Erlang Loss Models for the Static Deployment of Ambulances

Mateo Restrepo [1]
Center for Applied Mathematics,
Cornell University, Ithaca, NY 14853, USA
tel. 607-255-7760, fax. 607-255-9860
mr324@cornell.edu

Shane G. Henderson and Huseyin Topaloglu
School of Operations Research and Information Engineering
Cornell University, Ithaca, NY 14853, USA
{sgh9,ht88}@cornell.edu

April 15, 2007

Acknowledgements: We thank Armann Ingolfsson for the data used for the numerical experiments in Section 5. This research was partially supported by National Science Foundation grants DMI 0400287 and DMI 0422133.

[1]Corresponding author

Erlang Loss Models for the Static Deployment of Ambulances

August 8, 2007

Abstract

We consider the problem of finding a static deployment of ambulances in an emergency medical service system. The goal is to allocate a given number of ambulances among a set of bases to minimize the fraction of calls that are not reached within a time standard. We present two models. Our first model estimates the performance of the system by using the Erlang loss function and embeds this function in an optimization problem to find a desirable ambulance allocation. This model captures the dependence between the ambulances that are deployed at the same base, but it assumes *a priori* knowledge of the amounts of demand served by the different bases. Our second model uses the Erlang loss function to construct a set of equations that characterize the performance of the system. This model is aimed at obtaining a good estimate of the performance of the system for a fixed ambulance allocation, but it does not require *a priori* knowledge of the amounts of demand served by the different bases. Computational experiments provide useful insights.

Keywords: ambulance allocation, static deployment, logistics, optimization, Erlang loss, simulation, ambulance base, lost calls, performance measures, spatial queues.

1 Introduction and Relevant Literature

Emergency medical service providers face the problem of allocating a fixed number of ambulances among a set of bases with the objective of ensuring reasonable response times. The response time of a call is the elapsed time from when the call is received to when an ambulance arrives at the scene. Providers are usually interested in minimizing the fraction of calls with response times longer than some threshold, which is generally around ten minutes. In this paper, we study the static deployment problem, where ambulances return to their assigned bases whenever they are available to serve calls. We use the Erlang loss formula to construct models that add insight, and provide a computational tool to assess the performance of a given ambulance allocation.

Static ambulance deployment problems have received a great deal of attention in the literature. We present a brief survey here to give a feel for the primary approaches. For more detailed surveys, we recommend the excellent reviews by Swersey [1994], Brotcorne et al. [2003], Goldberg [2004] and Green and Kolesar [2004].

Static deployment methodology can be broken down into two primary areas. *Prescriptive* methods search over all possible ambulance allocations to identify a subset of allocations that merit further study and they typically use simplified performance measures to allow for efficient search procedures. Important early work on prescriptive models includes Toregas et al. [1971] and Church and ReVelle [1974]. The more recent model by Daskin [1983] attempts to capture some of the stochastic aspects of the problem under the assumption that the ambulances are statistically independent. Batta et al. [1989] build on this model to relax the independence assumption, whereas ReVelle and Hogan [1989] and Marianov and ReVelle [1994] extend the earlier models by adding constraints on the minimum number of ambulances required in various zones. The last two papers are related to our work in the sense that the minimum number of ambulances are obtained from the Erlang loss formula. More recently, Erdogan et al. [2006] incorporate medical outcomes into the objective function by concentrating on the lengths of the response times rather than the fraction of calls with response times below a certain time standard. All of the papers mentioned in this paragraph thus far generally use integer programming formulations, but nonlinear models have also been used by several authors. For example, Goldberg and Paz [1991] embed a heuristic search procedure into the nonlinear approximation scheme proposed by Jarvis [1975]. The underlying theme of this work is to use tractable approximations for the performance of the system to guide a search procedure.

Descriptive methods complement prescriptive methods by providing the ability to carefully evaluate a proposed ambulance allocation. These models provide more accurate estimates for quantities such as the utilizations of the ambulances and the probability that the response time is longer than some time standard. This area is dominated by the hypercube method proposed by Larson [1974] and its extensions, which include the work by Jarvis [1975], Jarvis [1985], Berman and Larson [1982], Brandeau and Larson [1986], Goldberg and Szidarovsky [1991] and Budge et al. [2006]. Another important descriptive approach is, of course, simulation. Simulation has been used in a large portion of ambulance deployment studies either to analyze the performance of other models or as a tool in and of itself. For an overview, we refer the reader to Henderson and Mason [2004].

In this paper, we present two models for static ambulance deployment. One of our models is prescriptive and the other one is descriptive. Both use the Erlang loss function to explicitly capture the uncertainty in ambulance availability and the dependence between ambulances.

We begin with a prescriptive model that is mostly intended to provide managerial insight into the ambulance deployment problem. This model is an integer program that allocates a fixed number of ambulances among a set of bases. The objective is to minimize the number of calls that do not find an available ambulance when they are received, and we capture this fraction by using the Erlang loss function. The solution to the two-base version of this model provides useful insights. In particular, it turns out that it is not optimal to allocate ambulances to bases in proportion to the call loads experienced by the bases. Instead, bases with lighter call loads receive more than a proportional share of the ambulances, and allocating ambulances in proportion to the call loads can result in practically significant increases in the fraction of calls that do not find an available ambulance when they are received.

The main simplifying assumption of this model is that it does not allow collaboration between the different bases, or put in a slightly different way, this model assumes *a priori* knowledge of the amounts of demand served by the different bases. Of course, bases always interact to some extent, but there are situations where bases cannot interact as much as one would like owing to natural barriers such as the sea harbors in Auckland, New Zealand or traffic congestion on arterial roadways. The results for our first model should also be useful for a centralized planner who is responsible for allocating the emergency medical service assets between different cities that do not share resources operationally.

This first model adds insight, but the assumption that the bases do not interact is unfortunate. Our second model relaxes this assumption, though at the expense of loss of tractability. The second model estimates, for a given ambulance allocation, the fraction of calls whose response times are longer than some time standard. It can also be used to estimate other performance measures, although we do not emphasize that generality here. The main idea is to use the Erlang loss formula with carefully chosen parameters to approximate the probability that all ambulances at one or more bases are busy. These probabilities are then used to compute the probability that a call will be answered by an ambulance at a particular base. The approach is essentially a fixed point scheme for a nonlinear system of equations and it is similar in spirit to the work of Larson [1974], Jarvis [1975] and Budge et al. [2006]. However, an important distinction of our work is that the earlier work formulates fixed point equations that involve a sophisticated version of the so-called Q factors, whereas our model completely avoids the need for Q factors by using the Erlang loss function. Although we do not provide a proof of convergence, the computational results indicate that there exists a unique fixed point and the fixed point can be computed extremely quickly. Therefore, our model can be used to prescreen a large number of ambulance allocations, after which the best allocations can be studied through simulation.

This paper is organized as follows. Section 2 introduces our first model, which is prescriptive. Section 3 studies the two-base version of this model in detail. Section 4 describes our second model, which is descriptive. Section 5 compares the estimates of the performance measures obtained by this model with those obtained through a simulation model. We conclude in Section 6 with a discussion of

static allocation schemes and how they compare with dynamic allocation procedures.

2 A Prescriptive Model for Static Ambulance Deployment

In this section, we present a prescriptive model for the static ambulance deployment problem. We have N ambulances to allocate among a set of bases to minimize the number of calls that do not find an available ambulance when they are received. The set of bases is $\mathcal{B}$ and we can allocate at most c_b ambulances to base b. An ambulance deployed at base b can serve the calls with rate μ_b and the arrival rate of the calls that are offered to base b is λ_b. We assume that we have *a priori* knowledge of the demand served by the different bases so that $\{\lambda_b : b \in \mathcal{B}\}$ are known parameters that do not depend on the allocation. We relax this assumption later in Section 4.

If we allocate n_b ambulances to base b, then this base operates as an $M/G/n_b/n_b$ queue. This implies that we can compute the probability that an arriving call offered to base b finds all n_b ambulances busy by using the Erlang loss formula

$$\mathcal{E}(n_b, \lambda_b/\mu_b) = \frac{[\lambda_b/\mu_b]^{n_b}/n_b!}{\sum_{j=0}^{n_b}[\lambda_b/\mu_b]^j/j!}.$$

In this case, we can solve the problem

$$\begin{aligned}
\min \quad & \sum_{b\in\mathcal{B}} \lambda_b\, \mathcal{E}(n_b, \lambda_b/\mu_b) && && (1)\\
\text{subject to} \quad & \sum_{n\in\mathcal{B}} n_b = N && && (2)\\
& n_b \le c_b && b \in \mathcal{B} && (3)\\
& n_b \in \mathbb{Z}_+ && b \in \mathcal{B} && (4)
\end{aligned}$$

to decide how many ambulances should be allocated to each base. In the problem above, $\mathcal{E}(n_b, \lambda_b/\mu_b)$ is the probability that a call offered to base b is lost when we allocate n_b ambulances to this base. Therefore, $\lambda_b\, \mathcal{E}(n_b, \lambda_b/\mu_b)$ is the expected number of lost calls offered to base b per unit time. It is well known that $\mathcal{E}(n_b, \lambda_b/\mu_b)$ is a convex function of n_b, which implies that the objective function of Problem (1)-(4) is a separable convex function; see Harel [1988]. Consequently, Problem (1)-(4) can be solved efficiently through simple marginal analysis; see Fox Fox [1966]. (One takes a greedy approach, adding ambulances one at a time, respecting the base capacities but otherwise choosing the base that leads to the greatest reduction in the objective function.)

Several remarks are in order for Problem (1)-(4). As mentioned before, the objective function assumes that if we allocate n_b ambulances to base b, then this base operates as an $M/G/n_b/n_b$ queue. The implicit assumption here is that a call that is offered to base b but cannot be immediately served by an ambulance deployed at this base is immediately lost, or equivalently, handled by an alternative agent such as the fire department or a competing ambulance organization. This is an appropriate assumption when the proportion of lost calls is small, as is the case in many emergency medical service systems. We also note that using the Erlang loss formula in the objective function amounts to assuming that the response times for successive calls are independent, but this is not strictly correct, because the locations

of the ambulances are affected by the locations of previous calls. This assumption is reasonable when a large portion of the calls are served by ambulances that are already waiting at the bases or when the time spent by an ambulance at the scene is much longer than the travel times. Again, both of these assumptions are usually approximately satisfied in practice.

The most unfortunate assumption in Problem (1)-(4) is that we have *a priori* knowledge of what portions of the demand are served by the different bases. In reality, the total rate of call arrivals into the system is known, but the proportions of the calls served by the different bases emerge as a result of the ambulance dispatch policy. Problem (1)-(4) assumes that fixed and known portions of the calls arriving into the system are offered to the different bases and the bases do not help each other even if they are not able to cope with their offered call loads. The primary reason for this assumption is mathematical tractability. In addition, as mentioned in the introduction, there are occasions where the bases cannot interact due to geographic barriers or traffic congestion. We also note that the ambulance allocations provided by Problem (1)-(4) can be very useful for a centralized planner who is responsible for allocating ambulances between different cities that do not share resources operationally. It is in situations such as these that we believe the insights generated by this model will be of most use.

One can attempt to overcome the assumption that $\{\lambda_b : b \in \mathcal{B}\}$ are known parameters by turning these parameters into decision variables in Problem (1)-(4). Without going into the details, this idea roughly amounts to solving a problem that looks like

$$\begin{aligned} \min \quad & \sum_{b\in\mathcal{B}} \lambda_b \, \mathcal{E}(n_b, \lambda_b/\mu_b) \\ \text{subject to} \quad & (2), (3), (4) \\ & \sum_{b\in\mathcal{B}} \lambda_b = \lambda \\ & \lambda_b \geq 0 \qquad b \in \mathcal{B}, \end{aligned}$$

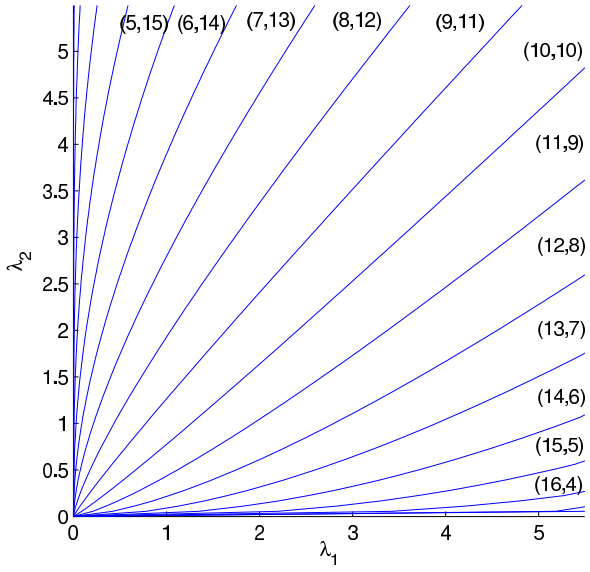
where λ is the total rate of call arrivals into the system, and $\{n_b : b \in \mathcal{B}\}$ and $\{\lambda_b : b \in \mathcal{B}\}$ are the decision variables. The problem above finds an allocation of the calls and ambulances among the bases to minimize the number of lost calls, and it mimics the natural load balancing sought after by dispatchers in practice. Unfortunately, $\lambda_b \, \mathcal{E}(n_b, \lambda_b/\mu_b)$ is not a jointly convex function of (n_b, λ_b). This new problem therefore has many local optima and finding a global optimum is almost as difficult as enumerating all possible values of $\{n_b : b \in \mathcal{B}\}$ that satisfy Constraints (2)-(4). In Section 4, we revisit this problem of allocating the total call volume among the bases and come up with a fixed point scheme for this purpose.

Despite its limitations, Problem (1)-(4) provides useful insights, especially in the two-base case that is examined in some detail next.

3 Insights from the Prescriptive Model

In this section, we consider problem (1)-(4) for the case where we have two bases so that $\mathcal{B} = \{1, 2\}$. The most important observation we want to convey here is that allocating the ambulances in proportion

Figure 1: Optimal solutions to Problem (1)-(4) for different values of (λ_1, λ_2).



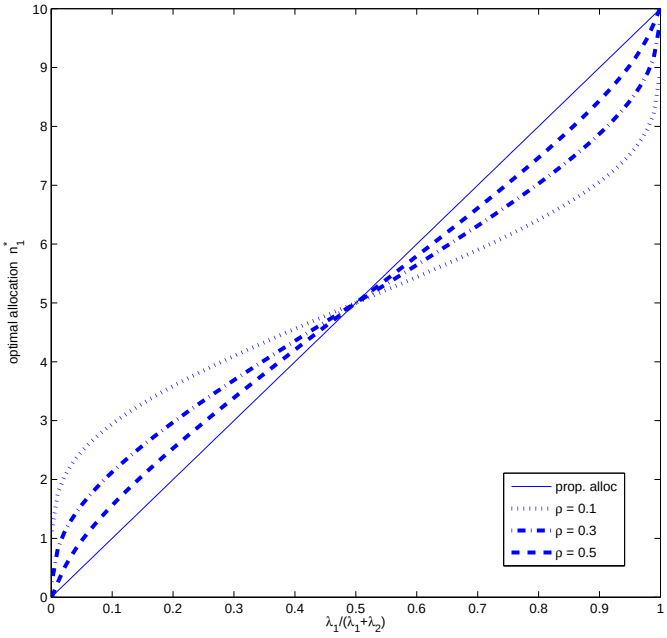
to the call loads offered to the different bases can be highly suboptimal. We also explore how the suboptimality of the proportional allocations changes with the system utilization.

We first let $(\mu_1, \mu_2) = (1, 1)$ and $N = 20$, and solve Problem (1)-(4) for different values of (λ_1, λ_2). We choose $(\mu_1, \mu_2) = (1, 1)$ only for convenience and other service rates can be captured by simply scaling our results. The regions and the associated labels in Figure 1 show the optimal solutions to Problem (1)-(4) for different values of (λ_1, λ_2). For example, if $(\lambda_1, \lambda_2) = (3, 2)$, then the optimal solution to Problem (1)-(4) is $(n_1, n_2) = (11, 9)$. In Figure 1, it is easy to see that allocating the ambulances in proportion to the call loads offered to the different bases is almost never optimal. For example, if $(\lambda_1, \lambda_2) = (5, 2.5)$, then we have $\lambda_1/\lambda_2 = 2$, but the optimal solution to Problem (1)-(4) is $(n_1, n_2) = (12, 8)$ and $n_1/n_2 = 1.5$.

An interesting observation from Problem (1)-(4) is related to the effect of the system congestion on the suboptimality of proportional allocation. To illustrate, we let $(\mu_1, \mu_2) = (1, 1)$ and $N = 10$. Since the service rates are equal to 1, the system utilization is $\rho = (\lambda_1 + \lambda_2)/N$. In Figure 2, we plot n_1 in the optimal solution $(n_1, N - n_1)$ to Problem (1)-(4) as a function of $\lambda_1/(\lambda_1 + \lambda_2)$ for different values of ρ. Notice that the ratio $\lambda_1/(\lambda_1 + \lambda_2)$ gives the proportion of the total demand $\lambda_1 + \lambda_2$ that is offered to the first base, and ρ is a measure of congestion.

If proportional allocation was optimal, then we would have $n_1/N = \lambda_1/(\lambda_1 + \lambda_2)$ and the optimal solutions to Problem (1)-(4) for different values of $\lambda_1/(\lambda_1 + \lambda_2)$ would lie on the straight line in Figure 2 irrespective of the system utilization. However, we observe that the optimal solutions differ substantially from proportional allocation, especially when the system utilization is low. When system utilization is high, proportional allocation is close to optimal. However, this requires utilization values as large as 0.5, whereas a more typical value in practice is 0.3. We also note that if $\lambda_1/(\lambda_1 + \lambda_2) < 0.5$, then the number of ambulances allocated to the first base is larger than that suggested by proportional allocation. In

Figure 2: Value of n_1 in the optimal solutions $(n_1, N - n_1)$ to Problem (1)-(4) for different values of $\lambda_1/(\lambda_1 + \lambda_2)$.



other words, bases with lighter loads receive more than their proportional share of the ambulances.

Figure 3 gives the loss rates for proportional allocation and the optimal allocation as a function of $\lambda_1/(\lambda_1 + \lambda_2)$. The results indicate that the cost of using proportional allocation can be very large. For example, if $\lambda_1/(\lambda_1 + \lambda_2) = 0.1$, then proportional allocation loses about 7% of the calls, whereas the optimal solution loses about 2% of the calls. This is certainly a practically significant difference.

In producing Figures 2 and 3 we used a continuous version of the Erlang loss formula (Jagers and van Doorn [1986]) that allows us to compute the Erlang loss formula for fractional ambulance allocations and thereby obtain smooth plots.

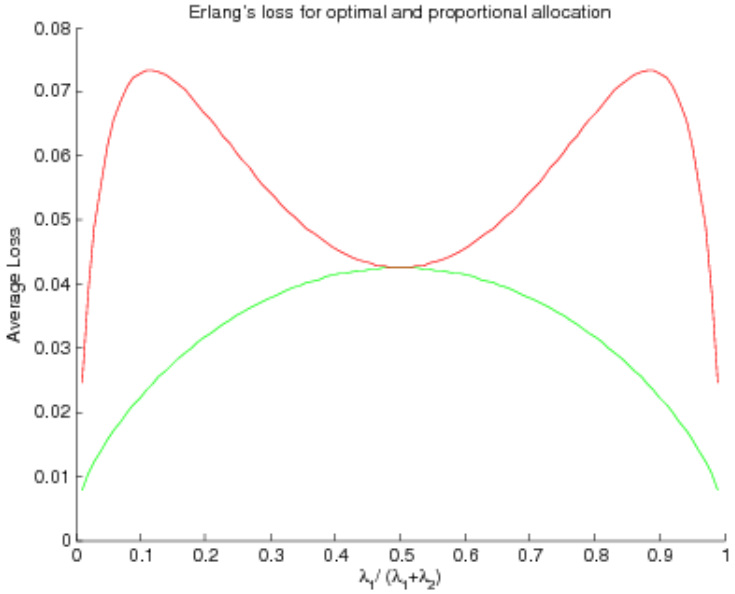
4 A Descriptive Model for Estimating Performance Measures

In this section, we propose a descriptive model that estimates performance measures for a given ambulance allocation. This model relaxes some of the critical assumptions in Section 2. In particular, it does not require the assumption that we know the call loads offered to different bases, nor the assumption that bases do not cooperate.

Instead of assuming that we *a priori* know what portions of the total demand are offered to the different bases, we now assume that there are a number of demand nodes indexed by $j = 1, \ldots, J$. Each demand node j generates a Poisson arrival process of calls with rate d_j and the processes corresponding to different demand points are independent. The set of bases is $\mathcal{B} = \{1, \ldots, B\}$.

Each demand node j has a fixed priority ranking of all bases $j(1), \ldots, j(B)$. A call originating at

Figure 3: Rate of lost calls for proportional allocation and the optimal allocation.



node j is first offered to an ambulance based at $j(1)$, and if no ambulance is available there, then it is offered to an ambulance based at $j(2)$, and so forth. If no ambulance is available at any base, then the call is lost and assumed to be answered by some alternative agent. The term "offered" emphasizes that some calls contribute to the load on a base, even if they are not actually served by that base. We say that a base is busy if all ambulances stationed at this base are busy.

For all $k = 1, \ldots, B+1$, let $\mathcal{A}_{jk}(t)$ denote the event that the first $k-1$ bases in demand node j's ranking of bases are busy at time t. In other words, $\mathcal{A}_{jk}(t)$ is the event that none of the bases preferred to the k-th base in demand node j's list has an available ambulance at time t. The event $\mathcal{A}_{j1}(t)$ has probability 1 by definition. If a call is generated at demand node j at time t, then on the event $\mathcal{A}_{jk}(t)$, the call is offered to base $j(k)$. We write $\mathcal{A}_{jk}$ for $\mathcal{A}_{jk}(0)$ and let $a_{jk} = P(\mathcal{A}_{jk})$ be the stationary probability of the event $\mathcal{A}_{jk}(t)$. (Here we assume that $P(\cdot)$ is a probability measure under which the system is stationary, so that our calculations are for a system in steady state.) We can interpret a_{jk} as the probability that a call originating at demand node j will be offered to base $j(k)$. We note that we are implicitly using a Poisson-arrivals-see-time-averages argument here to ensure that the Poisson arrivals at demand node j see the time-average behavior of the system. We also use the fact that the time-average behavior corresponds with the stationary behavior. For background on both of these topics, we refer the reader to Wolff [1989]. Similarly, we let $\mathcal{S}_{jk}(t)$ denote the event that the first available ambulance in the priority list of bases $j(1), \ldots, j(B)$ at time t is at $j(k)$, and write $\mathcal{S}_{jk}$ for $S_{jk}(0)$. In this case, $s_{jk} = P(\mathcal{S}_{jk})$ is the probability that a call originating at node j will be answered by an ambulance from base $j(k)$. We extend the definition of s_{jk} to include $k = B+1$, so that $s_{j,B+1}$ is the probability that no ambulance is available anywhere.

Suppose we are able to obtain approximations for these probabilities. In this case, it is possible to compute an approximation for the fraction, $r(j)$, of the calls originating from demand node j that meet

a time standard. In particular, if demand node j can be reached from, and only from, the first $N(j)$ bases in demand node j's list within the time standard, then we have

$$r(j) = \sum_{k=1}^{N(j)} s_{jk}.$$

Moreover, the overall fraction of calls that meet the time standard is

$$r = \frac{\sum_{j=1}^{J} d_j \, r(j)}{\sum_{j=1}^{J} d_j}. \tag{5}$$

We therefore turn to developing an approximation for the probabilities s_{jk} for all $j = 1, \ldots, J$, $k = 1, \ldots, B$.

We begin by noting that $\mathcal{A}_{j,k+1} \subseteq \mathcal{A}_{jk}$ and $\mathcal{S}_{jk} = \mathcal{A}_{jk} \backslash \mathcal{A}_{j,k+1}$ for all $k = 1, \ldots, B$. In other words, the first k bases on demand node j's list can be busy only when the first $k-1$ bases are busy, and a call is answered by the k-th base in demand node j's list if and only if the first $k-1$ bases are busy, but the k-th base is not. Hence, $s_{jk} = a_{jk} - a_{j,k+1}$, $\mathcal{A}_{jk} = \cup_{i=k}^{B+1} \mathcal{S}_{ji}$, and $a_{jk} = \sum_{i=k}^{B+1} s_{ji}$ for all $k = 1, \ldots, B$. We let A and S respectively be the matrix of values $\{a_{jk} : j = 1, \ldots J, \ k = 1, ..., B+1\}$ and $\{s_{jk} : j = 1, \ldots J, \ k = 1, ..., B+1\}$. Since we can recover either of these matrices from the other one, we focus on computing A through a fixed point equation.

Recall that $a_{j1} = 1$ for all $j = 1, \ldots, J$. To approximate a_{j2}, we first let λ_k be the "offered" load to base k, so that

$$\lambda_k = \sum_{j=1}^{J} d_j \, a_{jk}. \tag{6}$$

We then approximate a_{j2}, the probability that no ambulances are available at base $j(1)$, by

$$a_{j2} \approx \mathcal{E}(n_{j(1)}, \lambda_{j(1)}/\mu_{j(1)}). \tag{7}$$

(Note that the number of ambulances n_k and the service rate μ_k at base k are constants that do not change in our scheme.) It now remains to show how to compute approximations for a_{jm} for all $m > 2$. The nested nature of the events $\mathcal{A}_{jm}$ for all $m = 1, \ldots, B$ ensures that we have

$$a_{jm} = P(\mathcal{A}_{j2}) \, P(\mathcal{A}_{j3} | \mathcal{A}_{j2}) \cdots P(\mathcal{A}_{jm} | \mathcal{A}_{j,m-1}) \tag{8}$$

for $m > 2$. We approximate each factor of the form $P(\mathcal{A}_{j,k+1} | \mathcal{A}_{jk})$ by

$$P(\mathcal{A}_{j,k+1} | \mathcal{A}_{jk}) = \mathcal{E}(n_{j(k)}, \lambda(j,k)/\mu_{j(k)}) \tag{9}$$

for all $k = 2, \ldots, B$, where $\lambda(j,k)$ represents the conditional demand offered to base $j(k)$ conditional on the event $\mathcal{A}_{jk}$. By conditioning on $\mathcal{A}_{jk}$, the demand offered to base $j(k)$ may not be Poisson, but we ignore this issue in our approximation. In analogy with (6), the conditional demand can be written as

$$\lambda(j,k) = \sum_{i=1}^{J} d_i \, P(\mathcal{A}_{i,\{j(k)\}} | \mathcal{A}_{jk}), \tag{10}$$

where $\{j(k)\}$ is the index of base $j(k)$ in demand node i's list of bases. The probability in (10) computes the conditional probability that a call originating at demand node i will be offered to base $j(k)$, conditional on the first $k-1$ bases in demand node j's priority list being busy.

If all of the bases that appear before base $j(k)$ in demand node i's list also appear in demand node j's list before base $j(k)$, then we know that they are all busy, since we are conditioning on $\mathcal{A}_{jk}$. In that case, the conditional probability that the call originating at demand node i will be offered to base $j(k)$ is 1. If not, then we essentially need to compute the probability that a subset of the bases are busy, given that another subset of the bases are busy. In principle, it is possible to include such probabilities as separate variables and to write linking equations for them. However, this would cause a dramatic increase in the number of variables. To produce a tractable set of equations, we instead adopt the following approximation.

For given demand nodes i, j and a given base $j(k)$, we let n be such that $j(k) = i(n)$. In other words, n plays the role of $\{j(k)\}$ above, being the index of base $j(k)$ in demand node i's list. In this case, since $\mathcal{A}_{in} = \cup_{m=n}^{B+1} \mathcal{S}_{im}$, we have

$$P(\mathcal{A}_{in}|\mathcal{A}_{jk}) = \frac{P(\mathcal{A}_{in} \cap \mathcal{A}_{jk})}{P(\mathcal{A}_{jk})} = \frac{1}{P(\mathcal{A}_{jk})} P\left(\cup_{m=n}^{B+1} \mathcal{S}_{im} \cap \mathcal{A}_{jk}\right).$$

On the event $\mathcal{A}_{jk}$, the bases $j(1), \ldots, j(k-1)$ are busy. Therefore, if base $i(m)$ is contained in this list, then base $i(m)$ cannot answer the call originating at demand node i. In such a case, we have $\mathcal{S}_{im} \cap \mathcal{A}_{jk} = \emptyset$. Consequently, we define the set $M = \left\{m : n \leq m \leq B+1,\ i(m) \notin \{j(1), \ldots, j(k-1)\}\right\}$ of bases that are not necessarily busy on the event $\mathcal{A}_{jk}$. In this case, we have

$$P(\mathcal{A}_{in}|\mathcal{A}_{jk}) = \frac{1}{P(\mathcal{A}_{jk})} P\left(\cup_{m \in M} \mathcal{S}_{im} \cap \mathcal{A}_{jk}\right). \tag{11}$$

Finally, letting $c \wedge d = \min\{c, d\}$, we make the crude approximation that

$$P\left(\cup_{m \in M} \mathcal{S}_{im} \cap \mathcal{A}_{jk}\right) \approx \left(\sum_{m \in M} P(\mathcal{S}_{im})\right) \wedge P(\mathcal{A}_{jk}) = \left(\sum_{m \in M} s_{im}\right) \wedge a_{jk},$$

which amounts to neglecting the "higher order" correction to $P(\mathcal{S}_{im})$ that comes from intersecting with $\mathcal{A}_{jk}$. Taking the minimum with a_{jk} ensures that (11) yields a result in $[0, 1]$.

Combining (6)-(11), we obtain a system of equations for the entries of the matrix A that has the form $A = f(A)$ for an appropriately specified function f. Thus, A is a fixed point of the system, and once we determine A, we can compute performance measures as described earlier.

5 Computational Results for the Descriptive Model

In this section, we present computational results for the descriptive model. Our goal is to verify that the performance measure estimates computed through our descriptive model are accurate enough to reliably assist in identifying high-quality ambulance allocations.

Our input data are taken from Budge et al. [2006], describing the ambulance system operating in Edmonton, Alberta, Canada. The system is medium sized with 16 bases, 12 ambulances and 180

demand nodes. The total demand rate is about 5 calls per hour and the service rate is about 1.34 calls per hour, which corresponds to an average service time of about 45 minutes per call.

A simple computation shows that there are approximately 400,000 different ways of allocating 12 ambulances among 16 bases. In our experimental setup, we select a subset of 1,000 allocations among 400,000 possible ones. For each one of these ambulance allocations, we start with an arbitrary matrix of probabilities A^0 and generate a sequence of matrices $\{A^k\}_k$ by $A^{k+1} = f(A^k)$ to find the fixed point of the function f that is implicitly defined by (6)-(11). We stop when $\|A^{k+1} - A^k\|_1 \leq 0.1$ and use A^{k+1} as the fixed point of f. For our data set, this usually ensures termination within four iterations. Having obtained a fixed point of f, we use (5) to compute the fraction of calls that do not meet the time standard. This is the performance measure estimate obtained by our descriptive model.

For each one of the 1,000 ambulance allocations, we also use a detailed discrete-event simulation model to compute the fraction of calls that do not meet the time standard. The ambulance dispatching policy in this simulation model is to send the closest available ambulance to a call, and if there are no available ambulances, then the call is put in a first-in-first-out queue. Putting a call in a queue almost always implies that the call will not meet the time standard. We simulate the system for two weeks and make 10 runs to estimate the fraction of calls that do not meet the time standard.

The chart on the left side of Figure 4 shows a scatter plot of the estimates of the fractions of lost calls obtained from the descriptive model and the simulation model for the 1,000 ambulance allocations. Our descriptive model appears to predict the fractions of lost calls quite accurately. The coefficient of correlation between the predictions of the descriptive model and the simulation model is about 0.61. For more than 900 ambulance allocations, the estimate of the descriptive model is within 2% of the estimate of the simulation model. The ambulance allocation with the smallest fraction of lost calls as predicted by the descriptive model is marked by a "$\circ$" in Figure 4. The fraction of lost calls for this ambulance allocation is about 12.7%. The fraction of lost calls for the best ambulance allocation obtained by the simulation model is 10.9% and this ambulance allocation is marked with a "$*$."

The primary use we have in mind for the descriptive model is to quickly evaluate a large number of possible ambulance allocations. After this stage, we can evaluate a small set of the most promising allocations through a detailed simulation study. For our data set, we can indeed find the best ambulance allocation among the 1,000 possible allocations by simply evaluating all of these ambulance allocations through the descriptive model, and then, evaluating only the 10-20 most promising ones through the simulation model. For test problems with comparable sizes to our data sets, we can find a fixed point of f within a few seconds. Therefore, we can evaluate 1,000 ambulance allocations within minutes, a much shorter amount of time when compared with the time required to run the simulation model for 1,000 ambulance allocations.

To put the results of the descriptive model in perspective, we compare it with the results obtained by a naive assignment of demands, which is obtained by directing all demand from a demand node to the closest base. In this case, we know the call loads $\{\lambda_b : b \in \mathcal{B}\}$ offered to the different bases and we can estimate the fraction of lost calls by $\sum_{b\in\mathcal{B}} \lambda_b \, \mathcal{E}(n_b, \lambda_b/\mu_b)$. The chart on the right side of Figure 4

Figure 4: Fractions of lost calls estimated by the simulation model, the descriptive model and the naive demand assignment. The figures are percentages of the total rate of calls.

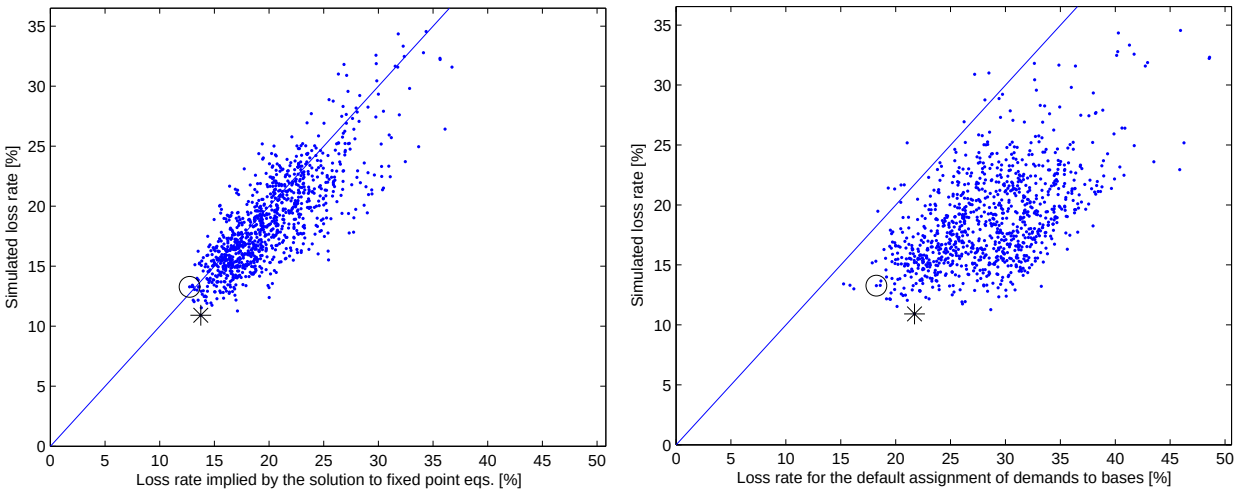
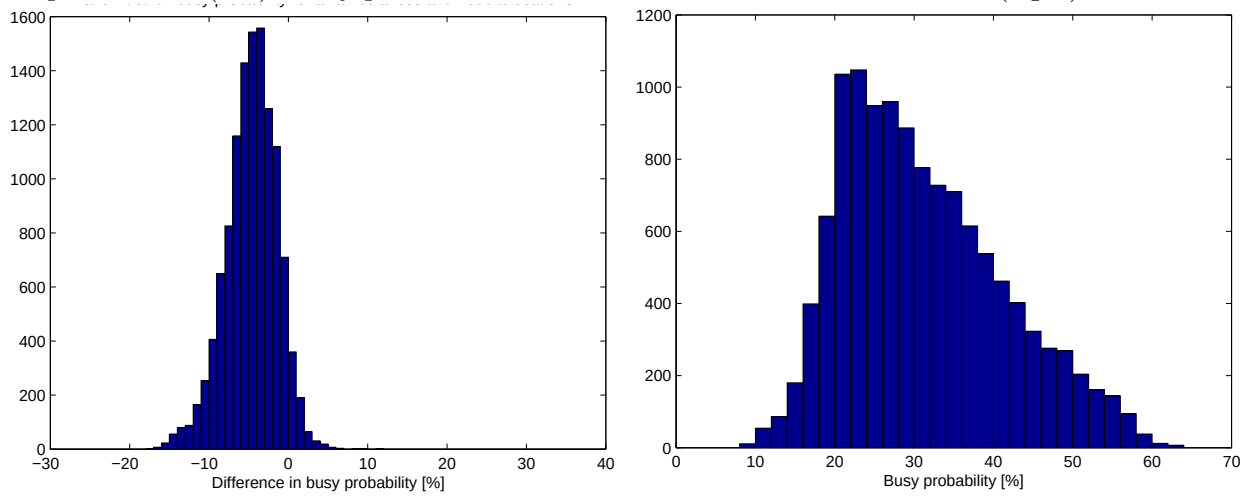


Figure 5: Differences between the busy probabilities obtained from the simulation model and the descriptive model (left). Busy probabilities obtained from the simulation model (right).



shows a scatter plot of the estimates of the fractions of lost calls obtained from the naive assignment of demands and the simulation model for the 1,000 ambulance allocations. The chart indicates that a naive assignment of call loads to the bases does not predict the fraction of lost calls well, and our descriptive model does a good job of assessing how the demand is shared between the bases.

Finally, the chart on the left side of Figure 5 shows a histogram for the absolute difference between the busy probabilities obtained from the descriptive model and the simulation model for the 1,000 ambulance allocations. The differences in busy probability range over $[-10\%, 0]$ for more than 90% of the cases. To give an idea about the magnitude of these differences, the histogram on the right side shows the busy probabilities themselves as predicted by the simulation model.

To close this section, we note that we tried generating the sequence of matrices $\{A^k\}_k$ by starting

with different initial matrices A^0. In every case, we ended up with the same fixed point. We conjecture that the function f implicitly defined by (6)-(11) has a unique fixed point, but we leave this question open for further research. We also note that we obtained similar results by running the same type of experiments with 16 ambulances and a total demand rate of 7 calls per hour. The coefficient of correlation between the predictions of the descriptive model and the simulation model was 0.59 in this case.

6 Conclusions and Future Research

We introduced two models for the static ambulance deployment problem. The models capture some of the essential queuing dynamics of an emergency medical service system while allowing efficient solution procedures. Our first model is particularly suitable when analyzing multi-region systems managed by a centralized planning agent. It illustrates that ambulances should not be allocated in proportion to the loads offered to the bases, especially when ambulance utilizations are low. Our second model estimates the loads offered to the different bases through a system of equations involving the Erlang loss formula. Whether this system of equations has a unique solution is open to further research. Computational experiments indicate that the predictions of our second model are quite accurate, and this model is particularly useful for evaluating a large set of possible ambulance allocations and selecting some allocations to be further evaluated through simulation.

Although we work within the context of emergency medical service logistics, our findings apply to a variety of systems where a set of resources are allocated to different uses under random demand. It would be interesting to investigate whether our models can incorporate the fact that the average service time for an ambulance stationed at a base is affected by the location of the demand assigned to it. Incorporating random travel times and checking whether this feature improves the accuracy of the predictions is another avenue for investigation.

Finally, we have found the insights in the present work extremely useful in tackling the more difficult problem of dynamically deploying ambulances. In this problem, an ambulance can be routed to any base depending on the locations of the other ambulances and the time of the day. Dynamic allocation is a potential strategy to provide better performance with existing resources and is the topic of an upcoming paper.

7 Acknowledgments

We thank Armann Ingolfsson for the data used for the numerical experiments in Section 5. This research was partially supported by National Science Foundation grants DMI 0400287 and DMI 0422133.

References

Batta, R., Dolan, J. and Krishnamurthy, N. [1989], 'The maximal expected covering location problem: Revisited', *Transportation Science* **23**, 277–287.

Berman, O. and Larson, R. [1982], 'The median problem with congestion', *Computers and Operations Research* **9**(22), 119–126.

Brandeau, M. and Larson, R. C. [1986], Extending and applying the hypercube model to deploy ambulances in Boston, *in* A. Swersey and E. Ignall, eds, '*Delivery of Urban Services*', North Holland, New York.

Brotcorne, L., Laporte, G. and Semet, F. [2003], 'Ambulance location and relocation models', *European Journal of Operations Research* **147**(3), 451–463.

Budge, S., Ingolfsson, A. and Erkut, E. [2006], Approximating vehicle dispatch probabilities for emergency service systems.
Available at `http://www.bus.ualberta.ca/aingolfsson/documents/PDF/Disp%20Prob.pdf`.

Church, R. and ReVelle, C. [1974], 'The maximal covering location problem', *Papers of the Regional Science Association* **32**, 101–108.

Daskin, M. [1983], 'A maximal expected covering location model: Formulation, properties, and heuristic solution', *Transportation Science* **17**, 48–70.

Erdogan, G., Erkut, E. and Ingolfsson, A. [2006], Ambulance deployment for maximum survival.
Available at `http://www.bus.ualberta.ca/aingolfsson/documents/PDF/Survival.pdf`.

Fox, B. [1966], 'Discrete optimization via marginal analysis', *Management Science* **13**(3), 210–216.

Goldberg, J. B. [2004], 'Operations research models for the deployment of emergency services vehicles', *Emergency Medical Services Management Journal* **1**(1), 20–39.

Goldberg, J. B. and Paz, L. [1991], 'Locating emergency vehicle bases when service time depends on call location', *Transportation Science* **25**(4), 264–280.

Goldberg, J. and Szidarovsky, F. [1991], 'Methods for solving nonlinear equations used in evaluating emergency vehicle busy probabilities', *Operations Research* **39**(6), 903–916.

Green, L. V. and Kolesar, P. J. [2004], 'Improving emergency responsiveness with management science', *Management Science* **50**(8), 1001–1014.

Harel, A. [1988], 'Convexity properties of the Erlang loss formula.', *Operations Research* **38**(3), 499–505.

Henderson, S. G. and Mason, A. J. [2004], Ambulance service planning: Simulation and data visualisation, *in* M. L. Brandeau, F. Sainfort and W. P. Pierskalla, eds, '*Handbooks in Operations Research and Management Science*, Volume on *Operations Research and Health Care: A Handbook of Methods and Applications*', Kluwer Academic, Boston.

Jagers, A. A. and van Doorn, E. A. [1986], 'On the continued erlang loss function', *Operations Research Letters* **5**(1), 43–46.

Jarvis, J. [1975], Optimization in stochastic systems with distinguishable servers, Technical Report TR-19-75, Operations Research Center of MIT.

Jarvis, J. [1985], 'Approximating the equilibrium behavior of multi-server loss systems', *Management Science* **31**, 235–239.

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

Marianov, V. and ReVelle, C. [1994], 'The queuing probabilistic location set covering problem and some extensions', *Socio-Economic Planning Sciences* , 167–178.

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

Swersey, A. J. [1994], The deployment of police, fire and emergency medical units, *in* S. M. Pollock, M. Rothkopf and A. Barnett, eds, '*Handbooks in Operations Research and Management Science*, Volume on *Operations Research and the Public Sector*', North Holland, New York.

Toregas, G., Saqin, R., ReVelle, C. and Berman, L. [1971], 'The location of emergency service facilities', *Operations Research* **19**(6), 1363–1373.

Wolff, R. W. [1989], *Stochastic Modeling and the Theory of Queues*, Prentice-Hall, Englewood Cliffs, New Jersey.