GSJ: Volume 12, Issue 7, July 2024, Online: ISSN 2320-9186

www.globalscientificjournal.com

QUEUEING THEORY AND ITS APPLICATION TO BANKING SECTOR IN NIGERIA

BY

ILESANMI A.O[1]., ODUKOYA E.A[2]., ALADEJANA A.E[3]., AJEWOLE K.P[4]., POPOOLA O.E.[5].

[1,2,3] DEPARTMENT OF STATISTICS, EKITI STATE UNIVERSITY, ADO-EKITI, EKITI STATE, NIGERIA.

[4,] DEPARTMENT OF PHYSICAL SCIENCES, MATHEMATHICS PROGRAM, LANDMARK UNIVERSITY, OMU- ARAN.

[5,] MATHEMATICAL SCIENCES DEPARTMENT, BAMIDELE OLUMILUA UNIVERSITY OF EDUCATION, SCIENCE AND TECHNOLOGY, IKERE.

ABSTRACT

Objective: This paper investigates the application of queuing models to optimize customer service efficiency in banks and it also explores how these models can be used to analyse customer arrival patterns and service times.

Methodology: The observed data comprised daily records of the queuing in the First Bank PLC, Ado Ekiti over a week, examining customer arrival and service patterns from 8:00 AM to 3:00 PM.

Result: The results indicates that the arrival rate of customers per unit of time is 14.2071 which is greater than the service rate of 8.3571 customers per unit of time. This shows that queue will occur since the arrival rate is greater than the service rate in the banking hall. The expected inter-arrival time was calculated at 4.2232 minutes per customer, and the average service time per customer was approximately 7.1795 minutes. A probability of 0.10869 was determined for customers needing to wait for service upon arrival. To improve efficiency at First Bank PLC, Ado Ekiti, It is recommended to increase the number of servers handling arriving customers, as the current capacity is inadequate.

Keywords: Queue theory, Arrival rate, Service rate, Customers, Inter-arrival time, Utilization factor, Probability of waiting.

INTRODUCTION

Queuing theory employs mathematical models and performance metrics to evaluate and potentially enhance the movement of customers within a waiting system. It finds widespread use in service industries. A queue arises when available resources are insufficient to meet the immediate demand placed on them. This situation occurs when all service counters are occupied to the extent that incoming customers cannot receive immediate service. A queue can be described as a line of individuals or items awaiting some form of attention or service. Queueing theory is the mathematical study of waiting lines or queues, encompassing the analysis of the processes involved in the arrival, waiting, and servicing of customers [1]. Also, queuing theory can be defined as an essential tool for analysing and optimizing systems where congestion and waiting are significant issues, enabling better resource allocation and improved service efficiency [2]. Queueing theory was described as an integrating probability, statistics, and stochastic processes to develop strategies for effective queue managements [3]. Queue theory was simply defined as a waiting line [4]. On the other hand, it was described in a similar manner, highlighting two crucial components: the source of customers they can draw from, known as the population source, and the service system [5]. Queuing theory is the mathematical study of waiting lines or queues. Queueing theory is commonly regarded as a subset of operations research, as the insights it provides are frequently employed in resource allocation decisions for service provision [6]. Some researchers emphasized using queueing theory to banking sectors to optimize service levels and efficiently allocate resources such as tellers and service points. For instance, queueing theory helps banks to design their service systems to minimize waiting times for customers [7]. Application of queue models to ATM networks by predicting usage patterns and determining the optimal number of ATMs needed at various locations. This helps reduce wait times and ensures a smooth customer experience [8].

queueing models has been used to minimize wait times and optimizing service process based on queueing theory insights. They concluded that bank can foster positive customer experiences and loyalty essential for competitive advantage [9]. Queues are generated when customers, whether they are individuals or objects, are in need of service and must wait because their quantity surpasses the available number of servers [10]. Alternatively, queues can form when the service facility is not operating efficiently or when the time required to serve a customer exceeds the prescribed duration. In certain instances, customers find themselves waiting when the overall number of customers seeking service surpasses the quantity of service facilities available, while in other situations, some service facilities remain inactive due to an excess of service facilities in comparison to the number of customers requiring service. In Nigeria, banks commonly face issues of congestion, resulting in diminished customer satisfaction and prompting customers to switch between banks in search of services with minimal delays. A queuing model was developed to facilitate the analysis of queuing methods, aiming to address issues related to customer arrival rates. This model was designed to determine various key factors, including the actual service time, customer arrival rates, expected time spent by customers in both the queue and the system, system capacity, the probability of customers having to wait for service at the bank upon arrival, and the potential utilization factor. This paper aimed to analyse and optimise customer service efficiency in the banking sector using queue models.

MATERIALS AND METHOD

The observed data comprised daily records of the queuing in the First Bank PLC, Ado Ekiti over a week, examining customer arrival and service patterns from 8:00 AM to 3:00 PM. The banking system is characterized as a multiple queue, multiple servers system based on queuing theory. In this particular type of queuing system, customers arrive at the banking hall, which consists of multiple waiting lines, each assigned to a single server. Customers make their choice of which queue to join based on certain criteria, often opting for the shortest line. This system grants customers the freedom to select and join the queue they believe will provide them with the quickest service, and it may lead to behaviours such as switching lines, leaving the queue (reneging), or deciding not to join a queue at all (balking).

SINGLE SERVERS QUEUE WITH M|M|1 MODEL

The single-server model is designed to predict queue lengths and waiting times. This model, known as the Single-Channel Queuing Model with Poisson Arrivals and Exponential Service times (M/M/1), provides insights into queuing systems where arrivals follow a Poisson distribution, service times are exponentially distributed, and there is only one server. This model focuses on systems with only one service point or server. It is model used in forecasting how long queues will be and how much time customers or items will spend waiting before being serviced. The model assumes that arrivals occur randomly over time, following a Poisson distribution. This means the probability of a certain number of arrivals in a given time period can be predicted using the Poisson formula.

MULTIPLE SERVERS QUEUE WITH M|M|S MODEL

The M/M/S system describes a queuing model characterized by Poisson arrival patterns and multiple servers (S), each with independently distributed exponential service times that remain consistent regardless of the system's state. In this model, the arrival rate (λ) is independent of the number of customers (n), meaning it remains constant for all n. Similarly, each server's service time is independent and follows an exponential distribution. However, the number of servers actively attending to customers at any given time depends on the number of customers in the system. Therefore, the overall time required to process customers is influenced by the system's current state, as more servers will be utilized when more customers are present. This model helps in understanding and managing systems with multiple servers, optimizing resource allocation, and predicting wait times and service efficiency.

Parameters in Queuing Models

In the banking sector, efficient customer service is crucial for maintaining satisfaction and operational effectiveness. Queue theory, a mathematical study of waiting lines, provides valuable insights into optimizing service delivery in banks. This study focuses on applying

queue theory to analyse and optimise customer service efficiency in the banking sector. By examining key parameters such as arrival rates, service rates, and the number of available servers. This application of queue theory will not only help First Bank in Ado Ekiti to reduce customer waiting times but also improve overall customer experience.

Arrival Rate

The arrival rate refers to the average number of customers arriving at the bank per unit of time. It is denoted by λ.

$$\lambda = \frac{Total\ Number\ of\ customers\ arrived}{Total\ arrival\ time}$$

Service Rate

The service rate is the average number of customers that can be served by a teller or service point per unit of time and it is denoted by μ

$$\mu = \frac{Total\ number\ of\ customers\ served}{Total\ service\ time}$$

Traffic Intensity

Traffic intensity, often denoted by ρ is the ratio of the arrival rate λ to the service rate μ in a queueing system. It measures the load on the service system.

$$\rho = \frac{\lambda}{\mu}$$

More generally, in a situation where there are C servers; $\rho = \frac{\lambda}{\mu \times C}$

Where λ is the mean arrival rate, μ is the mean service rate, and c is the number of servers in an M/M/C queue.

The expected inter arrival time per hour

The expected inter-arrival time per hour is the average time between consecutive customer arrivals at the bank. It is the inverse of the arrival rate λ and is calculated by $\frac{1}{\lambda}$

The average time between service distribution per hour

The average time between service distributions per hour is the average time it takes to serve one customer, calculated as the inverse of the service rate μ. This metric is given by $\frac{1}{\mu}$

The probability of having n customers in the system

The probability of having n customers in the system in a queueing model can be determined using the Poisson distribution for the M/M/1 queue (single server) and it helps in assessing the likelihood of different numbers of customers being present in the system

$$p_0 = [\sum_{n=0}^{C-1} \frac{(C\rho)^n}{n!} + \frac{(C\rho)^n}{C!(c-\rho)}]^{-1} \; where \; \rho = \frac{\lambda}{\mu}$$

Expected number of customers waiting on the queue

The expected number of customers waiting in the queue, often denoted by L_q, is the average number of customers who are waiting to be served in a queueing system.

$$L_q = [\frac{\rho^{c+1}}{(C-1)!(C-\rho)^2}]P_0 \quad \text{or} \quad L_q = [\frac{C^C \rho^{c+1}}{C!(1-\rho)^2}]P_0$$

Expected number of customers in the system

The expected number of customers in the system, denoted as L_s, represents the average total number of customers, both being served and waiting in the queue, in a queueing system.

$$L_s = L_q + \frac{\lambda}{\mu}$$

Expecting waiting time of customers in the queue

The expected waiting time of customers in the queue, denoted as W_q, is the average time a customer spends waiting to be served in a queueing system.

$$W_q = \frac{L_q}{\lambda}$$

Average Time in the System

The average time in the system, denoted as W_s, is the average total time a customer spends in the queueing system, including both the waiting time in the queue and the service time.

$$W_s = w_q + \frac{1}{\mu} = \frac{L_s}{\lambda}$$

Average waiting time for an arrival not immediately served

The average waiting time for an arrival not immediately served, often denoted as W_a, is the average time a customer spends waiting in the queue when they arrive and find all servers busy.

$$W_a = \frac{1}{c\mu - \lambda}$$

Probability that an arriving customer or customers will have to wait for service at the bank

The probability that an arriving customer will have to wait for service at the bank, denoted as P_w, is a measure of the likelihood that a customer will encounter a queue upon arrival.

$$P_w = \frac{Wq}{W_a}$$

The system capacity

The system capacity in queueing theory refers to the maximum number of customers that the system can accommodate simultaneously, including those being served and waiting in the queue and it is denoted by $C\mu$.

ANALYSIS AND RESULTS

The table 1 below showed the descriptive analysis of the daily queuing analysis of the servers reflecting a comprehensive overview of the customer arrival total, arrival rates, service total and service rates at the First Bank Plc Ado Ekiti for five consecutive days with the four (4) available servers at the banking hall between the hours of 8:00AM and 3:00PM. The service totals and service rates indicate the efficiency and capacity of the servers in managing the customer flow showing how fast and efficient the servers are. This data is very vital for optimizing staff

allocation and improving customer wait times, ultimately enhancing service quality and operational efficiency.

Table 1: Daily Queuing Analysis of the Servers

		Server 1		Server 2		Server 3		Server 4	
Days		Arrival rate	Service rate	Arrival rate	Service rate	Arrival rate	Service rate	Arrival rate	Service rate
Day 1 (Monday)	Total	116	60	111	63	125	67	119	70
	Average	16.57143	8.571429	15.85714	9	17.85714	9.571429	17	10
Day 2 (Tuesday)	Total	101	58	99	54	111	58	105	55
	Average	14.42857	8.285714	14.14286	7.714286	15.85714	8.285714	15	7.857143
Day 3 (Wednesday)	Total	91	57	107	55	102	64	115	62
	Average	13	8.142857	15.28571	7.857143	14.57143	9.142857	16.42857	8.857143
Day 4 (Thursday)	Total	71	47	85	55	109	66	81	59
	Average	10.14286	6.714286	12.14286	7.857143	15.57143	9.428571	11.57143	8.428571
Day 5 (Friday)	Total	72	47	83	56	90	56	96	61
	Average	10.28571	6.714286	11.85714	8	12.85714	8	13.71429	8.714286

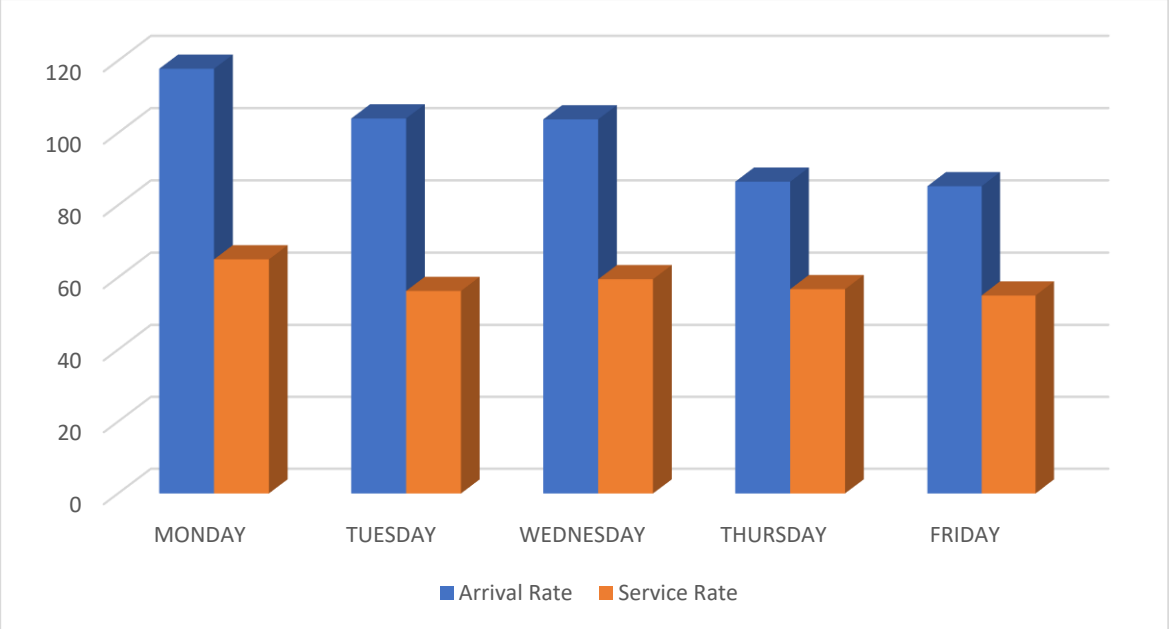

Fig 1.0: Multiple bar chart showing the customers' arrival rates and service rates.

Comment: The multiple bar chart above showed the average customers' arrival rates and service rates observed at First Bank PLC, Ado Ekiti for five consecutive days with the four available servers. The chart shows that the customers' arrival rates decreases on daily basis from Monday to Friday at the banking hall.

The mean arrival rate (λ) was calculated to be 14.2071425 by taking

$$\lambda = \frac{Total\ mean\ arrival}{The\ number\ of\ arrival\ time}$$

The overall mean service rate (μ) was calculated to be 8.3571 by taking

$$\mu = \frac{Total\ mean\ service}{The\ number\ of\ service\ time}$$

The results are shown below;

λ = 14.2071425 - Mean arrival rate of customers per hour

μ = 8.3571 – Mean service rate of customers per hour

c = 4 – Number of servers

The potential utilization factor of service facility

$$\rho = \frac{\lambda}{C\mu}$$

$$\rho = \frac{14.2071425}{4(8.3571)} = \frac{14.2071425}{33.4284} = 0.4250 \text{ or } 42.50\%$$

The system utilization factor represents the proportion of time the service facility is being used to provide service relative to its capacity. In this case, a value of 0.4250 indicates that the service facility is being utilized at 42.50% of its capacity. A utilization factor below 1.0 (100%) suggests that the system is not operating at full capacity. It means that there is some idle time or underutilization of the service facility, which can be interpreted as there being room for additional work or customers to be processed efficiently within the system.

The idle time = 1 - ρ

= 1 – 0.4250 = 0.5750 or 57.50%

The term "idle time" in queue theory refers to the amount of time a service facility or server remain unoccupied and is not actively serving customers. The value of idle time is 0.5750, it means that the service facility is idle or unoccupied for 57.50% of the time.

The average time between service distribution (service time per hour)

$$\frac{1}{\mu} = \frac{1}{8.3571} = 0.1197\text{hour} = (0.1197 \times 60)\text{ minutes} = 7.1795\text{minutes}$$

This means that the service times in the system follow a distribution where the mean (average) service time is approximately 7.1795minutes.

Expected inter-arrival time per hour

$$\frac{1}{\lambda} = \frac{1}{14.2071425} = 0.0704hour = (0.0704 \times 60)minutes = 4.2232minutes$$

The expected inter-arrival time of 4.2232 minutes means that on average, you can anticipate a new arrival or event occurring approximately every 4.2232 minutes. It helps in predicting or modelling the timing of events or arrivals within a system or process.

Average number of customers waiting in the queue before being served (L_q)

$$L_q = \left[\frac{\rho^{c+1}}{(C-1)!(C-\rho)^2}\right] P_0 \text{ or } L_q = \left[\frac{C^C \rho^{c+1}}{C!(1-\rho)^2}\right] P_0$$

$$where\ P_0 = \left[\sum_{n=0}^{C-1} \frac{(C\rho)^n}{n!} + \frac{(C\rho)^n}{C!(c-\rho)}\right]^{-1},\ \rho = \frac{\lambda}{\mu}$$

$$\rho = \frac{\lambda}{\mu} = \frac{14.2071}{8.3571} = 1.7000$$

Where C = 4

$$P_0 = \left[\left(\frac{\rho^0}{0!} + \frac{\rho^1}{1!} + \frac{\rho^2}{2!} + \frac{\rho^3}{3!}\right) + \left(\frac{(C\rho)^n}{C!(c-\rho)}\right]^{-1}$$

$$P_0 = \left[\left(\frac{1.7000^0}{0!} + \frac{1.7000^1}{1!} + \frac{1.7000^2}{2!} + \frac{1.7000^3}{3!}\right) + \left(\frac{(1.7000)^4}{4!(1-0.4250)}\right]^{-1}$$

$$P_0 = \left[\left(\frac{1}{1} + \frac{1.7000}{1} + \frac{2.8900}{2} + \frac{4.913}{6}\right) + \left(\frac{8.3521}{24(0.5750)}\right]^{-1}$$

$$P_0 = \left[(1 + 1.7000 + 1.4450 + 0.8188) + \frac{3.20114}{24(0.5750)}\right]^{-1}$$

$P_0 = (5.56902)^{-1} = 0.17956$

Therefore, $P_0 = 0.17956$

Also,

$$L_q = \left[\frac{C^C \rho^{c+1}}{C!(1-\rho)^2}\right] P_0$$

$$L_q = \left[\frac{4^4(0.4250)^{4+1}}{4!(1-0.4250)^2}\right] \times 0.17956$$

$$= \frac{0.6374}{7.9350} = 0.0803$$

$$\therefore L_q = 0.0803$$

An average number of customers waiting in a queue of 0.0803 implies that on average, there are approximately 0.0803 customers in the queue at any given time. The value is typically used in queuing theory to estimate the expected queue length or the number of people waiting for a service or assistance in a system over a certain period. It also indicates a relatively low level of congestion or waiting in the queue on average.

The expected waiting time in the queue (expected waiting time of customers in the queue)

$$W_q = \frac{L_q}{\lambda}$$

$$W_q = \frac{0.0803}{14.2071}$$

= 0.005652hour = (0.005652 × 60) minutes = 0.3391 minutes

The expected waiting time in the queue of 0.03391 minutes means that on average, a customer can expect to wait approximately 0.3391minutes before receiving service or assistance. This value is commonly used in queuing theory to estimate the average amount of time customers spend waiting in line or in a queue before they are being served.

The expected waiting time for an arrival not immediately served

$$W_a = \frac{1}{c\mu - \lambda}$$

$$W_a = \frac{1}{4 \times 8.3571 - 14.2071}$$

$$= \frac{1}{33.4284-14.2071} = \frac{1}{19.221}$$

$$= 0.0520\text{hour}$$

$$W_a = 0.0520\text{hour } or\ 3.1215\ minutes$$

This is a key metric in queuing theory and service systems analysis helping to assess and optimize service performance and customer satisfaction. The expected waiting time for an arrival not immediately served when we obtained it value to be 3.1215 minutes, indicate that the average amount of time a customer can expect to wait before being served.

Probability that an arriving customer or customers will have to wait for service at the bank

$$P_w = \frac{\text{Wq}}{W_a}$$

$$P_w = \frac{0.005652}{0.0520} = 0.10869$$

The probability that an arriving customer will have to wait for service represents the likelihood or chance of the customers waiting for service. It suggest that there is a 10.869% probability of an arriving customers waiting for service in the system.

The system capacity = Cμ

= 4 × 8.3571 = 33.4284

The system capacity when its value is 33.4284 typically refers to the maximum number of customers that the system can handle simultaneously or at its peak capacity. It suggest that the system is designed or capable of accommodating up to 34 customers concurrently.

The expected number of customers in the system (waiting in line or being served) is given by;

$L_s = L_q + \frac{\lambda}{\mu}$

$L_s = 0.0803 + \frac{14.2071}{8.3571}$

= 0.0803+ 1.7000 = 1.7803

The expected number of customers in the system when its value is 1.7803 indicates that on average, there are approximately 1.7803 customers within the entire system at any given time.

The average time spends in the system

$W_s = w_q + \frac{1}{\mu} = \frac{L_s}{\lambda}$

$W_s = 0.005652 + \frac{1}{8.3571}$

= 0.005652 + 0.1197 = 0.125352hour

Therefore, $W_s = 0.125352$ = (0.125352 × 60) minutes = 7.52112 minutes

The average time spent in the system when its value is 7.52112 minutes represents the average total time a customer spends within the system. This includes both the time spent waiting in the queue and the time spent actively receiving the service. With an average of 7.52112 minutes, it means that on the average, a customer can expect to spend 7.52112 minutes from the moment they enter the system (e.g. start waiting in a queue) until they finish receiving services.

DISCUSSION OF RESULTS

This research focused on application of queueing theory to banking sector. The table 1 shows a detailed analysis of customer arrivals, arrival rates, service totals, and service rates on the

data collected at First Bank Plc, Ado Ekiti, over five consecutive days during the banking hours from 8:00 AM to 3:00 PM, utilizing the four available servers at the banking hall. The system utilization factor of 0.4250 found in the study indicates that the service facility at First Bank Plc, Ado Ekiti, is operating at 42.50% of its maximum capacity during the observed period. A value of idle time at 0.5750 indicates that the service facility remains idle or unoccupied for 57.50% of the observed period and the mean (average) service time of approximately 7.1795 minutes indicates the typical duration customers spend being served. The expected inter-arrival time of 4.2232 minutes signifies the average time between successive customer arrivals into the banking hall and the average number of customers waiting in a queue of 0.0803 implies that on average, there are approximately 0.0803 customers in the queue at any given time. The expected waiting time in the queue of 0.03391 minutes means that on average, a customer can expect to wait approximately 0.3391minutes before receiving service or assistance and the probability that an arriving customer will have to wait for service represents the likelihood or chance of the customers waiting for service. Overall, with a system capacity estimated at 33.4284 customers for the four servers, these insights provide valuable metrics for optimizing service efficiency and customer satisfaction within the banking environment.

CONCLUSION

This study showed that the bank will experience congestion and longer waiting times for customers during peak hours because the arrival rate of 14.2071 customers per unit of time is greater than the service rate of 8.3571 customers per unit of time which indicates that the system is operating with an arrival rate greater than the service rate. It is recommended to increase the number of servers attending to the arriving customers since the current capacity is insufficient, leading to long wait times.

ACKNOWLEDGMENT

We thank the Department of Statistics, Ekiti State University, Ado Ekiti for support the received during the course of this research.

REFERENCES

1. Havrkort, T. (2020). *Queueing Theory and Stochastic Teletraffic Models*. Springer.
2. Cochran, J.J., & Bharti, A. (2020). A Practical Guide to Queuing Theory. *Operations Research*, 68(2), 123-135.
3. Kulkarni, V.G. (2016). *Modeling and Analysis of Stochastic Systems*. CRC Press.
4. Pei-Chun, S., Yang, S.C., & Cheng, Y.C. (2006). *Introduction to Queuing Theory*. MIT Press.
5. Tian, N., Zhang, Z., & Cao, J. (2011). *Stochastic Models in Queueing Theory*. Springer.
6. Wikipedia, (2008) Queueing Theory. Retrieved from https://en.wikipedia.org/wiki/Queueing_theory
7. Cochran, J.J., & Bharti, A. (2020). A Practical Guide to Queuing Theory. *Operations Research*, 68(2), 123-135
8. Bolch, G., Greiner, S., De Meer, H., & Trivedi, K.S. (2020). *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*. Wiley-Interscience.
9. Oladipupo, T.O., & Akinleye, O. (2019). Applying Queueing Models to Minimize Wait Times in Banking Sector. *Journal of Banking and Finance*, 45(3), 215-229.
10. Kasum, A.S., Dogo, A.S., & Musa, A. (2006). *Fundamentals of Queueing Theory*. Prentice Hall.