Journal of the Nigerian Association of Mathematical Physics
Volume 53 **(November *2019 Issue*), pp83 – 88**

QUEUING THEORY APPROACH IN IMPROVING THE QUANTUM OF SERVICE DELIVERY IN COMMERCIAL BANKING SECTOR

Abdulmajeed Inya Khadijah and Abubakar Yahaya

Department of Statistics, Ahmadu Bello University Zaria, Kaduna State

Abstract

Congestion is a major problem in many banks especially at ATM where customers are faced with serious problems of queues or delay in service delivery. Hence, queuing theory is very suitable in banking sector since it is linked with waiting line where customers who cannot be served immediately have to queue for service. In this research, data was obtained from the ATM counter and measurement were taken on the arrival times and service times of the customers at the ATM service point within morning hours of 9:00am-11:00am and evening hours of 4:00pm-6:00pm. The Kolmogorov-Sminorv goodness of fit test was used to establish the distribution of inter-arrival and service times. The G/G/s queuing model was found to describe the ATM queuing system across the bank's ATMs. The G/G/s model having general distribution of inter-arrival times, general distribution of service times and parallel servers is employed in modeling the ATM queuing system across the bank's ATMs. The result reveals that problem of long waiting time and ATMs over-utilization was common to the bank if the system is viewed as a multiple-server system each with separate queues. The result of the analysis shows that, the performance measures were greatly improved when customers obliged to maintain a single queue served by multiple ATMs than when customers observed single queue for each ATMs.

***Keywords*:** Kolmogorov-Sminorv, G/G/s model, queue, congestion, service delivery.

1.0 Introduction

Queuing Theory (QT), as the name suggests, is a mathematical study of waiting lines (queues) and is an entire discipline within the field of operations management. It's a popular theory used largely in the field of operational and retail analytics; because it uses probabilistic methods to make predictions in the field of operational research, computer science, telecommunications and traffic engineering [1]. QT was first implemented in the beginning of 20th century to solve telephone calls congestion problems and was believed to be originated in the early 1900s with the work of Erlang, (1878-1929) of the Copenhagen Telephone Company who derived several important formulae for telephone traffic engineering that today bear his name [2]. Waiting queue models are effective in service areas. Analysis of queues in terms of average waiting time, average number of customers and other aspect helps us to understand service systems such as automated teller machine stations. Automated teller machine (ATM) also known as cash machine, is an electronic telecommunications mechanism that provides customers of commercial organization with access to business operation in a public place without the need for a bank teller. The use of ATM is protected and suitable where the ATM has made settlement of bills in Nigerian banking system simple and saver. These benefits have resulted into incident growth in number of ATMs in Nigerian and the growth spread from 83% in 2006 to 289% in 2007 [3]. Kendall-Lee devised the notation of the queuing system having six characteristics denoted as 1/2/3/4/5/6 where the first characteristics specifies the inter-arrival time distribution and the second characteristics specifies the service time distribution, the third characteristics is the number of parallel servers, the fourth characteristics portray the queue discipline, the fifth characteristics specifies the maximum allowable number of customer in the system and the sixth characteristics gives the size of the population from which

Corresponding Author**:** Abdulmajeed I.K., Email: abdullahikhadeejah104@gmail.com, Tel: +2348130725887

customers are drawn [4]. For example an illustration of this notation M/M/s represents exponential inter-arrival times, exponential service times, with s parallel servers. The letter M stands for Markovian or Memoryless where the inter-arrival time distribution and the service time distribution are determined by goodness of fit test. Application of the analytic queuing models in findings and analyzing ATM waiting queues with the likelihood of decreasing this unhealthy event in the banking organization have been shown in [5-8]. A good number of researchers in the past have assumed the Poisson arrival distribution, the exponential service time distribution and automatically applied the M/M/1 or the M/M/c queue models in solving the problem of long waiting time of customers and server over utilization not minding whether the inter-arrival time/service time is exponentially distributed or not exponentially distributed. The collective suggestion made by these researchers is that the number of ATMs should be increased thereby acquiring new price of procurement, installation and maintenance of the machines not considering the rate of the machines.

2.0 Materials and method

This section presents the method of data collection, distribution fits, and the mathematical details of the G/G/s queue model employed in the study.

2.1 Method of Data Collection and distribution fits

The method used during the data collection was direct observation of customers recorded from three key required quantities (customers arrival times, service start and completion times). Data were collected for each ATM facility across the two banks over the period of two weeks including weekends for four hours each day (9am-11am and 4pm-6pm). The first week involves taking the data based on multiple queues, multiple servers while the second weeks involves taking the data based on single queue, multiple servers. It is important to mention that the congestion at the ATM during festive period and month endings was not captured in the data. This is because the period of data collection does not include those periods. The Kolmogorov Sminorv goodness of fit test was employed to establish the distribution of inter-arrival and service times to the collected data under the initial hypothesis that the inter-arrival and service times are exponentially distributed at 5% significance level.

2.2 Kolmogorov-Smirmov Test

To determine if the arrival time and service time data comes from a specified distribution; the kolmogorov–smirnov one sample test was performed on collected data. The Kolmogorov-Smirnov one-sample test is a test of goodness of fit. It compares a given dataset having a known distribution with another dataset of one unknown distribution and let one knows if both datasets are generated from the same distribution. Briefly, the test involves specifying the cumulative frequency distribution which would occur under the theoretical distribution $F_{O(X)}$ and comparing it with the observed cumulative frequency distribution$S_N(X)$. The theoretical distribution represents what would be expected under the initial hypothesisH_O. Under H_O we would expect the differences between $S_N(X)$ and $F_{O(X)}$ to be small and within the limits of random errors. The Kolmogorov-Smirnov test focuses on the largest of the deviations between the theoretical distribution and the observed cumulative frequency distribution. The largest value of $\left|F_{O(X)} - S_N(X)\right|$ is called the maximum deviation, D. That is

$$D = Max\,|F_O(X) - S_N(X)| \quad (1)$$

Where;

$F_O(X) =$ The cumulative frequency distribution of the hypothesized distribution.

$S_N(X) =$ The cumulative frequency distribution of the observed data.

Table1: One Sample Kolmogorov Test

Bank	ATM identification number	D_{max} (IAT)	D_{max} (ST)	p-value	Exponential	
UBA	I	0.2830	0.4399	2.2e-16	No	
UBA	II	0.3575	0.4399	2.2e-16	No	

IAT- Inter-arrival time ST- Service time UBA- United bank for Africa

The results of the K-S test above shows that the distribution of inter-arrival times is non- exponential and likewise the distribution of service times. Since the underlying assumptions for a Markovian queueing system do not hold, we can also see that the very small which tells us that the data does not match the specified distribution (not necessary exponential).

2.3 Assumptions for a Markovian Queuing System

i. Arrival pattern of customers are independent such that first arrival has no effect on the second arrival and so on.
ii. Past arrival does not affect future arrival
iii. Bulk arrivals cannot occur.
iv. If the arrival rate is stationary, then the number of customers follows a possion distribution

For that purpose, we proceed to fit appropriate distributions to the inter-arrival times and the service times. To fit the appropriate model for the inter arrival time and service time, the **packages fitdistrplus and actuar** in R was used.

Table 2: Summary of Distribution Fits for Inter-arrival and Service Time

Bank	ATM Identification number	Distribution of inter-arrival time	Parameter	Distribution of service time	Parameter
UBA	I	Gamma	$\alpha = 0.21335$	Weibull	$\alpha = 1.280335$
			$\beta = 0.08622$		$\beta = 1.857746$
UBA	II	Gamma	$\alpha = 0.26052$	Log-logistic	$\alpha = 2.79265$
			$\beta = 0.09987$		$\beta = 1.74491$

2.4 The G/G/s/FCFS/∞/∞ Queuing Model

Most queuing models have assumed exponential inter-arrival times (Poisson input process) and exponential service times [9]. However the assumption of exponential inter-arrival times implies that arrivals occur randomly and the service times are also random which a reasonable approximation in many situations. This assumption is violated if the arrivals are scheduled or regulated in some way that prevents them from occurring randomly and even the service time distribution frequently deviate greatly from the exponential form especially when the service requirement of the customers are quite similar [10]. For this reason such Queuing models in which both arrivals and service times do not follow the Poisson distribution are complex. In general, it is advisable in such cases to use simulation as an alternative tool for analyzing them. However there are few non-Poisson queues for which analytic result can be available and the only necessary thing to estimate is the mean $\frac{1}{\mu}$ and variance $\sigma^2$of both the inter-arrival/service times. The G/G/s/FCFS/∞/∞ model assumes that both inter-arrival times and service times do not follow an exponential distribution. The first **G** indicates that inter-arrival times follow some general distribution (but not necessary exponential), second **G** indicates that service times are governed by some general distribution (but not necessary exponential), there are **s** servers (s >1), the queue discipline is first come first served, the system capacity and the population are both infinite. Assuming there are n numbers of customers in the queuing system at any point in time and s is the number of servers. Then if $n < s$ (number of customers in the system is less than the number of servers) then there will be no queue and the combined service rate will be $\mu_n = n\mu$ and if $n \geq s$ (number of customers in the system is more than or equal to the number of servers) then all servers will be busy and the maximum number of customers in the queue will be n – s. Then the combined service rate will be $\mu_n = s\mu$. If the maximum mean service rate $s\mu$ exceeds the mean arrival rate λ then a queuing system fitting this model will eventually reach a steady state condition. But if the $\lambda \geq s\mu$ (i.e. the mean arrival rate exceeds the maximum service rate) then the queue grows without bound so, the steady state solution are not applicable. For this kind of queuing model, we do not have an exact result for the performance measures but fortunately, the Allen-Cunneen approximation often gives a good approximation [11].

Table 3: Formula for computing performance measures of the queuing system

Performance measures	Single server (G/G/1)	Multiple server (G/G/s)
Average server utilization in the system (ρ)	$\frac{\lambda}{\mu}$	$\frac{\lambda}{s\mu}$
Average number of customer In the system (L_s)	$\lambda\left[\frac{\rho^2(1+\gamma_s)\,(\gamma_\alpha+\rho^2\gamma_s)}{\lambda[2(1-\rho)\,(1+\rho^2\gamma_s)]}+\frac{1}{\mu}\right]$	$\lambda\left[\approx W_q^{M/M/c}\,\frac{\gamma_\alpha+\gamma_s}{2}+\frac{1}{\mu}\right]$
Average number of customer In the queue (L_q)	$\approx \frac{\rho^2(1+\gamma_s)\,(\gamma_\alpha+\rho^2\gamma_s)}{2(1-\rho)\,(1+\rho^2\gamma_s)}\,,\rho<1$	$\lambda\left[\approx W_q^{M/M/c}\,\frac{\gamma_\alpha+\gamma_s}{2}\right]$
Average time a customer spends in the system (W_s)	$\frac{\rho^2(1+\gamma_s)\,(\gamma_\alpha+\rho^2\gamma_s)}{\lambda[2(1-\rho)\,(1+\rho^2\gamma_s)]}+\frac{1}{\mu}$	$\approx W_q^{M/M/c}\,\frac{\gamma_\alpha+\gamma_s}{2}+\frac{1}{\mu}$
Average time a customer spends in the queue (W_q)	$\frac{\rho^2(1+\gamma_s)\,(\gamma_\alpha+\rho^2\gamma_s)}{\lambda[2(1-\rho)\,(1+\rho^2\gamma_s)]}$	$\approx W_q^{M/M/c}\,\frac{\gamma_\alpha+\gamma_s}{2}$

Where;

$$\gamma_s = \frac{\sigma_s^2}{(1/\mu)^2} \quad (2)$$

$$\gamma_\alpha = \frac{\sigma_\alpha^2}{\left(1/\lambda\right)^2} \quad (3)$$

γ_α = Coefficient of variation of inter-arrival time.

γ_s = Coefficient of variation of service time.

σ_α^2 = Variance of the inter-arrival time and σ_s^2 = Variance of service time.

3.0 Results and Discussions

The data obtained on arrival and service times at the ATM machine of the selected banks were used to fit probability distributions of inter-arrival and service times which enhance the selection of the appropriate queuing model. Model results were eventually used in computing the values of the queue performance measures for each ATM. Computed Values of arrival rate and service rate with mean and variance of inter-arrival time and service time distribution is displayed in table 4 and 5 for each ATM, while Performance measures of the current state of each ATM and the proposed state of each ATM is shown in table 6 and 7.

Table 4: computed values of arrival and service rate

Bank	ATM Indentification Number	Arrival rate λ	Service rate μ
UBA	I	0.4041	0.5073
UBA	II	0.3833	0.4770
UBA	I & II	0.3936	0.4918

From table 4, we can see that the maximum mean service rate $s\mu$ exceed the mean arrival rate λ which tells us that the system is in a steady state with combine service rate of $s\mu$.

Table 5: computed statistics of inter-arrival and service time distribution

Bank	ATM indentification number	Mean of inter-arrival time distribution (mins)	Variance of inter-arrival time distribution (mins)	Mean of service time distribution (mins)	Variance of service time distribution (mins)
UBA	1	2.0123	8.2468	1.971	2.6164
UBA	2	2.0690	6.0309	2.0962	2.2633
UBA	1 & 2	2.0404	7.1426	2.0332	2.4429

Table 6: Performance measures for multiple server ATMs where each ATM is considered as a separate queuing system.

Performance measures	UBA	
	ATM 1	ATM 2
ATM utilization (%)	79.65	80.36
L_S	5.45	4.97
L_q	4.67	4.17
$W_s\ (mins)$	13.5	13.00
$W_q(mins)$	11.6	10.90

Where; L_S: Average number of customers in the system, L_q: Average number of customers in the queue. W_s : Average time spent waiting in the system and in the system, W_q: Average time spent waiting in the system and in the queue.

Based on the analysis displayed in table 5 & 6, result of the ATM 1 UBA shows that customers spent 1.97 minutes on an average in service after waiting for 12 minutes on an average in the queue at service rate of 0.51 customers per minute. It was also found that an average of 5 customer wait on the queue while the ATM utilization is 79.65%. Result of the ATM 2 UBA shows that customers spent 2.1 minutes on an average in service after waiting for 11 minutes on an average in the queue with service rate of 0.50 customers per minute. It was also found that an average of 4 customer wait on the queue while the ATM utilization is 80.36%. The result also shows that a high ATM utilization can eventually cause machine breakdown and customers do spent much time waiting in queue for service but less time in service. The bank adopts a multiple server system and first come first serve is not ensured because there is high possibility of customer finishing service far ahead of those who arrived earlier. Also some of the customers may decide to create multiple possibly short lines in front of each ATM and receive serve in order of arrival within each queue. The FCFS queuing discipline is mostly implemented to ensure fairness in service delivery.

Table 7: Performance measures for multiple server ATMs where all ATMs are considered as a single queuing system.

Performance measures	UBA
	ATM 1 & 2
ATM utilization (%)	39.98
L_S	1.06
L_q	0.261
$W_s\ (mins)$	2.7
$W_q(mins)$	0.664

Where; L_S: Average number of customers in the system, L_q: Average number of customers on the queue, W_s : Average time spent waiting in the system and in the queue, W_q: Average time spent waiting in the system and in the queue

Based on the analysis displayed in table 7, result of the ATM 1& 2 UBA shows that customers spent 2.03 minutes on an average in service after waiting for 3 minutes on an average in the system at service rate of 0.49 customers per minute. It was also found that an average of 1customer wait on the system while the ATM utilization is 39.98%.

From the performance measures displayed both in Table 6 & 7, we see that the average waiting time both on queue and system, the average number of customers in the system and on queue are dramatically minimized when the queuing system is viewed as a single queue with two servers than when viewed as two separate queuing systems with single servers each. In addition, multiple-queue generates longer waiting time and result in poor system performance. A system that adopt single queue would perform better than one in which multiple queues are formed and fed strictly into each available ATM. Therefore for UBA to obtain an optimal service level of wait time not exceeding 5 minutes based on the manager's criterion, for FCFS to be ensured and to avoid congestion at the ATM counter customers should oblige to a system of single queue formation with multiple servers.

4.0 Conclusion

The queuing models for United Bank for African of Samaru-Zaria, Kaduna State is a G/G/s model, when no restriction are imposed on what the inter-arrival times and services times distribution should be alongside with the uncertainty of queuing discipline of FCFS system. From the performance measures, it was observed that the average waiting time both on queue and system, the average number of customers in the system and on queue were improved when the queuing system is viewed as one with two servers than when viewed as two separate queuing systems with single servers each. In addition, multiple-queue generates longer waiting time and result in poor system performance. A system that adopt single queue would perform better than one in which multiple queues are formed and fed strictly into each available ATM because single queue out performs the multiple queues even when more than one server is used. To ensure fairness in the system, customers should oblige to single queue so that FCFS can be ensured.

Reference

[1] Tavish, S. (2016). How to Predict Waiting Time Using Queuing Theory. *An Article of Business Analytics*.

[2] Shanmugasundaram, S. and Punitha, S. (2014). A Study on Multi- Server Queuing Simulation. *Journal of Science and Research, 3*(7), 1519-1521.

[3] Adeoti, J.A. (2011). Automated Teller Machine (ATM) Frauds in Nigeria, the way out. *Journal of Social Sciences, 27*(1), 53-58.

[4] Kendall, D. (1951). Some Problems in the Theory of Queues. *Journal of the Royal Statistical Society, 13*, 151-185.

[5] Ogunwale, O.D. and Olubiyi, O.A. (2010). A Comparative Ananlysis of Waiting Time of Customers in Banks. *Global Journal of Science Fontier Research, 10*(6), 97-99.

[6] Vasumathi, A. and Dhanavanthan, P. (2010). Application of Simulation Technique in Queuing Model for ATM Facility. *International Journal for Applied Engineering Research, 1*(3), 469-482.

[7] Famule, F.D. (2010). Analysis of M/M/1 Queuing Model with Application to Waiting Time in Banks. *Global Journal of Computer Science and Technology, 10*(13), 28-34.

[8] Al-jumaily, S.A. and Al-jabori, K.T. (2012). Automatic Queuing Model for Banking Application. *International Journal for Advanced Computer Science Application, 2*(7), 11-15.

[9] Denardo, E. (1982). Dynamic Programming: Theory and Applications. Englewood Cliff N.J: Prentie Hall.

[10] Hiller, F.S .and Liberman G.J. (2007). *Introduction to Operation Research* (7th ed.). Mc-Graw Hill.

[11] Tanner, M. (1995). *Practical Queuing Analysis.* New York: McGraw-Hill.