Note: The original publication is/will be available at www.springerlink.com

Optimal Ambulance Location with Random Delays and Travel Times

Armann Ingolfsson, Susan Budge,
University of Alberta School of Business, Edmonton, Alberta, Canada
Erhan Erkut,
Ozyegin University, Istanbul, Turkey

Abstract

We describe an ambulance location optimization model that minimizes the number of ambulances needed to provide a specified service level. The model measures service level as the fraction of calls reached within a given time standard and considers response time to be composed of a random delay (prior to travel to the scene) plus a random travel time. In addition to modeling the uncertainty in the delay and in the travel time, we incorporate uncertainty in the ambulance availability in determining the response time. Models that do not account for the uncertainty in all three of these components may overestimate the possible service level for a given number of ambulances and underestimate the number of ambulances needed to provide a specified service level. By explicitly modeling the randomness in the ambulance availability and in the delays and the travel times, we arrive at a more realistic ambulance location model. Our model is tractable enough to be solved with general-purpose optimization solvers for cities with populations around one Million. We illustrate the use of the model using actual data from Edmonton.

Key words: emergency medical services; ambulance location; facility location; dispatch delays

Acknowledgments: This research was supported in part by the Natural Sciences and Engineering Research Council of Canada. We thank anonymous referees for several comments that led to improvements in the paper.

Introduction

The design of emergency medical service (EMS) systems involves several interconnected strategic decisions, such as the number and locations of ambulance stations, the number and locations of the vehicles, and the dispatch system used. In this paper we focus on the allocation of vehicles to a set of (existing or planned) ambulance stations with known locations. An overriding issue when designing an EMS system is the "coverage" provided, and a common performance target is to respond to (or cover) a fraction α of all calls in δ minutes or less (for example 90% in under 9 minutes). Coverage is determined by the response time to calls. The most obvious and significant component of response time is the travel time between the ambulance station and the demand location (we will use "location" and "node" interchangeably to refer to a neighborhood that has been aggregated to a single point). Almost all of the existing operations research literature on ambulance location focuses on travel times, but this is not the only component of the response time, which is generally defined as the time from when a call for ambulance service arrives until paramedics reach the patient. Therefore, the response time includes any delays prior to the trip. Such delays can include time spent on the phone obtaining the address and establishing the seriousness of the call, time spent deciding which ambulance to dispatch, time to contact the paramedic crew of that ambulance, and time for the paramedic crew to reach its ambulance and start it. Further, the response time for a given emergency call can be affected by the availability of the ambulances because sometimes the closest ambulance will be busy and another ambulance must respond. Queueing delays (when no ambulances are available) can also occur, but in our experience with real systems, they occur infrequently. In the rare situations when all ambulances are busy, incoming calls are typically responded to using some type of backup system, such as supervisor vehicles or fire engines.

This paper focuses on ambulance location in urban areas. In rural areas, ambulance availability will be less of a concern and geographical coverage will be of greater concern. Our work is motivated by three real-world ambulance location projects—one for the City of Edmonton [21], a second (ongoing project) for the City of Calgary, and a third for the City of St. Albert [12], all in Alberta, Canada. We gained two important insights from these projects. First, we noticed that delays are significant and highly variable, and travel times between a given pair of points are highly variable. Second, we noticed that the estimated coverage depends on the way that ambulance availability, delays, and travel times are modeled, and we were unable to find any location model in the literature that provided a realistic enough estimate of coverage for our purposes. Next we provide some data for the first insight followed by an example to illustrate the relevant issues for the second insight.

Motivating Data

In this section we use data from the St. Albert study to illustrate the significance of the pre-travel delays, as well as the variability in that component and in the travel time component of the response time. We analyzed data from approximately 6,997 EMS calls serviced in over 4 years in St. Albert. Figure 1 displays the empirical distribution of pre-trip delays, which is well approximated by a lognormal distribution. The delays ranged from 20 seconds to 20 minutes, with an average of 175 seconds and a standard deviation of 95 seconds. The average delay of almost 3 minutes is a very substantial fraction of the 9-minute response time standard, and the variation in the delay is too large to ignore (the standard deviation is more than 50% of the mean). Anyone that has experience with real emergency service systems will be aware of the presence of such delays, and several past researchers have mentioned them (see, for example, several of the chapters in [33] and [3]) suggesting, in some cases, that such delays are negligible,

and in other cases that they can be incorporated in existing models by adding the average delay to the average travel time.

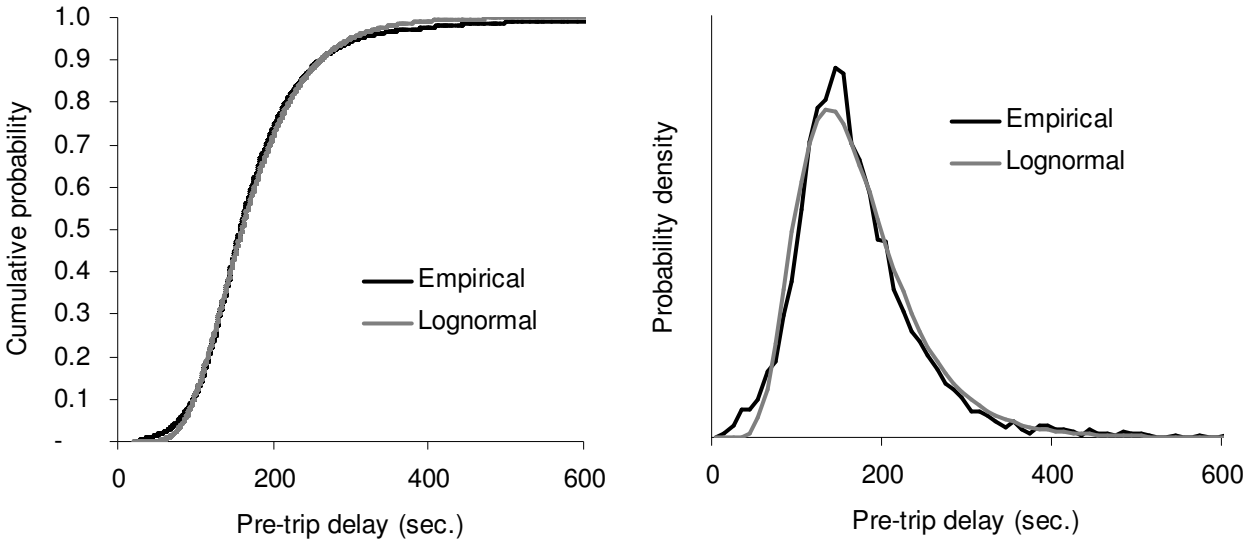

Figure 1: Empirical distribution of pre-trip delays for 6,997 EMS calls serviced in St. Albert, and a fitted lognormal distribution.

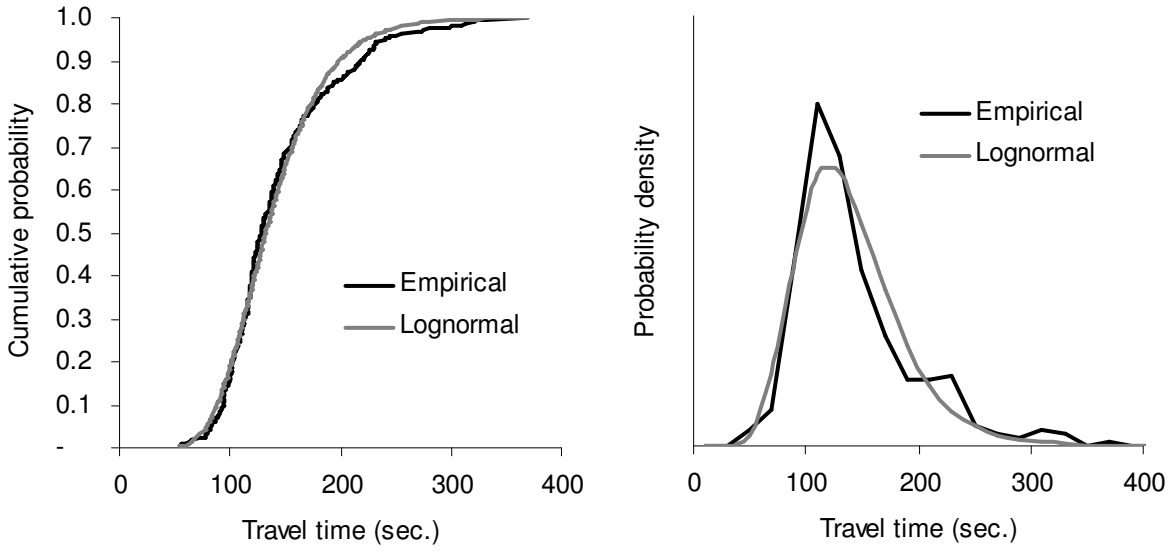

Figure 2: Empirical distribution for travel times between a particular station and demand point pair for a total of 352 trips, together with a fitted lognormal distribution.

The St. Albert dataset contains multiple trips to several locations, which allows us to analyze travel time distributions. Figure 2 shows the empirical travel time distribution for 352 trips from a particular station to the same multiple-resident demand point. The trip times range from 55 seconds to 370 seconds, with an average of 143 seconds and a standard deviation of 52 seconds. We analyzed a total of nine locations with multiple trips and found that the standard deviation was always considerable (on average 40% of the mean). Reporting on a project for locating emergency vehicle bases in Tucson, Arizona, [14] also found substantial variation in empirical travel times for given base-demand zone pairs. This variation can be due to variability in the effective travel speed, or due to randomness in the location of the incident (demand aggregation).

Motivating Example

A small town has a single ambulance station, a response time standard of 9 minutes, and three demand locations D1, D2, and D3, that are expected to generate 100 calls each in a given future time period. Travel times between the station and the three demand locations have means of 5.5, 7.5, and 9.5 minutes, and standard deviations equal to 40% of the means. The pre-trip delay is independent of the travel time and has a mean of 2.5 minutes and a standard deviation of 1 minute. Assume that the total response time (composed of the pre-travel delay and the travel time) follows a lognormal distribution. For simplicity, assume that an ambulance is always available when a call arrives. Table 1 lists six different ways to model pre-trip delays and travel times and shows the probability of coverage for a call from each demand location, as well as the total expected number of covered calls, and the overall expected coverage.

If we ignore the pre-trip delay and use average travel times to determine coverage (Model A), then we would characterize the first two demand locations as “covered,” the third one as “not covered,” and credit 200 calls to the coverage offered by the station when computing the

performance measure (that is, expected coverage = 200/300 = 0.667 or 67%). However, depending on whether and how each of the components is modeled, the expected number of covered calls for each demand node and for the system as a whole varies widely.

Table 1: Six ways to model pre-trip delays and travel times, with summary of probabilities of responding to calls from the three demand locations for each model used, the resulting expected number of covered calls, and expected coverage.

Model	Travel time	Delay time	Probability of responding to a call at a demand location within 9 minutes			Exp. no. of covered calls	Expected coverage
			D1	D2	D3		
A	Deterministic	Not modeled	100.0%	100.0%	0.0%	200.0	67%
B	Stochastic	Not modeled	92.9%	74.7%	52.1%	219.7	73%
C	Deterministic	Deterministic	100.0%	0.0%	0.0%	100.0	33%
D	Stochastic	Deterministic	73.4%	42.9%	21.4%	137.8	46%
E	Deterministic	Stochastic	85.7%	12.9%	0.0%	98.5	33%
F	Stochastic	Stochastic	70.8%	42.6%	22.9%	136.3	45%

Table 1 illustrates several differences between the six models:

- Comparison of models A and B (or C and D, or E and F) demonstrates that using constant as opposed to probabilistic travel times can result in considerable over- or underestimation of coverage at specific demand locations. Although over- and underestimates at individual demand locations may cancel each other to some extent when computing the overall expected coverage, the inaccuracy in this system performance estimate can be quite significant (the fully deterministic Model C gives 45% – 33% = 12% lower expected coverage than the fully probabilistic Model F).
- As one would expect, ignoring delays entirely results in severe overestimation of coverage. For example, Model D has 27% (73% – 46%) lower coverage than Model B, because Model D includes (constant) delays whereas Model B does not include delays.

- When one models randomness in travel times, ignoring randomness in the duration of delays causes less inaccuracy than ignoring delays altogether. The direction of the change in probability of coverage when one incorporates randomness in delay durations is not always the same, as one can see by comparing Models D and F: the constant delay model (Model D) overestimates the probability for D1 by 2.6% and underestimates the probability for D3 by 1.5%.

We believe that these inaccuracies can influence decisions adversely when every percent counts in trying to reach a coverage target.

We conclude that a convolution of the delay and travel time distributions is needed to obtain an accurate response time distribution, assuming travel time and delay are statistically independent—an assumption that is supported by the data that we worked with. Situations where the travel time and delay are dependent can be handled as well, as we will demonstrate. In this paper we introduce new methodology that incorporates randomness in both pre-travel delays and travel times and models the availability of ambulances in a more precise way than many previous ambulance location models and is therefore free of the inaccuracies demonstrated in the preceding example. The incorporation of the uncertain components is important not only in order to provide an accurate estimate of the coverage offered by specific ambulance configurations and to ensure an optimal distribution of ambulances to stations (based on the coverage provided), but also in order to predict the consequences of more or less uncertainty. We illustrate this in the Computational Results section by focusing on pre-travel delays. We show that models that ignore delays, or randomness in delays, may severely overestimate the coverage achieved with a given number of ambulances and, conversely, underestimate the

number of ambulances needed to meet a specified coverage objective (see Figures 5 and 6 in the Computational Experiments section).

In the remainder of the paper, we discuss the relevant literature, and then describe the problem data, our problem formulation, some useful properties of the formulation, the results of computational experiments, and further research that we intend to undertake to extend and experiment with the model.

Literature

See [33, 31, 29, 4] for reviews of literature on locating emergency service facilities. In addition, Berman and Krass [2] review the literature on facility location with stochastic demands, much of it motivated by emergency service applications. In this section we survey selected papers with an emphasis on those that are most relevant to our research.

Models such as the set-covering location problem (SCLP), first formulated by Toregas et al. [32], the maximal covering location problem (MCLP) of Church and ReVelle [8], and Daskin's [9] maximum expected covering location problem (MEXCLP), which accounts for the potential unavailability of ambulances using a single, system-wide busy probability, can be considered the most influential ones on subsequent research, since most other models are extensions of these. While many of these models can be solved to global optimality with reasonable effort, they suffer from simplifying assumptions.

An important model that provides more realism (i.e., uses fewer simplifying assumptions) is Larson's [26] hypercube model and subsequent approximate versions of that model [27, 24]. This model allows busy fractions to vary between ambulances and can accommodate ambulances responding to calls outside their assigned districts. Larson [28], and Brandeau and Larson [3]

describe applications and extensions of the hypercube model. We use an extension of the approximate hypercube model that allows multiple servers at a station [6]. Discrete event simulation can be used when even greater realism is needed [e.g., 19, 20, 21].

We extend the ambulance location modeling paradigm by incorporating uncertainty in ambulance availability and in response times, without sacrificing the ability to use general-purpose optimization solvers. All of the covering models that we listed above use deterministic (average) travel times. While delays are usually not explicitly mentioned in papers dealing with coverage models, it is easy to incorporate a constant (average) delay into all coverage models by simply subtracting the delay from the specified maximum response time. (For example, [11] uses MCLP with a 5-minute travel time, which may have been part of an 8-minute response time with an average delay of 3 minutes.)

The assumption made by early covering models is that if (and only if) an ambulance is available within a specified maximum distance of a demand point, then the demand point is covered. EMS systems typically measure performance based on the fraction of calls responded to within a specified time standard. However, for a given ambulance location and a demand point, it is not possible to know with certainty whether the call will be responded to within the time standard—it depends on the pre-trip delay and the travel time as well as the availability of the ambulance, none of which can be predicted with certainty. Our model does not rely solely on average response times, and hence, it is not limited by the resulting strict classification of demand points as covered or not covered. It allows incorporation of randomness in pre-trip delays and travel times, and computes an expected coverage for each demand point, given the ambulance locations and estimates of the ambulance availabilities. Hence, we increase model realism by replacing the 0-1 consequences implied by solutions of traditional covering models for demand points by

real numbers, which are better estimates of the fraction of calls emanating from different demand points that can be reached within the specified time standard.

In the remainder of this section, we focus on covering ambulance location models that incorporate response time variability. Although non-covering models have been developed for ambulance location (such as models that minimize mean response time) and many of these models include stochastic components, we do not discuss such models here because most EMS systems measure their performance based on the fraction of calls reached within some time standard, rather than by the mean response time. We know of three articles that include travel time variability in covering models. Marianov and ReVelle [30] assume travel time from station i to node j is normally distributed with known mean and variance. Then they define a node j to be covered by station i if the average travel time plus K standard deviations is less than a specified constant. While they acknowledge the variability in travel times, they do not use the distributions directly in the model. This model is more conservative (for $K > 0$) than a coverage model that uses only average travel times. However, it is still a traditional covering model in the sense that a demand point is either covered or not.

Daskin [10] presents a model that includes the probability P_{ij} that an ambulance at station i can travel to a call at demand node j within a response time standard. The focus of Daskin's model is the integration of location and routing, taking into account that some calls may require two vehicles to respond. This model does not account for ambulance unavailability.

Perhaps the paper that is most relevant to ours is Goldberg and Paz [16], which is inspired by a case study reported in [14, 15]. They formulate an emergency facility location model that models random travel times similarly to [10]. The quantities P_{ij} are used to calculate expected coverage in the objective function of their optimization problem. Goldberg and his co-workers

used an approximation related to the hypercube model to estimate the busy probabilities of the vehicles, and included an upper bound on the number of stations. They used regression to estimate average travel times as a function of distance along roads of various types, and compute the P_{ij} values using this mean and the standard deviation of the residuals, assuming normal distribution of path travel times.

While the way we model expected coverage is similar to that of [16], there are several differences between their work and ours. Perhaps the most significant modeling difference is the inclusion of pre-trip delays in our model. Also, we treat the calculation of the busy probabilities for the vehicles, and the computation of coverage probabilities for demand points in different ways. We consider dispatch policies as given, rather than including them as decision variables. Note that we do not believe that treating these dispatch policies as given is a limiting assumption for real applications, since in our experience with real systems the operators have indicated that there would be no alternative (to "dispatch the closest available ambulance") that would be acceptable in practice. An advantage of our approach is that the continuous relaxation of the problem we will formulate (P1) is a convex optimization problem, under certain assumptions, as we will show. For all of these reasons, our model is more compact and tractable and we are able to solve problems of realistic size using off-the-shelf solvers, while [16] propose pairwise interchange heuristics for their model.

Problem Data

We assume that the following data are available:

- A set S of m station locations, indexed by i, and a set N of n demand nodes, indexed by j.

- A positive arrival rate λ_j for each demand node *j*. We assume independent Poisson arrival processes at the nodes. We denote the system wide arrival rate with $\lambda \equiv \sum_{j \in N} \lambda_j$ and the fraction of the total demand from demand node *j* by $h_j \equiv \lambda_j / \lambda$.

- A dispatch order for each demand node *j*, i.e., a list of the *m* stations in order of preference for dispatching to a call originating from node *j*.

- Parameters δ and α which specify the coverage objective that calls should be responded to in at most δ time units with probability of at least α.

- The probability that the response time is less than or equal to δ time units, given that the call arrives to node *j* and the *i*th station in node *j*'s dispatch order responds, denoted w_{ij} $= \Pr\{ R_{ij} \leq \delta\}$, where R_{ij} is the response time.

- The average on-scene time, and average time spent traveling to and remaining at a hospital, denoted $\mathrm{E}[T_{\text{on scene}}]$, and $\mathrm{E}[T_{\text{hospital}}]$, respectively.

- The "busy fraction" ρ_i for ambulances at station *i*, i.e., the probability that an ambulance at station *i* is not available to respond to calls, and correction factors Q_{ij} for each station-node pair, to approximately account for the dependence in the busy fractions between servers. We assume that $\rho_i \in (0,1)$ and $Q_{ij} > 0$.

The last assumption, that the busy fractions and correction factors are exogenous input to the model, is obviously a limiting one. We discuss how to overcome this assumption later.

The best way to calculate the probabilities w_{ij} depends on the availability of data and the context. We now outline three possible methods. First, if detailed data for a sample of

individual calls is available, then one could estimate w_{ij} as the ratio k_{ij}^{δ} / k_{ij}, where k_{ij} is the total number of calls in the sample where an ambulance from station *i* responded to a call from node *j* and k_{ij}^{δ} is the number of such calls that had a response time less than or equal to δ.

Second, suppose that the distribution function $H_{ij}(t)$ of the travel time T_{ij} from the i^{th} station (in node *j*'s dispatch order) to node *j* as well as the distribution function $F(t)$ for the delay are available, and that it is reasonable to assume that the travel time and the delay are independent random variables. Then one can use convolution to calculate the probabilities, i.e.,

$$w_{ij} = \int_{x=0}^{\delta} H_{ij}(\delta - x) dF(x) \tag{1}$$

In practice, the pre-trip delay is often viewed as being the sum of the time spent on call taking and dispatching (call delay) and the time from when a crew is dispatched until the vehicle starts moving (chute delay). If data is collected separately for these two components, as is typically the case, then one could estimate separate distribution functions for each delay and use convolution to obtain the distribution function *F*(*t*) for the total delay. This approach is useful if one wants to separately consider strategies for reducing call delay and strategies for reducing chute delay.

Third, suppose that both travel times and pre-travel delays depend on call priority, but that for a given priority level, these two random variables are independent. Adding a superscript *p*, for priority level, and using v_j^p to denote the probability that a call from node *j* is of priority *p*, then the calculation in (1) would be adjusted as follows:

$$w_{ij} = \sum_{p} v_j^p \int_{x=0}^{\delta} H_{ij}^p(\delta - x) dF^p(x)$$

The first method is the most general in that it requires no independence assumptions, but it has two limitations: (1) the sample size k_{ij} might be small or even zero for some station-node pairs, even if the overall sample is large, and (2) the method is silent about how to predict the consequences of changes to the pre-travel delay distribution. The second and third methods assume independence, but they do not suffer from the two limitations just mentioned.

Note that the w_{ij} are conditional probabilities – they assume that the call comes from demand node *j* and is responded to by the *i*-th preferred station. Higher system congestion makes it more likely that less preferred stations respond to calls, and this can induce dependence between pre-travel delays and travel times. Our model captures such dependence by combining the conditional probability, w_{ij}, with the probability $f_{ij}(x)$ that the *i*-th preferred station responds to a call from node *j*, as shown below.

We emphasize that the calculation of w_{ij} is done for all station-node pairs, before solving the optimization model that we pose in the next section. That model requires no information about the probability distributions of travel times or delays other than the probabilities w_{ij}.

We will assume that the dispatch order for each node *j* is such that:

$$w_{1j} \geq w_{2j} \geq \ldots \geq w_{mj} \tag{2}$$

That is, the stations are arranged in descending order of the likelihood of responding to a call from node *j* in less than δ time units. Although dispatching the closest available unit is not always optimal [see, e.g., 28], studies such as [23] indicate that this policy is generally near-optimal. As previously mentioned, our experience with real EMS systems indicates that deviating from closest-available-unit dispatching would be difficult in practice. The formulation

that we present in the next section is valid without this assumption, but the concavity property that we discuss later requires it.

Problem Formulation and Properties

Let x_i be the number of ambulances located at station *i*, and let x_{ij} be the number of ambulances at the i^{th} preferred station for demand node *j*. The vector $(x_{1j}, x_{2j}, \ldots, x_{mj})$ is a permutation of $(x_1, x_2, \ldots, x_m)$, for each *j*. Similarly, let ρ_{ij} be the busy probability for the i^{th} most preferred station for demand node *j*. The optimization problem is:

$$\text{(P1)} \quad \text{maximize} \quad s(x) \equiv \sum_{j \in N} h_j s_j(x)$$

$$\text{subject to} \quad z(x) \equiv \sum_{i \in S} x_i = b \tag{3}$$

$$x_i \geq 0, \text{ integer, for all } i \in S \tag{4}$$

where

$$s_j(x) = \sum_{i \in S} f_{ij}(x) w_{ij}, \text{ for all } j \in N \tag{5}$$

and

$$f_{ij}(x) = Q_{ij}\left(1 - \rho_{ij}^{x_{ij}}\right) \prod_{u=1}^{i-1} \rho_{uj}^{x_{uj}}, \text{ for all } i \in S, j \in N \tag{6}$$

Problem (P1) maximizes the expected coverage $s(x)$, subject to a constraint on the total number of ambulances $z(x)$ being equal to *b*. For the moment, we assume *b* to be given, but in the algorithm in the next section, *b* will become a decision variable. The system-wide coverage $s(x)$ is a weighted combination of the coverages for individual demand nodes, and the coverage $s_j(x)$

for demand node *j* is calculated in (5) by conditioning on which station sends an ambulance to respond to a call from node *j*. The calculation of the node *j* coverage requires the "dispatch probability" $f_{ij}(x)$, the probability that a call from node *j* is responded to by an ambulance from its i^{th} preferred station. This probability is calculated, as shown in (6), as the product of the probabilities that all ambulances at the $i-1$ more preferred stations are busy, at least one ambulance at the i^{th} preferred station is free, and a correction factor Q_{ij}, to approximately account for the dependence between servers. Setting the correction factors to 1 is equivalent to assuming that the probability of an ambulance being busy is statistically independent of the status of all other ambulances in the system.

Concavity Result

Proposition 1: If $w_{1j} \geq w_{2j} \geq \ldots \geq w_{mj}$ for all $j \in N$, and Q_{ij} and ρ_j are invariant with *x* (recall that these are assumed to be exogenous input to the model) for all $i \in S, j \in N$, then the system-wide coverage is a concave function of *x*.

Proof: See Appendix A**.**

Proposition 1 implies that (P1) has a concave objective function and linear constraints. Consequently, the continuous relaxation of (P1) is a convex program, and a local optimum is also global. The proposition also implies that the coverage $s_j(x)$ for each demand node *j* has the following properties:

- An increase in the number of ambulances at any station increases the coverage for each demand node.
- As the number of ambulances at a particular station increases, the marginal increase in coverage from adding ambulances to this station decreases

Busy Fractions and Correction Factors

The assumption that the busy fractions ρ_i and correction factors for dependence Q_{ij} are exogenous input is not realistic, as they will depend on the number and distribution of ambulances between stations. To overcome this limitation, we propose iterating between solving (P1) and estimating the busy fractions and correction factors. Although we solve (P1) optimally, the iterative procedure is not guaranteed to provide a globally optimal solution to the more general problem that views the busy fractions and correction factors as functions of the allocation of ambulances to stations. In other words, the iterative procedure is a heuristic approach to solving this more general problem.

If all ambulances are assumed to have the same busy fraction, then a relatively simple estimation procedure can be used (refer to Appendix B for details). If all ambulances are not assumed to have the same busy fraction, then a more complicated estimation procedure is necessary. We use a generalization of the approximate hypercube model, detailed in [6], that allows for multiple vehicles at a station. This procedure evaluates the busy fractions ρ_i, the correction factors Q_{ij}, and the expected coverage. We will use $s^{AH}(x)$ to denote the expected coverage evaluated with the approximate hypercube model, to distinguish it from the expected coverage $s(x)$ as computed in formulation (P1).

In the original hypercube model [26], service times (the time an ambulance is tied up with a call) are assumed to be exponentially distributed. The pre-travel delay and the travel time are part of the service time, and if these components are lognormally distributed, then the service times will be far from exponentially distributed. Fortunately, one can expect the loss-version of the approximate hypercube model (which we use) to be relatively insensitive to the shape of the

service time distribution, as argued by Jarvis [23]. The related insensitivity property of the *M*/*M*/*s*/*s* loss system is discussed, for example, in [18].

We propose the following iterative algorithm to overcome the assumption of the busy fractions and correction factors being exogenous inputs:

Step 1: Choose an initial value for the total number of ambulances, *b*.

Step 2: Attempt to maximize coverage with *b* ambulances, as follows:

Step 2a: Set the busy fractions ρ_i^{in} to an initial estimate of the busy fraction, set all correction factors Q_{ij}^{in} equal to 1, and set $x^{0,*} = 0$. Set $n \leftarrow 1$ and choose a smoothing parameter $\gamma \in (0,1)$.

Step 2b: Solve (P1), using busy fractions ρ_i^{in} and correction factors Q_{ij}^{in}. Find the solution $x^{n,*}$ that maximizes $s(x)$ subject to, $x_i \geq x_i^{n-1,*} - 1, i \in S$, (3), and (4). If the convergence criterion is satisfied, go to Step 3.

Step 2c: Estimate the busy fractions ρ_i^{out} and correction factors Q_{ij}^{out} that result from the solution $x^{n,*}$. Set $\rho_i^{\text{in}} \leftarrow \gamma\rho_i^{\text{out}} + (1-\gamma)\rho_i^{\text{in}}$ for all stations, $Q_{ij}^{\text{in}} \leftarrow \gamma Q_{ij}^{\text{out}} + (1-\gamma)Q_{ij}^{\text{in}}$ for all station-node pairs, and $n \leftarrow n+1$. Go back to step 2b.

Step 3: Evaluate the expected coverage $s^{AH}(x)$ for the final solution(s), using the approximate hypercube model. Adjust the total number of ambulances *b* based on whether the highest coverage among the final solutions is less than or greater than the target of α. When it has been determined that the current total number of ambulances is the smallest one that will achieve the target coverage, then stop. Otherwise, return to step 2.

The algorithm includes an outer loop, which is a one-dimensional search (such as bisection search) for the smallest total number of ambulances needed to provide the required coverage, and an inner loop, which iterates between solving (P1) and estimating the busy fractions and correction factors. The expected coverage for each solution that is returned by the algorithm is evaluated using the approximate hypercube model, thus avoiding the simplifying assumptions made in formulation (P1), that the busy fractions and correction factors are exogenous inputs.

The constraints $x_i \geq x_i^{n-1,*} - 1$ are added in Step 2b to prevent the allocation of ambulances to stations from changing too much from one iteration to the next, recognizing that the busy fractions and correction factors depend on the allocation of ambulances to stations.

The convergence criterion for the inner loop could be expressed in terms of the sequence of solutions $\{x^{n,*}\}$, the estimated busy fractions $\{\rho_i^{\text{out}}(x^{n,*})\}$, or both. The inner loop algorithm is not guaranteed to converge to a unique solution. Indeed, we have sometimes observed convergence to a cycle of two or more similar solutions. In such cases, planners could be presented with multiple good solutions, which could be compared in terms of the values that they give for the coverage (as estimated by the busy fraction estimation procedure) or for other performance measures.

Call Priorities

When EMS calls are received, they are usually triaged into one of several priority classes. In the systems that we have experience with, all calls are responded to by the closest available ambulance, regardless of priority class. However, we have observed that delays and travel times are typically longer for lower priority calls, because calls above some priority threshold are viewed as warranting a “lights and siren” response, while calls below the threshold are not.

Different priority classes may have different expected coverage targets, but the target for the highest-priority calls is the most important one. Based on these observations, we use the following approach to incorporate call priorities in our model:

- When estimating the busy fractions and correction factors, we use response times, on-scene times, and hospital times that are averaged over all priority classes, to accurately estimate the system workload.
- When estimating the probabilities w_{ij} that are used to compute the expected coverage, we use only high priority (lights and siren) calls. In other words, the expected coverage in the model is only for the high priority calls.

This approach reflects the fact that the coverage target for the high priority calls is the most important one. In practice, we expect that if the high priority coverage target is met, the coverage targets for the lower priority calls will also be met, but this is not guaranteed.

Computational Experiments

In the instances of (P1) that we solved, based on data from Edmonton EMS, we used deterministic travel times in order to isolate the effect of randomness in delays. We estimated travel times using a simplified road network, with acceleration at the beginning of each trip and deceleration at the end of the trip (see [21] for details). The dispatch orders satisfied assumption (2). These instances had 10 stations and 180 demand nodes. We solved (P1) to optimality in at most a few minutes per instance with a standard branch-and-bound algorithm (the Premium Solver Platform for Microsoft Excel) that calls a nonlinear programming algorithm (the generalized reduced gradient solver) to solve the continuous relaxations. To overcome the assumption of exogenous busy fractions and correction factors, we used the algorithm described

in the last section. Figure 3 shows an example of how ρ_i^{in} and ρ_i^{out} evolved over 6 iterations for one problem instance based on Edmonton data, with the total number of ambulances equal to 16. In this instance, γ was set to 0.9, and ρ_i^{in} and ρ_i^{out} converged in about 3 iterations with an average after convergence of about 33%.

Our implementation of the algorithm was relatively inefficient: we solved the problems in a spreadsheet, with the computational overhead that that entails, and we used complete enumeration for the total number of ambulances *b*, rather than, say, bisection search. Thus, there are ample opportunities to solve the problem more efficiently.

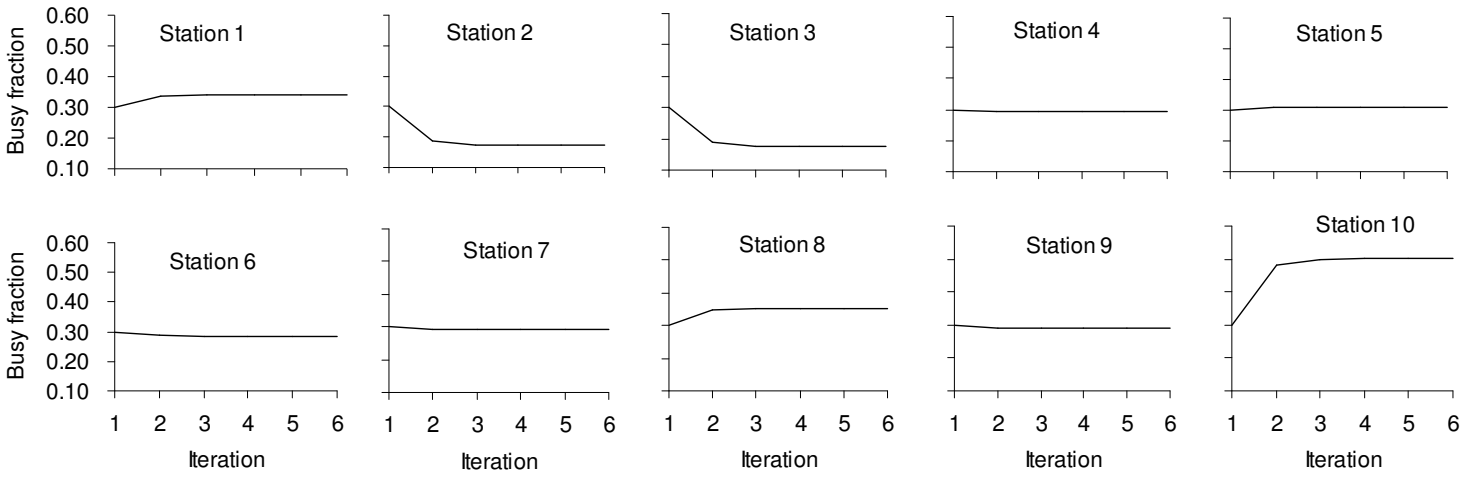

Figure 3: An example of iterating on the busy fractions ρ_i, where the initial input busy fraction was set to 0.3 for each station, and a smoothing constant of 0.9 was used.

We used the model to empirically explore the impact of varying the parameters of the delay distribution. Figure 4 shows how the minimum total number of ambulances needed to provide the specified coverage (90% reached in 9 minutes) changes when the mean and standard deviation of the delay distribution vary. We tried values that were 0%, 50%, 100%, 125%, and

150% of the current value for the mean (2.6 minutes) and for the standard deviation (1.3 minutes), except for combinations of parameters that made it impossible to meet the coverage goal. In all cases we used 0.3 for an initial estimate of the busy fraction at each station, and a smoothing constant, γ, of 0.9. Although we tried other values for these parameters (for example we tried 0.5 for the initial busy fractions estimates and a number of different values for γ), we found that these values worked well in our experiments. We will refer to the combination where both the mean and the standard deviation equal their current values as the *base case*. As Figure 4 shows, the total number of ambulances needed changes considerably when the parameters of the delay distribution are varied. The dramatic impact of ignoring the delay is illustrated by comparing the case when the delay is assumed to be zero to the base case. In the former case, only 11 ambulances are needed, while in the base case, 16 are needed.

Comparison of the case where the delay is assumed deterministic and equal to the current mean (i.e., the standard deviation is assumed to be zero) and the base case results in a less dramatic difference, of course: the number of ambulances needed increases from 15 to 16. However, the impact of ignoring the variability in delays would be far greater if the mean delay were higher. For example, if the mean delay were to increase by 25% (from 2.6 minutes to 3.25 minutes), while the standard deviation stayed the same, then 21 ambulances would be needed to reach the coverage goal. In this case, if the delay variability were ignored, then the model predicts that only 18 ambulances would be needed to reach the coverage goal. Hence, a model that incorporates delays but treats them as deterministic would underestimate the number of ambulances needed to provide the target coverage by (21-18)/21 = 14%.

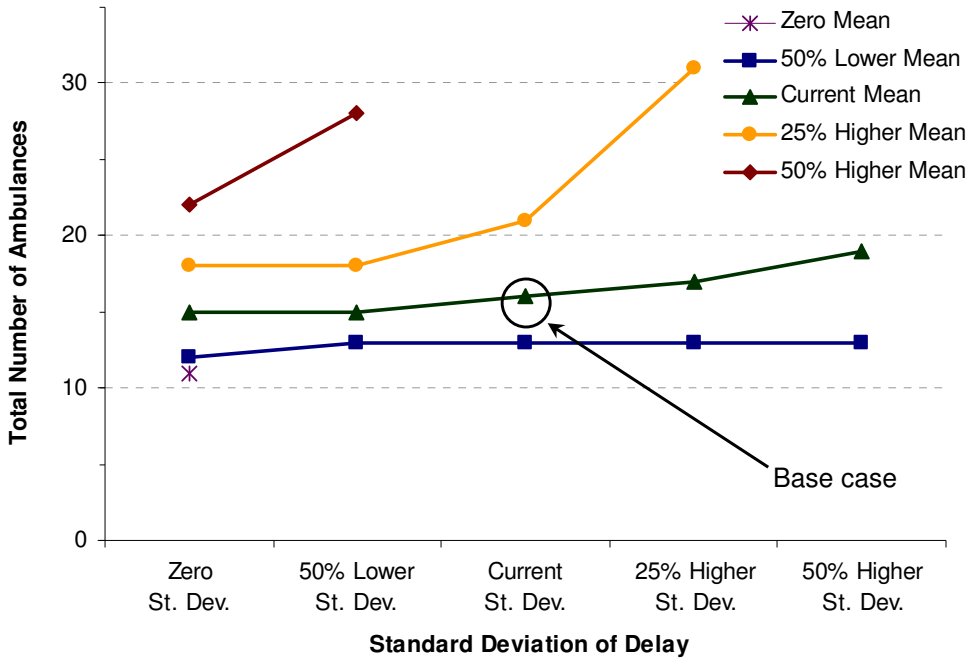

Figure 4: Sensitivity of the minimum total number of ambulances needed to provide the coverage goal to the mean and standard deviation of the delay distribution.

Figure 5 gives the complementary perspective and provides additional insight into the impact of the delay standard deviation. It demonstrates how the system wide coverage varies when the parameters of the delay distribution are varied in the same way as for the results in Figure 4, with the total number of ambulances fixed at 16. From Figure 5, we see that if the variability in the delay is not considered, the estimated coverage is about 92%, compared to just over 90% if the variability in the delay is incorporated. When the standard deviation is increased 25% from the base case, the coverage drops to about 89%. The results are magnified as the average level of the delay increases.

These results illustrate the importance of accounting for delays, and specifically the randomness in the delays, in order to obtain accurate estimates of the coverage and of the resources required to attain a specified coverage. They also illustrate the value of controlling the call-taking and

dispatching processes to ensure that delays do not increase (but preferably, decrease). This last point is important because delays can be far easier and less costly to reduce than travel times. It might be possible to reduce delays through simple process changes, such as dispatching an ambulance before the seriousness of the call has been established (thereby performing two activities in parallel rather than in series), or through the integration of 911 and EMS call centers (thus eliminating hand-off time from one call center to the other), whereas reducing travel times usually requires adding ambulances or stations. Our model can help compare the costs and benefits of actions to reduce delays versus actions to reduce travel times. This is valuable for decision-makers who are interested in the least-costly way of reaching service standards. As far as the response time standard is concerned, 30 seconds saved are 30 seconds saved, regardless of which component of the response time these savings come from.

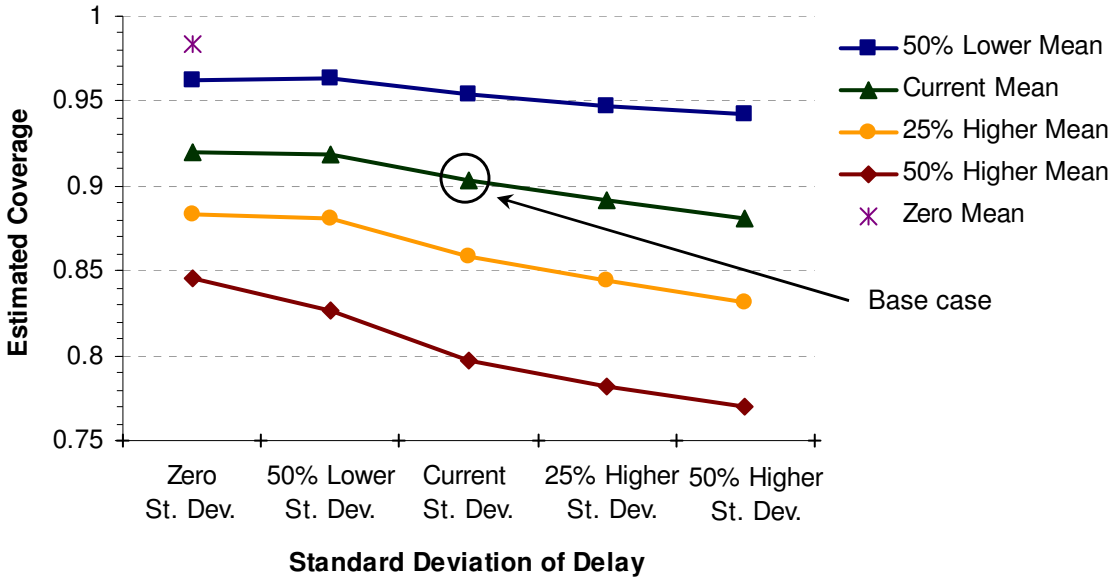

Figure 5: Sensitivity of the system wide service to the mean and standard deviation of the delay distribution, when the total number of ambulances is fixed at 16.

Discussion

This section outlines several possible avenues for further research involving exploration of the optimization model (P1), its properties, solution approaches, and insights from its application. First we discuss three extensions of the model that are fairly straightforward, and then we discuss some avenues for further research.

Model Extensions

One can add a constraint to (P1) to ensure that the probability that at least one ambulance is available is above some threshold β, as follows (assuming independence between ambulances):

$$1 - \prod_{i \in S} \rho_i^{x_i} \geq \beta \tag{7}$$

The constraint can be linearized by isolating the product of the busy fractions on one side of the inequality and taking logarithms of both sides, resulting in:

$$\sum_{i \in S} (-\ln(\rho_i)) x_i \geq -\ln(1-\beta) \tag{8}$$

Note that the coefficients $-\ln(\rho_i)$ and $-\ln(1-\beta)$ will be positive. Preliminary experiments using data from Edmonton indicated that the expected coverage target of reaching 90% of all calls in 9 minutes or less was tighter than constraint (8) for $\beta \leq 0.99$.

In addition to maximizing the system-wide coverage, one could add constraints on the coverage for each demand node, of the form

$$s_j(x) \geq \alpha_j, \text{ for all } j \in N \tag{9}$$

where α_j is the target coverage for demand node *j*. These constraints could be used to impose a common minimum coverage for all demand nodes or some demand node subset.

One can also add variables and constraints to decide which stations to open and to limit the number of ambulances at each station. Specifically, let y_i be a binary indicator variable for whether station *i* is opened; let c_i be the fixed cost of opening station *i*; let d_i be the variable cost of locating one ambulance at station *i*; and let b_i be the maximum number of ambulances at station *i*, if it is opened (if there are no such limits, then one can set $b_i = B$ for some sufficiently large number *B*). Upon replacing constraint (3) on the total number of ambulances with a budget constraint, the extended problem formulation becomes:

$$\text{(P2)} \quad \text{maximize} \quad s(x) \equiv \sum_{j \in N} h_j s_j(x)$$

$$\text{subject to} \quad \sum_{i \in S} (c_i y_i + d_i x_i) \le \text{budget}$$

$$(8), (4)$$

$$x_i \le b_i y_i, \text{ for all } i \in S$$

$$y_i \in \{0,1\}, \text{ for all } i \in S$$

The continuous relaxation of (P2) is a convex programming problem, by Proposition 1, but (P2) is more difficult to solve than (P1) because it has more integer variables.

Incorporating Time-Varying Demand

Demand for emergency medical services can vary considerably by time of the day and day of the week (for example see [7]). Ideally, the number and location of ambulances should vary with time to match such demand patterns. The model in this paper can be used as a building block in developing crew schedules to match time varying demand, as follows. First, divide the time horizon into time periods (for example, 168 hours for a weeklong time horizon) and solve our

model for each time period, resulting in an "ambulance requirement" (the minimum number of ambulances needed to reach the coverage target) for each period. Then, use a shift or tour scheduling model (see [13] for a recent survey) to generate crew schedules that match the ambulance requirements. In such a model, the ambulance requirements could be viewed as lower bounds that cannot be violated, or one could penalize deviations above and below the ambulance requirement and minimize the total penalty.

Our model evaluates system performance in steady state, and therefore the approach just described is an approximation, as it assumes that the system will quickly reach steady state during each time period. The approach is similar to what [17] termed the SIPP (Stationary Independent Period by Period) approach for generating staff requirements. As they demonstrated, although the SIPP approach is often an adequate approximation, it is unreliable in certain situations, such as when average service times are relatively long. Ambulances are often tied up for an hour or more with a single call, and this suggests that models that account for transient effects are worth investigating in this setting. A simple way to do this would be to use the heuristics proposed in [17]. A more sophisticated approach would combine optimization of crew schedules and ambulance location with a stochastic model that captures transient effects, either via simulation (as in [1]) or via numerical solution methods (as in [25] and [22]).

Future Research

Incorporation of uncertainty in availability and in delays and travel times may influence not only the total number of ambulances needed to provide a given level of service, but also how ambulances are distributed through the system. We plan to perform experiments to generate insight into whether this happens and how. In order to do further computational testing of the model, data from a city of similar size to Edmonton, but which is aggregated into many more

(smaller) zones and has up to 40 potential locations for ambulances will be used. We also hope to use the model to estimate the impact of various changes to the operation of an ambulance system. For example, it may be possible to reduce delays by performing activities in parallel rather than in series, but such a change may increase ambulance workload, if it results in more false alarms. Therefore, we would like to explore the trade-off between reducing delays and increasing busy fractions.

Estimation of the travel time distribution functions $H_{ij}(t)$ is likely to be challenging. We are working on developing procedures to estimate these functions, and have obtained detailed travel time data from a number of cities that we will use to validate such procedures. Preliminary results are reported in [5].

Although we can solve instances of our formulation involving Edmonton data in reasonable time, it is conceivable that problem instances for cities with more stations and ambulances will require the development of different heuristics to generate near-optimal solutions.

Conclusions

We have presented an optimization model for allocating a specified number of ambulances to stations so as to maximize system-wide expected coverage. The model differs from previous related work in that the variation in pre-travel delay is considered (in addition to the variation in travel time and uncertainty in ambulance availability) when calculating the coverage. Data from recent projects with the town of St. Albert and the City of Edmonton indicate that pre-travel delays are important and highly variable (with a standard deviation of about 40% of the mean). Our computational experiments demonstrate that the inclusion of the variability of such delays has a substantial impact on the solution that the model prescribes. Our formulation is sufficiently

tractable that problems with 180 demand nodes and 10 ambulance stations can be solved with reasonable effort with general-purpose solvers.

References

1. J. Atlason, M. A. Epelman, and S. G. Henderson (2007). Optimizing Call Center Staffing using Simulation and Analytic Center Cutting-Plane Methods. *Management Science*, published online before print Dec 11, 2007 , DOI: doi:10.1287/mnsc.1070.0774.
2. O. Berman and D. Krass (2001). Facility Location Problems with Stochastic Demands and Congestion. In *Location Analysis: Applications and Theory*, eds. Z. Drezner and H.W. Hamacher. Springer Verlag.
3. M. Brandeau and R.C. Larson (1986). Extending and Applying the Hypercube Model to Deploy Ambulances in Boston. In *Delivery of Urban Services*, eds. A. Swersey and E. Ignall. North Holland, New York.
4. L. Brotcorne, G. Laporte, F. Semet (2003). Ambulance Location and Relocation Models. *European Journal of Operational Research* **147** 451-463.
5. S. Budge (2004). Emergency Medical Service Systems: Modelling Uncertainty in Response Time. Ph.D. Dissertation. Department of Finance and Management Science, University of Alberta, Edmonton.
6. S. Budge, A. Ingolfsson, and E. Erkut (2007). Approximating Vehicle Dispatch Probabilities for Emergency Service Systems with Location-Specific Service Times and Multiple Units per Location. *Operations Research*, forthcoming.
7. N. Channouf, P. L'Ecuyer, A. Ingolfsson, and A. N. Avramidis (2007). The Application of Forecasting Techniques to Modeling Emergency Medical System Calls in Calgary, Alberta. *Health Care Management Science* **10** 25-45.
8. R. Church and C. ReVelle (1974). The Maximal Covering Location Problem. *Papers of the Regional Science Association* **32** 101–120.
9. M.S. Daskin (1983). A Maximum Expected Covering Location Model: Formulation, Properties, and Heuristic Solution. *Transportation Science* **17** 48–70.
10. M.S. Daskin (1987). Location, Dispatching, and Routing Model for Emergency Services with Stochastic Travel Times. In *Spatial Analysis and Location-Allocation Models*, eds. A. Ghosh and G. Rushton. Van Nostrang Reinhold Company, New York, 224–265.
11. D. J. Eaton, M.S. Daskin, D. Simmons, B. Bulloch, and G. Jansma (1985). Determining Emergency Medical Service Vehicle Deployment in Austin, Texas. *Interfaces* **15** 96–108.
12. E. Erkut, R. Fenske, S. Kabanuk, Q. Gardiner, and J. Davis (2001). Improving the Emergency Service Delivery in St. Albert. *INFOR* **39** 416–433.
13. A.T. Ernst, H. Jiang, M. Krishnamoorthy, and D. Sier (2004). Staff Scheduling and Rostering: A Review of Applications, Methods and Models. *European Journal of Operational Research* **153** 3–27.

14. J. Goldberg, R. Dietrich, J. M. Chen, M. G. Mitwasi, T. Valenzuela, and E. Criss (1990). A Simulation Model for Evaluating a Set of Emergency Vehicle Base Locations: Development, Validation, and Usage. *Socio-Economic Planning Sciences* **24** 125–141.

15. J. Goldberg, R. Dietrich, J. M. Chen, M. G. Mitwasi, T. Valenzuela, and E. Criss (1990). Validating and Applying a Model for Locating Emergency Medical Vehicles in Tucson, AZ. *European Journal of Operational Research* **49** 308–324.

16. J. Goldberg and L. Paz (1991). Locating Emergency Vehicle Bases when Service Time Depends on Call Location. *Transportation Science* **25** 264–280.

17. L. V. Green, P. J. Kolesar, and J. Soares (2001). Improving the SIPP Approach for Staffing Service Systems that have Cyclic Demand. *Operations Research* **49** 549–564.

18. D. Gross and C. M. Harris (1998). *Fundamentals of Queueing Theory*, Third Edition. Wiley, New York.

19. S. G. Henderson and A. J. Mason (2000). Development of a Simulation and Data Visualisation tool to assist in Strategic Operations Management in Emergency Services. School of Engineering Technical Report 595, University of Auckland, January 2000.

20. S. G. Henderson, and A. J. Mason (2004). Ambulance Service Planning: Simulation and Data Visualisation. Operations Research and Health Care: A Handbook of Methods and Applications, , eds. M. Brandeau, F. Sainfort, , and W. Pierskalla, Springer.

21. A. Ingolfsson, E. Erkut, and S. Budge (2003). Simulating a Single Start Station for Edmonton EMS. *Journal of the Operational Research Society* **54** 736–746.

22. A. Ingolfsson, E. Cabral, and X. Wu (2007). Combining Integer Programming and the Randomization Method to Schedule Employees. Working paper, available from http://www.business.ualberta.ca/aingolfsson/publications.htm.

23. J. Jarvis (1981). Optimal Assignments in a Markovian Queueing System. *Computers and Operations Research* **8** 17–23.

24. J. Jarvis (1985). Approximating the Equilibrium Behavior of Multi-Server Loss Systems. *Management Science* **31** 235–239.

25. P. J. Kolesar, K. L. Rider, T. B. Crabill, and W. E. Walker (1975). A Queuing-Linear Programming Approach to Scheduling Police Patrol Cars. *Operations Research* **23** 1045–1062.

26. R.C. Larson (1974). A Hypercube Queueing Model for Facility Location and Redistricting in Urban Emergency Services. *Computers and Operations Research* **1** 67–95.

27. R.C. Larson (1975). Approximating the Performance of Urban Emergency Service Systems. *Operations Research* **23** 845–868.

28.R.C. Larson (1979). Structural System Models for Locational Decisions: An Example Using the Hypercube Queueing Model. *Operational Research '78, Proceedings of the Eighth IFORS International Conference on Operations Research,* ed. K. B. Haley. North-Holland Publishing Co., Amsterdam, Holland.

29. V. Marianov and C. ReVelle (1995). Siting Emergency Services. *Facility Location: A Survey of Applications and Methods*, ed. Z. Drezner, Springer.

30. V. Marianov and 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.

31. A. J. Swersey (1994). The Deployment of Police, Fire, and Emergency Medical Units. *Handbooks in Operations Research and Management Science, Vol. 6: Operations Research and the Public Sector*, eds. S.M. Pollock, M.H. Rothkopf and A. Barnett, North-Holland.

32.C. Toregas, R. Swain, C. ReVelle, and L. Bergman (1971). The Location of Emergency Service Facilities. *Operations Research* **19** 1363–1373.

33. T. R. Willemain and R. C. Larson, eds. (1977). *Emergency Medical Systems Analysis*. Lexington Books, Lexington, MA.

Appendix A: Proof of Proposition 1

Recall that the system–wide coverage $s(x) = \sum_{j \in N} h_j s_j(x)$ is a convex combination of the coverages $s_j(x)$ for each demand node *j*. To prove that *s*(*x*) is concave, it suffices to prove that the coverage $s_j(x)$ for a particular node *j* is concave, since the weights h_j are positive. Therefore, we assume without loss of generality that there is only one demand node and we drop the demand node subscript *j* in the proof to simplify notation.

By assumption we have $\Delta w_i = w_{i+1} - w_i \le 0$ for all *i*. We can express the probability $f_i(x)$ as:

$$f_i(x) = Q_i\left(1-\rho_i^{x_i}\right)\prod_{u=1}^{i-1}\rho_u^{x_u} = Q_i\left(\prod_{u=1}^{i-1}\rho_u^{x_u} - \prod_{u=1}^{i}\rho_u^{x_u}\right) = g_{i-1}(x) - g_i(x)$$

where $g_i(x) = Q_i\prod_{u=1}^{i}\rho_u^{x_u}$ and $g_0(x) = 1$. Consequently,

$$\begin{aligned} s(x) &= \sum_{i \in S} f_i(x) w_i = \sum_{i=1}^{m} g_{i-1}(x) w_i - \sum_{i=1}^{m} g_i(x) w_i \\ &= \sum_{i=0}^{m} g_i(x) w_{i+1} - \sum_{i=1}^{m} g_i(x) w_i = w_1 + \sum_{i=1}^{m} g_i(x) \Delta w_i \end{aligned}$$

with the understanding that $w_{m+1} = 0$.

The gradient of *s*(*x*) with respect to *x* has the following entries:

$$\frac{\partial s}{\partial x_k} = (\ln \rho_k) \sum_{i=k}^{m} g_i(x) \Delta w_i$$

The entries in the Hessian matrix *H* are (assuming $k \le l$):

$$h_{kl} = \frac{\partial^2 s}{\partial x_k \partial x_l} = (\ln \rho_k)(\ln \rho_l) \sum_{i=l}^{m} g_i(x) \Delta w_i$$

Recalling that $Q_i > 0$, $\rho_i \in (0,1)$ and $\Delta w_i \le 0$, we see that $\partial s / \partial x_k$ is non-negative for all *k*, and $\partial^2 s / \partial x_k \partial x_l$ is non-positive for all *k* and *l*.

Consider the quadratic form $y^T H y$ where *y* is an arbitrary column vector with *m* elements. This quadratic form can be expressed as:

$$y^T H y = \sum_{k=1}^{m} \sum_{l=1}^{m} y_k y_l h_{kl} = \sum_{l=1}^{m} y_l^2 h_{ll} + 2 \sum_{k=1}^{m} \sum_{l=k+1}^{m} y_k y_l h_{kl}$$

Substituting the expression for h_{kl} we get:

$$y^T H y = \sum_{l=1}^{m} y_l^2 (\ln \rho_l)^2 \sum_{i=l}^{m} g_i(x) \Delta w_i + 2 \sum_{k=1}^{m} \sum_{l=k+1}^{m} y_k y_l (\ln \rho_k)(\ln \rho_l) \sum_{i=l}^{m} g_i(x) \Delta w_i \qquad \text{(A.1)}$$

By changing the order of summation, the double sum in (A.1) can be expressed as:

$$\sum_{l=1}^{m} y_l^2 (\ln \rho_l)^2 \sum_{i=l}^{m} g_i(x) \Delta w_i = \sum_{i=1}^{m} g_i(x) \Delta w_i \sum_{l=1}^{i} (\ln \rho_l)^2 y_l^2$$

Similarly, the triple sum in (A.1) can be expressed as:

$$\begin{aligned} \sum_{k=1}^{m} \sum_{l=k+1}^{m} y_k y_l (\ln \rho_k)(\ln \rho_l) \sum_{i=l}^{m} g_i(x) \Delta w_i &= \sum_{k=1}^{m} \sum_{i=k+1}^{m} g_i(x) \Delta w_i \sum_{l=k+1}^{i} y_k y_l (\ln \rho_k)(\ln \rho_l) \\ &= \sum_{i=2}^{m} g_i(x) \Delta w_i \sum_{k=1}^{i-1} \sum_{l=k+1}^{i} y_k y_l (\ln \rho_k)(\ln \rho_l) \end{aligned}$$

Substitution in (A.1) results in:

$$y^T Hy = \sum_{i=1}^{m} g_i(x)\Delta w_i \left\{ \sum_{l=1}^{i} (\ln \rho_l)^2 y_l^2 + 2\sum_{k=1}^{i-1} \sum_{l=k+1}^{i} (\ln \rho_k)(\ln \rho_l) y_k y_l \right\}$$
$$= \sum_{i=1}^{m} g_i(x)\Delta w_i \left(\sum_{l=1}^{i} (\ln \rho_l) y_l \right)^2$$

We see that each term in the outer summation is non-positive (because $g_i(x) \geq 0$, $\Delta w_i \leq 0$, and the squared summation is non-negative) and therefore $y^T Hy \leq 0$ for all y. Consequently, H is negative semi-definite and $s(x)$ is concave. ■

Appendix B: Estimating the Average Busy Fraction

The average fraction of time that an ambulance is busy (not available to respond to calls) is $\lambda\tau / z$, i.e., the average server utilization for a z-server queueing system, assuming that the number of calls "lost" due to queueing is negligible. The average "service time", τ, (during which an ambulance is tied up with a call) can be broken down into the following components: average travel time to the call, average on-scene time, and average time spent traveling to and remaining at a hospital, denoted $\mathrm{E}[T_{\text{to call}}]$, $\mathrm{E}[T_{\text{on scene}}]$, and $\mathrm{E}[T_{\text{hospital}}]$, respectively. Consequently, the average busy fraction can be expressed as $\lambda(\mathrm{E}[T_{\text{to call}}] + \mathrm{E}[T_{\text{on scene}}] + \mathrm{E}[T_{\text{hospital}}])/z$. The arrival rate λ as well as two of the three components of the average service time, the average on-scene time and the average time spent traveling to and being at a hospital, are exogenous input. The average travel time to a call can be expressed as $\mathrm{E}[T_{\text{to call}}] = \sum_{j \in N} h_j \sum_{i \in S} f_{ij}(x)\,\mathrm{E}[T_{ij}]$, where T_{ij} is the travel time from i to j. This leads to the following formula for approximating ρ as a function of x:

$$\rho(x) = \frac{\lambda}{z(x)} \left\{ \sum_{j \in N} h_j \sum_{i \in S} f_{ij}(x) \, \mathrm{E}[T_{ij}] + \mathrm{E}[T_{\text{on scene}}] + \mathrm{E}[T_{\text{hospital}}] \right\}$$

The derivation of this formula required some approximations. In particular, we excluded the time spent traveling back to a station from the hospital from the average service time since the ambulance is available to respond to incoming calls during this time. On the other hand, our expression for $\mathrm{E}[T_{\text{to call}}]$ assumes that all calls are responded to from an ambulance at a station.