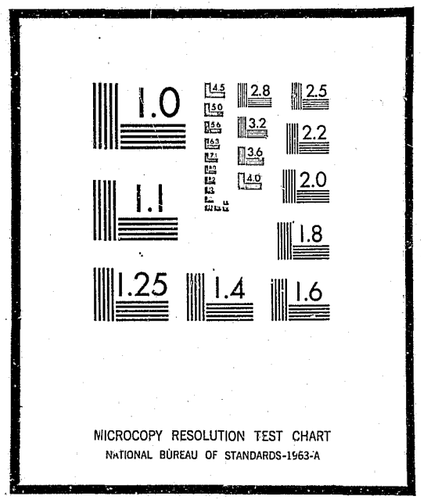
NCJRS



U.S. DEPARTMENT OF JUSTICE
LAW ENFORCEMENT ASSISTANCE ADMINISTRATION
NATIONAL CRIMINAL JUSTICE REFERENCE SERVICE
WASHINGTON, D.C. 20531

November 1974
Revised, February 1975

PP-03-74

APPROXIMATING THE PERFORMANCE OF URBAN EMERGENCY SERVICE SYSTEMS

by

Richard C. Larson

Preprint -- to appear in
Operations Research

"Innovative Resource Planning in Urban Public Safety Systems"

National Science Foundation Grant GI38004
Research Applied to National Needs
Division of Advanced Productivity, Research, and Technology

Operations Research Center
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

This paper presents an approximate procedure for computing selected performance characteristics of an urban emergency service system. Based on a recently developed hypercube queuing model , the procedure requires for N servers solution of only N simultaneous equations, rather than 2^N as in the exact model. The procedure relies on the theory of M/M/N queues in which servers are selected randomly and without replacement until the first available (free) server is found. The underlying model is intended for analyzing problems of vehicle location and response district design in urban emergency services, includes interdistrict as well as intradistrict responses, and allows computation of several point-specific as well as area-specific performance measures.

Recent analytical and simulation studies [2,3,12,13] have suggested ways of modeling certain spatially distributed emergency-service systems such as police, ambulance, and fire. While significant progress has been made in developing computer algorithms that calculate the performance characteristics of these systems from such models, there remains a strong need for more approximate methods that could be carried out either by hand calculation or by an easy-to-program computer algorithm.

The model presented in this paper can be used to analyze a number of resource allocation problems of urban emergency services, including (1) the "districting problem," (2) the "location" problem, and (3) the "workload balancing problem."

Given a region with a certain spatial distribution of demands for service and given N response units that are spatially distributed throughout the region, the districting problem is often stated as follows: "How should the region be partitioned into areas of primary responsibility (districts) so as to best achieve some level or combination of levels of service?" In the context of a spatially dispersed emergency ambulance service, a district for a particular ambulance would consist of a region in which calls for ambulance service would be handled by that ambulance, providing it is available when the call is received. If the district's ambulance is unavailable, then an out-of-district ambulance would be assigned by the ambulance dispatches. If all N ambulances should be simultaneously busy, then the dispatcher either enters the call in queue for later dispatch or refers the call to a back-up service (e.g., police department, other ambulance service).

In some cases more than one response unit may share the same district, thereby dividing the workload of the district; this occurs, for instance,

if several ambulances are garaged at the same location.

In the case of police patrol, our use of the word district applies to a police car's "sector" or "beat," which is the area that the car patrols while not responding to calls for service. The police car's sector may or may not correspond to the region in which the car has primary responsibility for responding to calls for service. Additionally, some cities provide "backup" cars to the regular sector cars, and these cars handle calls for service only when all the sector cars are simultaneously busy. Having no region of primary dispatch responsibility, the backup cars (which may be sergeant's cars or police wagons or other specialty units) often patrol a region covering several regular sectors.

The location problem, which is obviously closely related to the districting problem, is often stated as follows: "How should the N response units be located or positioned while not responding to calls for service?" In ambulance applications, it is usual to have one ambulance located in each of the N districts. Each location is fixed, corresponding to a garage, fire station house, point on a street, etc. In the case of shared disticts, two or more ambulances may be stationed at the same location. For police patrol, the "location" of each unit is mobile, corresponding to the areas that the unit patrols in its sector. In order to specify statistically the unit's location, one must know the relative amounts of time that it spends in various parts of its sector.

The workload balancing problem, which is in turn related to the districting and location problems, is as follows: "How should the units be positioned and selected for dispatch in order to balance (equalize) the workloads among units?" In effect, workload balancing may serve as an objective, perhaps one of several objectives, for the districting and location problems. Due to cross-district dispatches, it is important to note that workloads are not necessarily balanced by designing districts with equal internally generated workloads.

Thus, in urban emergency services, the analysis of districting, location, and workload balancing problems should include the possibility of overlapping (as well as disjoint) districts and mobile (as well as fixed) locations. Moreover, due to the dispatcher's desire to avoid delaying calls in queue, any analysis of these systems should include cross-district (or interdistrict) dispatches as well as intradistrict dispatches.

As applied to urban emergency services, most previous models and analyses [1,7,8,9,10,18,21,22] of districting and/or location problems have suffered from at least three deficiencies (see Note 1). First, most have focused solely on intradistrict responses of units, while ignoring interdistrict responses and design issues that relate to interdistrict response. Second, most previous studies have focused on only one performance measure (usually mean region-wide travel time or a closely related measure), thereby ignoring many other performance measures that characterize the operational effectiveness of these systems. Third, most previous studies have failed to incorporate the probabilistic nature of an urban emergency service system which is due to the Poisson nature of the call arrival process and the variability in service times.

Some recent work for small numbers of units has overcome many of the objections associated with the more traditional methods. Carter, Chaiken and Ignall [3] analyze the case of two fixed-position response units and rigorously derive the optimal districts for the two units, assuming a very general distance metric and a simplified form of interdistrict cooperation. The probabilistic and interdistrict behaviors of the system are fully

incorporated in their model. In addition, two measures of effectiveness are treated simultaneously: mean travel time and workload imbalance. Concurrent work by Larson and Stevenson [15] investigated several insensitivities of these and other location and districting models. But in all of this analytical work it has been very difficult to obtain results or define objectives for $N \geq 3$ units. Thus, although these models have provided useful insights into certain aspects of location, districting and workload balancing problems, they have not addressed computational problems that arise in practical situations with many response units.

The "hypercube" model represents a different approach to these problems [13]. Here, the multi-server queuing model employed in Refs. [3] and [15] that facilitates the study of probabilistic phenomena and interdistrict interaction is extended for up to $N = 15$, and the accompanying 2^N steady-state equations are solved numerically on a computer. Then, in an iterative user-interactive spirit similar to that represented in Refs. [19] and [20], the user can examine the numerical results and relocate and/or redistrict accordingly. In this iterative fashion, a very reasonable set of locations and districts can be found, incorporating a rich mixture of performance criteria.

There are many situations, however, in which a set of approximate solutions could suffice. For instance, data inaccuracies may not justify use of a highly precise model. Or the system planner may not have access to a sophisticated computer system necessary to perform calculations with the exact model. Or certain nonquantifiable concerns, perhaps involving political, legal, spatial, or administrative constraints, may play an important role in system design, thereby making precise estimates of quantifiable performance measures unnecessary.

In larger cities a system may exist having more than 15 cooperating units serving the city or a part of the city. In these cases the exact model would require very large amounts of computer storage and execution time, making costs of computation too great for most applications. The techniques presented in this paper would still be applicable, however, especially since it appears that the accuracy of the approximation improves (or at least does not degenerate) with increasing numbers of units. We find that the calculations in this paper are particularly convenient for the purpose of balancing workloads among units. Because the values of many performance measures, in addition to those for workloads, can be estimated with this method, it would seem that more complex applications (perhaps involving the reduction of inequities in the distribution of service) might be feasible.

The purpose of this paper is to present one simple iterative procedure for approximating the performance characteristics of such systems. The method had two advantages as compared to the exact analytical model it can replace: with N servers, it requires only N equations, rather than 2^N as is necessary in the exact model; and the calculations are often simple enough to be performed manually with the aid of an electronic calculator.

The measures of performance computed by the model include the following: region-wide: mean travel time, workload imbalance, and fractions of dispatches that are interdistrict dispatches; response unit specific: workload (measured in fraction of time busy servicing calls), mean travel time, fraction of responses of each response unit that are interdistrict, district specific: fraction of responses into each district that are interdistrict, mean travel time; point specific: mean travel time, fraction of calls handled by response unit n, $n = 1,2,\ldots,N$. This mixture of performance

measures allows one to focus simultaneously on several region-wide objectives while assuring that spatial inequities in the delivery of service are maintained at an acceptable minimum. Previously, values of these performance measures were available only from simulations, which are much more costly to execute than analytical procedures and are more difficult to interpret by decision makers due to problems of sample size and random statistical fluctuations.

The following are the main features of the approximation procedure:

1. One assumes that the dispatcher has a rank-ordered list of preferred units to dispatch to calls from each geographical unit (cell or atom) of the region and that he always dispatches the most preferred available (free) unit.
2. In addition, one assumes that the probability of dispatching the *jth* preferred unit to a call from a particular atom can be approximated to be proportional to the product of the utilization factors (or "workloads") of the first (j-1) preferred units and the availability factor (see Note 2) of the *jth* preferred unit.
3. The constant of proportionality depends on j and is determined by considering the simple M/M/N queuing model, assuming a situation in which j servers are selected randomly without replacement from the M/M/N system.
4. Given features 2 and 3, one can generate N simultaneous nonlinear equations relating the N unknowns (the utilization factors) to the dispatch policy and the call rates from the various geographical atoms.
5. The N simultaneous equations are solved iteratively, thereby yielding estimates of the workloads of the units.

6. If one desires other performance measures of the system (for instance, the mean travel time to each geographical atom or the fraction of dispatches that are cross-district), then the values of the utilization factors found in feature 5 may be used to estimate the fraction of dispatches that send unit i to atom j, for all i and j. These fractions are then entered into simple equations (detailed in Ref. [13]) to obtain estimates of the values of the desired performance measures.

To illustrate the ideas, a simple 3-server example is worked out in detail. Often the calculations are simple enough to be carried out by hand with the assistance of an electronic calculator.

The paper concludes with a general discussion of the observed error characteristics of the procedure. For most performance measures, the values estimated by the approximation procedure are within 2 percent of the exact values as derived by the hypercube model.

Reference [14] contains mathematical details relating to sampling servers without replacement in an M/M/N system. To assist in hand calculations, Reference [14] also contains tables of the values of the constant of proportionality.

A computer program, written in PL/I, which implements both the exact hypercube model and the approximation procedure, is documented in Ref. [15], and duplicate card decks are available from the M.I.T. Operations Research Center. Reports by Chelst [4] (in New Haven, Connecticut) and Jarvis and McKnew [4,17] (in Arlington and Wellesley, Massachusetts) focus on validity tests and implementation of these models. A preliminary case example using the hypercube model in Boston is reported by Larson [16].

I. REVIEW OF MODEL ASSUMPTIONS

Here we briefly review the assumptions of the model under consideration. A more extended discussion may be found in Ref. 13.

We assume that the system provides service to a certain geographical region that is broken down into K *cells* or *geographical atoms*. The fraction of region-wide workload generated from within atom k is $f_k (\sum_{k=1}^{K} f_k = 1)$. The mean travel time from atom i to atom j is denoted by τ_{ij}.

There are N response units that provide service to the region. The conditional probability that response unit i is located in atom j while available is $\ell_{ij} (\sum_j \ell_{ij} = 1)$. The ℓ_{ij}'s can be used to depict the location of mobile units, such a police patrol cars, in which case for each car i several ℓ_{ij}'s are likely to be nonzero, corresponding to the atoms in the car's patrol sector or beat. An ℓ_{ij} set equal to unity depicts a unit whose location, while available, is fixed in atom j (perhaps a firehouse or ambulance garage).

From a queuing point of view we assume that customers (*calls for service*) are generated from within the region in a Poisson manner, at a mean rate λ per hour, with each atom k acting as an independent Poisson generator with mean rate $\lambda f_k \equiv \lambda_k$.

If one is not concerned with the identity of busy servers, the queuing system is simply the M/M/N system, with either *zero-line capacity* (M/M/N/0) or *infinite-line capacity* (M/M/N/∞). The following assumptions are implied by the M/M/N model:

- Exactly one response unit is assigned to every call that is serviced;

- The service time of any response unit for any call for service has a negative exponential distribution with mean $1/\mu$ (see Note 3);
- The service time is independent of the identity of the server, the location of the customer, and the history of the system;
- For the zero-line capacity case, any call for service that arrives while all N response units are busy is either lost or (more likely in practice) serviced from outside the region or by special reserve units from within the region;
- For the infinite-line capacity case, any call for service that arrives while all N response units are busy is entered at the end of a queue of calls which is depleted in a *first-come, first-served* (FCFS) manner.

Given the geographical atom of the call, the dispatcher's selection policy is assumed to be one of "*fixed preference*." For such a policy one can always say that some unit i, if available, would be the first preference to dispatch to atom k, unit j would be the second preference, unit ℓ the third preference, etc. The dispatcher always selects the most preferred *available* unit.

Given the above assumptions, one can characterize the system as a continuous-time Markov process with 2^N states, corresponding to all combinations of servers busy and idle; in addition, if the system has infinite-line capacity, then the state space is augmented by an "infinite tail." Obtaining the steady-state probabilities of the system requires the solution of 2^N simultaneous linear equations, a formidable task even for many modern digital computers.

The approximation procedure derived here relies on the fact that the hypercube model is simply an M/M/N queuing system with a more finely

structured state space. In order to develop the detials of the procedure, we must first investigate certain properties of the simple M/M/N model when the identities of busy and idle servers are not required.

A summary of frequently used symbols is given in Table 1.

Table 1

SUMMARY OF FREQUENTLY USED SYMBOLS

N	Number of servers or response units.
K	Total number of cells or geographical atoms within the region being modeled.
f_k	Fraction of region-wide workload generated from within atom k ($\sum_k f_k = 1$).
τ_{ij}	Mean travel time from atom i to atom j.
ℓ_{ij}	Conditional probability that response unit i is located in atom j while available.
λ	Mean rate at which calls for service are generated from within the region.
λ_k	Mean rate at which calls for service are generated from atom k; $\lambda f_k = \lambda_k$.
μ^{-1}	Mean service time for any call for service.
S_k	State of an M/M/N queuing system indicating that exactly k servers are busy.
ρ	Equals $\lambda/N\mu$; called *utilization factor* for infinite-capacity system.
r	Fraction of time that each server is busy, averaged over all servers ($r = \rho = \lambda/N\mu$ for case of the infinite-capacity queuing system).
P_k	Probability that exactly k servers are busy, $k = 0,1,\ldots,N$.
B_j	Event that j*th* server selected is busy, $j = 1,2,\ldots,N$.
F_j	Event that j*th* server selected is free, $j = 1,2,\ldots,N$.
$Q(N,\rho,j)$	"Correction factor" for computing probability that the $j+1$*st* selected server is the first available server; given a total of N servers and a utilization factor ρ, $j = 0,1,\ldots,N-1$.
ρ_i	Fraction of time that unit i is busy servicing calls.

Table 1 -- Continued

W_i	Event that unit i is working (servicing a call).
R_i	Effective rate at which unit i is assigned to calls, given that unit i is idle.
R_i^T	Total rate (assignments per unit time) at which unit i is assigned to calls.
λ_D	Equals zero for zero-line capacity system; equals $\lambda P_N/N$ for infinite-line capacity system.
ρ_{ik}	Fraction of dispatches which send unit i to geographical atom k.
α_k	Exponential damping factor used in estimating the ρ_{ik}'s.

II. SAMPLING SERVERS WITHOUT REPLACEMENT IN AN M/M/N SYSTEM

M/M/N : INFINITE-LINE CAPACITY, FIRST-COME, FIRST-SERVED (FCFS), QUEUING SYSTEM

Consider the general M/M/N queuing system operating in the steady state. The following development assumes that the system has an infinite-line capacity, i.e., is an M/M/N/ system. Later, the analogous results are obtained for a zero-line capacity system (i.e., the M/M/N/0 system).

If the state S_k indicates the *exactly k servers are busy*, then the steady-state probabilities are given by

$$P\{S_k\} \equiv P_k = \frac{N^k \rho^k}{k!} P_o \qquad k = 1, 2, \ldots, N-1 ,$$

$$P\{S_N\} \equiv P_N = \frac{N^N \rho^N}{N!} \frac{1}{1-\rho} P_o , \tag{1}$$

$$P\{S_o\} \equiv P_o = \frac{1}{\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N \rho^N}{N!} \frac{1}{1-\rho}} ,$$

where, for the infinite-line capacity system, we assume

$$\rho \equiv \frac{\lambda}{N\mu} < 1 .$$

Defining r to be the fraction of time that each server is busy, averaged over all servers, for the M/M/N/∞ system we have $r = \rho$.

Now suppose we start randomly sampling *servers* in the system until we find the first server who is available or free (if there is one). Let

$B_j \equiv$ event that j*th* server selected is busy (not available).

$F_j \equiv B_j^c$ = event that j*th* server selected is free (or available).

$P\{B_1 B_2 \ldots B_j F_{j+1}\} \equiv$ probability that the first free server is the $j+1$*st* server selected.

The server selection process here is a strictly random sampling without replacement. We wish to derive an expression for $P\{B_1 B_2 \ldots B_j F_{j+1}\}$ that will motivate an approximation procedure for the more complicated hypercube model in which servers are not alike (see Note 4).

By laws of conditional probability we can write

$$P\{B_1 B_2 \ldots B_j F_{j+1}\} = \sum_{k=0}^{N} P\{B_1 B_2 \ldots B_j F_{j+1} | S_k\} P_k .$$

But

$$P\{B_1 B_2 \ldots B_j F_{j+1} | S_k\} = P\{F_{j+1} | B_1 B_2 \ldots B_j S_k\} P\{B_j | B_1 B_2 \ldots B_{j-1} S_k\} \ldots P\{B_1 | S_k\} .$$

Consider the conditional probability $P\{B_1 | S_k\}$. This is the probability that the first randomly selected server will be busy, given that a total of k servers are busy. Clearly,

$$P\{B_1 | S_k\} = \frac{k}{N} .$$

Given that the first selected server is found to be busy and that there are k busy servers,

$$P\{B_2 | B_1 S_k\} = \frac{k-1}{N-1} .$$

In general,

$$P\{B_i|B_1B_2\ldots B_{i-1}S_k\} = \frac{k - (i - 1)}{N - (i - 1)} \qquad i = 1, 2, \ldots, k + 1 .$$

Similarly,

$$P\{F_{j+1}|B_1B_2\ldots B_jS_k\} = \frac{N - k}{N - j} \qquad j = 0, 1, \ldots, k .$$

Combining these results we have the desired probability

$$P\{B_1B_2\ldots B_jF_{j+1}\} = \sum_{k=j}^{N-1} \frac{k}{N}\frac{k-1}{N-1}\cdots\frac{k-(j-1)}{N-(j-1)}\frac{N-k}{N-j}P_k ,$$

$$j = 0, 1, \ldots, N - 1 . \quad (2)$$

Rewriting Eq. (2), we have

$$P\{B_1B_2\ldots B_jF_{j+1}\} = \sum_{k=j}^{N-1} \frac{k}{N}\frac{k-1}{N-1}\cdots\frac{k-(j-1)}{N-(j-1)}\frac{N-k}{N-j}\frac{N^k\rho^k}{k!}P_o \quad (3)$$

$$= \left[\sum_{k=j}^{N-1} \frac{(N-j-1)!(N-k)}{(k-j)!}\frac{N^k}{N!}\rho^{k-j}\right]\rho^j(1-\rho)\left[\frac{P_o}{1-\rho}\right].$$

or,

$$P\{B_1B_2\ldots B_jF_{j+1}\} = Q(N,\rho,j)\rho^j(1-\rho) , \quad (4)$$

where

$$Q(N,\rho,j) \equiv \frac{\sum_{k=j}^{N-1} \frac{(N-j-1)!(N-k)}{(k-j)!}\frac{N^k}{N!}\rho^{k-j}}{(1-\rho)\left[\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i\right] + \frac{N^N\rho^N}{N!}} , \qquad j = 0, 1, \ldots, N - 1 . \quad (5)$$

It is convenient to isolate the term $Q(N,\rho,j)$ for the following reason: One may argue that by randomly selecting servers in a sequential manner, without replacement, the probability of each being busy is simply ρ and thus the probability that the $j + 1st$ is the first available server is simply $\rho^j(1 - \rho)$. Such an argument assumes independence among servers. The factor $Q(N,\rho,j)$ indicates the extent to which the result of the independence argument must be "corrected" in order to obtain the exact result.

Since by conditional probability

$$P\{B_1B_2\ldots B_jF_{j+1}\} = P\{F_{j+1}|B_1B_2\ldots B_j\}P\{B_j|B_1B_2\ldots B_{j-1}\}\cdots P\{B_1\} ,$$

we can write

$$Q(N,\rho,j) = \left[\frac{P\{F_{j+1}|B_1B_2\ldots B_j\}}{1-\rho}\right]\left[\frac{P\{B_j|B_1B_2\ldots B_{j-1}\}}{\rho}\right]\cdots\left[\frac{P\{B_1\}}{\rho}\right] . \quad (6)$$

Each of the terms in this product can be considered to be a "correction factor" indicating the relative amount by which ρ or $1 - \rho$ overestimates (or underestimates) the respective conditional probabilities of being busy or free. Checking the function Q for a limiting case, direct computation shows that $Q(N,\rho,0) = 1$, indicating that the probability that the first selected server is free is exactly $1 - \rho$. To investigate the case of the second, third, and in general the $j + 1st$ selected server, we require the following inequality:

$$P\{F_{j+1}|B_1B_2\ldots B_j\} < P\{F_j|B_1B_2\ldots B_{j-1}\} \leq 1 - \rho$$

$$j = 1, 2, \ldots, N - 1 , \quad (7)$$

where the right-hand inequality is an exact equality only for the case $j = 1$. This result, which is proved in Reference 14, states that as more servers are found to be busy, the chance that the next selected server will be found to be free becomes less and less. Intuitively, the conditioning event that the first j selected servers are busy provides

information that the entire M/M/N system is in a relatively congested state. Obviously, for the complementary event, there is an analogous inequality:

$$P\{B_{i+1}|B_1B_2\ldots B_i\} \cdot P\{B_i|B_1B_2\ldots B_{i-1}\} \geq \rho$$

$$i = 1, 2, \ldots, N - 1 , \qquad (8)$$

where the right-hand inequality is an exact equality only for the case $i = 1$.

Combining the results of Eqs. (7) and (8) for the case of the *second selected server*, we have

$$Q(N,\rho,1) < 1 , \qquad (9)$$

reflecting the facts that $(1 - \rho)$ is an overestimate of $P\{F_2|B_1\}$ and ρ is an exact expression for $P\{B_1\}$, thereby making $\rho(1 - \rho)$ an overestimate of $P\{F_2B_1\}$. Some additional insight may be gained here by examining the revised state probabilities, given B_1. Direct calculation using conditional probabilities yields

$$P\{S_k|B_1\} = \frac{k}{N\rho} P_k \qquad k = 0, 1, 2, \ldots, N . \qquad (10)$$

Those familiar with the theory of random incidence in renewal processes will notice that the biasing toward states with greater numbers of servers busy is equivalent to the biasing one observes in a random incidence situation toward interrenewal gaps with greater durations. Thus, since the system is more likely to be in a relatively busy state, the second selected unit is less likely to be free, hence $P\{F_2|B_1\} < 1 - \rho$.

Continuing the above reasoning, we may be tempted to think that $Q(N,\rho,j)$ would be a monotonically decreasing function of j. However, this may not always be the case. Examining Eq. (6), we note that $Q(N,\rho,j)$ is a product of $j + 1$ terms, one equal to unity, another less

than unity, and the remaining $j - 1$ all greater than unity. If the "less-than-unity" term always dominates, then $Q(N,\rho,j)$ is indeed a monotonically decreasing function of j; if not, then $Q(N,\rho,j)$ is a unimodel function of j, reaching a minimum for some value of j, say j^o, and then increasing for all $j > j^o$.

The test for unimodality, which can be proved by examining first differences of $Q(N,\rho,j)$, is

$$\rho \lesseqgtr 1 - \frac{2}{N} . \qquad (11)$$

If $\rho < 1 - \frac{2}{N}$, then $Q(N,\rho,j)$ is unimodal. If $\rho > 1 - \frac{2}{N}$, then $Q(N,\rho,j)$ is monotonically decreasing.

Reference 14 contains a table of values of $Q(N,\rho,j)$ for N up to 15 for the M/M/N/∞ queuing system. For illustrative purposes, plots of $Q(8,\rho,j)$ are given in Fig. 1. Note in Fig. 1 that $Q(8,0.7,j)$ is a unimodal function of j, whereas $Q(8,0.8,j)$ is not. The test shown by Eq. (11) indicates that the critical value of ρ in this case is $\rho = 0.75$.

M/M/N/0: ZERO-LINE CAPACITY QUEUING SYSTEM

We now consider the case of the M/M/N/0 system. The line of reasoning is directly parallel. The steady-state probabilities of the M/M/N/0 system are given by

$$P\{S_k\} = P'_k = \frac{N^k \rho^k}{k!} P'_o \qquad k = 0, 1, \ldots, N \qquad (12)$$

$$P\{S_o\} = P'_o = \frac{1}{\sum_{i=0}^{N} \frac{N^i \rho^i}{i!}}$$

where

$$\rho = \frac{\lambda}{N\mu} < +\infty .$$

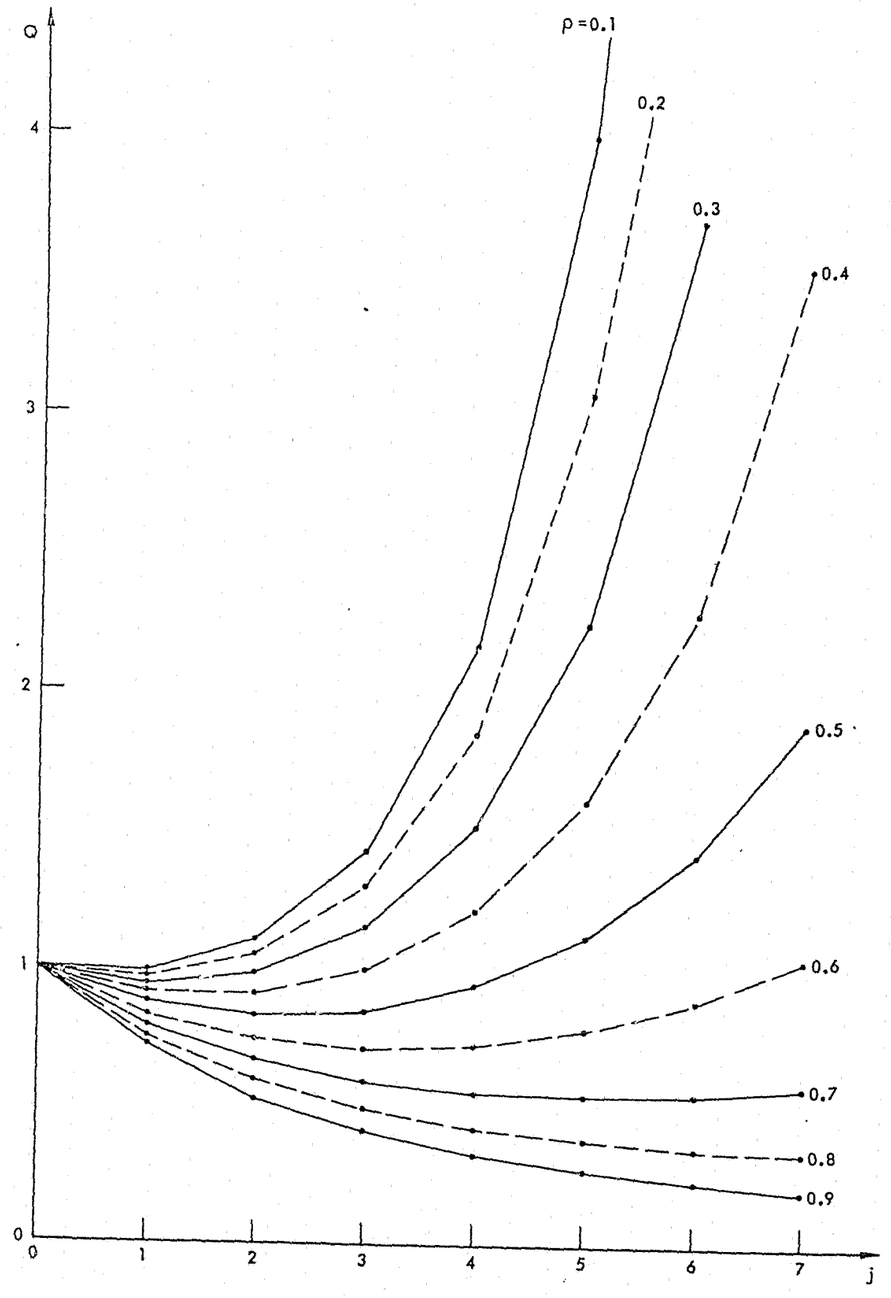


Fig. 1 — Graphs of Q (8, ρ, j)

In the M/M/N/0 system the actual fraction of time r that each server is busy is less than $\rho = \lambda/N\mu$, because of the fact that calls which arrive when all N servers are simultaneously busy are lost. In fact, the expression for r is easy to compute,

$$r = \frac{1}{N} \sum_{k=0}^{N} k\, P'_k = \rho(1 - P'_N) \ . \tag{13}$$

In this case we would like to develop a correction factor $Q'(N,\rho,j)$ that, when multiplied by $r^j(1 - r)$, gives the exact probability $P\{B_1B_2\ldots B_jF_{j+1}\}$ for the M/M/N/0 system.

Following the same reasoning that was used for the M/M/N/∞ system, we arrive at an expression for $P\{B_1B_2\ldots B_jF_{j+1}\}$ that is directly analogous to that obtained in Eq. (3). The result is

$$P\{B_1B_2\ldots B_jF_{j+1}\} \tag{14}$$

$$= \left\{ \left[\sum_{k=j}^{N-1} \frac{(N - j - 1)!(N - k)}{(k - j)!} \frac{N^k}{N!} \rho^{k-j} \right] \frac{P'_o}{1 - \rho} \left(\frac{1}{1 - P'_N} \right)^j \frac{1}{1 + \frac{\rho P'_N}{1 - \rho}} \right\} r^j(1 - r) \ .$$

Thus, if we define $Q'(N,\rho,j)$ as

$$Q'(N,\rho,j) = \frac{P\{B_1B_2\ldots B_jF_{j+1}\}}{r^j(1 - r)} \ , \tag{15}$$

then

$$Q'(N,\rho,j) = Q^*(N,\rho,j) \left(\frac{1}{1 - P'_N} \right)^j \frac{1}{1 + \frac{\rho P'_N}{1 - \rho}} \tag{16}$$

where $Q^*(N,\rho,j)$ is equal to $Q(N,\rho,j)$ as given in Eq. (3) but with P_o replaced by P'_o.

The function Q'(N,ρ,j), which is tabulated in Reference 14, has properties similar to those of Q(N,ρ,j). For illustrative and comparative purposes, plots of Q'(8,ρ,j) are given in Fig. 2.

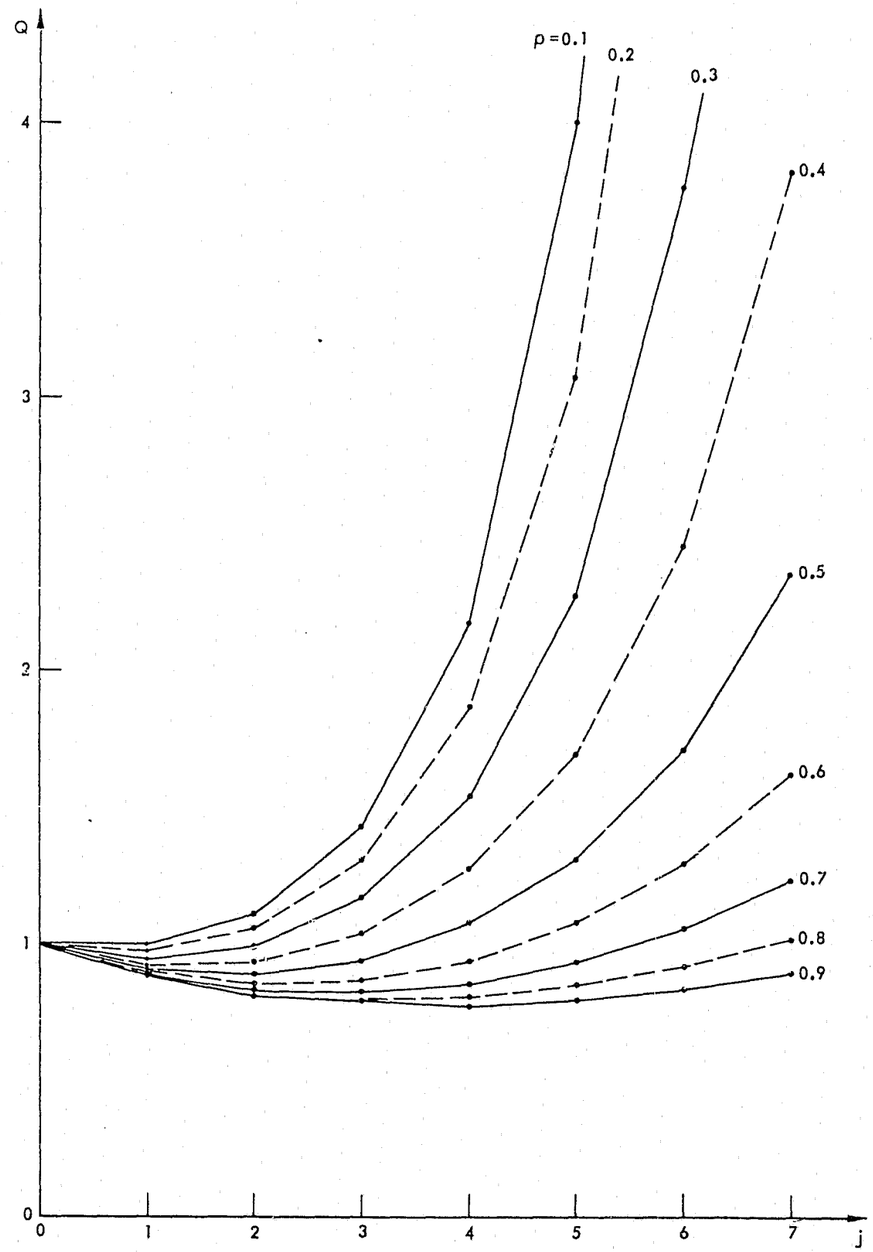


Fig. 2 — Graphs of Q'(8,ρ,j)

III. THE ITERATIVE PROCEDURE FOR ESTIMATING WORKLOADS

We can now derive an iterative procedure for estimating workloads of units. A second procedure using the workload estimates is then developed for estimating travel times, frequencies of cross-district dispatches, and other performance measures that can be obtained from the exact hypercube model. Both procedures rely on estimating dispatch probabilities as products of utilization and availibility factors and the appropriate correction terms as derived in the preceding section.

Let

$\rho_i \equiv$ fraction of time that unit i is busy servicing calls,

$i = 1,2,\ldots,N.$

We call ρ_i the *workload* of unit i. Define

$W_i \equiv$ event that unit i is working.

Clearly,

$$P\{W_i\} = \rho_i$$

$$P\{W_i^c\} = P\{\text{unit i is idle}\} = 1 - \rho_i \ .$$

For convenience we set $\mu = 1$, thereby equating the unit of time to the mean service time. Then for unit i we have the state transition diagram shown in Figure 3.

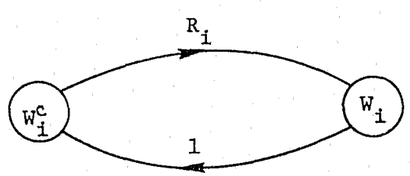


Figure 3: State Transition Diagram for a Single Unit

where

R_i = effective rate at which unit i is assigned to calls, given that unit i is idle.

(The "one" on the branch from state W_i to state W_i^c reflects the mean service rate of $\mu = 1$.) The difficulty in analyzing this 2-state process is that it is not a Markov or even a semi-Markov process. This is due to the fact that assignments of unit i from state W_i^c do *not* constitute a Poisson process with rate R_i.

However, by steady-state arguments,

$$P\{W_i\} = \rho_i = \frac{R_i}{1 + R_i} \ . \tag{17}$$

Thus the problem of determining ρ_i is reduced to the problem of determining R_i.

It is convenient to derive our approximation procedure in terms of

$R_i^T \equiv$ the total rate (assignments per unit time) at which unit i is assigned to calls,

and then, by recognizing that unit i is available for assignments only a fraction of time $(1 - \rho_i)$, to use the relation

$$R_i^T = R_i(1 - \rho_i) \ . \tag{18}$$

Let

$G_i^k \equiv$ set of geographical atoms for which unit i is the k*th* preferred dispatch alternative, i, k = 1, ..., N.

n_{ij} = identification number of j*th* preferred response unit for atom i.

Now an exact expression for R_i^T can be written as follows:

$$R_i^T = \sum_{j\in G_i^1} \lambda_j P\{W_i^c\} + \sum_{j\in G_i^2} \lambda_j P\left\{W_{n_{j1}} W_i^c\right\} + \sum_{j\in G_i^3} \lambda_j P\left\{W_{n_{j1}} W_{n_{j2}} W_i^c\right\}$$

$$+ \ldots + \sum_{j\in G_i^N} \lambda_j P\left\{W_{n_{j1}} W_{n_{j2}} \cdots W_{n_{j(N-1)}} W_i^c\right\} + \lambda_D , \qquad (19)$$

where the term λ_D accounts for delayed dispatches from a queue:

$$\lambda_D = \begin{cases} 0, \text{ for zero-line capacity case} \\ \frac{\lambda}{N} P_N , \\ \text{for infinite-line capacity case (see Note 5).} \end{cases}$$

The approximation we now make in order to simplify Eq. (19) is that the required dispatch probabilities can be estimated as products of utilization or availability factors and the appropriate correction term. For instance, we approximate

$$P\{W_3 W_6 W_5^c\} \approx Q(N,\rho,2)\rho_3\rho_6(1 - \rho_5) .$$

Given this assumption, Eq. (19) can be rewritten,

$$R_i^T = \sum_{j\in G_i^1} \lambda_j(1 - \rho_i) + \sum_{j\in G_i^2} \lambda_j Q(N,\rho,1)\rho_{n_{j1}}(1 - \rho_i)$$

$$+ \sum_{j\in G_i^3} \lambda_j Q(N,\rho,2)\rho_{n_{j1}}\rho_{n_{j2}}(1 - \rho_i) \qquad (20)$$

$$+ \ldots + \sum_{j\in G_i^N} \lambda_j Q(N,\rho,N - 1)\rho_{n_{j1}}\rho_{n_{j2}} \cdots \rho_{n_{j(N-1)}}(1 - \rho_i) + \lambda_D .$$

Using Eqs. (17), (18), and (20), we can now write the desired relationship,*

$$1 - \rho_i = \left[1 + \sum_{j\in G_i^1} \lambda_j + \sum_{j\in G_i^2} \lambda_j Q(n,\rho,1)\rho_{n_{j1}} + \sum_{j\in G_i^3} \lambda_j Q(N,\rho,2)\rho_{n_{j1}}\rho_{n_{j2}}\right.$$

$$\left. + \ldots + \sum_{j\in G_i^N} \lambda_j Q(N,\rho,N - 1)\rho_{n_{j1}}\rho_{n_{j2}} \cdots \rho_{n_{j(N-1)}} + \lambda_D/(1 - \rho_i)\right]^{-1}$$

$$i = 1, 2, \ldots, N . \qquad (21)$$

Equation (21) represents a set of N simultaneous nonlinear equations in the ρ_i's that can be solved iteratively.

A solution algorithm for the M/M/N/∞ model is given below. It depends on a convergence parameter ε which must be specified. The same algorithm can be used for the M/M/N/0 model, provided that $\{P_i\}$ is replaced with $\{P_i'\}$, the function Q is replaced with Q′, and λ_D is set to zero in Eq. (21).

Step 0: Initialization

a. Compute from the M/M/N queuing model the exact value for

$$r \equiv \text{average utilization factor} = \frac{1}{N}\sum_{i=1}^{N} \rho_i = \frac{1}{N}\sum_{k=1}^{N} k P_k .$$

$$= \begin{cases} \lambda/N \quad \text{for the M/M/N/}\infty \text{ system} \\ (\lambda/N)(1 - P_N') \text{ for the M/M/N/0 system.} \end{cases}$$

b. Set n = 0.

c. Define $\hat{\rho}_i(n) \equiv$ estimate of ρ_i at n*th* iteration.

Set $\hat{\rho}_i(0) = r, \quad i = 1, 2, \ldots, N.$

Step 1: Iteration

a. $n \leftarrow n + 1$.

b. For all i = 1, 2, ..., N compute $\hat{\rho}_i(n)$ from Eq. (21), using $\hat{\rho}_j(n-1)$ for ρ_j on the right side of the equation.

Step 2: Normalize (so that $\frac{1}{N}\sum_{i=1}^{N}\hat{\rho}_i(n) = r$)

a. Compute

$$\Gamma = \left[\frac{1}{N}\sum_{i=1}^{N}\hat{\rho}_i(n)/r\right]^{-1} .$$

b. $\hat{\rho}_i(n) \leftarrow \Gamma\,\hat{\rho}_i(n)$.

Step 3: Convergence Test (see Note 7)

$\underset{i}{\text{MAX}}|\hat{\rho}_i(n) - \hat{\rho}_i(n-1)| > \varepsilon$?

If yes, return to Step 1.

Otherwise, STOP.

This algorithm is very easy to program on a digital computer, and for small or moderate values of N and K, can be carried out manually with the aid of an electronic calculator.

IV. ESTIMATING OTHER PERFORMANCE MEASURES

All the remaining performance measures of the hypercube model can be calculated when the $\hat{\rho}_i$'s computed above are used to estimate

$\rho_{ik} \equiv$ fraction of dispatches which send unit i to geographical atom k.

For instance, the region-wide fraction of dispatches which are interdistrict dispatches is

$$= \sum_{i=1}^{N} \sum_{k \notin \text{district } i} \rho_{ik} .$$

The other algebraic formulas for travel times, cross-district dispatch frequencies, etc., are given in terms of the ρ_{ik}'s in Ref. 1.

To estimate ρ_{ik}, we use the same approximation we have used previously: dispatch probabilities can be approximated as products of utilization factors, availability factors, and correction factors.

THE INFINITE-LINE CAPACITY SYSTEM

Examining the M/M/N/∞ system first, we require an estimate for

$\rho_{ik}^{[1]}$ fraction of dispatches which send unit i to geographical atom k *and* incur no queue delay.

(The analogous term for dispatches that do incur a queue delay is $\rho_{ik}^{[2]} = f_k P_N/N$; see Note 5 and Ref. 13.) To estimate the ρ_{ik}'s, we use the ordering of dispatch preferences, and initially set

$$\hat{\rho}^{[1]}_{n_{kj}k} = f_k\, Q(N,\rho,j-1)\left[\prod_{\ell=1}^{j-1}\rho_{n_{k\ell}}\right]\left(1-\rho_{n_{kj}}\right). \qquad (22)$$

To calibrate the $\rho_{ik}^{[1]}$'s, we can obtain a set of normalization conditions from the M/M/N/∞ model,

$$\sum_{i=1}^{N}\hat{\rho}^{[1]}_{ik} = (1-P_N)\,f_k , \qquad k = 1, 2, \ldots, K , \qquad (23)$$

a set of equations which is not automatically satisfied by applying Eq. (22). There are numerous ways to accomplish the normalizations implied by Eq. (23), and the author has tried three.

1. The first is simply to scale, for each atom k, the results found by applying Eq. (22) so that Eq. (23) is satisfied. This results in larger-than-necessary errors in estimates of cross-district dispatch frequencies; the magnitude of the error, however (usually not greater than 0.05 in absolute value), may be acceptable for certain applications.
2. The second is to retain values for $\hat{\rho}^{[1]}_{n_{k1}k}$ found from Eq. (22) for all k = 1, 2, ..., K, and then simply to scale the remaining N - 1 values of $\hat{\rho}^{[1]}_{n_{kj}k}$ for each atom k. This results in very accurate estimates of cross-district dispatch frequencies, but slightly greater-than-necessary errors in travel time estimates.
3. The third is to find a factor α_k so that, if

$$\hat{\rho}^{[1]}_{n_{kj}k} = f_k \, Q(N,\rho,j-1) \left[\prod_{\ell=1}^{j-1} \alpha_k^{\ell} \, \rho_{n_{k\ell}} \right] \left(1 - \rho_{n_{kj}} \right),$$

then $\sum_{i=1}^{N} \hat{\rho}^{[1]}_{ik} = (1 - P_N) \, f_k$.

The quantity α_k is an "exponential damping factor," which, if greater than unity, will damp out at an accelerated rate the higher-order terms in Eq. (22); if less than unity, use of α_k will slow the geometric rate of decay of the higher-order terms in Eq. (22). The numerical value of α_k can be computed by a converging trial-and-error process. The author has found this method of normalization the most preferred in terms of minimizing approximation errors, but least preferred in terms of computational ease.

One's choice of a normalization method depends on balancing the demands for accuracy, on the one hand, with computational expediency, on the other.

THE ZERO-LINE CAPACITY SYSTEM

For the M/M/N/0 system we require an estimate for

ρ_{ik} = fraction of assignments that send unit i to geographical atom k (see Note 8).

By definition, none of the dispatches in an M/M/N/0 system incur any queue delay. Our estimate for ρ_{ik}, denoted $\hat{\rho}_{ik}$, is initially determined by applying Eq. (22) with Q replaced with $Q'/(1 - P'_N)$. However, we now have the normalization conditions,

$$\sum_{i=1}^{N} \hat{\rho}_{ik} = f_k , \qquad k = 1, 2, \ldots, K .$$

Each of the three normalization methods described above for the M/M/N/∞ system can be used for the M/M/N/0 system and, by and large, the same comments regarding accuracy and computational ease apply to the M/M/N/0 system.

V. EXAMPLE

To illustrate the calculations, consider the simple 3-district region illustrated by Fig. 4. The region consists of seven point geographical atoms served by three response units. Unit 1's district comprises atoms 1, 2, and 3; unit 2's district consists only of atom 5; and unit 3's district comprises atom 4, 6, and 7. While available, unit 2 is always prepositioned at atom 5, while units 1 and 3 are mobile; unit 1 is equally likely to be at atom 1 or 3 while unit 3 is equally likely to be at any of its district's three atoms. To summarize, the ℓ_{ij} matrix is given as follows:

$$\|\ell_{ij}\| = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

Regarding the distribution of calls for service, we assume that atom 5 generates 25 percent of all calls, while the remaining calls are uniformly distributed among the other 6 atoms. In other words, $f_5 = 0.25$ and $f_j = 0.125$ for $j \neq 5$.

We assume a strict center-of-mass dispatch selection policy [12], which yields the estimated travel distances shown in Table 2, assuming a right-angle or Manhattan distance metric. The fixed preferences implied in Table 2 yield the G_i^k sets given in Table 3.

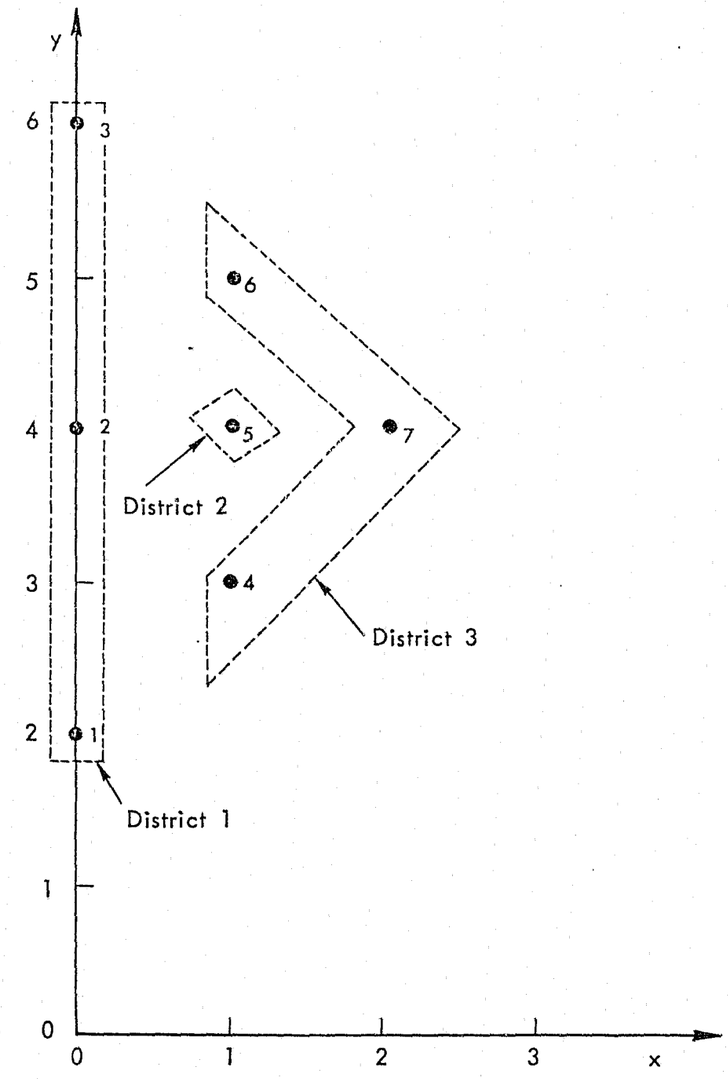


Fig. 4—A 3-district region

Table 2

MATRIX OF ESTIMATED TRAVEL DISTANCES:
STRICT CENTER-OF-MASS DISPATCHING

Atom Number	Unit Number		
	1	2	3
1	0.00	1.00	1.33
2	0.00	1.00	1.33
3	0.00	1.00	1.33
4	1.33	0.33	0.00
5	1.00	0.00	0.33
6	1.33	0.33	0.00
7	1.33	0.33	0.00

To conclude the description of this example, we assume that the system is a zero-capacity queue with $\lambda = 1.2$ (i.e., 1.2 calls per service time unit), or equivalently, $\rho = \lambda/3\mu = 1.2/3 = 0.4$ (see Note 9).

We are now ready to carry out the algorithm developed in Section III for estimating workloads of the individual units. First, note that District 1 and District 3 each generate 37.5 percent of the region-wide workload, while District 2 generates 25 percent. Thus, a procedure which claimed that ρ_i is proportional to the workload of District i would set $\rho_1 = \rho_3 = 0.375 \cdot C$ and $\rho_2 = 0.25 \cdot C$ for some constant C.

Table 3

MATRIX OF G_i^k SETS

k: Preference Number	i: Unit Number		
	1	2	3
1	1,2,3	5	4,6,7
2	---	1,2,3,4,6,7	5
3	4,5,6,7	---	1,2,3

Key: Entry in box (i,k) shows G_i^k = set of atoms for which unit i is the k*th* preference.

Now we execute the algorithm:

Step 0: Initialization

a. From the M/M/3/0 queuing model,

$$r = (\lambda/3)(1 - P_3') = 0.4(1 - 0.0898) \approx 0.3641.$$

b. $n = 0$.

c. $\hat{\rho}_i(0) = 0.3641, \quad i = 1,2,3.$

Step 1: Iteration

a. $n = 1$.

b. From tables in Reference 14,

$$Q'(3,0.4,0) = 1,$$
$$Q'(3,0.4,1) = 0.862,$$
$$Q'(3,0.4,2) = 0.887.$$

Applying Eq. (21) we get

$$1 - \hat{\rho}_1(1) = [1 + 0.375 + 0 + 0.625(0.887)(0.3641)^2]^{-1},$$
$$1 - \hat{\rho}_2(1) = [1 + 0.25 + 0.75(0.862)(0.3641)]^{-1},$$
$$1 - \hat{\rho}_3(1) = [1 + 0.375 + 0.25(0.862)(0.3641) + 0.375(0.887)(0.3641)^2]^{-1}$$

or,

$$\hat{\rho}_1(1) = 0.3096, \ \hat{\rho}_2(1) = 0.3268, \ \hat{\rho}_3(1) = 0.3322.$$

Step 2: Normalize

a. $\Gamma = 1.128$.

b. $\hat{\rho}_1(1) = 0.3491,$
$\hat{\rho}_2(1) = 0.3685,$
$\hat{\rho}_3(1) = 0.3746.$

Step 3: Convergence Test

For any reasonably small ε, the convergence test fails and we return to Step 1.

For $\varepsilon = 0.00033$ the procedure converges in two more iterations, yielding final workload estimates $\hat{\rho}_1 = 0.351$, $\hat{\rho}_2 = 0.367$, and $\hat{\rho}_3 = 0.374$. The actual workloads as computed from the hypercube model are $\rho_1 = 0.3548$, $\rho_2 = 0.3650$, and $\rho_3 = 0.3724$.

The maximum estimation error, MAX $|\hat{\rho}_i - \rho_i|$, is 0.0036, corresponding to a percentage error of less than 1 percent. The average percentage error is about 0.7 percent.

Other performance measures of this 3-server system are shown in Table 4, as computed both by the hypercube model and by the approximation procedure of Sec. V (using the third described normalization method). The average error in the amount of interdistrict dispatches is approximately 0.007, corresponding to an average percentage error of approximately 1 percent, while the average percentage error in the average travel distances is approximately 0.6 percent.

The workload calculations were performed on an electronic hand calculator in approximately 2 minutes. Computation of the remaining performance measures using normalization methods 1 or 2 (see Sec. IV) requires an additional 3 or 4 minutes. The third normalization method usually requires computer assistance.

Table 4

COMPARISON OF RESULTS COMPUTED FROM HYPERCUBE FORMULATION AND FROM THE APPROXIMATION PROCEDURE

Unit Number	Average Travel Distance	Fraction of Dispatches Out of District	
1	2.218	0.182	(exact value)
	2.203	0.169	(approximate value)
2	0.792	0.478	
	0.795	0.483	
3	.1.422	0.242	
	1.414	0.245	

District Number	Average Travel Distance	Fraction of Interdistrict Dispatches
1	2.142	0.291
	2.139	0.288
2	0.491	0.302
	0.479	0.305
3	1.450	0.311
	1.433	0.311

Atom Number	Average Travel Distance	Fraction of Calls from Atom Serviced by Unit Number		
		1	2	3
1	2.318	0.71	0.21	0.08
	2.314	0.71	0.21	0.08
2	1.790	0.71	0.21	0.08
	1.790	0.71	0.21	0.08
3	2.318	0.71	0.21	0.08
	2.314	0.71	0.21	0.08
4	1.419	0.09	0.22	0.69
	1.404	0.09	0.22	0.69
5	0.491	0.09	0.70	0.21
	0.479	0.09	0.69	0.22
6	1.419	0.09	0.22	0.69
	1.404	0.09	0.22	0.69
7	1.513	0.09	0.22	0.69
	1.491	0.09	0.22	0.69

Key: In each cell of the table the top entry is the exact value computed from the hypercube model; the bottom entry is the approximate value.

VI. DISCUSSION

While in our analyses of the error characteristics of the approximation procedures of Secs. III and IV are far from complete, the following general observations seem to hold.

First, the accuracies of both the workload approximation method and the ρ_{ij} approximation method seem to increase with the number of servers N, with error often averaging less than 1 or 2 percent. As an example, in one typical set of calculations for an M/M/8/∞ system with $\rho = 1/2$, the average errors (calculated as percentages) were 0.59 percent for workloads, 1.54 percent for cross-district dispatch frequencies, 1.55 percent for travel times of the units, and 1.73 percent for average travel times to individual atoms. (This set of runs used the second described procedrue for normalizing the ρ_{ij}'s.)

From a practical point of view, greater accuracy for larger N is just what we wish, since for small and moderate N we can "solve" the hypercube model exactly. The approximation method is practical, however, at least for machine computations, for N = 20, 30, or even 100. The hypercube model, requiring the solution of 2^N simultaneous equations, is not readily solved for N greater than about 15. The increased accuracy of the method for larger N is perhaps due to the fact that the random process generating calls for service for unit i becomes more and more like a Poisson process with rate parameter R_i. This is because the often dominant component of the process is an exact Poisson process, generated from the set of atoms G_i^1. Also, the other components represent the "pooling" of several individual processes, and large poolings often converge to a Poisson process with rate parameter equal to the sum of the individual rate parameters [5].

Second, the method usually converges quite rapidly. For small N, typically 2 or 3 or perhaps 4 iterations have been adequate. For $N \approx 10$, 4 to 6 iterations is usual, even for quite stringent convergence criteria. Thus, the method is well-suited for hand calculation for small and moderate N, but requires computer assistance for larger values of N.

Third, the method tends to be more accurate for systems in which no units are markedly different from others--either in the amounts of service demands they face or in the amount of area they cover. In part, this is due to the fact that the theoretical underpinning for the "Q" factors assumed that the workload was distributed uniformly among servers. One should not conclude that the method will fail to reveal large workload imbalances or differences in travel times; it will reveal them, but the estimated values of performance measures in such cases tend to have relatively larger errors.

For the task of balancing workloads among units it would seem to be particularly appropriate to use the approximation procedures developed here. In other less homogeneous situations, the procedures would appear to be valuable for providing a "first-cut" set of approximations. This may be all that is required or reasonable if the data estimates or the model assumptions are no more accurate than the numerical iteration procedures.

NOTES

1. Also, see the location theory bibliography by Francis and Goldstein [6].

2. The utilization factor of a unit is the fraction of time that the unit is busy servicing calls; the availability factor is the fraction of time the unit is not busy servicing calls.

3. Thus variations in service time that are due to variations in travel times are ignored.

4. In the hypercube model under a fixed-preference dispatching policy, the dispatcher always assigns the most preferred *available* server. Thus, the desired probability is the probability that the first j preferred servers are busy and the j + 1*st* is free.

5. For the infinite linc cap case, $\lambda_D = \frac{\lambda}{N} P_N$, using the facts that (1) all servers are assigned an equal proportion of the dispatches from queue, due to the FCFS queue discipline and (2) the fraction of calls that incur a queue is equal to the fraction of time that all units are simultaneously busy.

6. Equation (21) represents one of several ways of displaying the final result. Another is found directly from Eq. (20) by recognizing that $\rho_i = R_i^T$ (since $\mu = 1$).

7. This form of the convergence test could also be replaced with a test of relative errors or any one of numerous other reasonable convergence tests.

8. Since there is a zero-line capacity, it is important to keep in mind that the total rate of assignments is not equal to the total rate of calls for service, some of which are lost when all units are busy simultaneously.

9. This ρ is not the utilization factor or average workload per unit because of lost calls.

ACKNOWLEDGMENTS

The author wishes to thank J. M. Chaiken, T. B. Crabill, and E. Ignall, all from the Rand Corporation, as well as J. P. Jarvis, K. A. Stevenson, and K. Chelst of the Massachusetts Institute of Technology for their useful comments on an earlier draft. All computer computations were performed at the M I.T. Information Processing Center. This work was supported in part by the U.S. Department of Housing and Urban Development under a contract to the New York City Rand Institute and in part by the National Science Foundation (RANN, Division of Social Systems and Human Resources) under a grant to the M.I.T. Operations Research Center.

REFERENCES

1. W. J. BAUMOL and P. WOLFE, "A Warehouse-Location Problem," *Opns. Res.* 6, 252-263 (1958).

2. GREGORY CAMPBELL, "A Spatially Distributed Queuing Model with Application to Police Sector Design," Master's Thesis in Operations Research, Massachusetts Institute of Technology, Cambridge, 1972.

3. GRACE M. CARTER, JAN M. CHAIKEN and EDWARD IGNALL, "Response Areas for Two Emergency Units," *Opns. Res.* 20, 571-594 (1972).

4. KENNETH CHELST, *Testing the Hypercube Approximation Model in New Haven, Connecticut,* New York City Rand Institute (to appear).

5. D. R. COX, *Renewal Theory,* Wiley & Sons, Inc., New York, pp. 77, 103 (1962).

6. R. L. FRANCIS and J. M. GOLDSTEIN, "Location Theory: A Selective Bibliography," *Opns. Res.* 22, 400-410 (1974).

7. S. GASS, "On the Division of Police Districts Into Patrol Beats," In *Proceedings of the 23rd National Conference of the Association for Computing Machinery,* Brandon/Systems Press, Princeton (1968).

8. S. HAKIMI, "Optimum Locations of Switching Centers and the Absolute Centers and Medians of a Graph," *Opns. Res.* 12, 450-459 (1964).

9. S. HAKIMI, "Optimum Distribution of Switching Centers in Communications Networks and Some Related Graph Theoretic Problems," *Opns. Res.* 13, 462-474 (1965).

10. HESS et al., "Nonpartisan Political Redistribution by Computer," *Opns. Res.* 13, 462-475 (1965).

11. J. P. JARVIS, M. MCKNEW and L. DEETJEN, *Data Collection and Computer Analysis for Police Manpower Allocation in Arlington, Massachusetts* (to appear as an IRP Technical Report), Innovative Resource Planning in Urban Public Safety Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts.

12. RICHARD C. LARSON, *Urban Police Patrol Analysis*, Massachusetts Institute of Technology Press, Cambridge (1972).

13. RICHARD C. LARSON, "A Hypercube Queuing Model for Facility Location and Redistricting in Urban Emergency Services," *Computers and Operations Research* 1, 67-95 (1974).

14. RICHARD C. LARSON, *Urban Emergency Service Systems: An Iterative Procedure for Approximating Performance Characteristics*, Rand Corporation Report R-1493-HUD (1974).

15. RICHARD C. LARSON, *A Hypercube Queuing Model for Emergency Service Systems: User's Manual,* Innovative Resource Planning in Urban Public Safety Systems, Cambridge, Massachusetts (to appear).

16. RICHARD C. LARSON, "Illustrative Police Sector Redesign in District 4 in Boston," *Urban Analysis* 2, 51-91 (1974).

17. R. C. LARSON and K. A. STEVENSON, "On Insensitivities in Urban Redistricting and Facility Location," *Opns. Res.* 20, 595-612 (1972).

18. M. MCKNEW, *Testing the Validity of Hypercube-Type Models in Wellesley, Massachusetts* (to appear as an IRP Technical Report), Innovative Resource Planning in Urban Public Safety Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts.

19. L. C. SANTONE and G. N. BERLIN, *A Computer Model for the Evaluation of Fire Station Location,* National Bureau of Standards Report, U.S. Department of Commerce, Washington, D.C. (1969).

20. J. B. SCHNEIDER, "Solving Urban Location Problems: Human Intervention versus the Computer," *J. Am. Inst. Planners* 37, 95-99 (1971).

21. J. B. SCHNEIDER and J. G. SYMONS, *Locating Ambulance Dispatch Centers in an Urban Region: A Man-Computer Interactive Problem-Solving Approach,* RSRI Discussion Paper Series 49, Regional Science Research Institute, Philadelphia (1971).

22. R. D. SMITH, *Computer Applications in Police Manpower Distribution,* Field Service Division, International Association of Chiefs of Police, Washington, D.C. (1961).

23. C. TOREGAS, R. SWAIN, C. REVELLE and L. BERGMAN, "The Location of Emergency Service Facilities," *Opns. Res.* 19, 1363-1373 (1971).

END