SIMULATION OF WAITING LINE SYSTEMS WITH QUEUE LENGTH DEPENDENT PARAMETERS

By

Irwin Greenberg

and

Susan Heimrath

New York University
Department of Industrial Engineering and Operations Research
Bronx, New York 10453

Abstract

A queueing system is simulated in which the arrival and service rates are functions of the number of customers in the system. This type of system has been suggested as being more representative of real queueing situations than is the usual "simple" queue with constant arrival and service rates.

A modification of the "time of next event" simulation method is required since an arrival to the queue changes the distribution of residual service time of the customer being served and the departure of a customer changes the distribution of the time of the next arrival.

The primary purpose of the simulation is to examine the statistical problems involved in estimating the system parameters.

I Purpose

The purpose of this study is to examine the suitability of an estimation scheme for the parameters of a queue with state dependent arrival and service rates. This scheme consists of obtaining maximum likelihood estimate of the arrival rate and service rate for each state and then making a least squares fit to the logarithms of the rates.

Because of the complexity of this scheme, it is not possible to obtain analytical results regarding the properties of the estimates. To overcome this the queuing system can be simulated and the results of the simulation used to estimate the known parameters. The behavior of the estimates can be studied as a function of the length of the simulation.

II General Background

The queuing system in which the arrival rates and/or service rates are dependent on the "state" that is on the number of customers present in the system at any instant, has been presented by Conway and Maxwell[1] for a single server and by Hillier, Conway and Maxwell[3] for several servers. The arrival rate, given that there are n customers in the system (being served and waiting for service), is

$$\lambda_n = \lambda \qquad , n \le S-1$$

$$= \left(\frac{S}{n+1}\right)^b \lambda \qquad , n \ge S$$

where S is the number of servers, λ the "normal" arrival rate (when an arriving customer sees that he will not have to wait) and b is a constant. If $b = 0$, the arrival rate is the same for all queue lengths. A value of $b > 0$ reflects the "balking" phenomenon: the reluctance of an arriving customer to join the queue when he will be forced to wait, the reluctance increasing with increasing queue length.

The departure rate with n customers in the system is

$$\mu_n = n\mu \qquad , n \le S$$

$$= \left(\frac{n}{S}\right)^c S\mu \qquad , n \ge S$$

where μ is the normal service rate (when there are no customers waiting to begin their service) and c is a constant. If $c = 0$ the service rate is the same for all queue lengths. A value of $c > 0$ reflects the tendency of a server to speed up when faced with a line of waiting customers. This tendency was noted in toll collectors by Edie[2] in his classic study "Traffic Delays at Toll Booths".

The terms "arrival rate" and "service rate" refer to the "birth-death" process. If there are m customers in the system at some time, the probability of an arrival in the very small interval of time Δt is

$$\lambda_n \Delta t + O(\Delta t)$$

where $o(\Delta t)$ represents terms of order greater than Δt. Similarly, the probability of a departure in Δt is

$$\mu_n \Delta t + o(\Delta t).$$

These assumptions are equivalent to assuming exponential inter-event distributions, or Poisson processes. If n customers are in the system, the time until the next arrival, given that it occurs before a departure, has the exponential distribution $\lambda_n \exp(-\lambda_n x)$. The time until the next departure, given that it occurs before an arrival, has the exponential distribution $\mu_n \exp(-\mu_n x)$.

The study by Hillier, Conway, and Maxwell[3] derives the probability of finding n customers in the system at some random instant in the steady-state:

$$P_n = P_o \frac{(\lambda/\mu)^n}{n!} \qquad , n \le S$$

$$= P_o \left(\frac{S^S}{S!}\right)^{1-b-c} \frac{\left(\frac{\lambda}{\mu S^{1-b-c}}\right)^n}{(n!)^{b+c}} , \quad n \ge S$$

with P_o obtained from the relationship

$$\sum_{n=o}^{\infty} P_n = 1.$$

Tables of P_o, of the average number in the system, and of the average number of customers in the system who have not yet begun their service are presented.

One of the major drawbacks to utilizing this model has been the difficulty in estimating the parameters, b, c, λ, and μ. Hillier and Lieberman[4] suggest the following method: for $n \geq S$, take logarithms of λ_n and μ_n to obtain

$$\log \lambda_n = b \log(\frac{S}{n+1}) + \log \lambda$$

$$\log \mu_n = c \log(\frac{n}{S}) + \log S\mu \ .$$

Observe the queuing system for time T. Let

T_n = amount of time that there are n customers in the system

A_n = number of arrivals which occur when n are in the system

D_n = number of departures which occur when n are in the system.

Obtain the maximum likelihood estimates.

$$\hat{\lambda}_n = A_n/T_n$$

$$\hat{\mu}_n = D_n/T_n \ .$$

Determine the estimates of b and log λ by making a least squares fit of the log λ_n equation, above, substituting $\hat{\lambda}_n$ for λ_n. The data for $n < S$ can be concentrated at $n = S-1$ with

$$\hat{\lambda}_{S-1} = \sum_{n=o}^{S-1} A_n / \sum_{n=o}^{S-1} T_n \ .$$

Similarly, the estimates of c and log Sμ are determined by making a least squares fit of the log μ_n equation, substituting $\hat{\mu}_n$ for μ_n. The data for $n < S$ can be combined with $n = S$ to obtain

$$\hat{\mu}_S = \sum_{n=1}^{S} D_n / \sum_{n=1}^{S} n\ T_n \ .$$

This is the estimation procedure examined by the simulation of the queue with state dependent parameters.

III The Simulation

The simulation technique used was a modification of the "next event" method. The modification was necessitated by the change in arrival and service rates that occur with each change in state. Thus, when a customer enters service, it is not possible to generate his departure time since it may depend on subsequent arrivals to the queue. The simulation was conducted as follows: immediately following a change in state to n, two random, exponential numbers were generated; the first with a mean of $1/\lambda_n$ and the second with a mean of $1/\mu_n$. If the first was the smaller of the two, this indicated that the next event was an arrival. This number was added to the clock time, the second number was discarded, and the state was changed to n+1. The value of T_n' was changed accordingly and A_n was increased by one. If the second random number was smaller,.the next event was a departure. The first number was discarded, the the clock was moved ahead by the appropriate amount, the state changed to n-1, T_n increased by the appropriate amount, and D_n increased by one.

This method of simulation can lead to situations which appear contrary to common sense. For example, consider an instant following a change in state. Random numbers are chosen yielding an inter-arrival time of 1.2 and a service time of 1.5. The clock is then moved ahead by 1.2 and the state is increased to n+1. Two new random numbers yield an inter-arrival time 1.8 and a service time 0.7. Hence, the departure occurs at 1.2+0.7 = 1.9 time units, measured from the first instant. This, in spite of the fact that the original "service.time" generated was 1.5 time units and the intervening arrival should speed up the server by increasing the service rate from μ_n to μ_{n+1}.

Despite this apparent anomaly, the simulation procedure is correct. This follows from the "loss of memory" property of the exponential distribution. An alternative method of performing the simulation would be to make use of a well known and easily derived rule for two Poisson processes. If two Poisson processes are operating simultaneously, the first with rate λ_n and the second with rate μ_n, then the probability that an event from the first process occurs before an event from the second is $\lambda_n/(\lambda_n+\mu_n)$. Thus, following each change of state,

$$q_n = \frac{\lambda_n}{\lambda_n+\mu_n}$$

could be calculated. A random number $0 \leq r < 1$ is chosen. If $r \leq q_n$, the next event is an arrival and the time of its occurrence is obtained from a random exponential variable with mean $1/\lambda_n$. If $r > q_n$, the next event is a departure and its time of occurrence is obtained from a random exponential variable with mean $1/\mu_n$.

Aside from this one feature, the remainder of the simulation program was quite simple and straightforward. It was written in FORTRAN and run on a Univac 1108 at the UHMC Computer Center at New York University. Summaries of A_n, D_n, T_n, and estimates of P_n were printed at times 100, 200, 300, ..., 900, 1000, 2000, 3000, ..., 10000, as were the estimates $\hat{\lambda}_n$ and $\hat{\mu}_n$, and the least squares estimates of b, c, λ and μ.

IV Some Results

Table 1 presents the results of a run of the simulation program for 10000 time units. The values of the parameters were b = 0.2, λ = 0.96, c = 0.2, μ = 0.4. A three-server queue (S=3) was simulated and hence Sμ = 1.2. There was nothing particularly significant in this choice of parameters, rather it was chosen as a "reasonable" set of values which one might expect to encounter in practice.

The results seem to indicate that the estimates of λ and Sμ are more consistent and accurate than the estimates of b and c. The estimate of c seems to indicate a bias on the high side while the estimates of λ, $S\mu$, and b seem to be biased on the low side, with b showing the worst performance. After seeing these results, it is possible to explain them although they were not predicted before the fact.

The errors in estimation are due, in large part, to the distortion caused by the logarithmic transformation of the data prior to the least squares fit. A data point lying beneath the fitted curve exerts a greater effect than a point lying above it. For example, assume that the regression curve passes through the point $x = 1$, $y = 1$. If, at $x = 1$, $y = 0.5$ and $y = 1.5$ were added as data points, the curve would not be disturbed as their effect would cancel each other out. If the curve and data was put on log-log paper, the $y = 0.5$ point would be further away from $y = 1$ than would the $y = 1.5$ point, resulting in a downward "pull" of the regression line. This accounts somewhat, for the consistently low estimates for λ and μ.

Table 1

Parameter Estimates as a Function of Time

$b = 0.2$ $\quad \lambda = 0.96$ $\quad S = 3$ $\quad c = 0.2$ $\quad \mu = 0.4$

	Estimates			
Time	b	λ	c	Sμ
100	1.227	1.117	.139	1.343
200	.791	.966	-.056	1.211
300	.568	.953	.117	1.253
400	.129	.908	.288	1.167
500	.150	.938	.257	1.164
600	.145	.931	.201	1.191
700	.083	.923	.223	1.189
800	.114	.948	.247	1.181
900	.160	.944	.293	1.143
1000	.163	.962	.324	1.135
2000	.218	.959	.387	1.114
3000	.142	.953	.304	1.151
4000	.120	.940	.195	1.195
5000	.152	.953	.196	1.195
6000	.157	.953	.208	1.186
7000	.148	.951	.222	1.180
8000	.144	.953	.230	1.163
9000	.121	.946	.210	1.166
10000	.140	.957	.198	1.180

The bias in the slope estimates can be attributed to the same phenomenon. In 10,000 units of time, approximately 78% was spent in state $n = 4$ or less. (The largest observed n was 17.) Since $\log \lambda_n$ varies with $\log \frac{S}{n+1}$ (and hence, small values of n correspond with large abscissa values of $\frac{S}{n+1}$), there is a heavy concentration of points at the upper portion of the log-log line. This concentration exerts the downward pull on the upper portion of the line yielding the low estimates for the slope b. Conversely, μ_n varies with $\log \frac{n}{S}$ and hence the heavy concentration at low values tends to depress the lower part of the line yielding high estimates for the slope c. The effect on c does not appear as strong as the effect on b.

V Conclusions

The results of the simulation indicate that the estimation procedure suggested by Hillier and Lieberman[4] for queues with state dependent parameters does not give accurate results. Substantial biases exist and the precision appears poor, even after a long period of time.

Unfortunately, the simulation did not suggest a better method of estimation. Although the queue dependent model obviously has validity, its utility is severely limited until the problems of statistical inference are solved. These problems-first, showing that dependence does exist, - second, showing that the dependence follows the assumed form, and third, estimating the parameters - are sufficiently complex to make an analytical solution most likely. The simulation model can act as the testing ground for whatever procedures can be suggested.

References

1. Conway, R. W. and Maxwell, W. L. (1961), "A Queueing Model with State Dependent Service Rate", Journal of Industrial Engineering, 12 pp. 132-136.

2. Edie, L. C. (1954) "Traffic Delays at Toll Booths", Journal of Operations Research Society of America, 2 pp. 107-138.

3. Hillier, F. S., Conway, R. W., and Maxwell, W. L. (1964), "A Multiple Server Queueing Model with State Dependent Service Rate", Journal of Industrial Engineering, 15 pp. 153-157.

4. Hillier, F. S. and Lieberman, G. J. (1967), Introduction to Operations Research, Holden-Day, San Francisco, page 327.