10.48047/jocaaa.2020.28.6.12

Application of Queuing Theory for Optimizing Customer Waiting Time and Service Efficiency in Banking Systems: A Comprehensive Review

Poonam Chawla
Assistant Professor
Department of Mathematics
Mata Gujri College,
Fatehgarh Sahib, Punjab.

Received: 05.04.2020 Revised: 15.05.2020 Accepted: 05.06.2020

Abstract:
This paper reviews the application of queuing theory in banking operations, focusing on its role in reducing customer waiting time and improving service efficiency. Through examination of multiple real-world studies and analytical models, it highlights how queuing models such as M/M/1 and M/M/s assist in evaluating system performance, predicting congestion, and determining optimal server allocation. The findings show that effective queue management enhances customer satisfaction and strengthens operational productivity within banking environments.

Keywords: Queuing Theory; Waiting Time; Banking Service; Customer Satisfaction; M/M/1 Model.

Introduction

Efficient management of waiting lines has long been recognized as a critical component of service delivery in both public and private organizations. From banking halls and ATM facilities to call centers, transportation hubs, and investment-based service systems, the ability to analyze, predict and control queue behaviour significantly determines customer satisfaction, operational efficiency, and organizational performance. Queuing theory, since its origin in the work of Erlang (1909), has evolved into a strong mathematical framework used to examine arrival patterns, service mechanisms, congestion, customer impatience, and system capacity across diverse real-world applications. The series of research studies reviewed in this work demonstrate the breadth, richness, and contemporary relevance of queuing models in addressing modern service challenges.A large body of empirical literature highlights the persistent problem of long waiting times in the banking sector, particularly in developing countries. Studies such as those by *Vasumathi [23]*, *Odirichukwu[19]*, *Murthy[16]*, *Mei[15]*, *Patel[20]*, *Augustine[4]*, and *Abiodun[2]* show that banks frequently experience congestion at teller counters and ATM facilities due to imbalanced arrival–service rates, poor resource allocation, and operational inefficiencies. Despite advancements in electronic banking, any authors highlighted thatlong queues remain a prevalent issue due to behavioural, infrastructural, and cultural factors such as customers' preference for cash, poor internet connectivity, cyber-security concerns, and inadequate management of digital channels. These empirical studies consistently demonstrate that queuing theory, especially M/M/1 and M/M/s models, provides a powerful analytical tool for determining optimal service capacity, reducing customer waiting time, and improving service satisfaction.

Beyond banking environments, more advanced theoretical works deepen understanding of customer behavior within queues. Classical models incorporate balking (customers refusing to join when queues are long) and reneging (customers leaving after waiting), as explored by *Haight*, *Ancker&Gafarian*, and extended later by *Kuila[13]*, who integrated Chi-square arrival and service

10.48047/jocaaa.2020.28.6.12

distributions into balking/reneging models. Complementing these*analysis by Finch[9]* analysis of deterministic impatience in GI/M/1 queues demonstrates how finite customer patience stabilizes systems even when overloaded, yielding elegant integral-equation solutions for limiting waiting-time distributions.Modern extensions incorporate additional real-world complexities. Vacation queues, as studied in *Zhang et al [26]*, integrate balking, reneging, and server vacations simultaneously within an M/M/1/N structure, providing a matrix-analytic solution and an operational cost-optimization model. Some authors introduce reverse balking and reverse reneging**,** capturing behavioural patterns in investment and insurance industries where larger customer bases increase reliability and attract even more customers—an inversion of classical queue behavior. These models broaden the applicability of queueing theory to behavioral economics, decision science, and financial systems.Several reviewed papers also apply simulation methodologies—such as Arena**,** MATLAB**,** and discrete-event simulation—to validate analytical results and analyze ATM utilization, service time variability, optimal teller staffing, and customer flow management. Moreover, comparative studies across M/M/1, M/M/s, and M/M/1/K models demonstrate how system performance changes under different capacity and service configurations.

Together, these research contributions form a comprehensive landscape of queuing theory applications, ranging from foundational theoretical models to practical service optimization frameworks in banking, telecommunications, transportation, and investment sectors. They collectively underscore the significance of modelling customer arrival patterns, service rates, capacity limitations, behavioral responses, and system control mechanisms. This integrated understanding not only guides operational decision-making but also provides the groundwork for developing replacement-based strategies, where inefficient, overloaded, or suboptimal service mechanisms are systematically replaced or upgraded using quantitative analysis.In this context, the present review draws upon the extensive literature reviewed to examine queue dynamics, customer behaviour and service optimization using queuing models. By synthesizing analytical, empirical, and simulation-based insights, this study aims to contribute to the ongoing effort to design efficient, reliable, and customer-centric service systems that minimize waiting, enhance satisfaction, and optimize resource utilization across various sectors.

Defintions:

1. Queuing Theory: Queuing theory is a mathematical approach used to study the behaviour and performance of waiting lines, analyzing customer arrival patterns, service mechanisms, queue discipline, congestion, and system capacity to optimize resource allocation and minimize waiting time.

2. Queuing System: A queuing system is a structured service environment in which customers arrive, wait if service is not immediately available, receive service from one or more servers, and depart after service completion. A typical queueing system is represented by arrival rate (λ), service rate (μ), number of servers (s or c), system capacity (N), population size, and service discipline such as First-Come-First-Served (FCFS).

3. Balking: Balking refers to the behaviour of customers who decide not to join a queueafter observing the queue length or waiting conditions. In traditional queuing models, the probability of balking increases as congestion increases. The customer refuses service when the perceived waiting cost is greater than expected value or utility.

10.48047/jocaaa.2020.28.6.12

4. Reneging: Reneging is the process by which a customer leaves the queue after joining it, due to impatience or excessive waiting time. It occurs when actual waiting time exceeds a customer's tolerance threshold. Reneging decreases system efficiency and increases customer dissatisfaction, making it important for service optimization**.**

5. Reverse Balking: Reverse balking is the behavior in which customers are more likely to join a queue when more customers are already present**,** due to increased confidence or perceived reliability associated with large participation. Unlike classical balking, reverse balking occurs in investment-based services such as insurance, banking funds, and cooperative markets, where demand increases with popularity and social proof.

6. Reverse Reneging: Reverse reneging is a behavioural condition where customers are less likely to abandon a queue as the number of waiting customer's increases**,** reflecting increased trust in system credibility or expectation of service completion. This phenomenon contrasts with traditional reneging and commonly appears in investment or financial systems.

7. Server Vacations: Server vacations represent a queuing mechanism where the server becomes unavailable for service for a random or scheduled period once the system becomes empty. During vacation periods, arriving customers either wait or engage in balking/reneging behaviour. Vacation models improve efficiency when servers can utilize idle time for maintenance or secondary tasks.

8. M/M/1 Queue Model: An M/M/1 queue is a single-server queuing system with Poisson arrivals, exponential service times, infinite calling population, and FCFS discipline. It is widely used to analyze performance in service systems such as banks, ATMs, hospitals, and call centers.

9. M/M/c Queue Model: An M/M/c queue is a multi-server system with Poisson arrivals and exponential service times, where c parallel servers process customers simultaneously. It helps determine optimal staffing levels and capacity requirements in high-volume service environments.

10. M/M/1/N Queue Model: The M/M/1/N queue is a finite-capacity system with a single server and maximum capacity N. Customers arriving when the system is full are blocked or lost. This model is suitable for telecommunication channels, ATM service limits, and restricted space operations.

11. Customer Impatience: Customer impatience is the tendency of customers to abandon or avoid queues when expected waiting time exceeds their acceptable limit. Impatience behavior includes both balking and reneging and directly affects queue performance, service utilization, and customer satisfaction.

12. Waiting Time: Waiting time refers to the duration a customer spends in the queue before receiving service. It is a key performance indicator of queuing systems and strongly influences perceived service quality and organizational efficiency.

13. System Utilization: System utilization is the ratio of service demand to service capacity, it measures the proportion of time the server is busy, determining system efficiency, congestion level, and resource planning requirements.

14. Blocking Probability: Itis the likelihood that an arriving customer is denied entry into the system because capacity is full. It affects customer loss rate and system throughput in finite queue models.

10.48047/jocaaa.2020.28.6.12

Fuzzy Numbers: Triangular fuzzy representations for both arrival ($\tilde{\lambda}$) and service ($\tilde{\mu}$) rates.
α-Cut Method: Used to derive crisp equivalents for fuzzy parameters.
Simulation via Arena Software: To validate analytical results and model dynamic real-life scenarios.

Literature Review

In 1960 [9]Finch investigated a GI/M/1 queueing system with a particular feature of customers with deterministic impatience, meaning they were willing to wait no more than a fixed time W before abandoning the system. The paper aimed to formulate the waiting-time dynamics when customers have a strict patience limit, model the sequence of waiting times as a Markov chain, derived the limiting waiting-time distribution as a solution to an integral equation. To solve the integral equation through Laplace-transform methods, to explore connections to the classical GI/M/1 queue without impatience and to extend the analysis to multi-server systems. The study provided a full worked example for Erlang-k inter-arrival times with solution expressed in explicit exponential components and reduced to known Poisson results when k=1. The study examined the case when patience became infinite if traffic intensity $\rho<1$ it was shown that the standard GI/M/1 limiting distribution existed and the limiting finite-W solution converged heuristically. If $\rho>1$the GI/M/1 distribution without impatience did not exist, but the impatience model did have a limiting distribution and the solution connected to the Pollaczek–Khinchine formula for the dual queue. The author extended the model to N-server systems, though only heuristically, waiting time evolution followed the same recursion but with service rate Nμ customers waited only if at least N customers were in system. The waiting-time distribution conditioned on waiting followed a similar form to the single-server case. Generalizing impatience models beyond Poisson arrival processes, providing a rigorous mathematical solutionusing Laplace transforms and Markov chains, connecting finite patience models to classical queuing results in a unified manner, offering insights into overloaded queues were some strengths of the study. Limited accessibility, multi-server results were not fully developed and real-world impatience might not be deterministic were some limitations. The study successfully derived the limiting waiting-time distribution, demonstrated theoretical connections to classical models, and extended the framework to multi-server systems.Nosek, and Wilson[18] in 2001provided a foundational review of queuing theory, its terminologyand its practical applicationsto service industries, with a particular focus on pharmacy practice. The authors emphasized that as waiting in line was a universal human experience and that queuing theory, as a branch of operations management, offered systematic methods for analyzing and optimizing waiting processes. The main purpose of the study was to explain the basic concepts and terminology of queuing systems, explore the relationship between queuing management and customer satisfaction, discuss real-world applicationsof queuing theory across various service sectors and to highlight that a pharmacy operations could utilize queuing models and technologies to improve service delivery, reduce waiting time and enhance patient satisfaction. It was asserted that in pharmacies particularly those in high-volume outpatient settings queuing systems significantly influence patient satisfaction, staff morale and overall operational efficiency. The article reviewed successful applications of queuing theory in multiple sectors as retail and Telemarketing used queuing optimization to reduce staffing costs by $500,000 annually. Departments of Motor Vehicles in Virginia and Arizona employed queuing systems to eliminate long lines and improve satisfaction. Virginia Mason Medical Center (Seattle)reduced triage time from 45 to 15 minutes and patient appointment waits from 42 to 13 days through queuing-based workflow redesign. Queuing models had been applied to optimize branch layouts and minimize both actualand perceived wait times. The review showed that pharmacy operations face increasing workloads and complex workflows. High prescription volumes and multiple service points often resulted in bottlenecks. Traditional methods

10.48047/jocaaa.2020.28.6.12

offer psychological distraction but fail to address systemic inefficiencies**.** The study emphasized use of Automated Queuing Technology (AQT). It was shown that effective queue managementcould increase throughput without hiring additional staff, enhance patient satisfaction**,** thus promoting loyalty and compliance and provide real-time operational data for continuous improvement. Not including quantitative modeling or empirical validation, not deriving mathematical equations of queuing models and behavioural and psychological dynamic**s** were not modelled. It was concluded that queuing theory is a powerful yet underutilized management tool in pharmacy operations. Zhang et al[26] in 2005 presented an analytical investigation of an M/M/1/N queuing systemthat simultaneously incorporated balking**,** reneging and server vacations three realistic behavioural and operational features commonly found in service systems. The objectives were to develop the steady-state probability equationsof this combined queuing model, derive a matrix-form solution using Markov process techniques, compute performance measures, formulate a cost optimization model and to provide numerical simulations. The study analyzed a Poisson arrival queue M/M/1/N under the assumptions that when an arriving customer sees n customers ahead, they join with probability b_n, and balk with probability $1 - b_n$ where $b_0 = 1$and b_n decreases with n. Customers who join may leave the queue after a random waiting time reneging time was exponentially distributed with rate α. Service times were exponential with rate μ and FCFS discipline was followed. When the system became empty, the server immediately began a sequence of vacations with vacation time exponential with rate $\eta$and if the server returned and system was still empty then another vacation began. The study proposed a cost function that incorporated server cost when busy (C_1), server cost during vacations (C_2), waiting cost of customers (C_3) and loss cost from balking/reneging (C_4). It was found that as arrival rate λ increased server busy probability also increased with increased waiting, increased customer loss and optimal μ decreased slightly then increased. As vacation rate η increased server vacations became shorter, system performance improved and optimal μ increased. As reneging rate α increased more customers left early, queue length decreased and system costs reduced. Strengths of the study included novel integration, rigorous mathematical treatment and showing how operational parameters affect total cost. Limitations were assuming balking probabilities, restricting vacation distribution to exponential and considering finite buffer only. Thus the study made a valuable contribution by constructing a comprehensive queuing model that was applicable to call centers, hospitals, manufacturing, telecommunications and customer-service units.Vasumathi and Dhanavanthan[23]in 2010 addressed one of the most practical operational challenges in the banking sector queuing ofcustomer at Automated Teller Machines (ATMs). The primary aim of this study was to apply simulation techniques in conjunction with queuing theory to analyze ATM utilization, waiting time and service efficiency across three ATMs State Bank of India (SBI)**,** Indian Bank**,** and Centurion Bank located at the Vellore Institute of Technology (VIT). The authors designed and implemented a simulation-based queuing model for the three ATMs located within VIT and collected data for two months of observation covering peak hours, off-peak (free) hours and weekends including the key variables customer arrival times, service duration, idle times of the ATMs and customer waiting behaviour, including balking, reneging and jockeying. The empirical and simulated data were compared for each ATM (Indian Bank, SBI, Centurion Bank) under varying conditions (weekdays, weekends, busy and idle periods) and it was found that SBI and Indian Bank ATMs experienced high customer traffic during peak hours, with utilization rates reaching unity (1.0) implying maximum load. Centurion Bank's ATM remained underutilized due to lower customer preference and frequent service downtime. The simulation model and analytical queuing model produced slightly differing waiting times (Ws, Wq) due to limited data, yet showed consistent relative performance trends. SBI ATM service emerged as the most reliable and preferred, with the lowest average waiting time and higher customer satisfaction levels. Integration of theory and simulation,

10.48047/jocaaa.2020.28.6.12

collection of data from actual ATM, decision-oriented modelling and evaluating across multiple banks were some strengths of the study. Using small sample size, theoretically defining waiting and service costs, not accounting for technical errors and using short observation period were some limitations of the study. It was recommended that installing additional ATMs, especially SBI machines in high-traffic areas could reduce congestion. By applying this model, banks can minimize customer waiting times, reduce idle time of ATMs, and enhance service efficiency.Al-Jumaily and Al-Jobori[3]in 2010 presented an innovative approach to manage customer queues in banks through an intelligent, automated queuing system. The primary objective of the study was to design a dynamic queue management systemcapable of analyzing real-time queue status and switching between multiple scheduling algorithms to achieve the best possible average waiting time for customers. The authors proposed an Automatic Queuing System (AQS)architecture composed of five interconnected components. Where first would be customer area where customers select the type of service they require through a ticket dispenserwith multiple buttons for example deposits, withdrawals, account opening. Second would queuing area where the system queues customers virtually. Third would be testing area to test the results of available algorithms. Fourth would be scheduling algorithms database to

stores various scheduling algorithms and their performance results and the last would be service area where customers were served according to their service type which varies in duration. The authors also executed two case studies case 1 consists of sample size of 20 customers with fixed transaction time, account opening time and balancing inquiry time and case 2 was consisted of 12 customers with varying arrival and service times. In both cases, average waiting time (AWT)was calculated using the standard queuing formulae. It was found that for case 1 ordinary system average waiting time was 28.45 units and for proposed automatic system it was 24.95 units. For case 2 average waiting times for ordinary FCFS system was 32.75units and for proposed automatic system it was 31.08 units. Integrating multiple scheduling algorithms, empirical testing through simulated customer arrival and service patterns, designing specifically forbanking operations and ensuring continuous system improvement by automatic testing area were some strengths of the study. Using limited algorithm set, simplifying simulation, using small dataset and prioritizing performance were some limitations of the study. Study proposed to include more scheduling algorithmsin the database (e.g., Priority Queue, Round Robin), to incorporate additional performance metrics, to apply real-time implementation in actual banking environments to validate simulation results and to develop an AI-driven testing area. The paper successfully introduced an Automatic Queuing Model that dynamically managed banking queues by alternating between scheduling algorithms to minimize customer waiting times.In 2010 Xiao and Zhang[25]focused on the application of queuing theory to analyze and optimize bank service systemsin order to reduce customer waiting time and improve service efficiency. The study applied mathematical modelling and queuing analysis to determine the best queuing structure(single queue vs. multiple queues) and the optimal service ratefor balancing operational costs and customer satisfaction. The proposed model was based on $M/M/s/\infty$ queuing systems, representing multiple servers, exponential service and inter-arrival times and unlimited queue capacity. The study first investigated whether it was more efficient to have one common queue serving multiple tellers (M/M/s), or separate queues for each teller ($s \times M/M/1$). For two separate queues, each teller received 16 arrivals per hour and had service rate of 20 customers per hour. It was found that average waiting time was .09 units, total time in system was .14 units and average queue length was 2.84 customers. At s=5, the waiting probability drops below 5%, suggesting that five service windowsensured an acceptable service level, the utilization rate of servers at this point was 32% busy. It was found that single queues outperform multiple queues in both fairness and service quality, the number of tellers (service windows) directly affected customer satisfaction and system performance. There existed an

10.48047/jocaaa.2020.28.6.12

optimal service rate that minimized total system costs and queue modelingallowed banks to forecast customer behaviour and adjust resources proactively. Strengths of the study were providing a clear mathematical foundationand step-by-step derivations, applying both M/M/1and M/M/s models for comparison, including quantitative examples and decision rules for practical implementation and bridging theoretical modeling with real-world banking operations. Limitations were lacking empirical data from actual banks, assuming Poisson arrivals and exponential service times, focusing only on teller operations, ignoring ATMs and digital banking channels and not consider customer behavioural factors. The study provided a comprehensive and mathematically rigorous application of queuing theoryto the banking sector by identifying optimal configurations forqueues**,** service windows andservice rates.

Abensur[1] in 2011 presented an integrated mathematical and computational model designed to optimize the composition of **A**utomatic Teller Machines (ATMs)in bank branches. The objective of the paper was to develop a decision-support modelcombining Queuing Theory and Genetic Algorithms, to define the optimal mix of ATMs (specialist vs. multifunctional) that minimized total costs (fixed + variable)and ensured that the average waiting timein queues remain within legislative limits and acceptable customer experience levels. Field data were gathered from a major Brazilian retail bank over one month, including occupancy times for various ATM transactions (withdrawals, deposits, statements, payments, etc.), hourly transaction volumes for peak and low demand periods and fixed and variable cost data for different types of ATMs. The analysis used both high-demand and low-demand branches to assess scalability. It was shown that the GA model provided a balanced configuration, keeping waiting time within 3 minutes and queues manageable with 12 people, the Linear Programming (LP) model achieved lower cost but violated quality constraints, producing unrealistically long queues (40 min) and the Break-Even Point (BEP) model ignored service quality and capacity limitations, thus unsuitable for ATM management. Specialist ATMs for example deposit-only or inquiry-only outperformed multifunctional ATMsin high-demand settings due to faster service rates and multifunctional ATMswere more effective in low-demand branches, reducing costs and space requirements. For 2 multifunctional machines μ was1.06 min and for 2 specialist machines μ was 0.87 min, the specialist configuration consistently produced lower waiting times and shorter queuesandvalidation against real data confirmed that GA solutions closely matched observed performance. Strengths of the study were integrating queuing theory and GA, using real-world ATM usage data enhanced credibility, incorporated customer waiting-time laws into the optimization framework and offered a scalable model adaptable to branches of varying demand. The study suggested that GA-based model helped bank to balance service quality and cost efficiencydynamicallyand could serve as a strategic planning tool for branch design, equipment allocation and resource optimization. Some limitations were need of broader validation, did not account for customer impatience or balking, assumed fixed arrival and service rates and GA required numerous iterations. The study recommended including more branches and to develop adaptive models and integrating simulation toolsto complement GA optimization. This study established a methodological benchmark for quantitative decision-making in banking operations by demonstrating that Operations Research techniques could directly improve service delivery and profitability.In 2011 Cleiton and Sergio[6] presented a novel and interdisciplinary applicationof queuing theory to the management of commercial bank excess reserves. The study aimed to determine the optimal level of bank excess reserves by using Erlang B and Erlang C models tools which were conventionally employed in telecommunications and traffic engineering. Their model allowed banks to balance liquidity and profitability by ensuring cash was available for customer withdrawals while minimizing idle reserves that could otherwise generate lending revenue. The authors adapted queuing variables to

10.48047/jocaaa.2020.28.6.12

represent banking liquidity flows and to fit the banking scenario, the authors defined the total number of channels C as the number of cash units a bank can dispense, where each withdrawal uses one or more channels, The goal was to determine the minimum Cthat keeps the blocking probability below a desired threshold for example 1%. In first example optimal Excess Reserves via Erlang B it was shown that the bank could maintain liquidity with <1% probability of withdrawal blocking, while freeing up $82.5 million for lending substantially improving profitability without endangering solvency. In example 2 using the Erlang C model it was shown that the model fine-tuned liquidity allocation by incorporating acceptable delay probabilities by balancing service quality and reserve optimization. The study quantified liquidity risk probabilistically rather than deterministically, linked customer behaviour tooptimal reserve levels and bridged banking and physics-inspired stochastic modelling. Some limitations acknowledged were assuming Poisson arrivals, bank runs or correlated withdrawal behaviour violated this assumption, leading to underestimation of systemic risk and equity and solvency dynamics were ignored. The study marked a pioneering integration of queuing theory into banking economics.In 2012 Patel and Bhathawala[20] presented a detailed application of queuing theory to analyze and improve the performance of bank ATM (Automated Teller Machine) services. The study identified a common issue of customer dissatisfaction caused by long waiting lines at ATMs in modern banking, particularly during peak hours such as Monday during banking periods. The main aim of the study was to apply queuing theory models**,** specifically the M/M/1 modeland Little's Theorem**,** to determine key performance metrics customer arrival rate (λ), service rate (μ), average waiting time, queue length and system utilization. By collecting observational datafrom a single ATM located at a main branch of bank. The ATM was operational during 10 a.m. to 5 p.m. it was found that on an average of 60 customers arrived per houron Mondays, average service time was 1.5 minutes, average queue length (Lq)was 1 customer, average waiting time (Wq) was 0.5 minutes, average number of customers in system were 1.5 and probability of empty ATM was .40. The study explored customer behaviour in queues, including balking and reneging and it was also assumed that customers begin to balk if they see more than three peoplealready waiting and the maximum tolerable queue length was ten people and customers leaving the queue was calculated to be approximately 7.55%. Thesimulation model of the ATM queue system was developed a using computational tools based on ARENA simulation references. It was found that number of arrivals were 125, service time distribution was exponential with mean 1.66 cpm, interarrival time were exponential with mean 1.0 cpm, average queue length (Lq)was .44, average number of customers in system were 1.07, average waiting time (Wq) was .46 minutes, average facility utilization was .63. The analysis revealed that the ATM was effectively used but not overloaded, waiting times were minimal and if queues exceeded 3–4 customers, banks risk losing clients. The study emphasized the practical benefits of applying queuing theory in banking services by enabling forecastingof waiting times and queue lengths, assisting incapacity planning**,** improvingservice quality by reducing customer frustration and by providing a data-driven frameworkfor ATM network expansion. Some limitations of the study were focusing on a single ATM results may not generalize to other locations with different traffic patterns, ignoring variability during lunch hours or evening rushes, not incorporating priority customers**,** machine downtimes, or multiple ATM networks and assuming Poisson arrivals and exponential service may not always match real transaction behaviour. The study highlighted that mathematical modelling could be translated into practical banking strategies, improving customer experience and operational effectiveness. In 2013 Augustine[4]explored that queuing theorycould be effectively applied as a technique to address customer waiting line problems in the United Bank for Africa (UBA) Plc, Gariki Branch, Enugu as a case study. The primary objective was to determine how banks could strike a balance between the cost of customer waiting and the cost of providing additional service capacity. The study aimed to identify that number of service counters a bank should operate to

10.48047/jocaaa.2020.28.6.12

minimize cost while ensuring customer satisfaction. The study employed a descriptive survey design by using direct observation, interviews and questionnaires administered to customers and staff by randomly selecting 246 customers and the arrival and service rateswere measured empirically using time-based observations. Results interpreted that the three-server systemperformed best, achieving thelowest total cost of ₦2,069.50and minimal waiting time of 1.97 minutes, two-server systemproduced excessive queues and waiting times over 50 minutes on average, resulting in high economic loss due to customer dissatisfaction and time wastage and four-server systemimproved service slightly but at a higher operating cost of ₦2,166.91, indicating diminishing returns beyond three tellers. It was found that the current two-server setupwas not optimal and increasing to three tellers could reduce both queue length and total cost significantly, going beyond three servers lead to marginal benefits but higher fixed costs and it was show that on an average, customers spend 12 minutesin the system under a 3-server configuration, compared to nearly one hour in the current system. The study also observed behavioural issues such as favouritism, unscheduled closing of teller windows and poor communication and service discipline, which exacerbate perceived waiting times. Limitations of the study were use of single-branch study, simplified assumptions, not accounting for time-varying demand and human impatience and perception of fairness were not quantitatively modelled. The research bridged the gap between operations research theory and bank management practice, proving that mathematical models could directly inform policy and improve customer service in developing economies.Sharma et al [22]in2013 provided a detailed conceptual and mathematical analysis of queuing theory which was the scientific study of waiting lines and its application across various sectors. The study aimed at demonstrating role of queuing models to help manage congestion, optimize service capacity and improve operational efficiency in real-world systems such as telecommunication networks, traffic control, healthcare services, banking and reservation systems and computer and distributed systems. The objective was to discuss the theoretical underpinningsand mathematical modelling techniques of queuing systems and to show that these models could be effectively applied to analyze, design and control service operationsin modern technological and social environments. The study noted that the total waiting time in queues represented a massive national productivity loss, not only in developing countries like India but even in industrialized nations such as the U.S., where billions of hours were lost annually due to queuing inefficiencies. A detailed example demonstrated an M/M/1 queue in a network gateway with arrival rate (λ) 125 packets per sec, service rate (μ) 500 packets per sec and utilization (ρ) 0.25. It was calculated that average number of packets (L) were 0.33, average waiting time (W) was 2.66 milliseconds and probability of buffer overflow ($P > 13$ packets)was 1.49×10^{-8}, it was shown by example that queuing models could predict performance bottlenecks in network systems. It was implicated that overstaffing reduced waiting but increased costs and understaffing saved money but harmed satisfaction, models helped in forecasting queue lengths and delays before system implementation and queuing models were crucial for computer performance analysis, network traffic design, manufacturing logistics, and telecommunication systems. The study also discussed the flexibility–utilization trade-off, showing how higher server utilization reduced costs but degraded service flexibility and speed. The study concluded that queuing systems were vital for optimizing service operations across multiple domains. By providing mathematical models that described how queues behave, organizations could design efficient systems with optimal service levels, balance operational costs and user satisfaction and identify and eliminate performance bottlenecks.Magnus et al[14]2013 investigated that queuing theory could be effectively applied to manage time and improve service delivery efficiency in money deposit banks, with particular reference to Zenith Bank Plc., Nigeria. The primary objectives of the study were to assess the extent to which queuing theory principles could help manage customers waiting time in banks, to evaluate that time management

10.48047/jocaaa.2020.28.6.12

affected customer satisfaction and organizational productivity and to suggest strategies for improving queue efficiency and reducing customer waiting time. The study employed a descriptive survey designsupported by mathematical analysis based on queuing theory principles. It combined quantitative modelling with qualitative observation of customer service flow in Zenith Bank. A sample size of 100 respondentscomprising customers and employees was chosen using random sampling**.** The average waiting timefor customers was found to be 5.8 minutes**,** exceeding the acceptable service benchmark of 3 minutesandutilization rate (ρ)was above 0.85, indicating that servers (tellers) were almost constantly busy, with very little idle time. It was observed that peak periodsshowed queue build up due to limited counters, service time variability was high among tellers, suggesting inconsistent efficiency. Customer dissatisfactioncorrelated strongly with waiting time. The model suggested that adding one more teller could reduce waiting time by approximately 40%**,** demonstrating the predictive power of queuing theory. It was also shown that overcrowding increased stress and decreased concentration, inadequate teller allocation caused service slowdowns and lack of digital queue management systems aggravated congestion. The study suggested that queue analysis should be integrated into daily operational planning**,** banks should adopt automated queue management systems (QMS)with ticket numbers and electronic displays, employee training should emphasize speed and accuracy in handling transactions. The study concluded that queuing theory applicationwas essential for effective time management in banks and adopting scientific models like M/M/1helped banks to make data-driven decisions on staffing and service window management.In 2013 Kuila[13] provided a concise but mathematically rich investigation into balking and renegingwithin a single-server queuing system. The study focused on deriving steady-state solutions for an ordered queue where customers may balk (refuse to join the queue) or may renege (leave after joining due to impatience). The study used Chi-square queuewith Poisson-type balking probabilities that depended on the number of customers already in the system and the service rate. The study contributed incorporated two important real-world behaviours of balking and reneging crucial in real service systems such as banks, hospitals, call centers and ticket counters. The study allowed one to compute average congestion and waiting behaviour given arrival, service, and reneging parameters. The derived formulas provided impact of customer impatience as reneging probability increased queue length decreased, server utilization dropped and system became unstable for service providers.High balking rates reduced congestion but also reduced throughput**,** which might be undesirable for businesses depending on customer volume. Using Chi-square allowed modelling systems with busy arrivals, high variability and unreliable service speeds more realistic than standard Poisson/exponential models. The study extended these earlier works by combining balking, reneging and Chi-square distributions in a single analytical framework. Clear mathematical derivationof steady-state probabilities, more general modelthan the classical M/M/1 with balking/reneging and capturing both customer impatience and queue aversion were some strengths of the study. Lacking numerical examples, complexity of Chi-square distributions and lacking simulation or empirical validation were some limitations of the study. The study provided a significant theoretical contribution to queuing theory by developing a steady-state solution for a queue with balking and reneging**,** usingChi-square arrival andservice distributions.In 2013 Dhar and Rahman[7]investigated the role of queuing theory to optimize bank ATM operations by analyzing waiting lines, customer behaviour and service efficiency. The authors focused on a single ATM of BRAC Bank, located in Chittagong city, Bangladesh. The main aim was to analyze customer waiting times and ATM performance metrics including arrival rate, service rate, utilization, and queue length using Little's theoremand M/M/1 Queuing Model. The goal of study was to determine whether the existing number of ATMs in each branch could adequately handle customer demand, particularly during peak hours. The study followed a quantitative and observational approach where data were collected manually

10.48047/jocaaa.2020.28.6.12

from BRAC Bank ATM, over a one-month period, during banking hours, daily customer arrivals were recorded from Saturday to Friday over four consecutive weeks. Observations showed that customer traffic was highest on Sundays and Mondays**,** with nearly 600 customers per day**,** and lowest on Saturdays**,** averaging 260 customers per day and each customer spent approximately 2 minuteson average at the ATM. It was found that ATM was highly utilized (70%)**,** particularly on Sundays and Mondays, customers waited an average of 2.5 minutes, the probability of customer loss occurred when queues exceed three people was estimated at 10%, buffer overflow occurred when customers left due to long queues. The study highlighted that higher utilization lead to longer waiting times and larger queues but if service rate increased or additional ATMs were installed, utilization was dropped, reducing waiting times and customer impatience significantly impacted service performance. Some strengths of the study included data collection from an operational ATM enhanced practical relevance, using established mathematical models, clear operational insight and focusing on improving customer experience and bank efficiency. Limited scope, assuming steady-state conditions and ignoring daily or seasonal fluctuations and not including reneging, priority customers, or machine downtime were some limitations of the study. The findings recommended installing additional ATMs in high-traffic branches to reduce waiting time and customer dropout rates. The study effectively demonstrated that mathematical models could guide decision-making for service optimization.Odirichukwu et al [19] in 2014 examined long customer queues and waiting times in banking halls which was one of the most persistent problems in the Nigerian banking industry**.** By focusing on automation and FIFO queue principles, the study aimed to create an organized, fair, and efficient service system that ensured customers were attended to in the order they arrive. The study used both mathematical modellingand software implementation to simulate a bank queue system. The proposed system was web-based, interactive and menu-driven consisting of a touchscreen interface where customers selected services and received a queue numberelectronically, a central database that stored all customer queue data, multiple service modules (Savings, Current, Fixed Deposit) with forms recording queue numbers, timestamps, and service type and an admin login for system management and performance reporting. It was proposed that by selecting their desired service on the screen a queue number was automatically assigned to customers and stored electronically, customers enforcing aFIFO discipline performance reported and statistics were generated at the end of each day. By applying analytical methods performance metrics for a single-server (M/M/1) and multiple-server system where arrival rate (λ) was 0.5 customers per minute, service rate (μ) was 4 customers per minute and number of servers (n): 1–3 were evaluated. The queuing equations applied to evaluate that average server utilization was .125, average number in queue were .0182, average number in system were 0.143, average waiting time in queue was 2.14 min and average time in system was 17.16 min. The findings suggested that using three serverssignificantly reduced congestion and improved service efficiency. The queuing system could therefore reduce waiting time, increase customer satisfaction, optimize employee workload and minimize service bottlenecks during peak periods. Strengths of the study were addressing a real and persistent problem in Nigerian banking, applying queuing theory equations to real-world service conditions, proposing a modern, automated system, using quantitative data to support conclusion and use of web-based interface. Limitations of the study were assuming steady-state conditions, using averaged data, using only FIFO discipline and simulated, not field-testing. The study concluded that the web-based FIFO queue systemeffectively reduced waiting time, enhanced service efficiency, and improved customer satisfaction in Nigerian banks.In 2014 Sahu and Sahu[21] focused their study on applying queuing theory to improve banking service efficiency, particularly in rural banks that experience long customer waiting lines. The main objective of the study was to model and analyze customer flow using a single-channel queuing model (M/M/1)**,** determine optimal performance metricssuch as waiting time, queue length and utilization rate,

10.48047/jocaaa.2020.28.6.12

optimize the service processto reduce congestion and improve customer satisfaction. The study used four weeks of empirical datacollected from Chhattisgarh Grameen Bank, Arjuni and it was found that average arrivals per day was 59.04 customers, average arrival rate was 11.08 customers/hour, average service rate was 15 customers/hour, average number of customers in the system were 4, average number of customers in queue were 3, average waiting time was 14.08 minutes, average time in system was 18.80 minutes, utilization rate was .787, probability of waiting was .787 and probability of having more than two customers was .48. Graphs were included to show daily customer arrival and the relationship between expected customers in the system and actual observed counts. It was indicated that single-channel queuing systems, though stable under current load, lead to noticeable customer delays during peak periods, utilization rate below 1 ($\rho < 1$)confirmed that the system will not indefinitely grow, but ρ 0.8 suggested the bank operates at near-saturation and long waiting times could lead to customer dissatisfaction and potential loss of clients, especially in branches offering multiple banking services through one counter. Strengths of the study included collection of real data over four weeks ensures practical applicability, focusing scope, evaluating both operational and probabilistic parameters and offering clear improvement strategies for queue management. Some limitations of the study were assuming single counter limitation, assuming constant arrival and service rates are assumed constant, ignoring hourly variations and selecting data from four weeks may not reflect seasonal or long-term trends. It was recommended that by introducing one or more service counters would decrease utilization (ρ) and waiting time drastically, to adopt queue management software, to increase teller availability during peak hours and to enhance teller speed and efficiency to increase μ. To apply M/M/cand simulation modelsfor multi-server systems for future. The authors emphasized that queuing theory acts as a decision-making tool for management to analyze customer flow and predict operational efficiency. Murthy et al. [16]2014 applied queuing theory to analyze and optimize service efficiency at bank ATMs by focusing on the problem of long waiting linesthat cause customer dissatisfaction and loss of business in banking operations. The main objective of the study was to evaluate the arrival rate, service rate and utilization rateof an ATM system, determine average queue lengthandaverage waiting time and use queuing models to suggest improvementsthat could reduce waiting time and enhance service quality. The observational study was conducted forone-monthat a bank ATM in a city, collecting daily customer counts during 10 AM to 8 PM. The data revealed that weekend traffic was significantly higher than weekdays and Saturdays experience the heaviest load, with more than 600 customers observed. It was evaluated that arrival rate was 1 customer per minute, average service time was 1.6 minutes per customer, average queue length was 3 customers, average waiting time was 2 minutes, ATM was busy 61.7% of the time, 8 out of every 100 customers (8%)were likely to leave without service due to long queues. Theoretically waiting time was calculated as 3 minutes showing close alignment between observed and modelled results, validating the M/M/1 assumption. The study highlighted that applying queuing theory enabled banks to predicted customer flow and queue build up under varying demand levels, estimated required service capacity (number of ATMs) for optimal performance, improve customer satisfaction by minimizing waiting times and enhanced operational planning through quantitative modelling. The study provided a valuable quantitative framework for decision-making and service design, especially in urban banking contexts. Some limitations of the study were using real-world systems experience fluctuations in arrival and service rates, not validating results through software modelling and behavioural factors were treated qualitatively rather than empirically. It was recommended that by installing additional ATMsin high-traffic branches, especially on weekends, using multi-server queuing models (M/M/c)to simulate real-world banking conditions, applying time-varying models to capture fluctuating arrival rates, implementing queue management technologies, such as token systems or SMS alerts and collecting longer-term datafor more accurate trend forecasting.

10.48047/jocaaa.2020.28.6.12

In 2015 Brempong and Antwi[5] applied queuing theory to analyze customer service efficiency at the Agricultural Development Bank (ADB) branch in Koforidua, Ghana. The study aimed to assess the causes and effects of persistent customer queues at ADB, determine the performance of the current queuing system using mathematical and empirical models and to recommend practical measures by adopting multi-server (M/M/s) queuing model. The research used a case study approach, combining mathematical modelling and customer surveys by collecting data of arrival and service rates over five days during busy and non-busy periods in January–February 2008. It was found that 66.7% of customers anticipated queues before visiting the bank, average of 12 customers waited in line, with a mean waiting time of 13.2 minutes and 87.8% probability that a customer would have to wait for service. 86.7% were concerned about queues and 26.7% considered switching banks due to delays, most queues occurred on Fridays and month-end periods and despite long waits, 88.3% of respondents were satisfied with core banking services. It was suggested that by adding an additional service point there was reduction of waiting time from 13.2 minutes to 1.13 minutes and the number of customers in line from 12 to 1, the probability of waiting dropped from 87.8% to 45%, illustrating the quantitative benefit of capacity expansion and during peak hours, the queue stability condition ($\lambda < s\mu$) was violated for s = 3, but restored with s = 4 or s = 5.The paper recommended the use of ATM cards, improved IT networks and specialized service windows for faster transactions to reduce congestion. The strengths of the study included offering actionable solutions, combining mathematical modelling with real customer data for robust analysis, reliability of queuing model outcomes and balancing customer satisfaction and operational costs. The limitations were using small sample size, limited representativeness across the entire ADB network, using time-bound data and assuming homogeneous servers and infinite queue length. This study made a valuable contribution by demonstrating how queuing theory can be used as a strategic tool to enhance service delivery in the banking sector.Mutingi et al.[17]in 2015 presented a detailed simulation-based analysis of banking service operations using queuing theory principles and process improvement techniques. The study aimed to evaluate**,** simulateand optimize the performance of a customer service process in bank to minimize waiting timeand enhance service quality**.** The study emphasized the critical role of customer satisfaction in the service sector particularly in banking. The study used a real-life case study from an unnamed bank branch, operating between 08:30 and 15:30 hoursconducted over 6 days for 30 customers per service type, collected expected waiting and service times. The study found that average waiting time (teller) was 492 seconds, average expected waiting time was 242 seconds, average service time (teller) was 224 seconds and average expected service time was 71 seconds. The simulation model was built in Microsoft Excel**,** using the M/M/s queuing model with Poisson arrivals with 42 customers/hour, exponential service time with 16 customers/hour per teller and 3 tellers. It was found that average waiting time were 7.73 minutes, average total time was 11.48 minutes, average number of customers in the system was 8.03. It was shown that average waiting time decreased by 46%, demonstrating that small operational improvements could yield large gains in customer satisfaction. The study concluded that simulation modeling was an effective decision-support tool for optimizing banking operations, average waiting timeand average queue length are critical determinants of perceived service quality, the PADEI frameworkprovided a systematic approach to continuous improvement in banking service systems. By integrating process analysis, simulation, and re-engineering**,** banks could substantially improve productivity and customer satisfaction.In 2015 Abiodun and Omosule [2]presented a comparative analysisof Wema Bank and Skye Banktwo Nigerian commercial banks located in Owo Local Government Area of Ondo State. The main objective was to measure and compare queuing performance parameters between the two banks, to identify causes of customer waiting lines and service delays and to suggest optimization strategies to enhance efficiency and customer satisfaction. The study employed a comparative,

10.48047/jocaaa.2020.28.6.12

analytical design, examining queuing behaviour of customers at two banks both banks operated multi-server queuing systems. Data were collected simultaneously over three consecutive days at both banks, recording the arrival rate (λ), the service rate (μ) and the number of servers (X). Comparative analysis of both banks showed that number of servers of Wema bank were 7, average arrival rate was 169, average service rate was 148, waiting probability was .019%, while number of servers of Skye bank were 11, average arrival rate was 342, average service rate was 251, waiting probability was .000022% which implied that high waiting probability and longer queue Wema Bank and negligible waiting probability and smooth service at Skye Bank. The results showed that Skye Bankoperated more efficiently than Wema Bank due to a higher number of servers and faster service rate, the waiting probability at Wema Bank was almost 900 times higher than that of Skye Bank and as the number of servers increased, both waiting time and queue length decreased, confirming classical queuing theory predictions. The analysis proved that service optimization through proper staffing directly influenced queue performance. Some strength were demonstrating the practical difference between two real banks, employing M/M/X queuing formulations effectively and offering actionable insights for service managers in developing economies. Limitations were short data collection period limited generalizability, assuming steady-state Poisson arrivals and exponential service times, which may not reflect real-world customer behaviour, not considering customer perceptionornon-quantitative satisfaction metricsand focusing operation on teller operations only. The study concluded that queuing theory provided a realistic, effective, and practical approachto analyze and improving bank service delivery systems.Mei and Cheng[15]in 2015 presented an innovative pedagogical approach for teaching queuing theory to pre-university and early undergraduate students using simulation-based learning. The primary objective of this study was to demonstrate how simulation could be used as an effective instructional tool, showed that simulation activities could help students to model real-world systems, like ATM and cinema queues, while developing understanding of stochastic processesand probability distributions and to encourage mathematical modelling and active learning in classrooms using accessible technology tools like Microsoft Excel and MATLAB. The paper outlined two-part experimental and simulation process for teaching queuing theory. In phase 1 student recorded arrival times, service start times and service completion times of customers at real-world locations such as university ATMsandcinema ticket counters, the inter-arrival times and service times were computed using Excel spread-sheets. Empirical data allowed estimation of parameters of number of customers like sum of inter-arrival times and number of customers sum of service times. The study developed MATLAB simulation programs for both M/M/1 and M/M/c queues and generated random inter-arrival and service times using the inverse transform method. In ATM Queue (M/M/1) 75 customers were monitored over 70 minutes and arrival rate was1.087 per minute and service rate was1.239 per minute. It was found that simulation closely matched real-world data, validating the use of M/M/1 model, theoretical values were higher, likely due to ideal assumptions (steady-state) not fully met in the real scenario and variance arisen from small sample size and short observation period. In cinema ticket queue (M/M/3) model there were 39 customers across three counters with arrival rate 1.898 per minute and service rates 0.637per minute. The study concluded that simulation-based approaches offered a powerful educational toolfor teaching queuing theory because they facilitated experiential and visual learning by linking theory to real data, allowed students to grasp stochastic behaviour through trial and observation rather than pure derivation and promoted problem-solving and modelling skills, bridging the gap between abstract probability and applied systems. Some strengths of the study were moving queuing theory from a theoretical to an experiential framework, effective use of simulation software and spread-sheets for teaching, real-world data strengthening of the educational model and comprehensive scope. Limitations of the study were using small sample sizes, using simplified models and emphasizing teaching rather than

10.48047/jocaaa.2020.28.6.12

operational optimization. The study highlighted that simulations could make abstract topics accessible, relevant, and enjoyable to students at earlier educational stages.Yusuf et al[24] in 2015 explored the application of queuing theory as a performance management and service improvement tool within the banking industry**,** focusing on First Bank Plc, Gwagwalada, Abuja Branch**.** The main aim of the study was to examine whether the application of queuing theory lead to customer satisfaction and improved performance and to assess whether bank managers attended to customers on timeand manage queues efficiently. The paper used regression analysis (via SPSS) to quantify the relationship between queuing practices and customer satisfaction, offering empirical insight into the operational efficiency of Nigerian banking systems. The study used random sample of 300 customers, 250 questionnaires distributed out of which 200 valid responses were analyzed a survey research design was derived by combining both primaryand secondary data. The key hypothesis tested was application of queuing theory in First Bank lead to customer satisfaction and improved performance. The R^2 value (0.023) indicated that 2.3%of the variation in customer satisfaction can be explained by the application of queuing theory. The p-value (0.019) was less than 0.05**,** indicating that a statistically significant relationship between queuing theory implementation and customer satisfaction with a weak correlation of 0.151 was possible. The findings confirmed that effective application of queuing models could improve service flow, reduce waiting time and enhance satisfaction, but this potential remained underutilized in Nigerian banks. The study showed that human resource constraints**,** poor attitude of staff and lack of technology hindered effective queue management. It was also shown that banks often prioritize cost reduction over customer experience, resulting in insufficient tellers and inefficient service processes and psychological impact of waiting played a crucial role in satisfaction levels. The study offered some implications at managerial level such as banks should use queuing analysis to determine optimal staffing levels, especially during peak hours, proper queue discipline (FIFO) and fairness should be maintained, implementation of automated queuing systems and queue lengths, wait times and service rates should be regularly monitored. The study concludes that queuing theory played a vital yet underused management toolin Nigerian banks**.**In 2015 Jabee and Malviya[11]provided an analytical study of performance measures for the M/M/C/N queuing model**,** a finite-capacity system with Poisson arrivals (M), exponential service times (M), C parallel servers and finite system capacity N. The study aimed to drive the steady-state probabilitiesof the system, to provide performance measuresincluding waiting time, queue length, blocking probability, system size and special limiting cases and to establish connections between M/M/C/N model and well-known queuing systems such as Erlang loss model (M/M/C/C)**. A**rrivals were assumed to follow a Poisson processwith rate $\boldsymbol{\lambda}$ and service time were exponentialwith rate μ and queue discipline was FIFO (First-Come-First-Served) with C servers available, system capacity was assumed to be **N** (including customers being served). If an arriving customer found **N** customers already present, they were blocked. The steady-state probabilities for different ranges of customers were derived. As special first case system with infinite capacity ($N \rightarrow \infty$) was assumed and the model approached the classical $M/M/C/\infty$ system and customers were either served immediately or blocked and waiting probability was 0. Strengths of the paper were use of clear mathematical derivations was ideal for beginners in queuing theory and operations research and covered important special cases. Limitations were not including numeric illustrations or charts, no simulation or empirical validation was done and absence of cost-based optimization. This paper provided a concise but systematic mathematical treatment of the M/M/C/N finite capacity queue**,** deriving major performance measures and discussing important special cases. The study also listed real-world applications of the model in call centers (Erlang models), traffic congestion queues, machine repair systems, communication lines and network request handling and customer service departments.Ehsanifar et al [8] in 2017 presented an innovative approach by combining fuzzy set theory and simulation modelling to analyze performance measures

10.48047/jocaaa.2020.28.6.12

in queuing systems characterized by uncertainty and subjectivity. The study utilized Zadeh extension principle and the α-cut method to transform fuzzy queues into families of crisp systems that could be analyzed mathematically which enabled more suitable models for real-world applications such as banking or service operations. The model was applied to a bank branch, where empirical data on arrival and service times were collected over multiple days. Using the arena simulation software, the closet probability distribution was determined and the exponential nature of data was verified by simulated customer flow across five service counters. It was found that there was a strong alignment between theoretical and simulated outcomes for average queue length (Lq) and average waiting time (Wq) however differences were observed between fuzzy analytical and simulation results were attributed to the finite number of replications in the simulation compared to the theoretical assumption of infinite iterations. Using α-cut analysis the study demonstrated variation of performance indices under different uncertainty levels. The study offered a practical tool for decision-makers to predict and manage queue performance under vague conditions. The simulation results confirmed that fuzzy modelling provided more realistic and adaptable predictions compared to deterministic approaches. The findings suggested that by integrating real-time simulation and fuzzy prediction, banks could make data-driven decisions such as opening additional counters to reduce congestion and to enhance service efficiency. The study contributed in extending the classical M/M/C queue to a fuzzy environment, providing a simulation-based validation of the fuzzy model, demonstrating the applicability of fuzzy queuing theory in practical service operations and offering a computationally efficient hybrid approach suitable for real-time systems. The study provided a framework for improving service efficiency, customer experience, and decision-making in uncertain environment.

Kaur and Singh[12] in 2017 explored the application of queuing theory and simulation modellingin analyzing and improving customer service efficiency within the banking system. The paper aimed to examine existing queuing systems used in banking and other service industries, identify the tools, techniques and simulation methods applied in queuing analysis between 2000 and 2017, classify previous studies by methodology, simulation tools and performance parameters and highlight research gaps to support future modelling of customer management systems in banks. The study employed a systematic literature review and simulation-based analysis using IEEE Xplore, ScienceDirect, ACM Digital Library, Wiley InterScience and IJSTR on the research work published between 2000 and 2017. The analysis was based on queuing models, simulation approaches and application domains. It was found that simulation models effectively replicated real queuing systems, allowing researchers to test different configurations without disrupting actual operations. The majority of reviewed studies emphasized on banking applications, confirming the industry's sensitivity to customer waiting time and satisfaction using M/M/S modelMonte Carlo simulation business process reengineering (BPR) models. The authors listed tools like TORA Optimization, Arena Simulation, VB.NET-based simulation, and QSIM, highlighting their relevance for different model complexities. The study suggested that effective queue management systems significantly improved service quality, reduced operational bottlenecks and enhanced economic performance. By comparing global studies, the paper highlighted that queuing theory and simulation models consistently yielded measurable improvements in reducing waiting time and congestion, improving service scheduling and staffing efficiency and balancing service quality and cost-effectiveness. Covering 17 years of global research across major digital repositories, organizing prior works, linking theoretical models to customer service improvements in banking and employing structured search criteria and explicit inclusion/exclusion processes were some strengths of the study. Lack of empirical validation, geographical bias and descriptive findings rather than statistical were some limitations. The authors proposed extending the review to develop a simulation-based queuing model for customer management in banks. The review

10.48047/jocaaa.2020.28.6.12

provided a solid theoretical and methodological foundation for future research and practical implementations in banking operations management, reinforcing the value of queuing theory as both a scientific and managerial decision-making framework.In 2019 Haruna et al[10] presented an empirical and mathematical investigation into the effectiveness of queuing theory for optimizing customer service management within the banking sector. The study specifically aimed to develop strategies for efficient service, delivery and optimal resource allocation at the First Bank Bauchi Main Branch. The study applied M/M/4 queuing modelmodel real-world bank operations and evaluated performance indicators such as the average number of customers in the queue (Lq), system utilization, waiting time (Wq), and probability of server idleness (p_0). By adopting a quantitative and observational research approach, data were collected from direct field observation of customer arrivals and service times over two working days between 8:15 AM and 10:45 AM, covering 138 minutes per session and were processed using Microsoft Excel and SPSS for statistical validation. It was found that the average number of customers in the queue was about six, with an average waiting time of approximately 16 minutes. The mean service time was around 3 minutes, suggesting rapid service delivery, the traffic intensity was very low, meaning servers were underutilized rather than overburdened and the probability of an idle system ranged from 0.84 to 0.87 showing sufficient server capacity to meet customer demand efficiently. The study recommended enhancing ICT and staff deployment aligning with global trends in banking digitization and service automation. Enhancing credibility by collecting real-world data, providing clear analytical approach, and practical relevance and highlighting the importance of technology in modern queue management were some strengths of the study. Using small sample period, small sample size limited generalizability, focusing only on the withdrawal section of a single branch and not exploring cost-benefit analysis of adding servers or ICT investments were some limitations of the study. The study successfully demonstrated that queuing models, particularly the M/M/4 configuration, were powerful tools for diagnosing and improving banking service operations.

Conclusion

The study concludes that queuing theory is a valuable tool for optimizing service delivery in banks. Proper application of queuing models reduces waiting times, improves resource utilization, and increases customer satisfaction. Banks that adopt data-driven queue management strategies can significantly enhance service performance and maintain competitiveness.

References:

1. Abensur, E. O.: Banking operations using queuing theory and genetic algorithms ,Produto&Produção, volume 12(2), jun. 2011,pp. 69 – 86.
2. Abiodun, R. O. and Omosule, N. I.: Queuing Model for Banking System: A Comparative Study of Selected Banks in Owo Local Government Area of Ondo State, AJER, Volume-4(8), (2015), pp-191-195, e-ISSN: 2320-0847 p-ISSN : 2320-0936
3. AL-Jumaily, A. S. A. and AL-Jobori, H. K. T.: Automatic Queuing Model for Banking Applications, (IJACSA), Volume 2(7), (2011).
4. Augustine, A. N.: Queuing Model as a Technique of Queue Solution in Nigeria Banking Industry, International Institute for Science, Technology and Education (IISTE): E-Journals, Volume 3(8),(2013) ISSN 2224-607X (Paper) ISSN 2225-0565 (Online)
5. Brempong, S.A. and Antwi, F.: Improving Customer Service at the Agricultural Development Banks in Ghana: An Application of the Queuing Theory, African Journal of Management Research, (2015).

10.48047/jocaaa.2020.28.6.12

6. Cleiton, T. and Sergio, D. S.: Queuing Theory Applied to the Optimal Management of Bank Excess Reserves, (2011) .
7. Dhar, S.K. and Rahman, T.: Case Study for Bank ATM Queuing Model IOSR Journal of Mathematics, Volume 7(1), (May. - Jun. 2013), pp. 01-05, e-ISSN: 2278-5728,p-ISSN: 2319-765X,
8. Ehsanifar, M., Hamta, N. and Hemesy, M. : A Simulation Approach to Evaluate Performance Indices of Fuzzy Exponential Queuing System (An M/M/C Model in a Banking Case Study),JIEMS, Volume 4(2), 2017, pp. 35-51, DOI: 10.22116/JIEMS.2017.54604.
9. Finch, P.D.: Deterministic Customer Impatience in the Queuing system GI/M/1,Biometrika (I960), 47, 1 and 2, p. 45.
10. Haruna, A., Hassan, U. and Aliyu, J.: Application of Queuing Model to the Customer Management in a Banking System, Journal of African Sustainable Development, Volume 17(7) , June (2019).
11. Jabee, S. and Malviya, V.: Performance Measures of M/M/C/N Queuing Model, International Journal of Education and Science Research Review, Volume-2(1), (February- 2015), ISSN 2348-6457.
12. Kaur, A. and Singh, W.: Analysis of Queuing to Customers Management in Banking System using Simulation, IJARSE, Volume 6(7),(2017).
13. Kulia, M.: Balking and Reneging in the Queuing System, IOSR Journal of Mathematics, e-ISSN: 2278-5728.Volume 6(1), (Mar. - Apr. 2013),pp. 35-37.
14. Magnus, U., Joseph, O. C., and Anthony, O. I.: The Application of Queuing Theory in the Effective Management of Time in Money Deposit Banks - A Study Of Zenith Bank PLC In Enugu Metropolis Ugwa, The International Journal (RJSSM), Volume 05(8), Number8, (December 2015).
15. Mei, S. W and Cheng, A.K. : Introducing Queuing Theory Through Simulations, The Electronic Journal of Mathematics and Technology, Volume 9(2), (2015),152-165.
16. Murthy, M.S. R, Latha, M. P. and Deepthi, K.G.R.: Minimising The Waiting Time at Bank Atm For Service with Queuing Model, International Journal of Scientific Research, (June 2012).
17. Mutingi, M., Mapfaira, H., Moakofi, N. P.K., Moeng, S. A. and Mbohwa, C.: Simulation and analysis of a bank queuing system, Proceedings of the 2015 International Conference on Industrial Engineering and Operations Management Dubai, UAE, (March 3 – 5, 2015)
18. Nosek, R. A. and Wilson, J.P.: Queuing Theory and Customer Satisfaction: A Review f Terminology, Trends and Applications to Pharmacy Practice, Hospital Pharmacy, Volume 36(3), (2001), pp 275–279
19. Odirichukwu, J.C. ,Lekara, T. and Odii, J.N.: Banking Queue System in Nigeria,Computing, Information Systems, Development Informatics & Allied Research Journal, Volume 5(1), (March, 2014), ISBN 978-2257-44-7 (Print) ISSN 2167-1710 (online).
20. Patel, B. and Bhathawala, P.: Case Study for Bank ATM Queuing Model, International Journal of Engineering Research and Applications Volume 2(5), (September- October 2012), pp.1278-1284, ISSN: 2248-9622 www.ijera.com
21. Sahu, C. and Sahu, S.: Implementation of Single Channel Queuing Modelt Enhance Banking Services, Best: IJMITE, Volume 2(7), July 2014, pp. 71-78, ISSN 2348-0513
22. Sharma, A.K, Kumar, R. and Sharma, G. K.: Queueing Theory Approach with Queueing Model: A Study,International Journal of Engineering Science Invention, Volume 2(2), (February. 2013),pp-01-11, ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726

10.48047/jocaaa.2020.28.6.12

23. Vasumathi, A and Dhanavanthan, P.: Application of Simulation Technique in Queuing Model for ATM Facility, International Journal of Applied Engineering Research, Dindigul, Volume 1(3), (2010).
24. Yusuf, M. O., Blessing, N. and Kazeem, A. O.: Queuing Theory and Customer Satisfaction: A Review of Performance, Trends and Application in Banking Practice (A Study of First Bank PlcGwagwalada, Abuja Branch), European Journal of Business and Management, Volume7 (35), (2015), ISSN 2222-1905 (Paper) ISSN 2222-2839 (Online).
25. Xiao, H. and Zhang, G.: The Queuing Theory Application in Bank Service Optimization, IEEE Conference Proceedings, (2010).
26. Zhang, Y., Yue, D. and Yue, W.: Analysis of an M/M/1/N Queue with Balking, Reneging and Server Vacations,International Symposium on OR and Its Applications, (2005).