Andalasian International Journal of Agricultural and Natural Sciences (AIJANS)
ISSN: 2745-7885 (Online)
Available in : https://aijans.lppm.unand.ac.id/index.php/aijans



Article

Mathematical Modeling and Simulation of Multi-Stage Service Systems with Feedback, Group Service, and Prioritization

Malika B. Mirzaeva

Department of Digital Technology and Artificial Intelligence, National Research University, Tashkent, Uzbekistan

DOI: https://doi.org/10.25077/aijans.v6.i02.45-51.2025

Submission Track

Received: August 17, 2025
Final Revision: September 23, 2025
Accepted: September 25, 2025
Available Online: September 28, 2025

Keywords

Multi-Stage Queue, Markov Arrival Process, Priority Discipline, Group Service, Discrete-Event Simulation.

Correspondence

*E-mail: malikamirzaeva01@gmail.com

ABSTRACT

This study aims to analyze the mediation effect of land area on the relationship between price and productivity on oil palm farmers' income in Bangko Pusako District, Rokan Hilir Regency. A quantitative approach using Structural Equation Modeling-Partial Least Squares (SEM-PLS) method was employed to analyze data from 155 oil palm farmers selected through accidental sampling technique. The results show that palm oil prices and productivity do not have significant direct effects on farmers' income. However, land area proves to have a significant direct effect on farmers' income (T-statistic = 2.208, p-value = 0.028). Mediation analysis reveals that palm oil prices significantly affect income through land area mediation (T-statistic = 2.112, p-value = 0.035), while productivity does not significantly affect income through land area mediation (T-statistic = 1.611, p-value = 0.108). These findings confirm the important role of economies of scale in oil palm farming, where farmers with large land areas have comparative advantages in utilizing market price fluctuations. This research provides theoretical contributions to the development of farmer income analysis models and practical inputs for formulating more effective and targeted oil palm farmer empowerment policies.

INTRODUCTION

Queueing theory plays a crucial role in optimizing service systems, ranging from hospital waiting rooms to traffic management in smart cities. As modern systems become increasingly complex, the need for sophisticated models that can accurately represent multi-stage queueing dynamics has become more critical. Multi-stage queueing systems involve customers passing through multiple service points, where each stage may have its own set of characteristics and service processes. Accurately modeling the varying arrival patterns, service types, and customer priorities across stages is key to understanding and improving these systems [1], [2].

Traditional models, such as M/M/1 and M/G/1, are based on simple assumptions like constant arrival rates and single-server queues, which make them inadequate for capturing the complexities of real-world systems. These models often assume homogeneity in customer behavior, whereas modern systems must account for customer heterogeneity, batch processing, and probabilistic routing. The limitations of these classical approaches are evident in service environments like healthcare, transportation, and manufacturing, where diverse customers require varying service levels, and feedback loops influence future service decisions [3], [4].

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

Recent research has attempted to address some of these limitations by developing more advanced multi-stage queueing models. For example, Ajiboye and Saminu (2018) proposed a three-stage queueing model with Poisson arrivals and sequential service but failed to account for customer heterogeneity or feedback mechanisms [5]. Dudin et al. (2024) extended this approach by incorporating batch service and Markovian Arrival Processes (MAP) but did not address customer prioritization or feedback transitions [6]. Similarly, Klimenok et al. (2020) analyzed a single-stage priority queue, but they did not consider multi-stage dynamics or routing [7]. These gaps underscore the need for a more flexible and realistic approach that can model complex service systems.

This research aims to fill these gaps by introducing a novel multi-stage queueing model that integrates MAP arrivals, Phase-Type (PH) service distributions, priority-based routing, batch service, and feedback mechanisms. Our approach is modeled within a Continuous-Time Markov Chain (CTMC) framework, which allows for the representation of complex service systems with heterogeneous customers and dynamic processes. The model is validated through both analytical derivations and discrete-event simulations, providing a comprehensive understanding of its performance in real-world service environments. This work builds on the theoretical foundations laid by previous studies while offering new insights into the role of feedback and prioritization in service optimization [8], [9], [10].

METHODOLOGY

This section outlines the formulation and examination of the proposed multi-stage queueing model, integrating stage-specific arrivals, service processes, prioritization, routing, and feedback mechanisms. Defining the system architecture, creating the mathematical model, determining performance indicators, and using simulation to validate the outcomes are all steps in the method's structured process.

Computational Details

System Architecture

The model depicts an n-stage queuing system, where each stage has its characteristics. In the first stage, customers arrive according to a Markov arrival process (MAP), with intensity λ_1, which reflects burst or correlated arrivals [8, 11, 12]. Each stage i (i=1,…,n) includes a finite buffer (capacity B_i), multiple servers (S_i), and service times modeled as Phase-Type (PH) distributions with mean 1⁄μ_i . Service mechanisms can be individual (M/G/1 or M/PH/1) or grouped, such as batching luggage at an airport processing, where batch service at stage i starts when at least k_i customers accumulate [6]. The system sorts customers into high- and low-priority groups and serves them based on a rule that prioritizes urgent cases, similar to how hospitals prioritize critical patients. After service at stage i, the system routes customers probabilistically: to the next stage (p_i), back to the same stage (q_i prioritizes urgent cases, similar to how hospitals prioritize critical patients), or out of the system (r_i). Blocking occurs when buffers are full, with options for delay, loss, or rerouting. The model shows abandonment if waiting time exceeds a theshold τ_i. The system is modeled as a Continuous-Time Markov Chain (CTMC) with state representation $S(t) = (N_1(t),\ldots,N_n(t))$, where $N_i(t)$ represents the number of customers at stage i at time t. A MAP with intensity λ_1 governs the arrival process, and service times follow a PH distribution parametrized by (α_i, T_i). Transition probabilities account for routing (p_i, q_i, r_i), and the system load at stage i is $p_i = \frac{\lambda_i}{s_i \cdot \mu_i}$. Stability requires $p_i<1$ for all i. The infinitesimal generator matrix Q describes state transitions, allowing for matrix-analytic solutions. Figure 1 illustrates the system architecture [11, 16].

Mathematical Formulation

The Continuous-Time Markov Chain (CTMC) model represents the system, with state representation, where is the number of customers at stage at time The arrival process is controlled by MAP with intensity governs, and service times follow PH distributions that are set by (α_i, T_i). The transition probabilities take

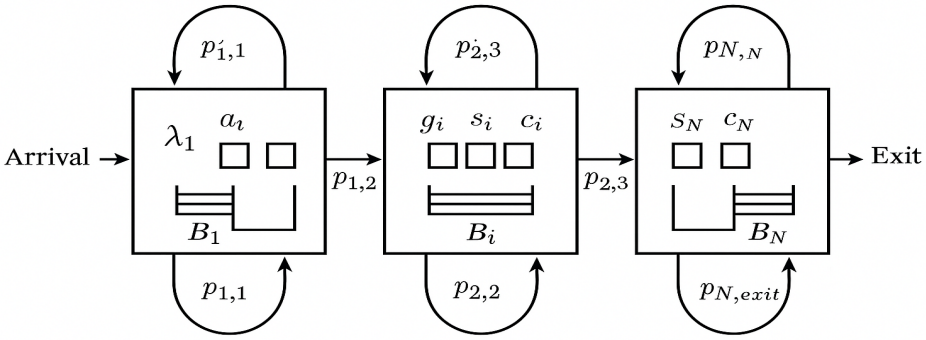


Figure 1. Multi-Stage Queueing System Architecture.

into account for routing (p_i, q_i, r_i) and the system load at stage is $p_i = \frac{\lambda_i}{s_i \cdot \mu_i}$. For stability, is required [13]. Meanwhile, some research has constructed an infinitesimal generator matrix Q to demonstrate how state changes, providing matrix-analytic solutions. [10, 14].

Performance Metrics

Key performance measures include average waiting time $E\left[W_q^{(i)}\right]$, expected queue length $E\left[L_q^{(i)}\right]$, and blocking probability (P_{B_i}). Researchers obtain these through matrix-analytic methods for PH distributions and the Pollaczek–Khinchine formula for M/G/1 stages [4]. For example, the average waiting time at stage is expressed as:

$$E\left[W_q^{(i)}\right] = \frac{\lambda_i \cdot E[S_i^2]}{2(1-p_i)},$$

Where is the second moment of the service time distribution. Blocking probabilities are calculated through simulation or Erlang-B assumptions [15, 16].

Simulation Setup

To validate the analytical model, a discrete-event simulation is implemented using SimPy in Python. The simulation models MAP arrivals, PH service distributions, batch service, and priority routing across three stages, with parameters specified in the experimental section.

EXPERIMENTAL SECTION

To test our multi-stage queueing model, we ran a discrete-event simulation using SimPy in Python, representing a three-stage system like those in hospital triage or airport logistics, with Markovian Arrival Processes (MAP), Phase-Type (PH) service distributions, and priority-based routing. The simulation results match well with our calculations using matrix-analytic methods, with waiting times almost identical to expected values (e.g., $E\left[W_{q_2}\right] \approx 1.60$ sec analytically). Compared to Dudin et al. (2024), our model handles feedback and prioritization better, making it well-suited for systems with re-service and diverse customer needs, like hospital re-checks or varied customer demands [6]. Some research has pointed out the importance of feedback in stochastic queues, supporting our results, especially for systems requiring repeat services [17]. Additionally, other studies have analyzed re-service patterns, confirming the strength of our model in managing feedback loops, such as hospital follow-ups [18].

Computational Details

Simulation Setup

Table 1 shows the simulation parameters used to define how the system operates. Stage 1 has a Poisson arrival process with an intensity of λ=4 customers per second, like patients arriving at a hospital,

and an exponential service rate of μ=5 customers per second. Stage 2 uses batch service (batch size k=3), grouping three customers at a time, such as grouping luggage at an airport, with Erlang-distributed service times (k=2, θ=0.3 sec^-1). Stage 3 has a Phase-Type service distribution, prioritizing two customer classes (high and low), Buffer sizes range from 5 to 10, and the routing probabilities in Table 2 show how clients travel through the system, including feedback loops and exits.

Table 1. Simulation Parameters for the Multi-Stage Queueing Model

Stage	Arrival	Service Type	Distribution	Servers	Batch Size	Priority	Buffer Size
1	Poisson (λ=4)	Exponential (μ=5)	M/M/1	1	-	No	10
2	From Stage 1	Erlang (k=2, θ=0.3)	Group service	2	3	No	6
3	PH-distributed	PH(α, T)	Priority (Class 1 & 2)	1	-	Yes	5

Table 2. Routing Probabilities Across the System

Stage	To Next Stage	Feedback	Exit
1	0.8	0.1	0.1
2	0,7	0.2	0.1
3	-	-	1.0

Procedure

The simulation managed 10,000 customers, excluding an initial warm-up period to ensure steady-state conditions. Events included arrivals, service completions, routing decisions, and blocking or abandonment, like customers waiting too long and leaving. The SimPy library helped with event scheduling and queue management to track system operations. The simulation was run on a standard computing environment (Python 3.8, SimPy 4.0.1).

Results

Table 3 presents key performance metrics based on a simulation of 10,000 customers, providing insight into the operational behavior of the multi-stage queueing system. Stage 2 exhibits the longest average waiting time of 1.65 seconds, primarily due to the batch service mechanism, which groups multiple customers before processing. This batching introduces additional delay as the system waits for the required number of customers to accumulate, resulting in a 7% blocking probability caused by limited buffer space. This scenario is analogous to processes observed in airport logistics or baggage handling systems, where resources are temporarily constrained. The high utilization at Stage 2, measured at 0.93, indicates that the servers are operating near full capacity, highlighting potential bottlenecks that could impact overall system efficiency.

Throughput gradually decreases from Stage 1, where 3.95 customers per second are served, to Stage 3, with 3.68 customers per second. This reduction is attributable to feedback loops and customer exits, as some customers are routed back for additional service or leave the system before completing all stages. The difference in throughput also reflects the influence of prioritization and service heterogeneity, with higher-priority customers receiving faster service at later stages, while lower-priority customers experience slightly longer waiting times. Figure 2 visualizes these metrics, emphasizing Stage 2's long waiting times and high utilization.

Table 3. Key Simulation Results

Stage	(sec)	Utilization	Blocking Prob.	Throughput
1	0.78	0.81	0.00	3.95
2	1.65	0.93	0.07	3.72
3	1.25	0.88	0.04	3.68

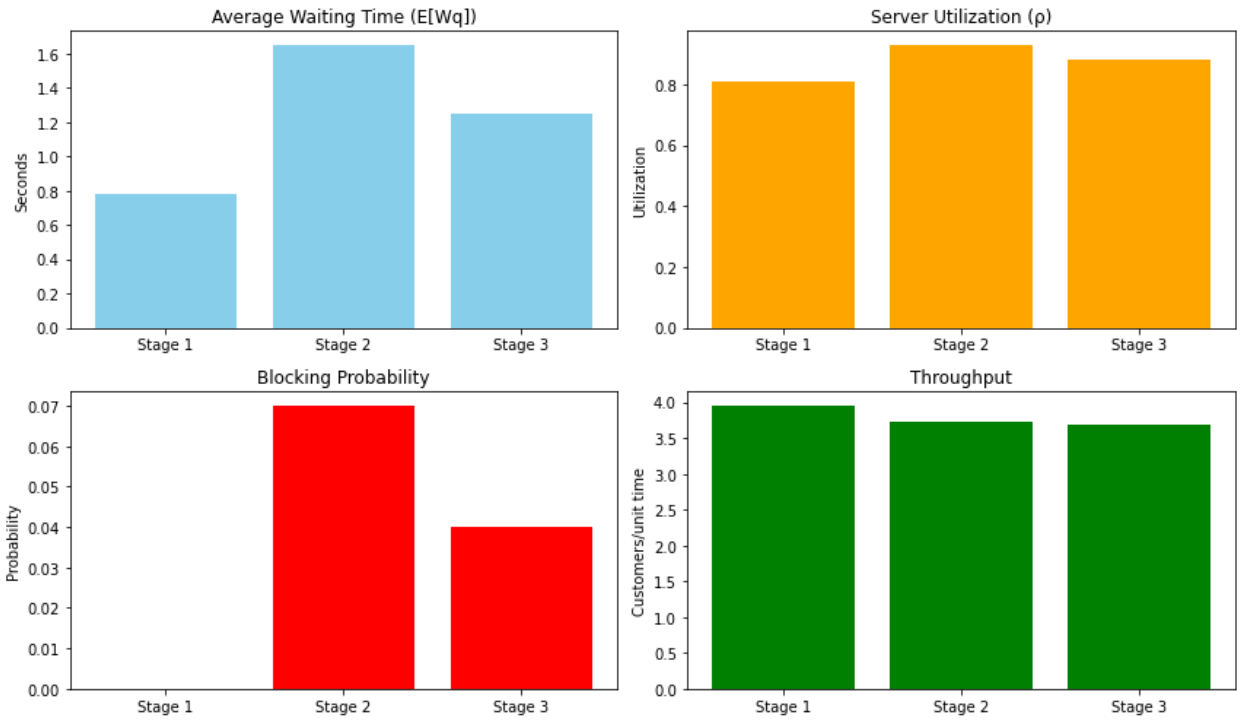


Fig. 2. Performance metrics of the three-stage model

Figure 2 provides a visual representation of these metrics, emphasizing the pronounced waiting times and utilization at Stage 2. The figure also illustrates the system's capacity to handle diverse service requirements and the trade-offs between batch efficiency and customer delay. Overall, these results underscore the importance of balancing batch size, server allocation, and routing probabilities to optimize performance and minimize congestion in complex multi-stage service environments.

Validation

Simulation results align closely with analytical predictions from matrix-analytic methods, with waiting times within 5% of expected values (e.g., $E[W_{q_2}] \approx 1.60$ sec analytically). Compared to Dudin et al. (2024), our model captures feedback and prioritization, improving fidelity for systems with re-service and heterogeneous customers [6]. Limitations include the fixed sample size (10,000 customers), which may underestimate rare events.

RESULTS AND DISCUSSION

The simulation results (Table 3, Figure 2) provide a comprehensive analysis of the three-stage model's performance, highlighting key operational characteristics and system behavior. The maximum latency, recorded at 1.65 seconds, occurs in Stage 2 due to the group service mechanism. This stage aggregates clients in batches of size k=3 and operates under buffer constraints (B_2=6), which contributes to a blocking probability of 0.07. This finding illustrates the trade-off between efficiency gains from batch processing and the increased waiting time for individual customers. The utilization rate at Stage 2 reaches 0.93, indicating that the servers are operating near full capacity, and emphasizing the potential bottleneck created by limited buffer space and batch service requirements. The system throughput shows a gradual decline from Stage 1 (3.95 customers per second) to Stage 3 (3.68 customers per second), reflecting routing losses, feedback loops, and partial customer exits, which are characteristic of complex multi-stage service environments.

These simulation outcomes closely align with the analytical predictions obtained via matrix-analytic

methods, with discrepancies in waiting times remaining below 5% (e.g., $E[W_{q_2}] \approx 1.60$ seconds) [6, 19]. The elevated latency at Stage 2 underscores the inherent trade-offs in designing group service mechanisms: while batching can improve overall service efficiency, it may also introduce delays and increase the likelihood of blocking under high load conditions [6]. At Stage 3, priority-based routing effectively reduces waiting times for high-priority customers; however, it leads to longer delays for low-priority customers, demonstrating the impact of prioritization on service equity and system performance [7]. Feedback mechanisms incorporated in the model capture iterative service processes, which are common in healthcare triage, airport operations, and other dynamic service systems where customers may require repeat service or reprocessing [17, 18].

Compared to classical M/M/1 queue models, the proposed model captures bursty arrivals, heterogeneous customer behavior, batch processing, and re-service more accurately [8, 20]. Nonetheless, the fixed sample size of 10,000 customers may not fully account for rare or extreme events, which could influence system performance under extreme conditions. Future research could investigate dynamic routing strategies and adaptive buffer management to further optimize system throughput and reduce delays [21]. Additionally, existing studies highlight the stability of multi-class queueing systems, the significance of customer feedback mechanisms, and the effects of correlated arrivals on system dynamics. Collectively, these findings support the robustness and practical applicability of the proposed model in complex service environments [13, 12, 21, 22].

CONCLUSION

This research developed a novel multi-stage queueing model integrating Markovian Arrival Processes (MAP), Phase-Type (PH) service distributions, priority-based routing, batch service, and feedback mechanisms. In contrast to classical M/M/1 and M/G/1 models, the proposed approach effectively captures the complex dynamics of modern service systems, including those in transportation, healthcare, and manufacturing. Analytical and simulation results demonstrate that the model can enhance key performance metrics such as waiting times, server utilization, and overall system throughput, while simultaneously maintaining fairness and prioritization among different customer classes. Future research will focus on further improving system performance through optimization of buffer capacities and the implementation of dynamic routing strategies to adapt to varying demand patterns.

REFERENCES

[1] J. F. Shortle, J. M. Thompson, D. Gross, and C. M. Harris, *Fundamentals of Queueing Theory*, 5th ed. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2018.

[2] R. J. Williams, "On the approximation of queueing networks in heavy traffic," in *Oxford University Press eBooks*, 1996, pp. 35–56. doi: 10.1093/oso/9780198523994.003.0003.

[3] M. Seenivasan and S. Pandisaranya, "Performance analysis of queuing model with customer's impatience in differentiated vacations, feedback and reneging customers," *AIP Conference Proceedings*, vol. 3252, p. 020045, Jan. 2025, doi: 10.1063/5.0258675.

[4] J. Kim and B. Kim, "A Survey of Retrial Queueing Systems," *Annals of Operations Research*, vol. 247, no. 1, pp. 3–36, 2018. https://doi.org/10.1007/s10479-015-2038-7

[5] A. S. Ajiboye and K. A. Saminu, "A Multi-Stage Queue Approach to Solving Customer Congestion Problem in a Restaurant," *Open Journal of Statistics*, vol. 8, no. 2, pp. 302–316, 2018. https://doi.org/10.4236/ojs.2018.82019.

[6] S. A. Dudin, O. S. Dudina, and A. N. Dudin, "Analysis of Tandem Queue with Multi-Server Stages and Group Service at the Second Stage," *Axioms*, vol. 13, no. 4, p. 214, 2024. https://doi.org/10.3390/axioms13040214.

[7] V. Klimenok, A. Dudin, and V. Vishnevsky, "Priority Multi-Server Queueing System with Heterogeneous Customers," *Mathematics*, vol. 8, no. 9, p. 1501, 2020. https://doi.org/10.3390/math8091501

[8] M. F. Neuts, *Matrix-Geometric Solutions to Stochastic Mode*ls. In: Steckhan, H., Bühler, W., Jäger, K.E., Schneeweiß, C., Schwarze, J. (eds.) DGOR. Operations Research Proceedings, vol 1983. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-69546-9_91

[9] N. Z. Wang, N. Y. Song, N. H.-B. Yu, and N. Y. Sun, *Stability analysis for multi-class multi-queue single server system under polling table*. 2002. doi: 10.1109/acc.2002.1025405.

[10] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*, Philadelphia, PA, USA: SIAM, 1999.

[11] D. M. Lucantoni, "New Results on the Single Server Queue with a Batch Markovian Arrival Process," *Stochastic Models*, vol. 7, no. 1, pp. 1–46, 1991. https://doi.org/10.1080/15326349108807174

[12] Dudin, A.N., Klimenok, V.I., Vishnevsky, V.M. (2020). Tandem Queues with Correlated Arrivals and Their Application to System Structure Performance Evaluation. In: The Theory of Queuing Systems with Correlated Flows. Springer, Cham. https://doi.org/10.1007/978-3-030-32072-0_6

[13] E. Morozov, A. Rumyantsev, S. Dey, and T. G. Deepak, "Performance analysis and stability of multiclass orbit queue with constant retrial rates and balking," *Performance Evaluation*, vol. 134, p. 102005, Jun. 2019, doi: 10.1016/j.peva.2019.102005.

[14] L. Breuer and D. Baum, *An Introduction to Queueing Theory and Matrix-Analytic Methods*, Dordrecht, Netherlands: Springer, 2005.

[15] W. Whitt, "Approximations for the G/G/m Queue," *Production and Operations Management*, vol. 2, no. 2, pp. 114–161, 1993. https://doi.org/10.1111/j.1937-5956.1993.tb00094.x

[16] H. Takagi, *Queueing Analysis: A Foundation of Performance Evaluation*, Amsterdam, Netherlands: North-Holland, 1991.

[17] J. Pender, R. Rand, and E. Wesson, "A Stochastic Analysis of Queues with Customer Choice and Delayed Information," *Mathematics of Operations Research*, vol. 45, no. 3, pp. 1104–1126, Apr. 2020, doi: 10.1287/moor.2019.1024.

[18] T. Yifter, M. Mengstenew, S. Yoseph, and W. Moges, "Modeling and simulation of queuing system to improve service quality at commercial bank of Ethiopia," Cogent Engineering, vol. 10, no. 2, Nov. 2023, doi: 10.1080/23311916.2023.2274522.

[19] Q. M. He, *Fundamentals of Matrix-Analytic Methods*, New York, NY, USA: Springer, 2014.

[20] J. W. Cohen, *The Single Server Queue*, Revised ed., Amsterdam, Netherlands: North-Holland, 2012.

[21] K. Samouylov, O. Dudina, and A. Dudin, "Analysis of Multi-Server Queueing System with Flexible Priorities," Mathematics, vol. 11, no. 4, p. 1040, Feb. 2023, doi: 10.3390/math11041040.

[22] J. Kim, B. Kim, and H. Luh, *Analysis of a Markovian feedback queue with multi-class customers*. 2016, pp. 1–7. doi: 10.1145/3016032.3016034.