Journal of the Nigerian Association of Mathematical Physics
Volume 52 (*July & Sept., 2019 Issue*), **pp59 – 66**
© J. of NAMP

DERIVATION OF A FINITE QUEUE MODEL WITH POISSON INPUT AND EXPONENTIAL SERVICE

[1]S.A. Ogumeyo and [2]C.C. Nwamara

[1]Department of Mathematics, Delta State College Of Education, Mosogar, Delta State, Nigeria.
[2]Department of General Studies, Petroleum Training Institute, Effurun, Delta State, Nigeria.

Abstract

Queues are waiting lines of customers waiting to be served. Queuing systems are prevalent in human activities such as we experience in filling stations, banks, toll gates, hospitals, e.t.c. Many queuing models in literature have dealt with several aspects of queuing theory ranging from single-server (M/M/1) to multiple-server model with Poisson input and exponential service (M/M/S). In these models, so far no restriction has be imposed on the total number of customers allowed in the systems at any time. In this paper a single-server model with Poisson input and exponential service (M/M/I) with finite queue length (restricted number of customers to wait in line at any instant) is developed. From the sensitivity analysis of different values of traffic intensity ρ and service rate μ, and arrival rate λ of the model, it is observed that as the traffic intensity ρ increases, the expected values of the number of customers, the line length, the time in the line, all increase rapidly. It is also observed that for a given service rate μ, when intensity ρ is small, most of the expected time a customer spends in the queuing system is due to average service time; but as the arrival rate λ increases, most of the expected time spent is due to waiting **in line.**

Keywords: **Queue, Customer, inter-arrival, departure.**

1.0 Introduction

According to Siagian [1] a queue is waiting line made up of customers requiring service from one or more facilities. Queue theory as stated in [2] was proposed and developed by Agner Kraup Erlang, a Danish Mathematician and Engineer in 1910 while working on the Copenhagen Telephone Company's problem of how to determine the number of people who needed telephone services and how many telephone operators were required to handle a certain volume on the call. Basic elements of queuing models depend on factors such as arrival distribution, service distribution, service facility, service discipline and queue size (number of customers in the queue) [3]. Taha [4] opined that a queuing system can be described by its input or arrival process, its queue discipline, and its service mechanism. The input process which is usually a description of the pattern of arrivals into the system is given by the probability distribution of time between successive arrival events, and the number of individuals or units that appear at each of these events as in case of a super-market, restaurant, petrol station etc. [5].

Subagyo [6] listed four basic structures of queuing models which are: single channel single phase, single channel multiple phase, multiple channel single phase and multiple channel multiple phase. According to Kakiay [7] there are four commonly used service disciplines: First Come First Served (FCFS), Services in Random Order (SIRO), and Priority Service (PS). Queues are situations we experience in the day-to-day activities of life. For instances, queues are being experience in hospitals, filling stations, banks, restaurants etc as customers wait to be served. As stated in [8], queues are formed when the demand for service exceeds its supply, and that wait time depends on the number of customers in the queue system, the number of servers attending to the customers and the amount of service time for each customer.

Corresponding Author: Ogumeyo S.A., Email: Simonogumeyo64@gmail.com, Tel: +2348052762209

Most elementary queuing models assume that the inputs (arriving customers) and outputs (leaving customers) of the queuing system occur according to the birth-and –death process [9]. However, in the context of queuing theory, the term birth refers to the arrival of a new customer into the queuing system and death refers to the departure of a served customer whereas the state of the system at time t(t > o), denoted by N(t), is the number of customers in the queuing system at time t [10]. According to Wagner [11], arrival and departure of customers in a queue system occur randomly, hence their mathematical models are formulated based on the following assumptions:

Assumption1: Given $N(t) = n$, the current probability distribution of the remaining time until the next arrival is exponential with parameter λ_n $(n= 0,1,2,...)$.

Assumption 2: Given $N(t) = n$, the current probability distribution of the remaining time until the next departure (service completion) is exponential with parameter μ_n $(n = 1, 2 .....)$.

Assumption 3: The random variable of assumption 1 (the remaining time until the next arrival) and the random variable of assumption 2 (the remaining time until the next departure) are mutually independent. The next transition in the state of the process is either $n \longrightarrow n + 1$ (a single arrival) or $n \longrightarrow n - 1$ (a single departure) depending on whether the former or later random variable is smaller.

The most common decisions that need to be made when designing a queuing system as contained in [12], include: (a) number of servers at a service facility (b) efficiency of the servers, (c) number of service facilities (d) amount of waiting space in the queue (e) any priorities for different categories of customers and the two primary considerations in making these kinds of decisions typically are (i) the cost of service capacity provided by the queuing system and (ii) the consequences of making the customers wait in the queuing system. Providing too much service capacity causes excessive costs, providing too little causes excessive waiting. Therefore, the goal is to find an appropriate trade-off between the service cost and the amount of waiting.

One basic approach available for seeking this trade-off is to establish one or more criteria for a satisfactory level of service in terms of how much waiting would be acceptable. For example, one possible criterion might be that the expected waiting time in the system should not exceed a certain number of minutes. Another might be that at least 95 percent of the customers should wait no longer than a certain number of minutes in the system. Similar criteria in terms of the expected number of customers in the system (or the probability distribution of this number) also could be used. The criteria also might be stated in terms of the waiting time or the number of customers in the queue system. Once the criterion or criteria have been selected, then it is usually straightforward to use trial and error to find the least costly design of the queuing system that satisfies all the criteria. The other basic approach for seeking the best trade-off involves assessing the costs associated with the consequences of making customers to wait [13].

Many authors have developed different queue models to address different situation involving queuing. For examples a queuing model on patient waiting time in ante-natal care clinic to determine the number of doctors required so that a given percentage of pregnant women do not exceed a given waiting time and the number of expectant mothers in the queue do not surpass a given threshold is developed in [14]. Nugraha [15] developed a queuing model on Toll Gate with the aim of decongesting traffic on the highways. In line with [12], which advocates minimum cost of waiting time of customers, we present a Finite Queue Model of Single-server with restricted number of customers allowed per time to wait in a queue system.

2.0 Model Assumptions and Mathematical Notations

The assumptions associated with the queuing model formation are as follows:

a) The system consists of serves n-customers arriving per time *t*.
b) A customer is assumed to join each line with equal probability regardless of the length of the several lines.
c) The expected number of customers in each system, is summed over all the separate system.
d) Customers are attended to on first come, first serve basis.

Mathematical Notations

1 λ = Average number of customers arriving per unit of time
2 μ = Average number of services per unit of time.
3. P = traffic intensity
4. T = time of arrival
5. h > o very small interval of time
6. n = number of customers in the system
7. Pn(T) = Probability that n customers are in the system at time T.
8. $\lambda e^{-\lambda t}$ = exponential inter arrival density
9. $\mu e^{-\mu t}$ = exponential service time density

The general standard symbols for the probability distributions are:

M ≡ Exponentially distributed inter-arrival service time (*M* is an abbreviation for Markovian)
D ≡ deterministic (or constant, regular) inter-arrival or service time
E_n ≡ Erlangian distribution of order n for inter-arrival or service time
GI ≡ general independent distribution of inter-arrival time
G ≡ general distribution of service time
The model of Poisson input, exponential service, and single server is denoted by *M/M/*1.

3.0 Model Description and Mathematical Formulation

In this model we impose a limitation on the total number of customers should be in the system at any time. *M is the total number of* customers allowed to be in the entire system, with not more than $M – 1$ persons permitted to wait in the queue line at any instant. If a customer arrives when *M* persons are already in the system, then the customer is restricted from entering, and is said to be lost from the system. The **finite-queue model** presented in this paper allows a statistical equilibrium to be reached for *any* value of the intensity ratio λ/μ.

The simplest single-server model having both input and service processes described probabilistically is one with exponential inter-arrival and service times (denoted by *M/M/*1). Specifically, assume

Exponential inter-arrival density = $\lambda e^{-\lambda t}$ …………………………………….(1a)

Exponential service time density = $\mu e^{-\mu t}$ ……………………………………(1b)

Define the number of customers n in the system at any point in time as including the persons waiting in line plus those in service. Let $t = 0$ represent the system's starting point in time, and define

$P_n(T)$ = probability that n customers are in the system at Time T……………(2)

Actually $P_n(T)$ depends on the number of customers in the system at Time 0, but this dependence is suppressed in the notation.

Let $h > 0$ be a *very* small interval of time. If there are $n > 0$ customers in the system at Time T + h, then we consider only the possibilities that there were either $n – 1$, n or $n + 1$ customers at Time T. Any other possibilities are of relatively insignificant probability. Consequently, for $n > 0$.

$P_n(T+h) = (\lambda h)(1 - \mu h)P_{n-1}(T) + (1 - \lambda h)(1 - \mu h)P_n(T) + (\lambda h)(\mu h)P_n(T) + (1 - \lambda h)$
$(1 - \mu h)P_{n+1}(T)$ *(h small)* ………………………………………………………..*(3)*

The first term on the right corresponds to the event of one arrival and no departure when $n – 1$ customers are in the system at Time T. The second and third terms refer to the events of no arrival and no departure, and of one arrival and a departure when n customers are in the system at Time T. and the final term relates to the event of no arrival and one departure when there are $n + 1$ customers in the system at Time T. As the symbol (=) indicates, the expression in (3) is approximate; it can be made exact by adding probability terms with coefficients h^k, where $k \geq 2$.

Bringing the term $P_n(T)$ to the left-hand side of (3), dividing by h, and letting $h \to 0$, yields

$$\frac{dP_n}{dT} = \lambda P_{n-1}(T) - (\lambda + \mu)P_n(T) + \mu P_{n+1}(T) \text{ for } n > 0 \ldots\ldots\ldots\ldots\ldots(4)$$

This expression is exact because the terms neglected in (3) become 0 as $h \to 0$.
Similarly, one can show that

$$\frac{dP_o}{dT} = -\lambda P_o(T) + \mu P_1(T) \quad \text{for } n = 0 \ldots\ldots(5).$$

With moderately advanced analysis it is possible to solve the system of linear differential equations (4) and (5) for each $P_n(T)$. (To do so, we must also state the number of customers *i* in the system at Time 0). The result is called the **Transient solution**, since it depends directly on the value of *T*.

Suppose, however, we examine the values of $P_n(T)$ as $T \to \infty$. If $P_n(T)$ approaches a limiting value, say, P_n, and if $E[n]$ is finite for this limiting distribution, then we will say that the system reaches **statistical equilibrium.** We denote the resulting $\boldsymbol{P_n}$ values as **equilibrium or stationary probabilities.** The label "stationary" derives from the property that if the number of customers in the system at any Time t is given according to the probability distribution $\boldsymbol{P_n}$, then for *any* $h > 0$, P_n is also the probability that *n* customers are in the system at Time t + h. The value of P_n can also be interpreted as the limiting fraction of an arbitrarily long period of time during which the queue contains *n* customers.

Provided

$$\rho \equiv \frac{\lambda}{\mu} < 1 \text{ (traffic intensity assumption),} \ldots\ldots(6)$$

the stationary probabilities P_n always exist; the symbol ρ (rho) in (6) is frequently called the **traffic intensity.**

We can find the equilibrium solution $P_n(T) = P_n$, for all *T*, by using the consequence that each dP_n/dT must equal 0, if the solution P_n is indeed independent of *T*. Hence to obtain P_n, all we need to do is set the time derivatives in (4) and (5) equal to 0, yielding

$$0 = \lambda P_{n-1} - (\lambda + \mu)P_n + \mu P_{n+1} \qquad \text{for } n = 1, 2, 3, \ldots\ldots(7)$$

$$0 = -\lambda P_o + \mu P_1 \qquad \text{for } n = 0. \ldots\ldots(8)$$

The system of **difference equations** (7) and (8) is easily solved recursively, starting with (8),

$$P_1 = P_o \left(\lambda/\mu\right) = P_0\rho \ldots\ldots(9)$$

And proceeding to (7) for $n = 1, 2, \ldots$,

$$P_n = P_0\rho^n \ldots\ldots(10)$$

It is easy to verify that P_n in (10) does satisfy (7). Given (6),

$$\sum_{n=0}^{\infty} P_n = P_0 \sum_{n=0}^{\infty} \rho^n = \frac{P_0}{1-\rho} = 1 \ldots\ldots(11)$$

and it follows that $P_o = 1 - \rho$, so that

$$P_n = (1 - p)p^n \quad \text{for } n = 0, 1, 2, \ldots \text{ (geometric distribution)} \ldots\ldots(12)$$

with

$$E\begin{bmatrix}\text{number of} \\ \text{customers} \\ \text{in system}\end{bmatrix} \equiv E[n] = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda} \qquad \text{Var}\,[n] = \frac{\rho}{(1-\rho)^2} \ldots\ldots(13)$$

$$P[n \geq N] = \rho^N$$

Notice that the probability distribution in (12) depends *only* on the traffic intensity ratio ($\lambda/\mu = \rho$). Since $\rho (= 1 - P_0)$ is also the fraction of time the server is busy, the quantity ρ is sometimes called the system's **utilization factor.** It is significant that this interpretation of ρ remains valid even when both the inter-arrival and service time distributions are general (that is, for the model *GI/G/1*).

- Letting P_n (T/*i*) denote the transient solution to (4) and (5) when *i* customers are in the queue at $t = 0$, it can be shown that

(i) $\quad |P_n(T \mid i) - P_n| \leq e^{-T(\lambda + \mu - 2\sqrt{\lambda\mu})}$

Consequently, $P_n(T \mid i)$ approaches P_n at least exponentially by a factor proportional to *T*. Note, however, that the proportionality factor, which can be written as $\left(\sqrt{\lambda} - \sqrt{\mu}\right)^2$ (approaches 0 as λ approaches μ. Hence *T* may have to be very large before $P_n(T \mid i)$ and P_n are nearly equal; this is especially so when ρ is large and *i* is small.

Operating characteristics. The expected line length can be found by noting that line length = $\begin{cases} \text{number in system if } n = 0 \\ \text{Number in system} - 1 \text{ if } n > 0, \end{cases}$ ……(14)

so that

$E[\text{line length}] = 0.\,P_0 + \sum_{n=1}^{\infty}(n-1)P_n = \sum_{n=0}^{\infty} nP^n - \sum_{n=1}^{\infty} P_n$

$= E[n] - (1 - P_0) = \dfrac{\rho^2}{1-\rho} = \dfrac{\lambda^2}{\mu(\mu-\lambda)}$ …………………………….(15)

Next consider the time intervals when the server is idle. Since these begin when a service terminates and end when a new arrival occurs, the length of the idle periods has the same distribution as the inter-arrival time, that is, exponential with mean $1/\lambda$. Let the period of length T be so long that we can safely utilize expected values. Then the server is idle for $[TP_0 = T(1 - \rho)]$ units of time, and $[T(1 - \rho)(1/\lambda) = \lambda T(1 - \rho)]$ is the number of separate idle periods during T. Because idle and busy periods alternate, $\lambda T(1 - \rho)$ is also the number of separate busy periods during T, and ρT is the total duration of all busy periods.

Consequently,

$E[\text{length of busy period}] = \dfrac{\rho T}{\lambda T(1-\rho)} = \dfrac{1}{\mu - \lambda}$ ……………………….*(16)*

and

$E\begin{bmatrix}\text{number of customers} \\ \text{served per busy period}\end{bmatrix} = \mu E\begin{bmatrix}\text{lenght of} \\ \text{busy period}\end{bmatrix} = \dfrac{1}{1-\rho}$ ………………(17)

The relations (16) and (17) are actually valid for any service time distribution.

Now turn to the probability density for the time a customer spends in the system, which is defined as the interval a customer waits in line plus the time in service. Suppose the system is in statistical equilibrium, so that when a new customer arrives, he finds n customers in the system ahead of him with probability P_n given by (12). Assume that the queue discipline is first come, first served. Then the total time the customer spends in the system is comprised of the sum of $n + 1$ independent and identically distributed exponential random variables, and has a gamma density.

[Total time the customer spends in the system]

$= \dfrac{\mu(\mu y)^n e^{-\mu y}}{n!}$ for $y \geq 0$,……………………………………….…(18)

as we already saw in (15). Hence the density of time that a customer who arrives at an arbitrary instant spends in the system is given by

$h(w) = \sum_{n=0}^{\infty}(1-\rho)\rho^n \left[\dfrac{\mu(\mu w)^n e^{-\mu w}}{n!}\right]$……………………………………(19)

$\mu = (1 - \rho)e^{-\mu(1-\rho)w}$ (exponential distribution),

with

$E[\text{time in line}] = E[w] = \dfrac{1}{\mu(1-\rho)} = \dfrac{1}{\mu - \lambda}$, …………………………(20)

E[time in line] = E[time in system] – E[service time] =

$\dfrac{1}{\mu}\left(\dfrac{\rho}{1-\rho}\right) = \dfrac{\lambda}{\mu(\mu-\lambda)}$.………………………………………………(21)

For fixed ρ, the expected times in the system and in line vary inversely with the service rate μ. Suppose you look at only the time that a customer waits in line, where we exclude the person's service time as well as that of any customers who arrive when the server is idle, then it can be proved that the *conditional* probability density of the *time spent in line* by those customers who *do* have to wait in line is also given by (19) for any inter-arrival distribution, that is, for the model *GI/M/*1. Consequently, *E[w]* in (20) is also the *conditional expected time waiting in line* given that a customer *does* have to wait in line.

The steady-state difference equations (7) and (8) still apply for $n = 0, 1,\ldots, M - 1$, but the equation for $n = M$ is

$0 = \lambda P_{M-1} - \mu P_M$. ………………………………………………….(22)

The corresponding solution for $n = 0, 1,\ldots, M$ *is*

$$P_n = \begin{cases} \left(\dfrac{1-\rho}{1-\rho^{M+1}}\right)\rho^n & \text{for } \lambda \neq \mu \\ \dfrac{1}{M+1} & \text{for } \lambda = \mu \end{cases}$$ ……………………..………………(23)

Of course, when $\lambda < \mu$ and $M \longrightarrow \infty$, P_n in (23) agrees with (12). By elementary calculations it can be shown that for $\lambda \neq \mu$

$$E\begin{bmatrix}\text{number of}\\ \text{customers}\\ \text{in system}\end{bmatrix} \equiv E[n] = \frac{\rho}{(1-\rho)}\left[\frac{1-(M+1)\rho^M + M\rho^{M+1}}{1-\rho^{M+1}}\right]$$

$$= \frac{\rho}{1-\rho} - \frac{(M+1)\rho^{M+1}}{1-\rho^{M+1}} \quad \text{for } \lambda \neq \mu. \qquad (24)$$

Observe when $\lambda < \mu$, the expected number of customers in this system is smaller than for the previous case of unlimited line length (13). Similarly, it can be proved that $\lambda = \mu$

$$E\begin{bmatrix}\text{number of}\\ \text{customers}\\ \text{in system}\end{bmatrix} \equiv E(n) = \frac{M}{2} \quad \text{for } \lambda = \mu \qquad (25)$$

A delicate question arises in *defining* the amount of time a customer spends in the system (and waiting in line). If a customer arrives when there are already *M* persons in the system, the customer does not enter and consequently literally spends no time in the system. Hence average time spent in the system can be defined so as to refer either to all customers who arrive, regardless of whether they enter, or only to those customers who are permitted to enter. We adopt the latter, since in most situations the interest in delay time is only for those who actually do enter the system. So, referring to a customer arriving at an arbitrary moment who *does* join the system, and given that the discipline is first come, first served, it can be shown that

$$E\begin{bmatrix}\text{time in}\\ \text{system}\end{bmatrix} \equiv E(w) = \frac{\rho}{\mu(1-\rho)}\left[\frac{1-\mu\rho^{M-1} + (M-1)\rho^M}{1-\rho^M}\right] + \frac{1}{\mu}$$

$$= \frac{1}{\mu(1-\rho)} - \frac{M\rho^M}{\mu(1-\rho^M)} \quad \text{for } \lambda \neq \mu, \qquad (26)$$

and

$$E\begin{bmatrix}\text{time in}\\ \text{system}\end{bmatrix} \equiv E[w] = \frac{1}{\mu}.\ \frac{M+1}{2} \quad \text{for } \lambda = \mu \qquad (27)$$

4.0 Sensitivity Analysis:

Various operating characteristics of this simple queuing system are displayed in Table 1 for different values of traffic intensity ρ and service rate μ.

Notice that as the traffic intensity ρ increases, the expected values of the number of customers, the line length, the time in the system, and the time in line [formulas (13), (15), (20),] and (21) above] all increase rapidly. Although these quantities can be made arbitrarily large for sufficiently large $\rho < 1$, the system may take an accordingly long time to reach steady-state equilibrium. For a given service rate μ, when intensity ρ is small, most of the expected time a customer spends in the system is due to the average service time $1/\mu$; but as intensity ρ increases (that is, as the arrival rate λ increases), most of the expected time spent in the system is due to waiting in line.

For the sake of illustration, suppose the unit of time in Table 1 is an hour (or 60 minutes) and that $\rho = .8$, then on the average, the server is idle .2 hour (or 12 minutes per hour) and there are four persons in the system. If $\mu = 10$, so that the service rate is 10 per hour (or at the rate of six minutes per customer), then the average time a customer spends in the system is .5 hour (or 30 minutes), and .4 hour (or 24 minutes) of this is due to waiting in line. If ρ remains .8 but both the arrival rate and service rate double, so that $\mu = 20$, then the average times spent in the system and in waiting are cut in half.

Table 1: Operating Characteristics of *M/M/1* System.

Traffic Intensity ρ	Probability of Server Idle = $1-\rho$	Expected Number in System = $\frac{\rho}{1-\rho}$	Expected Line Length = $\rho^2/1-\rho$	$\mu = 10$			$\mu = 20$		
				λ	Time in System	Time in Line	λ	Time in System	Time in Line
.1	.9	.11	.01	1	.11	.01	2	.06	.01
.3	.7	.43	.13	3	.14	.04	6	.07	.02
.5	.5	1.00	.50	5	.20	.10	10	.10	.05
.7	.3	2.33	1.63	7	.33	.23	14	.17	.12
.8	.2	4.00	3.20	8	.50	.40	16	.25	.20
.9	.1	9.00	8.10	9	1.00	.90	18	.50	.45
.95	.05	19.00	18.05	9.5	2.00	1.90	19	1.00	.95
.99	.01	99.00	98.01	9.9	10.00	9.90	19.8	5.00	4.95
.999	.001	999.00	998.00	9.99	100.00	99.90	19.98	50.00	49.95

λ = arrival rate per unit of time. <u>*Note:*</u> *E*[length of Busy Period]= *E*[time in System]

μ = service rate per unit of time $\rho = \lambda/\mu$ traffic intensity

5.0 Conclusions

Queuing systems are prevalent throughout society. The adequacy of these systems can have an important effect on the quality of life and productivity. In this paper we have derived a single-server model with Poisson input and exponential service. This model imposed a constraint on the number of customers expected to wait in line at any given time. From the sensitivity analysis of the model, it is observed that as the traffic intensity increases, the number the number of customers the line length, the time in the line increases rapidly.

REFERENCES

[1] Siagian, P. (1987). *Penelitian Operational: Teori dan Praktek.* Jakarta: Universitas Indonesia Press.

[2] Obamiro, J.K. (2003). Queuing Theory and Patients' Satisfaction, An overview of Terminology & Application in Ante-natal Care Unit.http://www.upg-bulletin-se.ro.

[3] Adele, M., Barry, S., (2005). Modeling Patients Flow in Hospitals Using Queuing Theory, Unpublished Manuscript.

[4] Taha, H.A. (1987). An Introduction to *Operations Research,* 4th Edition. New York: Macmillan.

[5] Wolff, R. W. (1989) Stochastic Modeling and the Theory of Queues, Prentice Hall, Englewood Cliffs, NJ.

[6] Subaagyo, P.,Marwan, A. (1992) Operations Research, Yogakarta: BPFE

[7] Kakiay, T.J. (2004). *Dasar Teori Antrian Untuk Kehidupan Nyata.* Yogyakarta: Andi.

[8] Hillier, F.S. and Lieberman G.J. (2007). 'Introduction to Operations Research' pp. 790-805. Mcgraw-Hill Companies.

[9] Walrand, J. (1988) An Introduction to Queuing Networks, Prentice Hall, Englewood Cliffs, NJ.

[10] Chao, X., Miyazawa, M. and Pinedo, M. (1999) Queuing Networks: Customers, Signals and Product Form. Wiley, New York.

[11] Wagner, H.M. (2001). 'Principles of Operations Research' pp. 854-865. Prentice-Hall of India.

[12] Nordgren, B. (1999) The Problem with Waiting Times: IIE Solutions, Pp 44-48.

[13] Chen, H., and Yao, D. D. (2001) Fundamentals of Queuing Networks: Performance, Asymtotics and Optimization, Springer, New York.

[14] Kembe, M. M, Onah, E.S., Lorkegh, S.A., (2012). A Study of Waiting and Service Costs of a Multi-server Queuing Model in a Specialist Hospital, International Journal of Scientific and Technology Research, 5(2): 2277-8616.

[15] Nugraha, Dedi. (2013). *Penentuan Model System Antrean Kendaraan di Gerbang Tol Banyumanik.* Skripsi, FSM, Statistika, Universitas Diponegoro.