ELSEVIER

European Journal of Operational Research 93 (1996) 110–120

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

Theory and Methodology

The Queueing Maximal Availability Location Problem: A model for the siting of emergency vehicles

Vladimir Marianov [a,*], Charles ReVelle [b]

[a] *Department of Electrical Engineering, Catholic University of Chile, Santiago, Chile*
[b] *The Johns Hopkins University, Baltimore, MD, USA*

Received November 1993; revised May 1995

Abstract

The Maximal Availability Location Problem (MALP) has been recently formulated as a probabilistic version of the maximal covering location problem. The added feature in MALP is that randomness into the availability of servers is considered. In MALP, though, it is assumed that the probabilities of different servers being busy are independent. In this paper, we utilize results from queuing theory to relax this assumption, obtaining a more realistic model for emergency systems: the Queueing MALP or Q-MALP. We also consider in this model that travel times or distances along arcs of the network are random variables. We show here how to site limited numbers of emergency vehicles, such as ambulances, in such a way as to maximize the calls for service which have an ambulance available within a time or distance standard with reliability α – using a queueing theory model for server availability. We also propose some extensions to the basic model. Formulations are presented and computational experience is offered.

Keywords: Emergency vehicles; Location; Congestion; Queueing theory

1. Introduction

Mathematical programming models for the siting of emergency services began their evolution, along with plant and warehouse location models, in the late 1960's and early 1970's. Among the early formulations was a group of models that either required the siting of a server within a distance or time standard of a demand node or sought the siting within the standard as a desirable goal. This is the lineage of models that is developed further in the present paper, but the interested reader can refer to ReVelle et al. (1977) for a fairly complete review of these early models.

In this paper, we develop a new probabilistic model for the siting of a single type of emergency service – of which the ambulance is the prototype. The model provided here is based on the well-developed ideas of queueing theory. This model represents a new and undoubtedly temporary endpoint in the gradual evolution of the original covering models. To understand the relationship between this model and its predecessors, we offer a brief verbal description of the constraints and objectives of the predecessor models.

The location set covering problem (LSCP) sought to position the least number of servers required to

* Corresponding author.

SSDI 0377-2217(95)00182-4

achieve coverage (at least one server within the time or distance standard) of all nodes of demand (Toregas et al., 1971). Since demand nodes were not differentiated by population or frequency of calls, and since all demand nodes needed coverage no matter how peripheral their position was, the number of servers required by the LSCP could well exceed the resources for servers available to a community or district. This condition motivated the development of the maximal covering location problem (MCLP) (White and Case, 1974; Church and ReVelle, 1974), a model which placed a limited number of servers in such a way as to maximize the population or calls which had a server stationed within the time or distance standard. Absent in these formulations was any consideration of congestion; the servers were assumed to almost always be available at the time that a call arrived.

During roughly the same period that these siting models were developed, Larson (1974, 1975) was developing a set of spatial queueing models, which calculated the steady-state busy fractions of servers on a network once their positions had been specified. Larson's hypercube and simplified hypercube models offered decision makers a way to assess the quality of any server placement scheme. Thus, a consideration of server congestion was begun in parallel with the line of research on server placement.

By the early 1980's, the problem of siting to cover demand begun to be extended to the issue of server congestion. Daskin and Stern (1981) restructured the LSCP to consider not simply needed first coverage but coverage in excess of the first coverage requirement. Their motivation in providing additional coverage within the standard was to increase the likelihood of a server continuing to be available within the standard even after one of the servers had responded to a call. Their model maximized the number of coverers in excess of first coverage subject to a constraint on the total number of servers available to the system. Eaton et al. (1986) suggested that weights, proportional to population, be placed on the number of redundant coverers for each of the demand nodes, thus placing emphasis on the presence of excess coverers for the most densely populated areas. Hogan and ReVelle (1986) showed how to maximize the population with *two or more* coverers given a requirement that all demand nodes have first coverage; that is, coverers beyond the first excess coverer were not counted as providing additional benefit. Eliminating the *requirement* for first coverage, they also showed how to trade off population which achieves two or more coverers against the population which achieves first coverage. All of these models focused on providing server availability even when one or more servers were busy, but they did so within a deterministic framework. All of these models were brought together and compared to each other in a review by Daskin et al. (1988).

Probabilistic models received their current impetus when Daskin (1983) suggested the maximum expected coverage location problem (MEXCLP). His model deployed a limited number of servers on the network in such a way as to maximize the expected population coverage; his assumption was that a uniform and calculable busy fraction existed for all servers in the system. His paper reviewed an unpublished probabilistic location covering model suggested by Chapman and White (1974); the model was essentially a concept which had never been fully implemented because of mathematical difficulties that arose from the problem statement. Specifically, the model utilized site-specific server busy fractions that could not be effectively estimated *a priori*. The problem formulation sought the least number and the positions of servers such that for each and every demand node the probability of a server actually being available from a location within the time standard was at least α.

ReVelle and Hogan (1989a) showed how that probabilistic location set covering model could be made operational by exploiting the structure of the probabilistic constraints and by assuming that the busy fraction was roughly constant in the local region around each demand node. Except for these two most important differences, the problem statement and conceptual constraints on reliability matched those of the original Chapman and White model. Subsequently, ReVelle and Hogan (1989b) converted the constraints on the reliability of server availability into desired rather than required conditions. In the same way that the maximal covering location problem distributed a limited number of servers in such a way as to maximize the population covered, the maximum availability location problem (MALP) al-

located a limited number of servers in such a way as to maximize the population with a server available within the time standard with α reliability. ReVelle and Marianov (1991) and Marianov and ReVelle (1992) extended the model framework to the fire protection system wherein two types of servers, engine companies and truck companies, are needed within their respective time standards for coverage to be achieved. Fire stations to house the vehicles need to be sited in addition. Their model seeks to distribute engines, trucks and stations in such a way as to maximize the population or calls for service which have an engine company available within an engine standard time or distance and a truck company available within a truck standard time or distance, either with independent availabilities α_E and α_T, or with a joint availability α.

In this paper, we show for the first time how queueing theory can be explicitly applied to the estimation of local-region busy fractions in the maximum availability location problem. The application of queueing theory to deriving these estimates of busy fractions within the siting model itself addresses a concern that was raised by the introduction of the MALP model (ReVelle and Hogan, 1989b) about the independence of the actual server availability. The new estimates of local busy fractions are then utilized to structure the constraints that define availability in a MALP-type model.

2. Review of MCLP and MALP models

The Maximal Covering Location Problem, MCLP (Church and ReVelle, 1974), was the first model that took account of both a limited number of servers and the fact that coverage of a node is achieved only if a server is positioned within a time or distance standard. The formulation seeks to site p facilities in such a way that the maximum population is covered at least by one of the facilities:

$$\text{Maximize} \quad Z = \sum_{i \in I} a_i y_i \tag{1}$$

subject to

$$y_i \leqslant \sum_{j \in N_i} x_j \quad \forall i \in I, \tag{2}$$

$$\sum_{j \in J} x_j = p, \tag{3}$$

$$x_j,\ y_i = 0,\ 1 \quad \forall j \in J,\ i \in I,$$

where:

$J =$ Set of eligible facility sites (indexed by j).

$I =$ Set of demand nodes (indexed by i).

$x_j = \begin{cases} 1 & \text{if a facility is located at node } j, \\ 0, & \text{otherwise.} \end{cases}$

$y_i = \begin{cases} 1 & \text{if node } i \text{ is covered,} \\ 0, & \text{otherwise.} \end{cases}$

$t_{ji} =$ Shortest time from potential facility site j to demand node i.

$S =$ The time or distance standard for coverage.

$a_j =$ Demand at node i.

$N_i = \{j \mid t_{ji} \leqslant S\}$; that is N_i is the set of nodes j located within the time or distance standard of demand node i, or the neighborhood of i. If a call for service originating at this node is answered by available servers stationed inside this neighborhood, it will be answered within the time or distance standard.

The objective (1) maximizes the sum of covered demands. Constraints (2) state that the demand at node i is covered whenever at least one server is located within the time or distance standard S. Constraint (3) gives the total number of facilities that can be sited. This last constraint can be relaxed to

$$\sum_{j \in J} x_j \leqslant p,$$

because an optimal value of the objective function will naturally push to the needed number of servers.

This model does not consider the possibility that a server may be busy at the time the call arrives, that is, the possibility of congestion. In order to improve this model by taking explicit care of congestion, ReVelle and Hogan (1989b) formulated the maximum availability location problem (MALP), which is the base for the model presented in this paper. In MALP, ReVelle and Hogan utilized the concept of a *local estimate of busy fraction*. By analogy to Daskin's estimate (1983) of a system-wide busy fraction, the local estimate of busy fraction in the service region around demand node i was given by

$$q_i = \frac{\bar{t} \sum_{k \in M_i} f_k}{24 \sum_{j \in N_i} x_j}, \tag{4}$$

where:

$x_j =$ Integer variable which takes the value m if m servers are positioned at location j, and zero otherwise.

$\bar{t} =$ Average duration of a single call, in hours.

$f_k =$ Frequency of calls for service at demand node k, in calls per day.

$M_i = \{k \mid t_{ki} \leqslant S\}$; that is the set of demand nodes located within S of node i.

If we assume that the neighborhood N_i is an isolated region with all its servers identical, the likelihood of any server in that region being busy is

$$q_i = \lambda_i/(\mu_i s).$$

The parameter λ_i is the region's arrival rate, $(1/\mu_i)$ is the single server's mean service time, and s is the number of servers in the region. The parameter λ_i can be represented as $\Sigma_{k \in M_i} f_k$, the parameter μ_i as $24/\bar{t}$, and s as $\Sigma_{j \in N_i} x_j$. The traffic intensity ρ_i is represented as

$$\rho_i = \lambda_i/\mu_i,$$

so that

$$q_i = \frac{\rho_i}{s} = \frac{\rho_i}{\sum_{j \in N_i} x_j}.$$

The probability that at least one server is available within time standard S when a call arrives from node i is simply one minus the probability that all servers within S of node i are busy. The MALP model utilized the binomial distribution to calculate this probability; that is,

$$1 - \mathrm{P}(\text{all servers of node } i \text{ are busy}) = 1 - q^{\Sigma_{j \in N_i} x_j}.$$

Hence, the requirement that the probability of at least one server being available within the time standard with reliability α is given by

$$1 - \left(\frac{\rho_i}{\sum_{j \in N_i} x_j}\right)^{\Sigma_{j \in N_i} x_j} \geqslant \alpha. \tag{5}$$

Although this probabilistic constraint does not have an analytical linear equivalent, it can be shown (ReVelle and Hogan, 1988) to be equivalent to

$$\sum_{j \in N_i} x_j \geqslant b_i, \tag{6}$$

where b_i is the smallest integer which satisfies

$$1 - (\rho_i/b_i)^{b_i} \geqslant \alpha.$$

In ReVelle and Hogan's 1989b MALP, coverage with reliability α is desired but not required, so Eq. (6) cannot be used. MALP is the probabilistic version of the MCLP; that is, its objective is, given a limited number of servers, to maximize the population with service available within the standard with a reliability α. ReVelle and Hogan presented two models. The first one (MALP I) utilized a system-wide busy fraction q, and the second one (MALP II), utilized the local busy fraction estimate described above. The formulation of MALP II is:

$$\text{Maximize} \quad Z = \sum_{i \in I} f_i y_{ib_i} \tag{7}$$

subject to

$$\sum_{k=1}^{b_i} y_{ik} \leqslant \sum_{j \in N_i} x_j \quad \forall i \in I, \tag{8}$$

$$y_{ik} \leqslant y_{i(k-1)} \quad \forall i,\ k = 2, 3, \ldots, b_i, \tag{9}$$

$$\sum_{j \in J} x_j = p, \tag{10}$$

$$x_j = \text{nonnegative integer} \quad \forall j \in J,$$

$$y_{ik} = 0, 1 \quad \forall i, k,$$

where the new variable y_{ik} is 1 if node i is covered by at least k servers within the time limit, and 0 otherwise. The variable y_{ib_i} is one if node i is covered by at least b_i servers. That is, it is one if node i is covered with reliability α, because the parameter b_i is the smallest number of servers that must be located within the service area of node i for node i to be covered with reliability α.

In objective (7), the number of calls covered at least b_i times is maximized, that is, covered with a reliability level of at least α. Constraint (8) implies that node i is covered b_i times only if at least b_i servers are stationed within the given time limits. Constraint (9) states that node i can not be covered k times if it is not covered $k-1$ times. Constraint (10) limits the total number of servers to p.

Note that the number of variables can be reduced by utilizing a variable z_i in place of y_{ib_i}. The model becomes

$$\text{Maximize} \quad Z = \sum_{i \in I} f_i z_i \tag{11}$$

subject to

$$\sum_{j \in N_i} x_j \geqslant b_i z_i \quad \forall i \in I, \tag{12}$$

$$\sum_{j \in J} x_j = p, \tag{13}$$

x_j = nonnegative integer $\forall j \in J$,

$z_i = 0, 1 \quad \forall i$,

where z_i is one if node i is covered by b_i servers. Numerical experience indicates, however, that this model has the tendency to give substantially more fractional results when solved using linear relaxation; consequently it will need more branching when branch and bound is utilized.

3. The Queueing Maximal Availability Location Problem

The principal and significant distinction between the model proposed here and the MALP model resides in the methodology for the calculation of the parameter b_i. As before, this parameter represents the minimum number of servers required to be stationed within the time or distance standard of node i. A second distinction between MALP and the present model lies in the treatment of travel distances or times. In this model we can view the distances/times as random and, as a consequence, derive possibly different sets of N_i, the set of the server positions eligible to serve node i by virtue of being within the time or distance standard of node i.

The parameter b_i is calculated here by treating arrival and service activities in the neighborhood around i as an M/G/s-loss (or M/G/s/s) queueing system (Poisson distributed call arrival rate, generally distributed service times, s servers in the neighborhood, and up to s calls being serviced at the same time). The assumption in the original MALP model of the independence of server busy fractions is thus avoided. Since we use queueing theory to arrive at busy fractions within the neighborhoods, we thus are able to account for the dependence between the probabilities of different servers being busy. The division of the region into neighborhoods also means that we do not have to track the state of each of the servers in the system, as Larson (1974) does with his queueing theory model.

Implicit in our model, as in MALP, is the assumption that the call rate in any neighborhood i does not differ to a significant extent from the call rate in the neighborhoods that border i. This suggests a rough equivalence between 1) the number of calls originating outside of N_i and requiring servers stationed inside N_i, and 2) the number of calls inside N_i which require servers to come from stations in adjacent, or nearby, neighborhoods. If we assume also that travel time within a neighborhood and to adjacent areas is small as compared to service time (an assumption born out in most urban systems), then there is little difference between the situation in which a server attends a call inside its area and the situation in which the server attends a call outside its neighborhood. It follows from these two assumptions on the spatial variation in call rate and the relative magnitude of travel time that the flows of servers into N_i and out of N_i are not too different, indeed approximately cancel each other. Such a situation would justify our treatment of each neighborhood as an isolated, independent unit whose demands and servers interact solely with each other.

This treatment is further supported by the conventional choice of the magnitude of the reliability level α. The value of α is always chosen close to one if the problem is at all meaningful. Thus, the number of servers chosen for placement within N_i after application of the model is greater than or equal to the number required to serve node i with α reliability. Nearly all calls from i will then find a server available within the neighborhood of the i-th node. That is, resort to servers from outside the neighborhood of i should occur only occasionally, making flow of servers across the boundary sufficiently small to be ignored.

These arguments and assumptions are made, as well, in previous models (Hogan and ReVelle, 1986; ReVelle and Hogan, 1988, 1989a,b), where they are also used to justify calculation of neighborhood-based busy fraction – but in a system that assumed independence of server availabilities. In the model presented here, we not only use region-specific busy fractions, but we also allow dependence between busy fractions at a local, neighborhood level. We thus offer an improvement over the total independence assumption. Further, these assumptions are far better than the next level of abstraction, the situation in which the probability of a server being busy is the same across the whole system (as in MEXCLP). We do not claim that the assumptions are perfect, but

they allow exact solution of an optimization model and they at least begin to approximate a spatial server system.

In addition to the above assumptions, we model calls for service (CFS) in neighborhood i as Poisson arrivals with intensity λ_i. General service time is also assumed with a mean rate of service completions per unit of time equal to μ_i, where the service time includes the travel time. In Section 2, we showed how these parameters are calculated. Each neighborhood is thus modeled as an M/G/s-loss system. When all the servers in a neighborhood are busy, new calls are presumed lost, relative to service in the local neighborhood (servers from outside the neighborhood take these calls).

Let s be the number of servers in the neighborhood. If we define the state k of the system as k servers being busy, the probability p_k of the system being in state k is computed by writing the following standard queueing theory steady-state equations:

$$\begin{aligned}&\text{P[getting into state } k]\\ &\quad - \text{P[getting out of state } k]\\ &= [\,p_{k-1}\lambda_i + (k+1)\mu_i p_{k+1}]\\ &\quad -[\,p_k\lambda_i + k\mu_i p_k] = 0 \qquad (14)\end{aligned}$$

for states 1, 2, 3, ..., s, and, for the state 0,

$$\mu_i p_1 - p_0\lambda_i = 0. \qquad (15)$$

Solution of these equations at steady-state yields the probability of all s servers being busy, p_s:

$$p_s = \frac{(1/s!)\rho_i^{b_i}}{1+\rho_i+(1/2!)\rho_i^2+\cdots+(1/s!)\rho_i^s}. \qquad (16)$$

This probability is a decreasing function of the parameter s. The recursive formula for p_s as a function of p_{s-1} illustrates this as the term in parentheses in the following equation is less than one:

$$p_s = \left(\frac{1}{p_{s-1}+s\mu_i/\lambda_i}\right)p_{s-1}. \qquad (17)$$

Now, the probability of at least one server being available in the region is $1-p_s$. For each neighborhood around demand node i and each value of s, we can compute the value of p_s, and if for that demand node, $1-p_s \geqslant \alpha$ or, equivalently, $p_s \leqslant 1-\alpha$, then we assume that node i will be covered with reliability α. As p_s is a decreasing function of s, there always exist a nonnegative integer b_i, such that for $s \geqslant b_i$, $1-p_s > \alpha$. This integer b_i represents, as in MALP, the minimum number of servers which must be located within the time or distance standard of node i, for that node to be considered as covered with reliability α. That is, b_i is the smallest integer that satisfies

$$\frac{(1/b_i!)\rho_i^{b_i}}{1+\rho_i+(1/2!)\rho_i^2+\cdots+(1/b_i!)\rho_i^{b_i}} \leqslant 1-\alpha.$$

Given a value for α and knowing the values of λ_i and μ_i, we can pre-compute, or determine exogenously to the optimization problem, this integer b_i. The value of b_i is calculated by determining $p_1, p_2, \ldots, p_s, p_{s+1}, \ldots$, etc. in sequence, and choosing as b_i the smallest value of s that satisfies the above inequation.

Analogously to MALP, to maximize the population or calls with α-reliable service, we maximize the population with b_i or more servers. Let $y_{ik} = 1, 0$; it is one if k or more servers are within S of demand area i, and it is 0 otherwise. Clearly, y_{ik} cannot be one unless $y_{i(k-1)}$ is also one, that is,

$$y_{ik} \leqslant y_{i(k-1)} \quad \forall i,\ k = 2, 3, \ldots, b_i$$

should be one of the constraints of the model.

To count coverers for each demand node, we write again constraint (8):

$$\sum_{k=1}^{b_i} y_{ik} \leqslant \sum_{j\in N_i} x_j \quad \forall i \in I.$$

The capacity C_j (in servers) of each location j is reflected in the model as an upper bound on the corresponding integer variable x_j.

There is an alternative formulation which considers only zero-one variables. Instead of integer variables x_j, variables x_{kj} are defined, which are one if a k-th server is located at site j, and zero otherwise. In this case, there are as many variables x_{kj} for each site j as possible places to be filled in the corresponding depot.

The full formulation in this case is

$$\text{Maximize} \quad Z = \sum_{i\in I} f_i y_{ib_i} \qquad (18)$$

subject to

$$\sum_{k=1}^{b_i} y_{ik} \leqslant \sum_{j \in N_i} \sum_{k=1}^{C_j} x_{kj} \quad \forall i \in I, \tag{19}$$

$$y_{ik} \leqslant y_{i(k-1)} \quad \forall i,\ k = 2, 3, \ldots, b_i, \tag{20}$$

$$\sum_{j \in J} \sum_{k=1}^{C_j} x_{kj} = p, \tag{21}$$

$$x_{kj},\ y_{ik} = 0, 1 \quad \forall i,\ j,\ k. \tag{22}$$

If instead of the weights f_i (call arrival rate from demand node i) we utilized no weights in Eq. (18), we would be maximizing the number of demand nodes which are covered by b_i or more servers, i.e. the number of demand nodes which have a server with reliability α. With the weights f_i added to the objective, we maximize the calls for service with service available with α reliability. Constraints (19) and (20) were already explained, and constraint (21) states that there are only p servers available to be located over the whole region. Constraint (22) forces all variables to be zero or one.

Without modifying the model presented in Eq. (18) to (22), an improvement might be introduced in the way N_i is computed, by considering the travel time or distance a random quantity. Particularly, we assume that travel times or distances along arcs of the network are random variables, and we choose the neighborhood of each node in such a way that, if a call for service originating at this node is answered by an available server located within the neighborhood, it will be answered within time standards with probability β. In this model we do not take into account the uncertainty regarding the identity of the first responding server (See Larson and Odoni, 1981, for an analysis of this subject).

As Daskin (1987), we assume that travel times are normally distributed. This assumption may be relaxed and any distribution used, provided that the inverse of its cumulative distribution function exists. We redefine our set N_i as

$$N_i\{j \mid P(t_{ij} \leqslant S) \geqslant \beta\},$$

that is, the set of possible server locations such that the probability of the travel time being within the time standard S is greater than or equal to some value β.

Let $\bar{t}_{ij}$ be the expected value and σ_{ij} be the standard deviation of t_{ij}. We can rewrite the condition for membership in the set N_i as a function of the zero mean, unit standard deviation variable $z = (t_{ij} - \bar{t}_{ij})/\sigma_{ij}$.

$$P\left(\left(t_{ij} - \bar{t}_{ij}\right)/\sigma_{ij} \leqslant \left(S - \bar{t}_{ij}\right)/\sigma_{ij}\right) \geqslant \beta,$$

that is,

$$P\left(z \leqslant \left(S - \bar{t}_{ij}\right)/\sigma_{ij}\right) \geqslant \beta,$$

or

$$F_z\left(\left(S - \bar{t}_{ij}\right)/\sigma_{ij}\right) \geqslant \beta, \tag{23}$$

where $F_z(x)$ is the normal cumulative distribution function. We can find the smallest value K_β such that $F_z(K_\beta) = \beta$ and, since $F_z(x)$ is a nondecreasing function, we can write the deterministic equivalent of Eq. (23):

$$\left(S - \bar{t}_{ij}\right)/\sigma_{ij} \geqslant K_\beta, \tag{24}$$

or

$$\bar{t}_{ij} + K_\beta \sigma_{ij} \leqslant S. \tag{25}$$

Eq. (25) becomes the new condition for membership in the set N_i, which is rewritten as

$$N_i = \left\{j \mid \bar{t}_{ij} + K_\beta \sigma_{ij} \leqslant S\right\}.$$

Given a value of β, we can compute the parameter K_β. Once the expected value and standard deviation of each travel time t_{ij} are known, we can determine which nodes j belong to N_i. Other forms of treating random travel times can be found in Larson and Odoni (1981).

The ability to calculate a neighborhood specific busy fraction, p_s, using queueing theory allows us to do more than specify the number of servers within N_i to achieve availability with α reliability. We can also constrain workload. Indeed p_s is itself workload. The constraint would be

$$p_{si} \leqslant w \quad \forall i \in I.$$

The smallest value of s (found iteratively as in the search for b_i) which achieves $p_{si} \leqslant w$, using Eq. (16), can be called g_i, where g_i is the minimum number of servers needed within N_i to assure that the workload in the neighborhood is less than or

equal to w. The constraint which enforces $p_s \leqslant w$ then is

$$\sum_{j \in N_i} x_j \geqslant g_i \quad \forall i \in I.$$

A constraint on workload in each neighborhood thus leads to a requirement on the number of servers stationed within the neighborhood.

Constraints also may be added which force all the demands to be covered at least by one server, or covered with a reliability $\alpha_1 (\leqslant \alpha)$. In other words, constraints may be appended which impose a minimum reliability standards for all call origins. These constraints, which look exactly like the workload constraints, would improve the degree to which equity is enforced in the model.

A workload constraint is likely to decrease the total number of calls which have a server available with α reliability. Alternatively, more servers could be needed to achieve the desired level of availability.

4. Computational experience

We used the 55-node test network of Swain, 1971, to test Q-MALP. We assumed that the server vehicles are ambulances. The population concentration at each node in the network was multiplied by a constant factor and used as an estimate of the number of calls per node per day. The resulting average of calls per node per day over the network is 0.4. An average duration of a single service, $1/\mu$ of 3/4 of an hour (45 minutes) was used. This figure was estimated considering the average of three cases: in the first one, the ambulance goes to the site of the call, stays there for some time, and then goes back to the depot. In the second case, the ambulance reaches the emergency site, takes a patient to a hospital and returns to its assigned depot. The third possibility is a false alarm, or the event that the emergency is over when the ambulance reaches the alarm site. The standard response distance was set at 1.5 miles.

With these values of the parameters, we first computed two values of b_i: using the method of ReVelle and Hogan's MALP and the methodology described in Section 3. Table 1 shows how many nodes have b_i equal to one, two, three or four, at varying levels of desired availability and using the two different forms of computing this parameter. It is interesting to note that the values of b_i in MALP and Q-MALP differ most significantly when preset availabilities approach very close to one (see the rows corresponding to availability 99%).

Table 1
Values of b_i for different levels of availability, for MALP and Q-MALP. $\lambda = 0.4$. Each entry shows the number of nodes which need the value of b_i in parentheses, when the availability of service for all nodes is forced to be at least the level indicated in the first column

Availability	MALP	Q-MALP
85	18 (1)	20 (1)
	37 (2)	35 (2)
90	12 (1)	14 (1)
	43 (2)	21 (2)
95	6 (1)	7 (1)
	26 (2)	24 (2)
	23 (3)	24 (3)
97	3 (1)	3 (1)
	24 (2)	22 (2)
	28 (3)	30 (3)
99	1 (1)	1 (1)
	19 (2)	17 (2)
	35 (3)	17 (3)
		20 (4)

Looking at the problem in a different way, we assumed that only one server is sited in each neighborhood, and then, computed for each neighborhood, the availabilities of servers when using MALP formulas and Q-MALP formulas. Table 2 shows how many nodes fall in each range of availability of servers, with only one server in each neighborhood, for MALP and Q-MALP.

Suppose we define m as the number of possible sites where stations may be located, and n as the

Table 2
Number of nodes in the different ranges of availability when only one server is located in each neighborhood, for Q-MALP and MALP

Availability	MALP	Q-MALP
under 0.45	1	–
0.45–0.55	20	–
0.56–0.65	7	1
0.66–0.75	2	28
0.76–0.85	7	6
0.86–0.95	12	13
0.96–1.00	6	7

total number of demand nodes. Despite the fact that the number of variables in Q-MALP is

$$\sum_{j=1}^{m} C_j + \sum_{i=1}^{n} b_j,$$

not all of them need to be declared (0, 1); i.e., need to be potentially branched on. In particular, if the variables y_{b_i} are declared (0, 1), and an upper bound of 1 is set on variables y_i, constraint sets (20) will force the variables $y_i (b < b_i)$ to be equal to 1 whenever the corresponding y_{b_i} is equal to 1. This is the only case of interest, because when y_{b_i} is equal to zero, we are not concerned about the value that those variables take.

The variables x_{kj} must be declared as (0, 1).

Thus, we need to declare as (0, 1) only $\sum_{j=1}^{m} C_j + n$ variables when solving Q-MALP.

Linear Programming Relaxation was utilized to solve the problems and Branch and Bound was applied when needed. A commercial linear and integer programming package (LINDO) was used on a VAX 750 computer. Lindo branches first on the integer declared variable with fractional value whose weight in the objective has the largest absolute value. Thus, the algorithm branches first on the variables y_{b_i} in decreasing weight order, and then on variables x_{kj} following the order given by the secondary objective

$$Z_2 = \sum_{j=1}^{m} \sum_{k=1}^{C_j} \frac{1}{k} x_{kj}.$$

This objective, included with a small weight, produces an ordering in the variables x_{kj}, that is, it makes the variable x_{1j} enter the solution before the variable x_{2j}, and $x_2 j$ before x_{3j}, and so on. Also, this objective was very helpful in decreasing the branching time of LINDO.

The parameter C_j was set as 3 in most of the runs, that is, each depot could house up to 3 vehicles.

Table 3 shows the set of results obtained. The numbers of vehicles that were used as input parameters to the runs are shown. Also shown are the server locations, and the percent of population (or calls) covered.

Figs. 1 and 2 show two of the solutions obtained, for 6 servers and availability 90%, and 4 servers and availability 85%, respectively.

Table 3
Computational experience with Q-MALP

Availability level (%)	Servers	Unconvered locations	Covered popul. (%)	Server locations
85	4	37, 43, 46, 50, 51, 52, 53	96.42	7, 22, 25, 43
	5	40, 51, 52	98.50	1, 11, 17, 18, 38 or 6, 18, 22, 38, 47
	6	–	100	7, 16, 18, 21, 43, 53
90	4	20, 37, 43, 46, 48, 50, 51, 52, 53, 55	94.20	7, 22, 25, 43
	5	37, 50, 51, 52, 53, 55	97.29	3, 22, 25, 43, 48 or 3, 22, 25, 43, 15 or 3, 22, 25, 43, 36
	6	40, 44	99.32	15, 16, 18, 37, 38, 53
	8	–	100	11, 14, 15, 18, 37, 43, 45, 50
95	4	12, 14, 16, 27, 28, 35, 38, 39, 40, 43, 46, 48, 49, 52, 54, 55	86.12	7, 7, 9, 37
	5	24, 26, 35, 39, 40, 46, 48, 49, 51, 52, 54, 55	92.51	9, 9, 11, 22, 22
	6	26, 35, 48, 51, 52, 55	96	9, 11, 22, 22, 25, 43
	8	–	100	16, 22, 23, 36, 43, 47, 49, 53

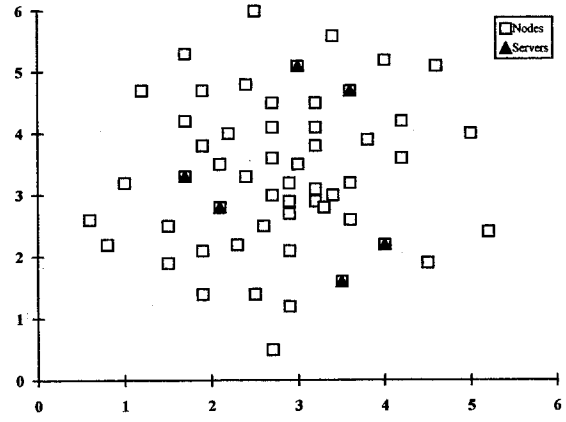

Fig. 1. Locations of servers relative to demand nodes for 6 servers, availability 90%. Coverage of calls = 99.32%. $C_j = 3$. Distances in miles.

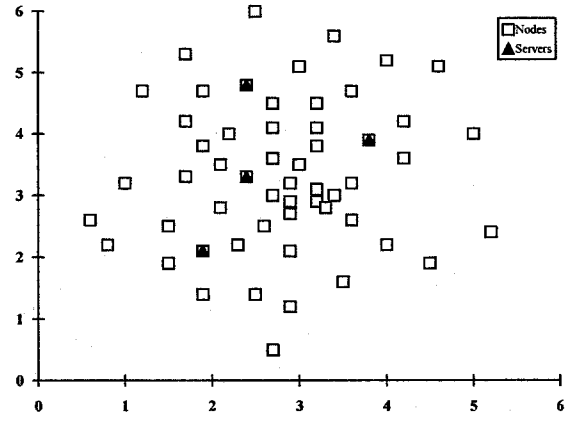

Fig. 2. Locations of servers relative to demand nodes for 4 servers, availability 85%. Coverage of calls = 96.42%. $C_j = 3$. Distances in miles.

Fig. 3 shows the trade-off between number of servers and percent of coverage.

5. Conclusions

A model is presented, in this paper, which seeks to maximize the population covered by emergency vehicles with availability α. A probabilistic formulation is structured in which availability is computed utilizing queueing theory. As opposed to former models in which the probabilities of different servers being busy in a neighborhood were considered as independent, in this formulation these probabilities depend on each other, which is achieved by the use of queueing theory. A method is shown to take into account the fact that travel times or distances are random, and computational experience is presented.

As an extension of this model, a hierarchical formulation can be easily developed to locate two or more types of servers. An example of application of this model would be the siting of Advanced Life Support (ALS) Ambulances and Basic Life Support (BLS) Ambulances.

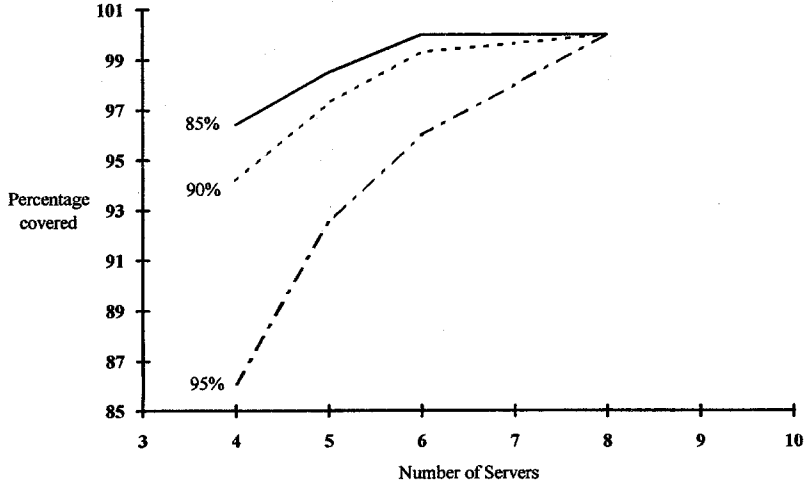

Fig. 3. Coverage of population vs. number of servers, for different availabilities.

From a practical point of view, it is interesting to comment on how the parameters of the model can be estimated from observation of the behavior of the system. The fact that the parameters λ_i and μ_i are random variables could be taken into account when the queueing differential equations are formulated and solved, by assuming that the arrival rates and service times are doubly stochastic processes (Snyder, 1975). However, unless the probabilistic distribution function of the random parameters is a simple one, the mathematical treatment of doubly stochastic processes is complicated, and it is most probably not justified in this case.

Acknowledgments

This research was funded by a NORTEL (External Research Department) grant, and by the Fondo Nacional de Ciencia y Technología, FONDECYT (Chilean National Science and Technology Fund).

References

Chapman, S.C., and White, J.A. (1974), "Probabilistic formulations of emergency service facilities location problems", Paper presented at the ORSA/TIMS Conference, San Juan, Puerto Rico.

Church, R., and Re Velle, C. (1974), "The Maximal Covering Location Problem", *Papers of the Regional Science Association* 32, 101–118.

Daskin, M.S. (1983), "A maximum expected covering location model: Formulation, properties and heuristic solution", *Transportation Science* 17, 48–70.

Daskin, M.S. (1987), "Location, dispatching and routing models for emergency services with stochastic travel times", in: A. Ghosh and G. Rushton (eds.), *Spatial Analysis and Location-Allocation Models*, Van Nostrand Reinhold, New York.

Daskin, M.S. and Stern, E.H. (1981), "A hierarchical objective set covering model for emergency medical service vehicle deployment", *Transportation Science* 15, 137–152.

Daskin, M.S., Hogan, and K. ReVelle, C. (1988), "Integration of multiple, excess, backup, and expected covering models", *Environment and Planning B: Planning and Design* 15, 15–35.

Eaton, D., Hector, M., Sanchez, V., Lantigua, R., and Morgan, J. (1986), "Determining ambulance deployment in Santo Domingo, Dominican Republic", *Journal of the Operational Research Society* 37, 113.

Hogan, K., and ReVelle, C. (1986), "Concepts and applications of backup coverage", *Management Science* 32, 1434–1444.

Larson, R.C. (1974), "A hypercube queuing model for facility location and redistricting in urban emergency services", *Computers & Operations Research* 1, 67–95.

Larson, R.C. (1975), "Approximating the performance of urban emergency service systems", *Operations Research* 23, 845–868.

Larson, R., and Odoni, A. (1981), *Urban Operations Research*, Prentice-Hall, Englewood Cliffs, NJ.

Marianov, V., and ReVelle, C. (1992), "A probabilistic fire-protection siting model with joint vehicle reliability requirements", *Papers in Regional Science* 71, 217–241.

ReVelle, C., and Hogan, K. (1988), "A reliability-constrained siting model with local estimates of busy fractions", *Environment and Planning B: Planning and Design* 15, 143–152.

ReVelle, C., and Hogan, K. (1989a), "The maximum reliability location problem and α-reliable p-center problem: Derivatives of the probabilistic location set covering problem", *Annals of Operations Research* 18, 155–174.

ReVelle, C., and Hogan, K. (1989b), "The maximum availability location problem", *Transportation Science* 23, 192–200.

ReVelle, C, and Marianov, V. (1991), "A probabilistic FLEET model with individual vehicle reliability requirements", *European Journal of Operational Research* 53, 93–105.

ReVelle, C., Bigman, D., Schilling, D., Cohon, J., and Church, R. (1974), "Facility Location Analysis: A review of the context-free and EMS models", *Health Services Research*, Summer.

Snyder, D. (1975), *Random Point Processes*, Wiley, New York.

Swain, R. (1971), "A decomposition algorithm for a class of facility location problems", Ph.D. Dissertation, Cornell University, Ithaca, NY.

Toregas, C., Swain, R., ReVelle, C., and Bergmann, L. (1971), "The location of emergency service facilities", *Operations Research* 19, 1363–1373.

White, J., and Case, K. (1974), "On covering problems and the central facility location problem", *Geographical Analysis* 6, 281.