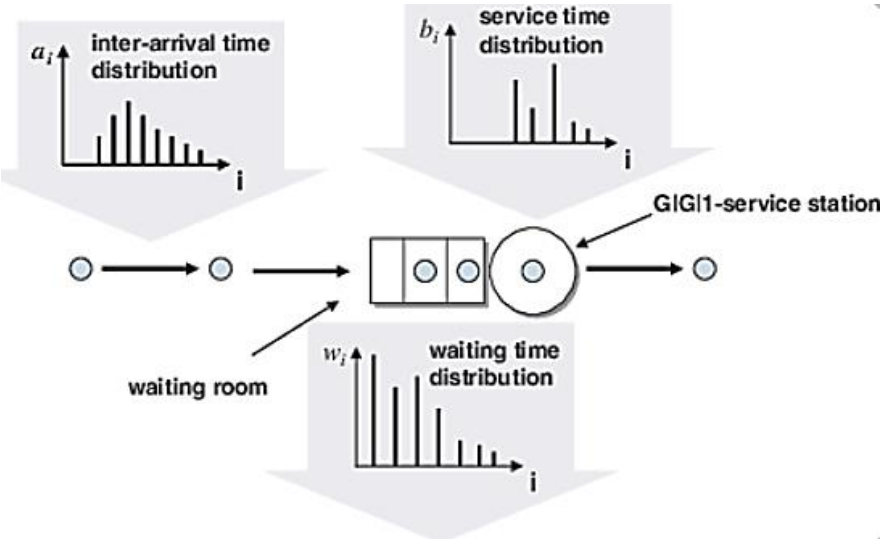

Figure 1: Flow Chart of the G/G/1 Queue Model

3.2 G/G/1 Performance Measures of HBL and UBL

Table 2
Descriptive statistics of HBL and UBL arrival and service time distributions

	HBL		UBL	
	Arrival Time (minutes)	**Service Time (minutes)**	**Arrival Time (minutes)**	**Service Time (minutes)**
Mean	1.6667	1.4830	1.7042	1.2832
Variance	3.3333	1.3860	4.0576	0.5072
CV	1.0954	0.7938	1.1819	0.5549

Table 2 shows the descriptive statistics of customers' arrival and service times in minutes who avail the services of the HBL and UBL. We observed that the HBL average service time of customers is 1.4830 minutes, while the UBL average service time of customers is 1.2832 minutes. This shows that the average service time of UBL is better than that of HBL. These results showed that the UBL has an efficient service than the HBL due to its smaller average service time. The average arrival time of customers of HBL is 1.6667 minutes and the UBL arrival time of customers is 1.7042 minutes. This indicates that the arrival time of HBL customers is more rapid than that of UBL. The variations of the HBL arrival time of customers and the UBL arrival time of customers are 3.3333, and 4.057, respectively. While the variations in service times of HBL and UBL are 1.386 0.5072 respectively. So, according to the variation in bank service times, we observed that UBL has a smaller variation in service times as compared to UBL. So, UBL is best. Similarly, the CV of the HBL service time of customers is 0.7938 and the CV of UBL is 0.5549 which is smaller than HBL. Based on the variation in service times of the two banks, we observed that the service of UBL is more consistent than that of HBL.

Table 3
Performance Measures of G/G/1 Queue Model to Compare the Service Performance of HBL and UBL

Measures	HBL	UBL
Arrival Rate	0.4001	0.4132
Service Rate	0.3257	0.2207
Average server Utilization. (ρ)	0.8896	0.7529
Average number of customers in the system. (L)	5.3478	2.3852
Average time (minutes) of customers in the system. (W)	8.9155	4.0649
Average number of customers waiting in the system (L_q).	4.5482	1.6322
Average waiting time (minutes) of customers in the system (W_q)	7.4324	2.7816

Table 3 presents a comprehensive comparison of the G/G/1 queue model performance metrics for HBL and UBL, shedding light on their service efficiency. The arrival rate for HBL is 0.4001 customers per minute, slightly lower than UBL's 0.4132, indicating that both banks experience comparable customer inflow, with UBL seeing marginally more frequent arrivals. However, the service rate reveals a stark contrast: HBL's service rate is 0.3257 customers per minute, while UBL's is 0.2207. Since the service rate is the reciprocal of the average service time, UBL's lower rate corresponds to a shorter service time (1.2832 minutes vs. 1.4830 minutes for HBL), demonstrating that UBL processes customers more speedily, enhancing throughput efficiency. Server utilization, a critical indicator of system stability, further underscores UBL's advantage. HBL's utilization is 0.8896, suggesting that the server is busy nearly 89% of the time, operating close to capacity and potentially prone to bottlenecks during peak periods. In contrast, UBL's utilization of 0.7529 indicates a more balanced system, with the server busy approximately 75% of the time, allowing greater resilience to demand fluctuations (Ravindran, 2008).

The average number of customers in the system is significantly higher for HBL (5.3478) compared to UBL (2.3852), a difference of approximately 124%, reflecting HBL's longer queues and higher overcrowding. This is validated by the average time customers spend in the system, which is 8.9155 minutes for HBL, more than double UBL's 4.0649 minutes, a reduction of about 54%. This substantial gap highlights UBL's ability to expedite the entire customer journey, from arrival to service completion. The average number of customers waiting in the queue further amplifies UBL's efficiency: HBL's queue averages 4.5482 customers, while UBL's is only 1.6322, a reduction of roughly 64%. This shorter queue length translates to a significantly lower average waiting time in the queue for UBL (2.7816 minutes) compared to HBL (7.4324 minutes), a decrease of

approximately 63%. These metrics collectively indicate that UBL's customers experience less delay and overcrowding, contributing to a more seamless banking experience. The consistency of UBL's performance across these measures suggests operational advantages, such as optimized workflows, better staff training, or technological enhancements, which HBL could emulate to improve service delivery (Sundari & Srinivasan, 2011). Overall, the G/G/1 model results affirm that UBL provides a superior service experience, characterized by faster processing, shorter queues, and greater system stability, positioning it as the more efficient bank compared to HBL.

4. DISCUSSION

The application of the G/G/1 queue model with Pareto distributions provides a nuanced understanding of bank service dynamics. Unlike previous studies using exponential-based models (e.g., M/M/1 by Famule and Daisi, 2010; M/M/C by Sundari and Srinivasan, 2011), our approach accounts for the heavy-tailed nature of arrival and service times, which is more realistic for banking environments. The Pareto distribution captures the variability in customer behavior, such as sporadic arrivals or prolonged service interactions, making the G/G/1 model particularly suitable.

UBL's advantages may stem from operational efficiencies, such as streamlined processes or better staff training. HBL, despite a comparable arrival rate, struggles with longer service times and higher variability, leading to extended queues. These outputs can guide bank managers in implementing targeted improvements, such as increasing service counters or adopting technology to expedite transactions (Sundari and Srinivasan, 2011).

The study's methodology is extensible to other sectors, such as healthcare or hospitality, where queue management is critical. Future research could explore multi-server models (e.g., G/G/C) or incorporate additional variables, such as customer type or service complexity, to further refine performance evaluations.

5. CONCLUSION

This study demonstrates the utility of the G/G/1 queue model with Pareto-distributed arrival and service times in evaluating bank service performance. By analyzing data from UBL and HBL, we find that UBL provides superior service, characterized by shorter waiting times, fewer customers in the queue, and greater system stability. These results underscore the importance of tailoring queueing models to the empirical distributions of customer data. The findings have practical implications for bank management. Reducing service times and queue lengths can enhance customer satisfaction and operational efficiency. The methodology can be extended to other sectors, such as healthcare or hospitality, where queue management is critical. Future research could explore multi-server models or incorporate additional factors like customer priority or service type.

REFERENCES

1. Amin, M., Akbar, A. and Ullah M.A. (2016). Quality variation minimizer: A new approach for quality improvement in textile industry. *Pakistan Journal of Scientific and Industrial Research Series A: Physical Sciences*, 59(2), 109-113.
2. Lewis, B.R. (1991). Service quality: An international comparison of bank customers' expectations and perceptions. *Journal of Marketing Management*, 7, 47-62.
3. Arasil, H., Katircioglu, S.T. and Mehtab-Smadi, S. (2005). A comparison of service quality in the banking industry: Some evidence from Turkish and Greek speaking areas in Cyprus. *International Journal of Bank Marketing*, 23(7), 508-526.
4. Oyetunji, E.O., Baguri, B.B. and Otis, A.E. (2009). Comparison of service quality of two commercial banks in upper east region of Ghana. *Proceedings of the 2014 International Conference on Industrial Engineering and Operations Management Bali, Indonesia,* January 7-9, 2014, 782-792.
5. Banerjee, N. and Sah, S. (2012). Comparative study of customer's perceptions of service quality dimensions between public and private banks in India. *International Journal of Business Administration*, 3(5), 33-44.
6. Ahmad, A., Kashif-ur-Rehman, and Saif, M.I. (2010). Islamic banking experience of Pakistan: Comparison between Islamic and conventional banks. *International Journal of Business and Management*, 5(2), 137-143.
7. Famule and Daisi, F. (2010). Analysis of M/M/1 queuing model with application to waiting time of customers in banks. *Global Journal of Science and Technology*, 10(13), 28-34.
8. Sundari, S.M. and Srinivasan, S. (2011). M/M/C queuing model for waiting time of customers in bank sectors. *International Journal of Mathematical Sciences and Applications*, 1(3), 1569-1575.
9. Kendall, D.G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the Imbedded markov chain. *The Annals of Mathematical Statistics,* 24(3), 338-354.
10. Cohen, J.W. (1982). *The Single Server Queue*, 2nd ed., North Holland, New York.
11. James, R., Langabeer, I.I., FHIMSS, C.M.A. (2009). *Performance Improvements in Hospitals and Health Systems*, HIMSS. U.S.A.
12. Ravindran, A.R. (2008). *Operations Research and Management Science* Handbook. CRC Press Taylor and Francis Group, New York.