Maximum Availability Models for Selecting Ambulance Station and Vehicle Locations: A Critique

Erhan Erkut (erkut@bilkent.edu.tr)[2]
Armann Ingolfsson (armann.ingolfsson@ualberta.ca)[1,3]
Susan Budge (sbudge@ualberta.ca)[3]

[1]Corresponding author

[2]Faculty of Business Administration
Bilkent University
Bilkent, Ankara, Turkey

[3]School of Business
University of Alberta
Edmonton, Alberta, T6G 2R6

June 2007

Abstract

Several researchers have employed the notion of reliability of coverage to extend the set covering and maximal coverage models. We discuss the suitability of these models in general for choosing ambulance station or vehicle locations. Then, we discuss one particular model (the queueing maximum availability location problem) in greater detail and describe difficulties encountered in applying it using realistic data.

Acknowledgments

This work has been partly supported by Discovery grants 25481 and 203534 from the Natural Sciences and Engineering Research Council of Canada.

1. Introduction

EMS systems typically have performance targets of the form "reach *x*% of patients in *y* minutes or less" (e.g., Fitch et al., 1993, Blackwell and Kaufman, 2002, Henderson and Mason, 2004). Many planning models for EMS station and vehicle location use an analogous concept of *coverage*, where a call is considered covered if there is an available ambulance within some time or distance standard. The most basic coverage models are the set covering (Toregas et al., 1971) and maximum coverage (Church and Revelle, 1974) models.

In this paper, we critique a class of coverage models that have been proposed for emergency service systems and are referred to as either maximum availability models (Revelle and Hogan, 1988 and 1989, Marianov and Revelle, 1996) or maximum reliability models (Ball and Lin, 1993). Although these models differ from each other in many ways, they share the following characteristic: the demand nodes are classified into ones where the probability of an ambulance being available within a coverage standard is at least α, i.e., are "covered with α-reliability," (α is a user-specified parameter) and ones where this is not true. Nodes in the former set contribute their demand to the objective function and nodes in the latter set do not. The primary basis for our critique of this family of models is that they measure performance in a fashion that is inconsistent with performance measures currently used in most real EMS systems.

We will not attempt a complete survey of maximum reliability and maximum availability models, as our purpose is to examine the fundamental assumptions that these models have in common rather than the modeling features that differentiate them. We refer the reader to Berman and Krass (2002) for an excellent review of location problems with stochastic demand and congestion which includes a thorough discussion of the family of models that we focus on. Brotcorne et al. (2003) provide another good review of ambulance location models.

The remainder of this paper is organized as follows: § 2 describes a probabilistic generalization of the set covering location problem, § 3 contrasts maximum availability or reliability models with the maximum coverage location problem, and § 4 considers a specific model from the maximum availability / reliability family of models, provides computational results, and discusses difficulties with using this model in practice.

2. The Probabilistic Set Covering Problem

The set covering model, originally proposed by Toregas et al. (1971), is one of the cornerstones of the facility location literature. If demand point *i* is within a distance standard of a service location, so that the service can be delivered within an acceptable period after a request for service from *i* is received, then *i* is considered to be covered. Defining *I* as the set of demand locations and *J* as the set of candidate station locations, and

a_{ij} = 1 if candidate location *j* can cover demand point *i*, 0 otherwise,

x_j = 1 if candidate location *j* is selected, 0 otherwise,

the set covering problem can be expressed as follows:

$$\text{(P0)} \quad \min \sum_{j \in J} x_j$$

$$\text{s.t.} \sum_{j \in J} a_{ij} x_j \geq 1, \; i \in I$$

$$x_j \in \{0,1\}, \; j \in J \; .$$

While the set covering model may be appropriate in some contexts, it is not suitable for locating emergency medical service vehicles. The coverage constraints require that every demand point be covered by a station. In most instances this requires the location of too many stations and the solutions produced are not economically viable.

When assessing the applicability of the set covering model in the context of emergency vehicle service location, researchers noted another shortcoming: the model assumes that vehicles are always available. In reality, ambulances are typically busy at least 30% of the time. Consequently, researchers attempted to incorporate vehicle availability into the set covering model (Revelle and Hogan, 1989, Marianov and Revelle, 1994). Defining x_j to be the number of ambulances located at site *j*, such probabilistic covering models can be represented as follows:

$$\text{(P1)} \quad \min \sum_{j \in J} x_j$$

$$\text{s.t.} \; P_i(x) \geq \alpha, \; i \in I,$$

$$x_j \in \{0,1,\ldots\}, \; j \in J \; .$$

Here, the coverage constraints have been replaced by chance constraints that ensure that the probability $P_i(x)$ that demand point *i* is covered is at least α, for all demand points.

The assumptions and the approach used to express the coverage probabilities $P_i(x)$ are crucial in determining the tractability of models in this class. The usual approach is to find a way to convert the chance constraints into linear deterministic equivalents of the form

$$\sum_{j \in J} a_{ij} x_j \geq b_i, \ i \in I ,$$

i.e., the number of vehicles that would cover demand point *i*, if available, is at least b_i. Instead of discussing the different approaches that have been taken to achieve this simplification, we make a simple observation: the stochastic set covering model (P1) is *more demanding* than the deterministic set covering model (P0), because achieving the specified probabilistic coverage may require multiple coverage of some demand points. If the deterministic set covering model produces solutions that are too expensive to implement, then its probabilistic generalization will produce solutions that are even less likely to be useful in practice.

Berman and Krass (2002) critique the simplifying assumptions made to arrive at a tractable formulation in these models. However they do not mention perhaps the most serious shortcoming: these models do not realistically represent the goals of EMS system designers. A common service level target in EMS system design in North America is the coverage of 90% of *all* urgent calls in a municipality within 8 minutes (for example, see Fitch et al., 1993, De Maio et al., 2003). In other words, the focus is on system wide coverage rather than coverage in individual neighborhoods. In our experience, EMS practitioners realize that it is financially prohibitive to provide the same response time performance for all areas of a city and they focus on the aggregate performance target. Although targets may be set for individual neighborhoods, these targets are not the drivers of system design. Yet (P1) forces coverage of every demand point with a certain probability.

A more realistic approach is to minimize the number of vehicles, subject to an *aggregate* coverage constraint. We will refer to this problem as (P2) and in the process of formulating it we will demonstrate that (P2) is a relaxation of (P1). Define d_i to be the demand at demand point *i*, and multiply both sides of each constraint in (P1) by d_i to obtain the constraint set $d_i P_i(x) \geq \alpha d_i, \ i \in I$. Then add these constraints to obtain a single constraint $\sum_{i \in I} d_i P_i(x) \geq \alpha D$, where *D* is the total demand in the system. Here, the left-hand-side equals the

expected number of calls covered in the entire system, and the right-hand-side equals a fraction α of the total demand. To summarize, (P2) is the following problem.

$$\text{(P2)} \quad \min \sum_{j \in J} x_j$$

$$\text{s.t.} \sum_{i \in I} d_i P_i(x) \geq \alpha D,$$

$$x_j \in \{0,1,...\}, \quad j \in J .$$

Given that (P2) is a relaxation of (P1), it follows that (P1) will require at least as many ambulances as (P2), and in most real-world instances, it will require many more. We now demonstrate, via an example, that the ratio of the two objective function values can be arbitrarily large.

Example: Consider a 5-node network; a central node and 4 peripheral nodes. Suppose 96% of the calls come from the central node and 1% of the calls come from each of the peripheral nodes. Assume that travel times between nodes are too high for cross-nodal ambulance service, so each node operates as a separate subsystem. Calls arrive according to a Poisson process with arrival rate $\lambda = 100$ calls per week and service times are distributed exponentially with average service time $1/\mu = 50$ minutes, which implies a service rate of $\mu = 202$ per week. Calls arriving to a subsystem while at least one subsystem server is idle are served, and calls arriving while all servers are busy are not served, which implies that each subsystem is an independent *M*/*M*/*s*/*s* queue. If we require coverage of 90% of all calls *for each subsystem* (Problem P1), then we would locate 2 servers at the central node and one server at each of the peripheral nodes, for a total of 6 ambulances. In contrast, if we require coverage of 90% of the *total* number of calls (Problem P2), then we would only locate 2 servers at the central node. Hence, $v(\text{P1})/v(\text{P2}) = 6/2 = 3$, where $v(\text{P}i)$ is the value of problem P*i*. We can make this ratio arbitrarily high by increasing the number of peripheral nodes while keeping the percentage of demand at the central node constant. Hence, the solution to (P1) can be arbitrarily poor when used as a solution for (P2).

Admittedly, this is a pathological example. Locating servers only at the central node amounts to service denial to the peripheral nodes and this is not a defensible way to design an EMS system. However, the example demonstrates the possible consequences of solving the wrong problem. While the example overstates the case, most real-world EMS systems provide better coverage in high-demand areas (such as the downtown core) than in regions of lower demand (such as suburbs). Response time standards are typically less strict and actual

performance is typically worse in rural than in urban areas in the US, UK, and Germany (Fitch, 2005, Felder and Brinkmann, 2002), indicating that standard setters have decided against equal access to ambulance service irrespective of location.

3. Maximum Availability Location Problems

The maximum coverage model addresses the problem that the set covering model may prescribe solutions that are too expensive. It does so by maximizing the amount of demand that is covered, given an upper limit on the number of stations. Like the set covering model, the maximum coverage model assumes that an ambulance is always available at every station. Revelle and Hogan (1989), Ball and Lin (1993), and Marianov and Revelle (1996) all present generalizations of the maximum covering model that fall into the following framework:

$$\text{(P4)} \quad \max \sum_{i \in C(\alpha,x)} d_i$$

$$\text{s.t.} \sum_{j \in J} x_j \le p$$

$$x_j \in \{0,1,...\}, \quad j \in J,$$

where $C(\alpha,x)$ is the set of demand nodes i for which $P_i(x) \ge \alpha$. Importantly, the objective function of this family of models measures something very different from the "fraction of calls covered" that is commonly used in EMS systems: it measures the amount of demand that is at demand points that will be "covered with α-reliability." A simple example illustrates the difference. Suppose that response times are deterministic. Demand node 1 has a demand of 100 and $P_1(x) = 0.8$ while demand node 2 has a demand of 10 and $P_2(x) = 0.95$. Then the expected number of calls that will be covered is $100 \times 0.8 + 10 \times 0.95 = 89.5$, while the objective function of (P4), using $\alpha = 0.9$, would count only 10 calls as being covered with 90% reliability. Galvao, et al. (2005) hint at this inconsistency between the objective function of maximum availability models and the performance measures that drive EMS system design.

In the next section, we describe the details of Marianov and Revelle's (1996) formulation, which is one instance of (P4).

4. Q-MALP: Implementation Issues and Computational Results

Marianov and Revelle (1996) presented the Queueing Maximal Availability Location Problem (Q-MALP), formulated as follows. Let x_{kj} equal one if station j has k or more vehicles and zero

otherwise, and let y_{ik} equal 1 if k or more vehicles can cover demand node i. The set of candidate locations that can cover demand point i is denoted $N(i)$. Suppose R_{ij} is the response time from station j (if it has an available ambulance) to a call at demand point i and suppose that the coverage time standard is S. Then station j is included in the set $N(i)$ if $\Pr\{R_{ij} \le S\} \ge \beta$. Note that in contrast to much of the literature on ambulance location models that tacitly assumes a one-to-one relationship between distances and response times, Marianov and Revelle recognize that response times can vary even for a fixed station-demand point pair.

Next, consider the set of all demand nodes that, if they had stations (with ambulances), could cover demand node i, i.e., $M(i) = \{k : \Pr\{R_{kj} \le S\} \ge \beta\}$. Define the arrival rate for this region as $\lambda_i = \sum_{k \in M(i)} d_k$, let $1/\mu_i$ be the average service time for ambulances within $N(i)$ that respond to calls within $M(i)$, and let $s_i = \sum_{j \in N(i)} \sum_{k=1}^{p} x_{kj}$ be the total number of vehicles in $N(i)$, with p the maximum number of vehicles in the system. Marianov and Revelle approximate the probability P_i that demand node i is covered by $1 - B(\lambda_i / \mu_i, s_i)$, where $B(r, s)$ is the Erlang B loss function, i.e., the probability that all servers are busy in an $M/G/s/s$ loss system with offered load r and s servers. Marianov and Revelle (1996) and Berman and Krass (2002) discuss the assumptions that are necessary to justify this approximation. Then $b(i) = \min\{s_i : 1 - B(\lambda_i / \mu_i, s_i) \ge \alpha\}$ is the minimum number of vehicles needed in $N(i)$ to ensure that demand point i is covered with α reliability. With all of these assumptions and definitions in place, we can present the integer program that Marianov and Revelle proposed:

(P5)

$$
\begin{aligned}
\text{Maximize} \quad & \sum_{i \in I} d_i y_{i,b(i)} \\
\text{Sujbect to} \quad & \sum_{k=1}^{b(i)} y_{ik} \le \sum_{j \in N(i)} \sum_{k=1}^{p} x_{kj}, i \in I \\
& y_{ik} \le y_{i,k-1}, i \in I, k = 2, \ldots, b(i) \\
& \sum_{j \in J} \sum_{k=1}^{p} x_{kj} = p \\
& x_{kj}, y_{ik} \in \{0,1\}, i \in I, j \in J, k = 1, \ldots, p
\end{aligned}
$$

We implemented this model for data from Edmonton, Alberta, with 180 demand nodes and 10 candidate station locations. Before presenting our computational results, we discuss some issues that we encountered when we tried to use the Q-MALP model.

First, one needs to determine values for the parameters α and β. This is problematic because the meaning of these parameters is unclear. As far as we know, these quantities are not measured, tracked, or discussed by EMS practitioners.

An appropriate benchmark value for β might be 50%, which corresponds to counting candidate locations with a median response time within the coverage standard as being able to cover a particular demand point. For α, values close to 100% would presumably be appropriate. Given that there is no obvious way to determine the "right values" for α and β, we attempted to solve Q-MALP parametrically with α = 80%, 90%, 95%, and 99%, and β = 35%, 50%, and 65%. (We encountered long computational times when α was set to 99% and failed to find a feasible solution for one of the instances.)

Second, one needs to determine the average service time $1/\mu_i$ for the region $M(i)$ around demand node i. The difficulty here is that average service times $1/\mu_{kj}$ for station – demand node pairs (k, j) where $k \in M(i)$ and $j \in N(i)$ can vary widely. The average service time $1/\mu_i$ should presumably be a weighted average of the average service times $1/\mu_{kj}$ over all station-demand node pairs, but the appropriate weights depend on the number of vehicles at each station, which are not known a priori. To circumvent this difficulty, we used a simple average.

We solved (P5) parametrically in α, β, and p (the maximum number of vehicles). To measure the quality of each solution to (P5) we used the approximate hypercube model, extended to allow for multiple vehicles at some stations (Budge et al., 2007a), to estimate the probability that a vehicle from station j is dispatched to a call from demand point i, for all station – demand point pairs. We then combined these dispatch probabilities with the probability that an ambulance from station j, if dispatched to respond to a call from demand point i, would reach the call within the response time standard, to arrive at the system-wide expected coverage. For comparison, we provide the expected coverage that is achieved by attempting to maximize expected coverage directly, using the heuristic described in (Budge et al., 2007b).

Table 1 presents our results. For each value of p we show the expected coverage for the optimal solution to Q-MALP, as a function of α and β. For comparison, we show the expected coverage obtained using the approach from Budge et al. (2007b). We refer to this as the "maximum expected coverage," but note that the procedure from Budge et al. (2007b) does not guarantee global optimality. The computation times shown for Q-MALP are totals over all

instances solved for a particular value of p. The computation times for maximum expected coverage correspond to running the algorithm from Budge et al. (2007b) for ten iterations.

We observe the following from Table 1:

- The Q-MALP solutions are quite sensitive to the values of α and β. Coverage differences of more than 20% are observed for different choices of parameter values.
- For every p that we tried, direct maximization of expected coverage achieves higher expected coverage than parametric solution of Q-MALP over α and β with differences ranging from 0.1% to 0.6%.
- The "best" values for α and β vary depending on the value for p. However, there is a consistent pattern that the highest value we tried for β results in the highest expected coverage.
- Solving Q-MALP parametrically takes two to six times as long as maximizing expected coverage directly, for the set of parameter values we chose.

For this data set, the most realistic values for p are in the range 16 to 20. With 16 ambulances, the system performance target of 90% expected coverage is just met, using the maximum expected coverage solution, which was found in 6.7 minutes. In contrast, parametric solution of Q-MALP indicates that the 90% performance target cannot be met with 16 ambulances, and it required about an hour of computation to reach this conclusion. The 0.6 percentage point difference in expected coverage between the best Q-MALP solution and the maximum expected coverage is of practical importance—in the vicinity of 90% coverage a one percentage point difference corresponds roughly to adding one ambulance to the system, which is a significant expense.

To summarize, the Q-MALP objective function is not equivalent to the expected coverage performance measure that typically drives EMS system design and the same is true for other maximum availability location models. One can solve Q-MALP parametrically in α and β, evaluate the expected coverage for the resulting solutions, and choose the best one. However, this procedure is not necessary because one can heuristically maximize expected coverage directly and, based on our computational results, this usually results in higher quality solutions in less time. Furthermore, it may be difficult to explain the concept of α-reliability and how it

differs from expected coverage to practitioners. Given these difficulties with maximum availability models, we propose the use of maximum expected coverage models for ambulance location problems instead.

Table 1: Expected coverage for Q-MALP solutions found by varying p, α, and β, and the maximum expected coverage, for each value of p. The highest expected coverage found by solving Q-MALP is shown in bold, for each value of p.

p = 16 ambulances

$\beta \setminus \alpha$	80%	90%	95%	99%	Max. exp. coverage
35%	85.7%	76.2%	76.8%	*	90.3%
50%	87.5%	85.7%	84.4%	70.5%	
65%	**89.7%**	87.1%	84.5%	77.1%	
Comp. time (min.):			29.9**		6.7

* No feasible solution found in 12 hours of computation.
** Excludes instance where no feasible solution was found

p = 20 ambulances

$\beta \setminus \alpha$	80%	90%	95%	99%	Max. exp. coverage
35%	90.9%	88.7%	77.8%	68.9%	93.1%
50%	91.2%	89.5%	91.2%	83.5%	
65%	**92.7%**	92.5%	90.0%	80.9%	
Comp. time (min.):			12.7		6.7

p = 25 ambulances

$\beta \setminus \alpha$	80%	90%	95%	99%	Max. exp. coverage
35%	92.1%	91.3%	86.3%	82.4%	94.4%
50%	91.7%	91.0%	92.6%	85.4%	
65%	93.4%	93.5%	**94.3%**	87.3%	
Comp. time (min.):			35.0		5.8

References

Ball, M.O., F. L. Lin. 1993. A reliability model applied to emergency service vehicle location. *Operations Research* 41:18-36.

Blackwell, T. H., J. S. Kaufman. 2002. Response time effectiveness: comparison of response time and survival in an urban emergency medical services system. *Academic Emergency Medicine* 9:288-295.

Brotcorne, L., G. Laporte, F. Semet. 2003. Ambulance location and relocation models. *European Journal of Operational Research* 147:451-463.

Budge, S., A. Ingolfsson, E. Erkut. 2007a. Approximating vehicle dispatch probabilities for emergency service systems with location-specific service times and multiple units per location. Working paper, available from http://www.business.ualberta.ca/aingolfsson/working_papers.htm.

Budge, S., A. Ingolfsson, E. Erkut. 2007b. Optimal ambulance location with random delays and travel times. Working paper, available from http://www.business.ualberta.ca/aingolfsson/working_papers.htm.

Church R, C. Revelle. 1974. The maximal covering location problem. *Papers of the Regional Science Association* 32:101-120.

De Maio, V.J., I.G. Stiell, G.A. Wells, D.W. Spaite. 2003. Optimal defibrillation for maximum out-of-hospital cardiac arrest survival rates. *Annals of Emergency Medicine* 42:242-250.

Felder, S., H. Brinkmann. 2002. Spatial allocation of emergency medical services: minimizing the death rate of providing equal access? *Regional Science and Urban Economics* 32:27–45.

Fitch, J. J., R. A. Keller, D. Raynor, C. Zalar. 1993. *EMS Management: Beyond the Streets*, 2nd edition. JEMS Communications, Carlsbad, CA.

Fitch, J. 2005. Response times: myths, measurement and management. *Journal of Emergency Medical Services* 30:46–56.

Galvao, R. D., F. Y. Chiyoshi, R. Morabito. 2005. Towards unified formulations and extensions of two classical probabilistic location models. *Computers & Operations Research* 32:15-33.

Henderson S.G., A. J. Mason. 2004. Ambulance service planning: simulation and data visualisation. *Operations Research and Health Care: A Handbook of Methods and Applications*, eds. F. Sainfort, M. Brandeau, and W. Pierskalla, Kluwer.

Marianov V., C. Revelle. 1996. The queueing maximal availability location problem: a model for the siting of emergency vehicles. *European Journal of Operational Research* 93:110-120.

Marianov V., C. Revelle. 1994. The queuing probabilistic location set covering problem and some extensions. *Socio-Economic Planning Sciences* 28:167-178.

Revelle C., K. Hogan. 1988. A reliability-constrained siting model with local estimates of busy fractions. *Environment And Planning B-Planning & Design* 15:143-152.

Revelle C., K. Hogan. 1989. The maximum availability location problem. *Transportation Science* 23:192-200.