地域硏究, 第10卷 第1號
1994年 6月 pp.127～135

Simulation of Emergency Service Unit Dispatching Problems : A Test of the Efficacy of SLAM Ⅱ

Cheol-Joo Cho
Department of Regional Development
Chongju University

1. Introduction

Today cities have become a predominant form of habitats for human beings as a large proportion of the citizens live in cities or city-oriented metropolitan areas. The well-being of the people living in these high-density areas depends, to a great extent, on effective provision of a variety of urban services. Then, urban service systems with spatially distributed nature over urban area include emergency services(e.g., police, fire, ambulance, emergency repair), door-to-door pickup and delivery services(e.g., mail delivery, solid waste collection), neighborhood service centres(e.g., outpatient clinics, little city halls, libraries, social work agencies), and transportation services(bus and subway services, taxicab services, dial-a-ride systems).

Urban service delivery has been a popular problem for those interested in the urban service system. In attacking this problem, computer simulation has been extensively used as a method. A few examples include models for allocating emergency service facilities(Chaiken and Larson, 1972;Fitzsimmons, 1973;Kolesar and Blum, 1973;Savas, 1969;Toregas et al., 1971;Uyeno and Seeberg, 1984) and models for defining police patrol zone boundaries(Carroll and Laurin, 1981). These examples are all characterized by server-to-customer system, in which the customers and/or servers are distributed over the city or part of the city, and the servers travel to the customers for service.

The primary focus of this paper is to develop a SLAM Ⅱ(Simulation Language for Alternative Modeling) model for appraising the server-to-customer system, thereby to demonstrate that SLAM Ⅱ can very usefully be applied in the field of urban and regional planning. Police car patrol problem or emergency unit dispatching problem, among others, is considered most suitable to be handled by the model discussed in the paper.

In constructing the model, data from a hypothetical urban setting are used, since this paper attempts to test applicability of SLAM Ⅱ. The fact that hypothetical information is utilized, and that servers are moveable within a service region makes analytical solution nearly impossible, and hence impose some faults on performing the system science paradigm as a method by which model verification and validation can be processed(Shoderbeck et al., 1985)

2. System Structure

The number of queuing models, even with indistinguishable servers, is enormous. Even apparently minor change in assumptions can bring about a host of new analytical problems, thus requiring new analysis. This problem is much more severe in a spatial setting where the spatial distribution of servers and/or customers adds to the already nearly limitless number of possible system configurations and operating rules. This limitation leads to the simulation approach.

The demand-responsive spatial queuing systems of emergency services, for instance, police or emergency repair, operate as(Larson and Odoni, 1981) :

1. Requests for service are generated as a Poisson process in time form throughout the city or part of city.

2. A mobile server, called service unit or response unit, is assigned or dispatched to travel to the customer and provide on-scene service.

Here requests for service are geographically distributed throughout the service region. The service units may be prepositioned at a fixed location where service is completed when idle or free. The service units are dispatched to service requests as they arrive, and queued requests are handled according to some queue discipline. Service time consists of time to travel to the scene of the service request, and on-scene service time. It should be noted that service to a call begins as soon as the trip of the service unit toward the location of that call begins.

In building the simulation model following the rules of system operation mentioned above, it is assumed that (1) the service region is a 2-by-1 mile rectangle as shown in Figure 1, (2) a right-angle metric is in effect, with directions of travel parallel to the sides of the region, (3) the effective travel speed of service unit is constant and 20 mph, (4) calls for service are generated in a Poisson manner at the rate of three per hour, which is equivalent to exponential distribution with the mean of interarrival time of 20 minutes, (5) locations of calls for service are independent and uniformly distributed over the service region, and (6) the amount of time on the scene needed for the emergency service has the triangular probability density function.

There is no considerable difference in system's structure among versions of the model discussed here, even though some parameters concerned with system operation change. However, the model versions are different in logic according to the parameters adopted, and hence the codes of computer programs show great difference. Specifically, the number of emergency units introduced into system, and the queue selection rule, influence the complexity of logic of model. Then, four versions of the model are constructed to examine the effect of changes in some parameters on system performance. These four versions of the model are discussed in the next section.

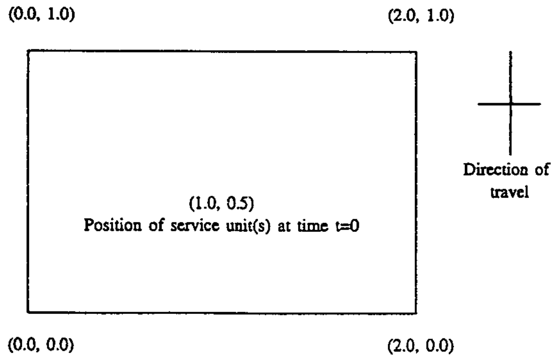

Figure 1. Service Region of Interest.

3. Parametric Model

As was discussed in the previous section, the parameters adopted determine the logic of each model version. The model is very complicated when two units for emergency service are operated and/or the shortest-travel-time queue discipline as a dispatching rule is adopted, compared to one unit and/or first-come-first-serve rule. This appears to be due to the characteristics intrinsic to the spatial queuing system, in which service units are moveable and the location of each call for service changes.

The four model versions are classified basically according to two dimensions shown in Figure 2. These versions are very effectively constructed through the SLAM Ⅱ discrete event modeling approach(Pritsker, 1986).

Figure 2. Classification of Model Versions

Criteria		The number of service units	
		1	2
Dispatching rule	First-come-first-serve rule	Model version I	Model version Ⅲ
	Shortest-travel-time rule	Model version Ⅱ	Model version Ⅳ

Version I is the most simplest version. In this version, single server is available, and if server is busy, calls for service queue up and are processed in a FCFS order. It is assumed that no calls are ever lost, no matter how long they have to wait. The index j is used to denote the call selected for service by the service unit's dispatcher. Associated with call j is the location of the call (xj, yj). The current location of the service unit is denoted by (xo, yo). Then the time needed to travel at 20 mph to a call at (xj, yj) once the unit completes service at a location (xo, yo) is given in minutes by

$$t = 3[(xj-xo) + (yj-yo)] \qquad (1)$$

Then, the total service time for each call is

$$\text{Service Time} = 3[(xj-xo) + (yj-yo)] + \text{TRIAG}(5., 10., 15., 1)_1 \qquad (2)$$

The SLAM Ⅱ logic for this version is simple to construct using equations (1) and (2).

Version Ⅱ is one in which one service unit is introduced and service is performed by following the shortest-travel - time rule which means that the call selected for service is one located nearest to the current location of service unit, when there are calls waiting for service. If the same notations as in Version I are used, then the travel time and the service time are expressed as in equation (1) and equation (2), respectively.

The SLAM Ⅱ logic for this version is much more complex, especially, end-of-service logic, because the program should be written to identify the call which is located nearest from the current location of service unit. This job is performed by making access to the values of attributes maintaining the locations of calls in waiting file, that is, ATRIB (2) and ATRIB (3)[2]. SLAM Ⅱ function subprogram, LOCAT(NRANK, IFILE)[3], is very useful for this task.

Version Ⅲ is different from the previous two versions in the sense that two service units are introduced into the system. Both units are at the center of the service region at time t = 0. From then on, whenever both units are available at times when calls for assistance are received, the unit closest to the location of the new call is dispatched there. When units are busy, arriving calls are placed in a waiting queue and served according to the FCFS discipline as in Version I. Concerned with the server selection rule, Version Ⅲ and Version Ⅳ, which will be discussed later, in a strict sense, share a common area rather than they are mutually exclusive. The travel time and the service time, defined in equation (1) and (2), respectively, are also applied here without any change.

Most complicated is Version Ⅳ. The complexity is due to the dispatching rule of service units as well as the number of service units used. The queue discipline always follows the shortest-travel-time rule, unlike in Version Ⅲ. Whenever both units are available at times when calls are received, the service unit nearest to the location of the new call is dispatched. When both units are busy, then arriving calls are filed into waiting queue. These rules are the same as used in Version Ⅲ. Then, the difference between the two versions comes when there are calls waiting for service and only one unit is available. Whenever this situation happens, the call closest to the service unit available is selected in Version Ⅳ, whereas the first call in the queue is chosen in Version Ⅲ. The job of identification of the call nearest to the idle unit, combined with identification of the available unit, makes the logic and computer program complicated. The travel time to the call for service and the service time are defined in the same manner as in the preceding versions. Four SLAM Ⅱ computer programs were written, one for each model version, based on the logic of the model versions.

4. Model Verification and Validation

As was seen so far, the model assumes that servers or service units are moving, service time is composed of moving time plus on-scene service time, which are randomly sampled from uniform distribution and triangular distribution, respectively. In these situations, the model can not be analyzed through classical queuing theory model. Especially, in the case that the nearest-point criterion of priority in service is employed, successive total service times are not independent and hence the analytical approach becomes impossible. To understand the interdependence of successive total service time, observe that, after a call which requires a long service time, it is more likely that the service unit will be dispatched to a relatively nearby point owing to the increased probability of several spatially distributed calls in queue.

The fact that the results produced by computer program can not be compared with the results solved by analytical method for a simple problem casts very serious problem in verifying models. Thus, unfortunately, verification through analytical solution can not be used. So another method is inevitably adopted. The method used here is like :

> Change the model parameter, the arrival rate of service calls, continuously, and then check whether the system performance produced accordingly, that is , the waiting time in the system, is reasonable, comparing those for the four versions considered.

Another parameter, on-scene service time, is considered to be changed in the process of experimental design discussed in the next section.

The results of this procedure are shown in Table 1. According to this table, there is not any evidence that the waiting time for each model version is not reasonable. Rather, it shows one interesting fact that for each model there exists a threshold in the arrival rate of service calls in terms of the waiting time. Specifically, when interarrival time reaches the mean of 12 minutes with the exponential distribution, then the response time tremendously jumps from 43.75 minutes to 474.77 minutes, for Version Ⅰ and Version Ⅱ. For Version Ⅲ and Version Ⅳ, this point comes at the rate of 6 minutes.

As the arrival rate per hour increases continuously beyond these points, computer program does not produce the values of response variable because the limit on spaces available in waiting file is reached during the run time. Only average waiting times in queue are obtained, as shown in Table 1. This means that the system with too high arrival rate is not possible to be simulated, using SLAM Ⅱ.

Table 1. Average Waiting Time in System

The number of calls/hour (Poisson)	1	2	3	4	5	6	10	15
Interarrival time(Exponent)	60 min	30 min	20 min	15 min	12 min	10 min	6 min	4 min
Version Ⅰ	15.02	17.75	23.00	43.75	474.77	554.34	–	–
Version Ⅱ	14.99	17.17	21.04	34.51	83.53	552.52	–	–
Version Ⅲ	10.12	10.12	10.13	10.33	11.34	12.85	221.48	–
Version Ⅳ	10.97	10.08	10.12	10.31	11.09	11.39	53.19	219.98

Notes : 1. The entries represent waiting time in queue.
2. Unit is minute.
3. On-scene time is sampled from TRIAG(5., 10., 15., 1)

Model validation, the procedure of checking whether the output from the computer simulation is consistent with the real world system, is not performed, because in this simulation study the historical data are not used. If real data could be obtained on the arrival rate of service calls, travel time, and on-scene service time, then it would be straightforward to measure the closeness of the probability distribution functions assumed in this study to the historical data. Square mean root, Chi-square, or visual comparison of histograms produced by the computer program with real data, could be used as criteria for the goodness-of-fit test(Kleijnen, 1974 ; Wilson, 1974).

5. Research Design and Experimentation

The four versions of the model, each of which is unique in the rule of system operation and in the number of service units, suggest some factors that should be considered in the designed experiment. That is, the service dispatching rule and the number of service units introduced into system are regarded as good candidates for factors in examining the sensitivity of system performance to policy changes. In addition to these two factors, on-scene service time is also chosen as a factor, since this time could be reduced or increased according to how many persons are employed for on-scene service.

For each of three factors selected, two levels are determined. The factors and levels selected for the experimentation are described in Table 2.

Table 2. Factors and Levels

Factors	Levels	Meaning
The number of service units	(1) 1	The number of service units introduced into system is 1.
	(2) 2	The number of service units introduced into system is 2.
Service unit dispatching rule	(3) FCFS	In the one unit system, selection of a call for service is done on the basis of the order of insertion in the waiting queue. However, in the two unit system, this means that when only one unit is available and also there are waiting both units are idle and no call is waiting, the more closely located unit from the call is dispatched.

Factors	Levels	Meaning
Service unit dispatching rule	(4) NCFS	Regardless of the one or two unit system, the call nearest from any available unit is chosen for the next service.
On-scene service time	(5) Low (Fast service)	The service time spent on the scene is a sample from a triangular distribution in the interval 5 min. to 15 min. with mode 10 min.
	(6) High (Slow service)	The service time on the scene is based on a triangular distribution in the interval 10 min. 50 20 min. with mode 15 min.

Using the factors and levels explained in Table 2, the number of cells for the experiment would be determined by the following formula

$$N = (q1)(q2)(q3) = 2x2x2 = 8$$

where

N = total number of cells in the experiment

q1 = the number of levels for the # of service unit

q2 = the number of levels for the service unit dispatching rule, and

q3 = the number of levels for on-scene service time.

The design matrix, which is showing the possible combinations of various levels and the policies resulting from those combinations, is expressed like in Table 3. The 2x2x2 design yields 8 possible policies or cells for evaluation.

Table 3. Policies and Their Components

Policy identification	Combination of levels(The numbers in the parentheses represent the level numbers in Table 2)
Policy 1	(1)-(3)-(5)
Policy 2	(1)-(3)-(6)
Policy 3	(1)-(4)-(5)
Policy 4	(1)-(4)-(6)
Policy 5	(2)-(3)-(5)
Policy 6	(2)-(3)-(6)
Policy 7	(2)-(4)-(5)
Policy 8	(2)-(4)-(6)

An important issue in experimental design is the problem of determination of run length and number of replications of the simulation. The use of a few long runs as opposed to many short runs generally produces a better estimates of the steady state mean because the initial bias is introduced with fewer times and less data truncated. However, the reduced number of samples corresponding to fewer replications may increase the estimate of variance of the mean(Kleijnen, 1974).

Considering the fact mentioned above, the run time of 7200 minutes(2 days) seems to be long enough to be justified. So relatively few replications are decided to be adopted in this experiment. Specifically, the last five of ten replications in each policy are chosen. The first five observations are thrown away because they are considered to be still in transient states. The output variable, that is, the response variable, is the average waiting time of a call spent in system.

A common random number stream is used for each run in an attempt to stabilize the variance between observations. This would be best known as a variance reduction technique.

Then, the data set, collected following

Table 4. Data Set

Policy	Replications				
(cell)	1st	2nd	3rd	4th	5th
Policy 1	23.00	23.99	22.85	19.35	24.56
Policy 2	57.44	58.61	67.38	63.99	64.45
Policy 3	21.04	22.46	20.54	20.26	22.54
Policy 4	54.88	47.67	54.60	65.81	53.12
Policy 5	10.13	10.37	10.21	10.46	10.29
Policy 6	15.84	16.47	15.31	16.24	16.21
Policy 7	10.12	10.38	10.19	10.40	10.29
Policy 8	15.65	16.40	15.27	16.19	15.82

Note : The entries represent average waiting time in the system(in minutes).

the rules outlined above, is summarized as in Table 4.

The one way ANOVA model is used to test simultaneously the differences between policies. The model in mathematical form is

$$X_{ij} = u + T_j + e_{ij}$$

where

X_{ij} = the average waiting time in the system of the i-th observation on the j-th policy

u = the unknown overall mean

T_j = the unknown policy j(treatment j) effect

e_{ij} = the random error present in the i-th observation of the j-th policy.

The hypotheses for test are

$H_0 : T_1 = T_2 = \cdots\cdots = T_8$

H_A : All T_j's are not equal.

The result of one way ANOVA shows that H_0 would be rejected. That is, there is no strong evidence that all treatment effects are equal at the significance level of .05. This implication, then, leads to the question of which policies are significantly different from others. This question is answered by performing the Duncan's multiple range test. According to the result of this test, 8 policies would be broken down into 5 homogeneous subsets of policies at .05 level.

Subsetl consists of policy5 and policy7, subset2 policy6 and policy8, subset3 policyl and policy3, subset4 policy4, and subset5 policy2, respectively, as shown in Table 5.

Table 5. Subsets and Their Constituent Policies

Subsets	Policies	Characteristics
Subset 1	Policy 5	2 units, FCFS dispatching rule, more service employees(fast on-scene service)
	Policy 7	2 units, NCFS dispatching rule, more service employees(fast on-scene service)
Subset 2	Policy 6	2 units, FCFS dispatching rule, less service employees(slow on-scene service)
	Policy 8	2 units, NCFS dispatching rule, less service employees(slow on-scene service)
Subset 3	Policy 1	1 units, FCFS dispatching rule, more service employees(fast on-scene service)
	Policy 3	1 units, NCFS dispatching rule, more service employees(fast on-scene service)
Subset 4	Policy 4	1 units, NCFS dispatching rule, less service employees(slow on-scene service)
Subset 5	Policy 2	1 units, FCFS dispatching rule, less service employees(slow on-scene service)

Among these subsets, subset4 and subset5, that is, policy4 and policy2, should be excluded from further consideration from the viewpoint that too long waiting time more than 50 minutes would be hardly justified in emergency situation. These two policies are characterized by one service unit and by slow on-scene service, which means that less service men are employed. In the systems in which one service unit and less employees are employed, dispatching rule seems to affect significantly the system performance.

Subsetl, composed of policy5 and policy7, appears to be most convenient to clients. However, 2 units and more service employees are needed. From the management point of view, these two policies are the most costly strategies. Service dispatching rule does not turn out to exert significant effects on system performance with 2 units and fast service system, when calls for service arrive at the rate of three per hour following the Poisson distribution.

Next, comparing subset2 with subset3, subset2 service units and less service men for on-scene service, whereas subset3 adopts one service unit and more service employees. These two subsets are in trade-off fo each other in terms of the number of service units and employees for on-scene service. In the long-run costs, the policies constituting these two subsets may be equalized. The system performance is, however, much more favorable for the policies in subest2(policy6 and policy8) than for the policies in subest3(policyl and policy3). Of course, in order for this claim to be justified, more long-range cost analysis, taking into account the initial investment and life time of service units, and wage rate, should be performed. As before, the dispatching rule does not have great influence on the response variable within each subset of policies. This means that the system performance is not so much sensitive to the dispatching rule as long as the arrival rate of service calls in maintained at 3 per hour with the Poisson distribution. If the arrival of calls becomes, however, more frequent, then things would not like this as was implied in the model verification and validation section.

Then, the shortest-travel-time rule needs more elaboration on the part of dispatcher of service units than the first-come-first serve rule does. Judging from limited information available, therefore, policy6 would be the recommended option rather than policy8.

6. Conclusions

So far, the emergency service system characterized by spatially distributed queues has been analyzed following the system science paradigm. Four alternative model versions are constructed according to the service unit resources available and dispatching rule by which the call for service is selected. Based on these versions, changes in the values of the response variable, the average waiting time in system, are monitored as employment policy for on-scene service varies.

Given the hypothesized information in this study, the system in which two service units are used, the call for service is chosen by first-come-first-serve rule, and relatively low level of employment for on-scene service is adopted, seems to be the most favorable option, taking into account the management point of view and simultaneously the convenience of clients.

By obtaining quite satisfactory results through the simulation on the system in question, the SLAM II discrete event modeling technique turns out to be a very effective tool for analyzing the service delivery problems occurring in urban areas. Once the data on the real world system would be obtained, and in addition, if the concept of probabilistic transportation network would be introduced in simulation the travel time, then more realistic simulation models could be created. This constitutes a promising research topic in the field of urban operations research in the future.

Endnotes

1. TRIAG(XLO, XMODE, XHI, IS) represents the SLAM II function from which samples are obtained in accordance with a triangular distribution in the interval XLO to XHI with mode XMODE using random number stream IS.
2. ATRIB(2) AND ATRIB(3) are the SLAM II variables. The vectors ATRIB(I) defines the attributes I of an entity as it flows through the network.
3. LOCAT(NRANK, IFILE) is a SLAM II subprogram to return the pointer to the

location of the entry whose rank is NRANK in file IFILE.

References

Carroll, J.M. and P.G.Laurin, 1981, "Using Simulation to Assign Police Patrol Zones," *Simulation*, 36.

Chaiken, J. and R.C.Larson, 1972, "Methods for Allocationg Urban Emergency Units : A survey," *Management Science*, 19.

Fitzsimmons, J., 1973, "A Methodology for Emergency Ambulance Deployment," *Management Science*, 19.

Kleijnen, J.P.C., 1974, *Statistical Techniques in Simulation*(Part1 and Part2), Marcel Dekker, New York.

Kolesar, P. and E.H. Blum, 1973, "Square Root Laws for Fire Engine Response Distances," *Management Science*, 19.

Larson, R.C. and A.R. Odoni, 1981, *Urban Operations Research*, Englewood Cliffs, N. J. : Prentice-Hall.

Pritsker, A. Alan B., 1986, *Introduction to Simulation and SLAM* Ⅱ, Systems Publishing, West Lafayette, Indiana

Savas, E., 1969, "Simulation and Cost-Effectiveness Analysis of New York's Emergency Ambulance Service," *Management Science*, 15.

Schoderbek, P.P., C.G. Schoderbek, and A.G. Kefalas, 1985, *Management System : Conceptual Considerations*(2nd ed), Business Publications, Plano, Texas.

Toregas, C., R.W.Swain, C.S.ReVelle, and L. Bergman, 1971, "The Location of Emergency Service Facilities," *Operations Research*, 19.

Wilson, A.G., 1974, *Urban and Regional Models in Geography and Planning*, John Wiley & Sons, New York.