Ambulance Location for Maximum Survival

November 2007

Erhan Erkut[1]

Armann Ingolfsson[2]

Güneş Erdoğan[3]

[1]Özyeðin University
Istanbul, Turkey

[2]School of Business
University of Alberta
Edmonton, Alberta, Canada

[3]Department of Industrial Engineering
Bilkent University
Ankara, Turkey

Acknowledgments: This research was supported in part by NSERC. We thank Dr. Philip Yoon, Associate Professor of Emergency Medicine at the University of Alberta, for insightful comments on an earlier version of this paper.

Abstract

This paper proposes new location models for emergency medical service stations. The models are generated by incorporating a survival function into existing covering models. A survival function is a monotonically decreasing function of the response time of an EMS vehicle to a patient that returns the probability of survival for the patient. The survival function allows for the calculation of tangible outcome measures—the expected number of survivors in case of cardiac arrests. The survival-maximizing location models are better suited for EMS location than the covering models which do not adequately differentiate between consequences of different response times. We demonstrate empirically the superiority of the survival-maximizing models using data from the Edmonton EMS system.

Keywords: ambulance location, covering models, survival function.

1. Introduction

The goal of this paper is to question a widely-used modeling construct ("coverage") and a related performance measure (fraction of calls reached within some time standard) for emergency medical service (EMS) systems, and offer a superior alternative which takes advantage of medical research on the relationship between response times and survival rates. We show how nonlinear survival functions can be incorporated into EMS location models and we offer computational evidence based on realistic data to support our claim of superiority.

The units of measurement matter for EMS performance measures, just like they matter for other organizations. Metrics in concrete, easily interpreted units, such as dollars or lives saved, get more attention and facilitate comparisons between competing uses of funds. Ideally, EMS planning would be driven by input-output relations linking resource allocation to patient outcomes, as argued by Willemain (1975). It also matters whether standards are set locally or nationally. As an example, for fire services, standards that are followed in the US and Canada are set by the National Fire Prevention Association. The insurance industry ranks fire departments based on adherence to such standards, and when they are not met, insurance rates may rise (Pedersen, 2002). Thus, there is a direct link between failure to meet standards and (monetary) outcomes. This is not the case for EMS coverage standards, which vary even between communities in close proximity to each other (for example, see Moeller, 2004). Davis (2003a) argues that emergency services in most U.S. cities "don't know how many lives they're losing, so they can't determine ways to increase survival rates." We will take as given that the real objective for an EMS system is to maximize the number of patients that survive and that coverage is used as a proxy for the real objective. There are other measures that matter, such as life expectancy and quality of life for survivors. However, it is unclear to what extent faster EMS response times influence such secondary measures. Yet there is clear evidence that faster response times can save lives of cardiac arrest patients.

The problem of selecting the locations of emergency medical service (EMS) vehicles has been quite popular among operations researchers. Such models typically focus on either *coverage* or *average response time*—two performance measures that were discussed in an early survey paper by Chaiken and Larson (1972). In one of the earliest papers on this topic, Toregas et al. (1971) developed a coverage model, which minimizes the number of facilities needed to serve a set of given demand nodes. The coverage concept utilizes a travel distance (or time) standard for service delivery. All demand points that are within this threshold distance to a service facility are considered to be served by the facility, i.e., covered. Hence, for a given set of facility locations and demand points, the covering model classifies the demand points into two sets: those that are covered and those that are not. The set cover model of Toregas et al. (1971) minimizes the number of facilities so that all demand points are covered, and the max cover model of Church and Revelle (1974) maximizes the demand covered with a given number of facilities.

Coverage models have been used frequently by researchers and practitioners for the following reasons:

- The concept is simple to communicate to decision-makers and the public (a call is either covered or not).
- Many EMS systems use the percentage of calls covered as a performance measure. Perhaps the most common EMS standard is to respond to 90% of all urgent calls within 8 minutes (De Maio et al., 2003).

- Deterministic coverage models typically result in integer programs that are easy to solve using standard optimization software.

Despite these advantages, the black-and-white nature of the coverage concept is an important limitation, and standard coverage models should not be used for EMS vehicle location. First, coverage can result in large *measurement errors* because of their limited ability to discriminate between different response times. Second, these measurement errors are likely to result in large *optimality errors* when one uses covering models to locate emergency facilities instead of a model that takes survival probabilities into account. The following example demonstrates that the optimality error can be arbitrarily large.

Example: Assume that demand locations A and B in Figure 1 are 18 minutes apart, and a station is located at X, halfway between them. A covering model with a covering radius of 9 minutes would count all demand at A and B as covered, so X is the optimal location, regardless of the relative magnitude of the demands. Suppose the demand at A is 10, the demand at B is 1, and the survival probability as a function of the response time t is $\exp(-t)$. Hence, if the emergency facility is located at X, then Pr{survival at A} = Pr{survival at B} = $\exp(-9) = 0.000123$, and the expected number of survivors in the system is $11 \times 0.000123 = 0.001358$. If the station is located at A instead, then the expected number of survivors increases to 10, which is over 7,000 times better. This ratio can be made arbitrarily large by increasing the demand at A. As this (admittedly pathological) example demonstrates, covering models can result in arbitrarily poor location decisions for emergency facilities.

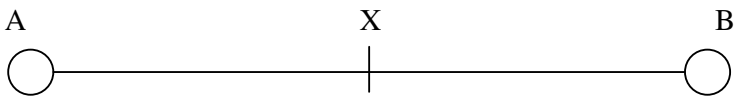

Figure 1: An example depicting the difference of "coverage" and "survival."

Although this example is artificial, it illustrates the real and important issue of whether EMS response time standards should be the same regardless of population density. Economies of scale make it less expensive to provide a given level of service in urban areas than in rural areas. Similarly, service can be provided more efficiently in the densely populated center of a city than in the more spread out suburban areas. The objectives of providing equal access to EMS versus maximizing the total number of survivors lead to different deployment patterns (Felder and Brinkmann, 2002). Response time standards and actual performance are typically different for urban and rural areas in the US, UK, and Germany (see Fitch, 2005, and Felder and Brinkmann, 2002), indicating that the standard setters have decided against equal access. According to these references, the most common standards in North America are to reach 90% of calls in 9, 15, and 30 minutes for urban, rural, and wilderness areas, respectively. In the UK, a national standard calls for reaching 75% of calls in 8 minutes, regardless of location. Furthermore, 95% of urban calls should be reached in 14 minutes in urban areas and 19 minutes in rural areas. German standards vary across the country, requiring that 95% of calls be reached within statutory response times that range from 10 to 15 minutes. These issues raise important ethical concerns, some of which are addressed by Felder and Brinkmann (2002). As they point out, although a policy of equal access seems difficult to criticize, such a policy implies that lives are valued differently in different areas, because the cost of saving a life can be much higher in sparsely populated rural areas than in urban centers.

To explicitly model the utility of response time for a patient, we need a function that maps the response time of an EMS vehicle to a patient, to the probability that the patient survives. However, there are many different types of EMS calls and the survival probabilities depend on the emergency. We focus on one type of emergency: out-of-hospital cardiac arrest. There are several reasons for our choice:

1) Cardiac arrest calls are of the highest priority, and the response time is crucial. Davis (2003b) refers to such calls as the most "saveable" and "the truest measure of emergency medical performance."
2) Current response time standards were derived from cardiac arrest survival studies (Fitch, 2005).
3) Medical researchers have studied the relation between survival probability and response time extensively (see next section).
4) These calls account for a sizeable portion of high priority EMS calls.

We have access to detailed call data for the Calgary, Alberta, EMS system. In 2004, Calgary EMS responded to 14,152 Priority 1 (or "Delta") calls. The leading Priority 1 call category was "chest pain – Delta" with 1,865 calls. The combined categories of "chest pain – Delta," "cardiac arrest – Delta," and "heart problems – Delta" added up to 2,463 (17.4% of all Priority 1 calls). Furthermore, the top ten Priority 1 categories not related to heart problems were the following: breathing problems, unconscious, traffic accident, building fire, house fire, fall, convulsions and seizures, hemorrhage and lacerations, traumatic injuries, and unknown problem (man down). While these categories are likely to have survival functions different from cardiac arrest, it can be safely argued that for each of these categories response time is important and the probability of recovery decreases gradually with time. These ten categories, combined with chest pain, cardiac arrest, and heart problems, add up to 11,187 calls, making up almost 80% of all Priority 1 calls.

In addition, there were a significant number of Priority 2 calls in critical categories, including 3,961 "chest pain – Charlie" calls, 3,570 calls with breathing problems, 2,278 calls in the unconscious category, and 1,395 stroke calls. While these were classified as Priority 2 calls, some may have been Priority 1 calls that were misclassified as Priority 2 calls, and some Priority 2 conditions may have deteriorated to Priority 1 during the response period. The number of calls in these four Priority 2 categories alone, where response time is critical, is 11,204. Hence, for a considerable number of EMS calls, faster is better and OR models for EMS system design should take this into account.

The rest of the paper is organized as follows: § 2 reviews relevant literature on ambulance location models and cardiac arrest survival probabilities, § 3 discusses how to model the probability of survival, § 4 introduces the maximal survival model and compares it to two models from the literature numerically, § 5 builds on the maximal survival model from § 4 and introduces coverage models with increasing realism and shows how survival functions can be incorporated into these more refined models, § 6 provides computational results, and § 7 offers concluding remarks.

2. Literature Review

The literature on ambulance location is quite rich. It has been reviewed thoroughly by Swersey (1994), Marianov and Revelle (1995), and more recently by Brotcorne et al. (2003) and Jia et al (2007). We limit our discussion of this literature to papers that are most relevant

to this paper, namely those that use generalizations of covering models where numbers other than 0 or 1 are used to quantify the quality of coverage, as well as approaches to minimize average response time. In this section we also discuss the literature on survival functions for EMS, which is central to our paper.

2.1 Generalizations of coverage models

Brotcorne et al. (2003) provide a recent survey of the EMS location literature, and identify 18 different models for ambulance location. All of these models use the concept of coverage. Deviations from a 0 or 1 for coverage occur for two reasons:

1) Incorporating the probability that a station may have no EMS vehicles to respond to a call: If the probability of having an idle EMS vehicle at a given station is p, then the expected coverage for a demand point with a demand of one unit within the coverage radius is not 1 but p (e.g., Daskin, 1983, Saydam and McKnew, 1985, Revelle and Hogan, 1989).
2) Incorporating response time uncertainty: If the probability of responding from a station to a demand point within the given time limit is q, then the expected coverage for a demand point within the coverage radius is q (Daskin, 1987).

In a model that incorporates both EMS vehicle availability and response time uncertainty, the expected coverage for a unit demand would be pq, assuming the two sources of uncertainty are independent.

While such models are more realistic than the basic coverage model, the deviation from zero or one in coverage is due to factors other than the time-dependent utility of the response time. All such models still use the covering concept with a fixed (and arbitrary) coverage radius. The central assumption is still the same: if the vehicle reaches the demand within a specified time limit then the call is covered (the patient is saved) and if not it is not covered (the patient is lost). Hence, while these generalizations of the covering model have different levels of sophistication in the way different uncertainties are incorporated, all suffer from the same shortcoming in the modeling of the patient's utility as a function of the response time. We discuss these models in more detail later in the paper.

We are aware of only three papers that take a critical view of the 0-1 coverage concept and attempt to generalize it. Church and Roberts (1983) suggest a piecewise linear step function to incorporate quality of service in a covering model. They show that the piecewise linear utility function may result in solutions that are significantly different from those obtained using the standard max cover model. While this is a step in the right direction, it is ad-hoc and rather limited. More recently, Karasakal and Karasakal (2004) and Berman, Krass, and Drezner (2003) independently introduced coverage models where coverage decays gradually with distance. Karasakal and Karasakal (2004) focus on algorithmic issues, and they design a Lagrangian heuristic to solve the problem. In their computational experiments, they assume that coverage changes from 1 (full coverage) to 0 (no coverage) in a narrower interval than would be appropriate for the context we focus on, where survival probability might decay gradually from around 30% to 5% when response time varies from 0 to 10 minutes. Berman, Krass, and Drezner (2003) present a structural result that allows one to limit candidate locations in a network to a finite set without loss of generality. Then, they show how the problem can be formulated as an uncapacitated facility location problem and they also provide an alternative and more efficient formulation. Neither of these papers discusses how one would quantify a coverage function, which is an issue that we emphasize. We also emphasize estimation of the benefit of using a finer graduation than 0-1 coverage. While Karasakal and Karasakal (2004) report on *spatial differences* between optimal solutions to

their model and corresponding instance of the standard max cover model, they do not address differences in the objective function value—the measure of quality of service.

2.2 Average response time minimization

In a deterministic setting, the *p*-median model can be used to locate ambulance stations to minimize average distance from demand nodes to the closest ambulance station. Larson's (1974, 1975) exact and approximate hypercube models can be used to estimate average response time, taking ambulance unavailability into account. Jarvis (1975) developed a locate-allocate heuristic that assigns ambulances to stations to minimize average response time, as evaluated by the approximate hypercube model (see also Larson, 1979).

2.3 Survival functions

Almost all of the published research we have found relating survival rates to EMS response times focuses on cardiac arrest. One notable exception is Cretin and Willemain (1979), who focus on survival rates after myocardial infarction (heart attack).

Eisenberg et al. (1990) reviewed published reports from 29 cities on survival rates after out-of-hospital cardiac arrest. They identified many factors besides response times that may influence survival rates, including system design (how EMS staff are trained; which procedures they perform), the consistency with which procedures are applied, physiological and demographic differences between regions, and inconsistencies in definitions used for terms such as "cardiac arrest" and "response time."

They present hypothetical survival curves from the time of collapse for five different EMS system types: EMS vehicles staffed by emergency medical technicians only (EMT), EMT with defibrillation capability (EMT-D), paramedic, EMT followed by paramedic, and EMT-D followed by paramedic. The hypothetical survival curves assume that without intervention, the survival rate begins at 100% at the time of collapse and decays linearly to zero after 10 minutes. When EMTs arrive and administer cardiopulmonary resuscitation (CPR), the slope of the survival curve is assumed to decrease, but remain negative. If EMTs administer defibrillation as well, then the slope is assumed to decrease further. The survival curve is assumed to stabilize (have a slope of zero) either when paramedics arrive and provide medication and intubation, or, for EMS systems with no paramedics, when the patient arrives in a hospital.

The authors suggest benchmark survival rates after stabilization ranging from 10% for EMT systems to 35% for EMT-D / paramedic systems. The benchmark values are close to values that have been achieved in King County, WA, where the EMS system has evolved from EMT, to EMT-D, to EMT/Paramedic, to EMT-D/Paramedic over time.

Perhaps the most convincing evidence that short response times improve survival rates of cardiac arrest patients comes from a study conducted in casinos (Valenzuela et al, 2000), where security officers were trained to administer CPR and defibrillation. The exact time of collapse was determined from security videos and times from collapse to CPR were typically under three minutes. This separates this study from most others, where the time of collapse is either subjectively estimated by bystanders or ignored and times from collapse to CPR are considerably longer. In this study, patients who received defibrillation within 3 minutes after collapse had a 74% survival rate, while those who received defibrillation later had a 49% survival rate.

Of the many studies that Eisenberg et al. (1990) surveyed, some did not report response times at all, while others reported only averages or percentiles, and a few reported response time distributions and estimated how survival depended on response time. We now discuss four relevant studies that estimated such survival functions.

The first study was conducted by Larsen et al. (1993). The authors used data from the cardiac arrest surveillance system of King County (Washington, US). Using multiple linear regression, they estimated the following equation for survival probability:

$$s(I_{\text{CPR}}, I_{\text{Defib}}, I_{\text{ACLS}}) = 0.67 - 0.023 I_{\text{CPR}} - 0.011 I_{\text{Defib}} - 0.021 I_{\text{ACLS}} \quad (1)$$

where

I_{CPR} = the duration from collapse to CPR,
I_{Defib} = the duration from collapse to defibrillation,
I_{ACLS} = the duration from collapse to Advanced Cardiac Life Support (ACLS),

and all three durations are measured in minutes. The authors reported that the interactions between the variables were insignificant, and the additive model was accurate. A comparison between the predicted and the observed survival rates revealed that the largest difference was observed when response time was very large. Where the model predicted a survival rate of 0%, observed survival rates ranged from 3% to 20%, depending on specific attributes of the system, e.g., whether the ambulances have defibrillators and are staffed by paramedics.

The second study is by Valenzuela et al. (1997), who used data from Tucson (Arizona, US) and King County (Washington, US) and logistic regression to construct a survival function. The function included many factors: age, manual CPR applied by bystanders, time interval from collapse to CPR, time interval from collapse to defibrillation, and manual CPR initiated by bystanders / collapse to CPR interval interaction. Notably, the authors found that the site (Tucson or King County) did not have a significant effect on survival after controlling for the aforementioned variables, i.e., the same survival function could be used for both urban areas. The authors then gave a second survival function which included only the time interval from collapse to CPR and the time interval from collapse to defibrillation. This second function, which quite accurately approximates their first function, is:

$$s(I_{\text{CPR}}, I_{\text{Defib}}) = \left(1 + e^{-0.260 + 0.106 I_{\text{CPR}} + 0.139 I_{\text{Defib}}}\right)^{-1} \quad (2)$$

In contrast with the previous study, the authors reported that the survival function overestimated the probability of survival when the response time was large.

The third study is due to Waaelwijn et al. (2001). This study used data from Amsterdam, Netherlands, and the surrounding region. Using logistic regression, three different survival functions were estimated, from the perspectives of the bystander, the first responder, and the paramedic. Many details were included in the last two functions such as the initially diagnosed heart rhythm and the necessity of advanced CPR. The first function had three variables: a binary variable to denote whether the collapse was witnessed by EMS staff or not, the length of the time interval from collapse to basic CPR, and the length of the time interval from basic CPR to the arrival of the EMS vehicle. Their first function is:

$$s(X_{\text{EMS}}, I_{\text{CPR}}, I_{\text{Response}}) = \left(1 + e^{0.04 + 0.7 X_{\text{EMS}} + 0.3 I_{\text{CPR}} + 0.14 (I_{\text{Response}} - I_{\text{CPR}})}\right)^{-1} \quad (3)$$

where X_{EMS} is 1 if the cardiac arrest was witnessed by EMS staff and 0 otherwise, and I_{Response} denotes the length of the response time in minutes.

The fourth study was conducted by De Maio et al. (2003), using data from several municipalities in Ontario, Canada. The authors used stepwise logistic regression to estimate

survival probability. The variables that remained in the final model were EMS response time, age, whether the collapse was witnessed, whether a bystander administered CPR, and whether fire or police administered CPR. They then used an ad-hoc procedure to average over the effects of all of the explanatory variables except response time, resulting in a function that predicts survival probability based solely on EMS response time, for people in the population they studied:

$$s(I_{\text{Response}}) = \left(1 + e^{0.679+0.262 I_{\text{Response}}}\right)^{-1} \tag{4}$$

In using the four estimated survival functions, it is important to consider how "survival" was defined, which cases were included, and the type of EMS system in the study region. All four studies defined "survival" to mean "survival until discharge from hospital." Larsen et al. (1993) and Valenzuela et al. (1997) limited their study to patients with ventricular fibrillation (a type of heart rhythm that is classified as "shockable") whereas Waaelwijn et al. (2001) and De Maio et al. (2003) included all cases that were treated for cardiac arrest by EMS personnel, regardless of whether the initial rhythm was shockable. The former two studies were done in regions with a two-tier EMS system, where first responders had EMT training and second responders were paramedics. The latter two studies were for single-tier EMS systems staffed by EMTs (De Maio et al., 2003) or personnel trained according to European standards (Waalewijn, et al., 2001).

The medical literature we have reviewed assumes, implicitly or explicitly, that EMS systems are driven by a coverage standard, such as a target to reach 90% of the highest priority calls in 8 minutes. One study (Blackwell and Kaufman, 2000) reaches the pessimistic conclusion that "there is little evidence … to suggest that changing … response time specifications to times less than current, but greater than 5 minutes, would have any beneficial effect on survival." Our contention is that the performance measure (coverage) should be questioned, and that if EMS systems are designed to directly maximize the expected number of survivors rather than using coverage as a proxy, then improvement is possible.

In the next section we compare the four survival functions introduced in this section, and discuss how one might deal with variables other than response time that appear as explanatory variables in the survival functions.

3. Modeling Probability of Survival

3.1 Modeling the Response Time

The survival functions that we surveyed differ from each other in many aspects. Some functions include additional explanatory variables besides response time. For example, (1) includes the length of the time intervals from collapse to CPR, defibrillation, and intensive care, and (3) requires knowledge of whether EMS staff witnessed the collapse or not. It is beneficial to include such additional explanatory variables because it allows local calibration of the functions, taking into account that the community where one wishes to use the location model may differ systematically from the community where the data that were used to estimate the survival function was collected. Such calibration involves, on the one hand, ensuring that the estimated survival function is based on data from an EMS system that is comparable to the one where one wishes to use the location model. For example, if the EMS system uses only emergency medical technicians, then one should not use a survival function that is based on an EMS system that uses paramedics. On the other hand, one should

"average over" behavioral explanatory variables such as whether a bystander administers CPR, leaving only the response time.

For illustration purposes, consider the following deterministic (and in some cases unrealistically optimistic) assumptions that one could use to eliminate all variables except the response time:

- The collapse of the patient is not witnessed by an EMS unit, i.e., $X_{EMS} = 0$ in (3).
- A call is placed to EMS as soon as the patient experiences cardiac arrest.
- CPR is performed by the responding EMS unit immediately upon arrival. CPR is not performed by a bystander. Together with the preceding assumptions, this implies that $I_{Response} = I_{CPR}$ in (1) – (3).
- All EMS units are equipped with defibrillators and staff who are trained to use them. Defibrillation is performed one minute after arrival, which implies $I_{Defib} = I_{Response} + 1$ in (1) and (2).
- ACLS is performed at the hospital which takes an average of 16 minutes to reach after the first response (i.e., $I_{ACLS} = I_{Response} + 16$ in (1)).

Considering we used assumptions in favor of the patient (EMS contacted immediately, immediate CPR upon arrival, defibrillation within one minute of arrival), Figure 2 clarifies two sobering messages about the consequences of a cardiac arrest where immediate response is not available and a call to EMS must be made.

1) All survival functions start well below 100%. This means a cardiac arrest is quite likely to result in death even if the response is almost instantaneous.
2) All functions show survival probabilities below 10% at 10-minute response times.

Figure 2 also makes it rather clear that the standard maximal covering model with a response radius of, say, 9 minutes is not likely to maximize the number of cardiac arrest survivors. The survival probability is about five times higher when responding immediately than when responding in 9 minutes, but a covering model does not differentiate between these two response times. Furthermore, response times of 9 and 10 minutes result in almost the same survival probability, while a covering model attaches a major difference to these two response times. Finally, the survival probability is nonzero for response times over 9 minutes while the covering model would place no value on responses over 9 minutes.

3.2 Modeling Explanatory Variables Other than Response Time

We now turn to incorporating the impact of explanatory variables besides response time. Let $\bar{s}(d)$ be the probability of survival as a function of distance d, for a patient at a particular location, assuming the responding EMS vehicle comes from a particular station. We fix the location and the responding station to simplify the notation in this section. The objective functions of the location models we present aggregate over demand locations and stations. The distance will determine the distribution for the response time R. The probability of survival will also depend on a vector of other explanatory variables, $\mathbf{O}$. Medical studies attempt to quantify the probability of survival as a function $s(R(d), \mathbf{O})$ of response time (which we show here as a function of distance) and other explanatory variables. To obtain appropriate input for location models, we need to "average over" both the response time and the other explanatory variables, i.e., $\bar{s}(d) = \mathrm{E}[s(R(d), \mathbf{O})]$.

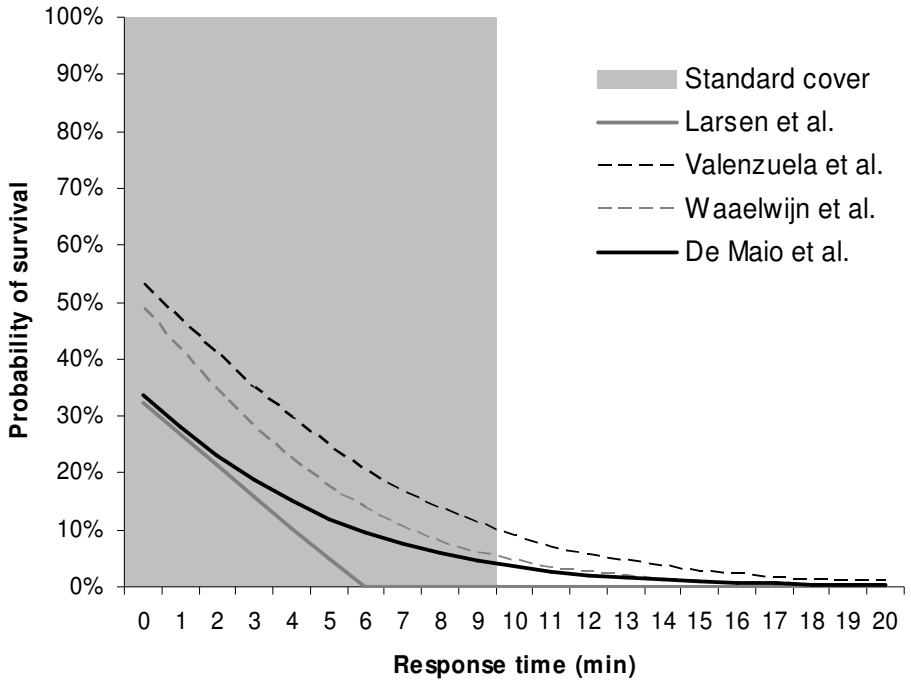

Figure 2: A comparison of the four survival functions discussed, plotted on the background of the step function of the standard cover with a 9-minute threshold.

The effort needed to quantify the variation in these other explanatory variables depends on the variable. Some are known system design features, e.g., whether ambulances are staffed by paramedics or emergency medical technicians. Others are measurable but typically not tracked, e.g., whether cardiac arrest is witnessed by a bystander and whether the patient receives CPR from a bystander. In the U.S., efforts by the Centers for Disease Control and Prevention and the American Heart Association are underway to facilitate the routine collection of such data in a cardiac arrest registry (Anonymous, 2006, pg. 66) and a recent survey (Williams, 2007) indicates that 25.4% of big-city EMS systems in the U.S. track rates of bystander CPR. Finally, some variables are difficult to measure and rough estimates must be used, e.g., for the time from when a patient collapses due to cardiac arrest until a phone call is placed to EMS. It is important to assess the sensitivity of the model to estimates for variables in this last category.

Assuming that one has information about the probability distribution for $R(d)$ and for $\mathbf{O}$, one could attempt to compute $\overline{s}(d)$ using exact or approximate closed-form relationships, numerical integration, or Monte Carlo simulation. We now elaborate on the Monte Carlo approach, because it is general and easy to implement. First, choose a set of representative distances (d_1, d_2, …, d_m). For each distance, simulate n EMS calls, and let R_{ij}, and $\mathbf{O}_{ij}$ be the values for the response time and other explanatory variables for the i-th call and the j-th distance. Then, one can use the function $s(R(d), \mathbf{O})$ to compute the probability of survival s_{ij} for the i-th call and j-th distance. The sample average $\sum_{i=1}^{n} s_{ij} / n$ provides an estimate for $\overline{s}(d_j)$. The estimates $(\overline{s}(d_1), \overline{s}(d_2), \ldots, \overline{s}(d_m))$ could then be used to approximate the survival function, possibly by fitting some parameterized function to them.

To illustrate the procedure, we will make the following assumptions, which are roughly consistent with data reported by Eisenberg et al (1979) for King County, WA. We focus on cases where cardiac arrest occurs before an EMS vehicle is called. The first time interval of interest is *access time* T_1, the time from when the patient collapses from cardiac arrest until a phone call is placed to EMS. Consistent with Eisenberg et al (1979), we assume that 61% of cardiac arrests are witnessed or heard by a bystander, and in those cases access time is exponentially distributed with a mean of 1.2 minutes. In the 39% of cardiac arrests that are not witnessed, we assume that access time is exponentially distributed with a mean of 30 minutes. This is obviously a rough estimate; we discuss the sensitivity of the estimated survival curve to it later. Second, we consider the time T_2 from the moment EMS is contacted until the patient receives CPR. CPR could be administered by a bystander or by EMS staff when they arrive. We assume that 64% of bystanders will perform CPR on the patient, and that the time until they do so, after contacting EMS, is an exponentially distributed random variable B with a mean of 1 minute. Thus, with 36% probability, T_2 will equal the EMS response time R and with 64% probability it will equal min(R, B). Third, we consider the time from beginning of CPR until the first EMS unit arrives, T_3. Consistent with our previous assumptions, T_3 will equal zero with 36% probability. Finally, let T_4 be the interval from arrival of an EMS unit until defibrillation, which we'll assume to follow an exponential distribution with a mean of 2 minutes.

To simulate the response time R, we assume that it consists of pre-travel delay that is independent of distance, and travel time, which depends on the shortest path distance d. We assume the pre-travel delay is lognormally distributed with a mean of 3 minutes and a standard deviation of 1.5 minutes (consistent with data from the City of St. Albert, as reported in Budge et al., 2007b). We also assume the travel time (in seconds) to be lognormally distributed, with a median and multiplicative standard deviation given as follows (based on Budge, 2004).

$$m(d) = \begin{cases} 5.42\sqrt{d} \text{ for } d \le 4400 \text{ m} \\ 180 + 0.041d \text{ for } d > 4400 \text{ m} \end{cases} \quad (5)$$

$$\sigma^*(d) = \begin{cases} \left(0.277 d^{0.123}\right)^{-1.483} \text{ for } d \le 4400 \text{ m} \\ \left(1.5 d^{0.123} / (180 + 0.041d)\right)^{-1.483} \text{ for } d > 4400 \text{ m} \end{cases} \quad (6)$$

For more on the modeling of travel times, see Kolesar et al. (1975), Carson and Batta (1990), and Campbell (1992).

We used these assumptions, together with the survival function (2) from Valenzuela et al. (2000) to estimate the survival probability as a function of distance, in increments of 500 m. This survival function has $I_{CPR} = T_1 + T_2$ and $I_{Defib} = T_1 + T_2 + T_3 + T_4$ as explanatory variables. The results are shown in Figure 3.

To illustrate how one could assess the impact of parameter estimates for the explanatory variables, consider perhaps the least reliable estimate, that of the mean access time for cardiac arrests that are not witnessed. When we decreased this estimate from 30 to 5 minutes, the survival curve shifted up by anywhere from 3.6 percentage points (at a distance of zero) to 0.2 percentage points (at a distance of 45 km). When we increased the estimate from 30 to 60 minutes, the survival curve shifted down by 0.7 to 0.02 percentage points. A more extensive sensitivity analysis would determine how much this variation in the survival curve impacts

the optimal solution to the models that will follow, and the resulting estimate of the total expected number of survivors.

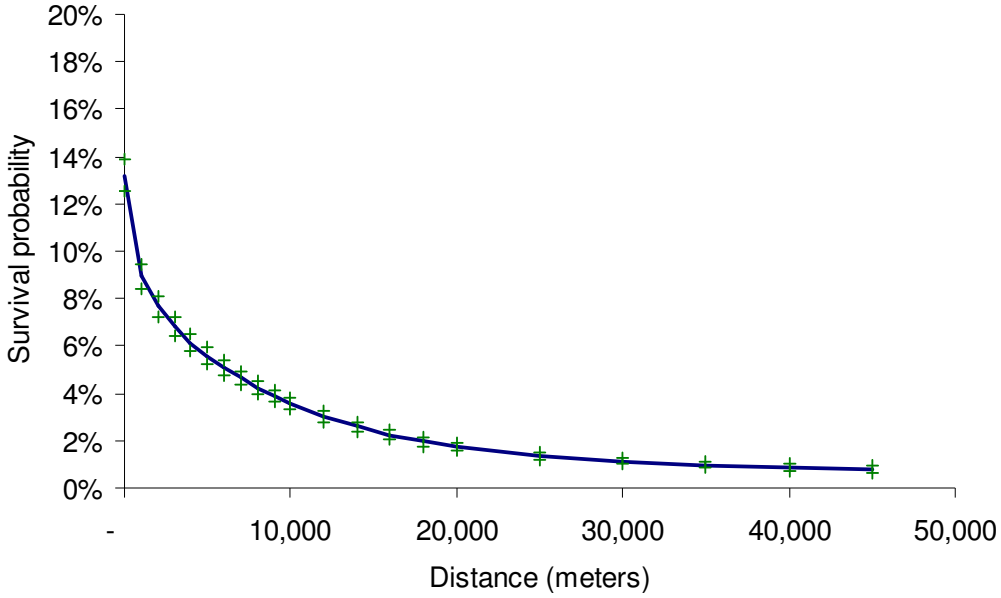

Figure 3: Estimated survival probability as a function of distance after averaging over the explanatory variables in the Valenzuela et al. (2000) survival function. The "+" signs show 95% confidence intervals around the estimated average survival probabilities.

This brief discussion demonstrates that the response time is the most important component of the survival functions, and the other parameters are unlikely to impact significantly the results of ambulance location studies. However, as we described above, it is possible to conduct a parametric analysis to fully assess the impact of the function parameters on the results. In the next section we introduce the maximal survival location problem and use a survival function similar to the one in Figure 3 to illustrate the benefits of incorporating survival functions into a standard max cover model.

4. The Maximal Survival Location Problem

4.1 Formulation

The maximal covering location problem (MCLP) and the q-median problem are the most basic models that one could use to locate ambulances (see the appendices for formulations). MCLP aims to maximize total covered demand with q facilities and the q-median problem aims to minimize average distance to the closest of q facilities.

We now formulate the maximal survival location problem (MSLP), where the objective is to maximize the expected number of patients who survive. Let p_{ij} denote the probability that a patient at demand node i survives and is served by an EMS vehicle from station j. We assume that every demand node is served by the closest station. Then the objective function is:

$$\max \sum_{i=1}^{m} d_i \sum_{j=1}^{n} p_{ij} \tag{7}$$

where m is the number of demand nodes, n is the number of candidate locations, and d_i is the demand at node i. In MSLP we need to keep track of which station serves which demand point, so we define decision variables y_{ij} to equal one if demand node i is served by an EMS vehicle at location j, and zero otherwise. Then

$$p_{ij} = \begin{cases} s(t_{ji} + t_d) & \text{if } y_{ij} = 1 \\ 0 & \text{if } y_{ij} = 0 \end{cases} = s(t_{ji} + t_d) y_{ij} \tag{8}$$

where t_{ji} is the travel time from candidate location j to demand node i and t_d is the pre-travel delay. We assume that the travel time and pre-travel delay are deterministic, but we relax this assumption later. Letting q be the number of facilities, and x_j be equal to one is candidate location j is selected (and zero otherwise), the formulation for the maximal survival problem (MSLP) follows:

MSLP:

$$\max \quad \sum_{i=1}^{m} d_i \sum_{j=1}^{n} p_{ij} = \sum_{i=1}^{m} d_i \sum_{j=1}^{n} s(t_{ji} + t_d) y_{ij} \tag{9}$$

$$\text{s.t.} \quad \sum_{i=1}^{m} y_{ij} \le m x_j \,, j = 1, \ldots, n, \tag{10}$$

$$\sum_{j=1}^{n} y_{ij} = 1, i = 1, \ldots, m, \tag{11}$$

$$\sum_{j=1}^{n} x_j \le q, \tag{12}$$

$$x_j \in \{0,1\}, j = 1, \ldots, n, \tag{13}$$

$$y_{ij} \in \{0,1\}, i = 1, \ldots, m, j = 1, \ldots, n. \tag{14}$$

4.2 An Empirical Comparison of MSLP to two Models from the Literature

We now compare the MCLP, the q-median, and the MSLP empirically. These three simple and deterministic models allow us to focus on the impact of replacing zero-one coverage or average response time with the survival probability. For this comparison we use data from Edmonton, Canada, with 180 demand points and 16 candidate locations for EMS stations, and only the demand for Priority 1 calls. Using CPLEX 8.11, we solved the MCLP, the q-median problem, and the MSLP on a Dell PowerEdge workstation with 1.13 Ghz CPU clock and 1 GB of RAM, and experienced run times of at most 1 CPU second. We use a survival function based on the same assumptions as in the previous secton (see Figure 3), except that uncertainty in response times was ignored, i.e., we used $\mathrm{E}[s(\mathrm{E}[R(d)], \mathbf{O})]$ instead of $\mathrm{E}[s(R(d), \mathbf{O})]$.

We solved the models to optimality for q (number of stations) ranging from 1 to 16. Figure 4 shows the expected number of survivors (evaluated using the approximate hypercube model) for the optimal solution of each model. The results demonstrate that using the optimal solution of the MCLP or the q-median can lead a decision-maker to select locations that are far worse than those which maximize the number of survivors. Using the MCLP solutions,

the expected number of survivors is up to 7.7% lower than it could be, and using the q-median solutions, the expected number of survivors is up to 23.5% lower than it could be.

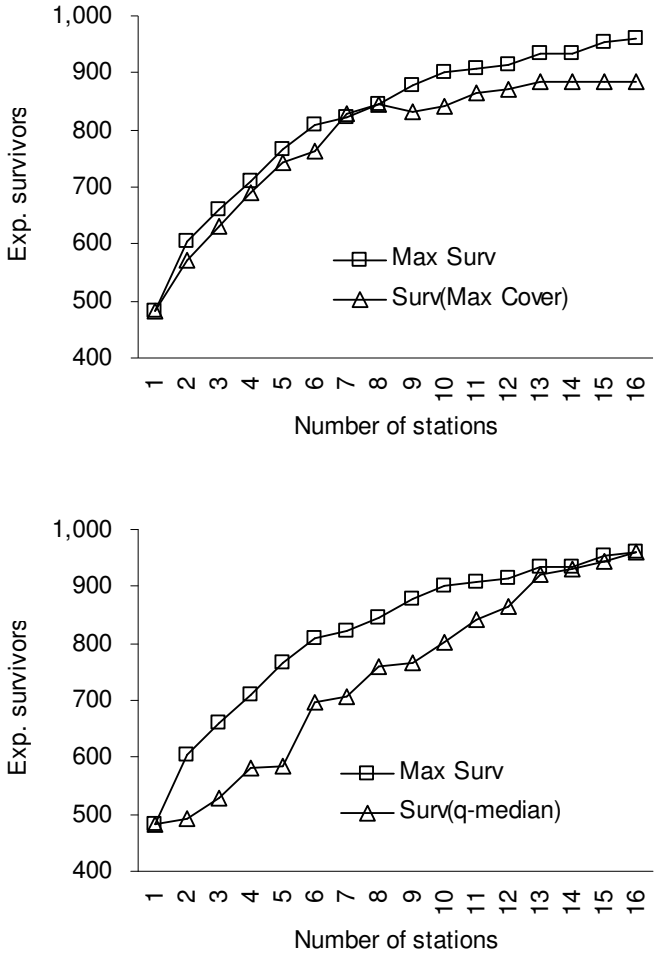

Figure 4: The expected number of survivors for the optimal solutions of MCLP and MSLP (top panel) and q-median and MSLP (bottom panel) for 1 through 16 stations.

When used as proxies for MSLP, both the MCLP and q-median have the weakness that when the number of stations increases, the expected number of survivors may *decrease*—see the MCLP results in Figure 4 when q increases from 8 to 9. For MCLP, this can happen when the model sees an opportunity to extend coverage to areas where the travel time from the closest facility is just below the coverage standard, and the survival probability is low. For the q-median, this can happen when the model sees an opportunity to reduce the longest response times at the expense of the shortest response times. The shortening of the long response times may not do anything to increase survivability, while lengthening the short response times can decrease survivability considerably.

We note one additional weakness of MCLP: With 13 stations, this model can cover all of the demand in the city. Hence, the solutions of the problems with $q > 13$ are all identical to the solution for $q = 13$. In contrast, MSLP (and the q-median, with the exceptions noted in the previous paragraph) is able to improve the expected number of survivors each time a new station is added.

Finally, the objective function value of MSLP is more meaningful than that of MCLP and q-median and it can be more useful in deciding how many stations an EMS system should have. For example, if a decision-maker is undecided between 9 and 10 stations, all one can say based on MCLP is that 10 stations will "cover" 23 more calls. In contrast, based on MSLP one can state that the 10^{th} station will save an average of 15 more lives per year in cardiac arrest cases.

We conclude that MCLP is a blunt tool for the task on hand. It lacks the sophistication to properly differentiate between different outcomes and oversimplifies the problem by classifying the population into two sets (covered and uncovered). It may be adequate for the design of non-emergency service systems where the response time is not critical. However, for EMS systems, MCLP is a poor model and MSLP is superior. The q-median has a different limitation—it sees a response time of 20 minutes as twice as bad as a response time of 10 minute, while in terms of survivability, there is little difference.

4.3 Sensitivity of the MSLP Results to the Shape of the Survival Function Used

We explored the sensitivity of our results to the shape of the survival function by solving MSLP with two other survival functions—one with higher survival probabilities and slower decay and the other with lower survival probabilities. Figure 5 shows all three survival functions. Recall that the base case survival function was the one from Figure 3, adjusted for the assumption of deterministic response times. The "high" survival function is the one from Figure 3 and the "low" survival function is the one from Figure 3, divided by two.

The solutions to MSLP were identical for the three survival functions, for all values of q. This provides us with some empirical evidence that the optimal locations are not sensitive to the parameters of the survival function.

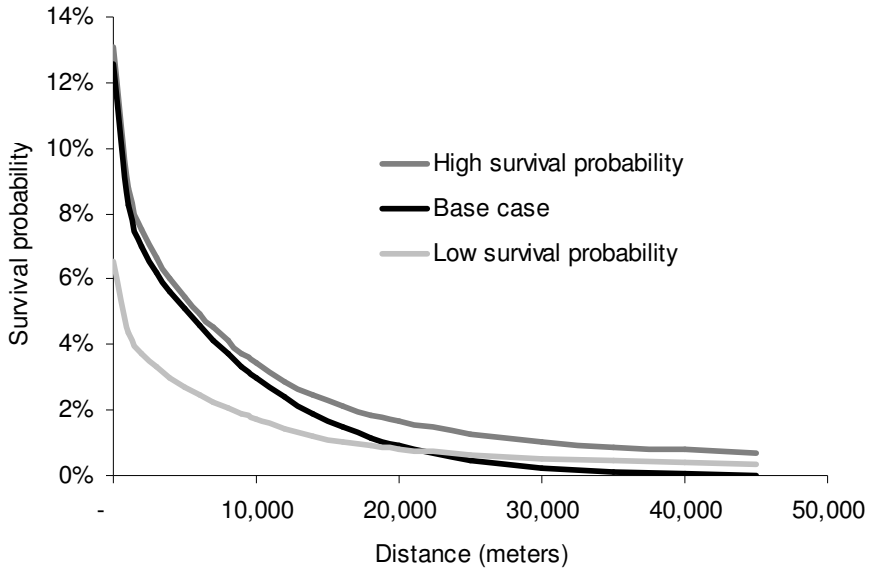

Figure 5: The survival functions used to compare the MCLP, MSLP, and q-median models.

In this section we have empirically demonstrated significant differences between MCLP, q-median, and MSLP and hopefully convinced the reader that MSLP is more suitable than

MCLP or *q*-median for EMS station location and that the exact shape of the survival function is not very important. We now present covering models with increased realism, each of which can be enhanced further by incorporating a survival function.

5. More realistic EMS location models

In Section 4 we focused on one weakness of MCLP; the lack of discrimination between distances that are within (outside) the coverage standard. There are other shortcomings of MCLP which have been addressed in the literature. MCLP ignores two significant sources of uncertainty:

1) It assumes there is always an EMS vehicle available at a station. However, in practice, EMS vehicles are busy 30% – 70% of the time and stations are regularly exposed (i.e., have no EMS vehicles to respond to a call) during the course of a day.
2) It assumes response times are deterministic. Yet actual travel times between an origin and a destination show lognormal distributions with fairly high coefficients of variation (Budge, 2004) and pre-travel delays are highly variable as well (Budge et al, 2007b).

The Maximal Expected Covering Location Problem (MEXCLP) and the Maximal Covering Location Problem with Probabilistic Response Times (MCLP+PR) have been proposed as extensions of MCLP to deal with these two types of uncertainty. Finally the Maximal Expected Covering Location Problem with Probabilistic Response Times (MEXCLP+PR) incorporates both types of uncertainty. The two sources of uncertainty can also be incorporated when minimizing average response time. Jarvis (1975) proposed a *locate-allocate* heuristic for this purpose; see also Larson (1979). This heuristic uses the approximate hypercube model to evaluate the average response time. The heuristic iterates between evaluating the average response time and dispatch probabilities (the allocation step) and moving ambulances to different stations so as to minimize average response time, assuming that the dispatch probabilities do not change (the location step). Appendix A contains formulations for MEXCLP, MCLP+PR, MEXCLP+PR, and a description of Jarvis's *locate-allocate* heuristic.

In the remainder of this section, we describe how survival functions can be incorporated into MEXCLP, MCLP+PR, and MEXCLP+PR models, replacing the maximization of expected coverage with maximization of the expected number of surviving patients. As in MSLP, let p_{ij} denote the survival probability of a patient at demand node *i*, when served by an EMS unit from station *j*. In what follows, we describe the computation of p_{ij} under different assumptions about response time variability and ambulance availability.

5.1 The Maximal Expected Survival Location Problem (MEXSLP)

MEXCLP does not differentiate between locations covering a demand node so long as they are within the radius of coverage. However, the use of the survival function necessitates a model which recognizes EMS units from different stations. The model for MEXCLP+PR, with its definitions of preference orders and the way it handles the busy probabilities, is suitable to integrate the survival function with the assumptions of this model. In this case,

$$p_{i,k(i,j)} = s(t_{ji} + t_d) \prod_{u=1}^{j-1} p^{z_{k(i,u)}} \left(1 - p^{z_{k(i,j)}}\right) \tag{15}$$

where p is the average fraction of time an EMS unit is busy, $k(i, j)$ is the jth preferred station for demand node i, and z_j is the number of EMS units allocated to station j. The formulation is:

MEXSLP:

$$\max \quad \sum_{i=1}^{m} d_i \sum_{j=1}^{n} p_{ij} = \sum_{i=1}^{m} d_i \sum_{j=1}^{n} p_{i,k(i,j)} \tag{16}$$

$$= \sum_{i=1}^{m} d_i \sum_{j=1}^{n} s(t_{ji} + t_d) \prod_{u=1}^{j-1} p^{z_{k(i,u)}} \left(1 - p^{z_{k(i,j)}}\right) \tag{17}$$

$$\text{s.t} \quad \sum_{j=1}^{n} z_j \le r \tag{18}$$

$$z_j \in \{0, 1, \ldots, c_j\}, \; j = 1, \ldots, n \tag{19}$$

5.2 The Maximal Survival Location Problem with Probabilistic Response Time (MSLP+PR)

For this case, the variable and constraint structure of MSLP is sufficient and the only modification required is an updated objective function. We can express p_{ij} as $p_{ij} = E[s(R_{ij})]y_{ij}$, resulting in the following formulation:

MSLP+PR:

$$\max \quad \sum_{i=1}^{m} d_i \sum_{j=1}^{n} p_{ij} = \sum_{i=1}^{m} d_i \sum_{j=1}^{n} E[s(R_{ij})]y_{ij} \tag{20}$$

s.t. (10) – (14)

The coefficients $E[s(R_{ij})]$ can be pre-computed for each demand node – station pair (i,j), using the methods discussed in Section 3. Interestingly, the formulation for MSLP+PR is structurally identical to that for MSLP, MCLP+PR, and the q-median problem. The only difference between these formulations is the constant that multiplies y_{ij} in the inner summation of the objective function.

5.3 The Maximal Expected Survival Location Problem with Probabilistic Response Time (MEXSLP+PR)

Similar to the previous case, the constraint structure of the original model (MEXCLP+PR) is sufficient and we only need to modify the objective function. In accordance with the assumptions of this model

$$p_{i,k(j)} = \prod_{u=1}^{j-1} \hat{p}_{k(i,u)}^{z_{k(i,u)}} \left(1 - \hat{p}_{k(i,j)}^{z_{k(i,j)}}\right) E[s(R_{i,k(i,j)})] \tag{21}$$

where $\hat{p}_j$ is the average fraction of time an EMS unit at station j is busy. The resulting model is:

MEXSLP+PR:

$$\max \quad \sum_{i=1}^{m} d_i \sum_{j=1}^{n} p_{ij} = \sum_{i=1}^{m} d_i \sum_{j=1}^{n} p_{i,k(i,j)} \tag{22}$$

$$= \sum_{i=1}^{m} d_i \sum_{j=1}^{n} E[s(R_{i,k(i,j)})] \prod_{u=1}^{j-1} \hat{p}_{k(i,u)}^{z_{k(i,u)}} \left(1 - \hat{p}_{k(i,j)}^{z_{k(i,j)}}\right) \tag{23}$$

s.t. (18) – (19)

As in the previous case, we pre-compute $E[s(R_{ij})]$ for each demand node – station pair (i,j).

6. Computational results

We used data from the Edmonton EMS system, as in Section 4. The data are available from http://www.bus.ualberta.ca/aingolfsson/data/. Table 1 compares the size of the different optimization models. The linear models were solved on a Dell PowerEdge workstation with 1.13 Ghz CPU clock and 1 GB of RAM, using CPLEX 8.11 in under 1 CPU second per problem. The nonlinear models were solved on a PC with 3.0 GHZ CPU clock and 1 GB of RAM, using the student version of GAMS 22.0, with runtimes ranging from 10 to 300 CPU seconds. On the same PC the runtimes for the *locate-allocate* average response time minimization heuristic were no more than 5 seconds.

Table 1: Sizes and characteristics of the optimization problems solved (for 180 demand nodes and 16 candidate locations).

Problem	Binary variables	Integer variables.	Constraints	Linear?
MCLP	196		181	Linear
MSLP	2,896		197	Linear
q-median	2,896		197	Linear
MCLP+PR	2,896		197	Linear
MSLP+PR	2,896		197	Linear
MEXCLP	180*r*	16	181	Linear
MEXSLP		16	1	Nonlinear
MEXCLP+PR		16	1	Nonlinear
MSLP+PR		16	1	Nonlinear

The algorithm of Budge et al. (2007b) to solve MEXCLP+PR is reported to either converge to a single solution, or to cycle between two solutions. In case of cycling, the objective function values of each of the two solutions are incorrect since they are computed with respect to busy probabilities associated with the other solution. To overcome this problem, at every iteration we computed the "real" objective function value associated with a solution by computing the corresponding busy probabilities, and we chose the solution with the higher "real" objective function value as the best solution in case of cycling.

When computing the expected number of survivors, we used only the arrival rate of urgent calls (about 29% of the total), i.e., we assumed that the probability of survival for all urgent calls varies similarly with response time as it does for cardiac arrest calls, while for non-urgent calls, the probability of death is negligible. We used the survival function from Figure 3. When computing busy probabilities, we used the arrival rate for all types of calls, because

all calls contribute to the workload of the EMS units. The same yardstick is used to compare all solutions, i.e., the expected number of survivors as evaluated with the approximate hypercube model.

6.1 Comparing Models Incorporating only Randomness in Response Time: MCLP+PR vs. MSLP+PR

We begin by considering models that incorporate randomness in response times but assume perfect availability, i.e., we compare MCLP+PR to MSLP+PR. Recall that in Section 4 we reported a comparison of MCLP and MSLP and found that the number of survivors resulting from ambulance locations generated by MCLP can be up to 7.7% lower than the number of survivors resulting from MSLP. When one incorporates random response times into MCLP, then the coverage value for a particular demand node, instead of being either zero or one, becomes a probability between zero and one. As Table 2 shows, after this has been done, the marginal benefit of incorporating a survival function is smaller than before (as reported in Section 4). In 8 of the 16 problems MCLP+PR and MSLP+PR provide the same expected number of survivors. The percent improvement due to the incorporation of the survival function can be as high as 5.3%, and it averages 0.9% over the 16 problems. As shown in Figure 6, the incorporation of probabilistic response time results in expected coverage decaying gradually with distance in a manner similar to the survival probability and this leaves less room for improvement due to the use of a survival function. Nevertheless, the survival function improves the performance of the model in the majority of the test problems.

Table 2: The expected number of survivors for solutions to MCLP+PR and MSLP+PR (i.e., the two models that incorporate probabilistic response times) for 1 through 16 EMS stations. The fourth column contains the percent deviation between the entries in columns two and three.

q	MCLP+PR	MSLP+PR	% Deviation
1	481.2	481.2	0.0%
2	572.0	604.3	5.3%
3	636.3	659.0	3.5%
4	686.6	708.8	3.1%
5	764.3	764.3	0.0%
6	809.1	809.1	0.0%
7	821.2	821.2	0.0%
8	843.8	843.8	0.0%
9	866.5	878.6	1.4%
10	901.5	901.4	0.0%
11	908.0	906.6	-0.1%
12	913.6	914.0	0.0%
13	930.1	932.7	0.3%
14	931.8	934.0	0.2%
15	944.8	953.4	0.9%
16	959.0	959.0	0.0%

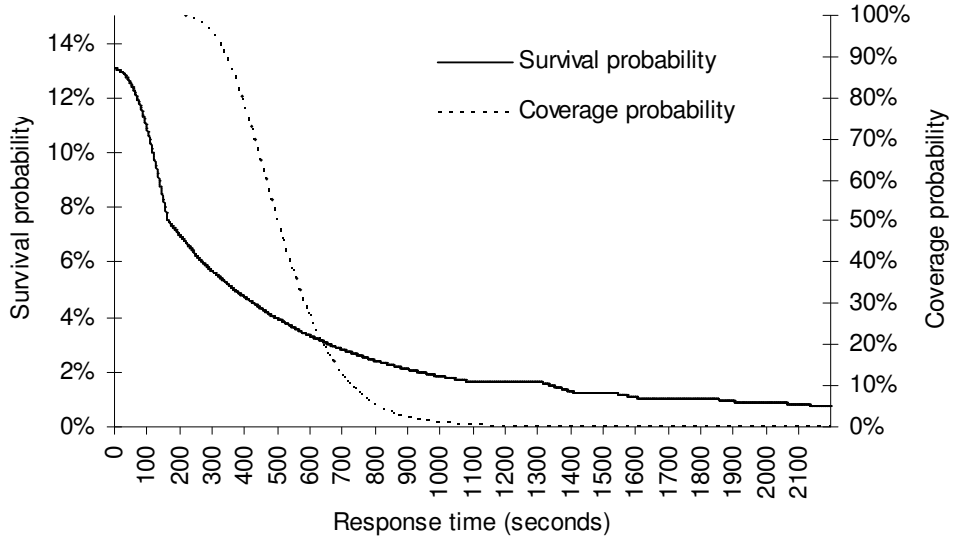

Figure 6: A comparison of the expected survival probability E[$s(R)$] using (4) and the probability of coverage Pr{$R \leq t_c$}, as a function of expected response time $t \equiv$ E[R].

6.2 Comparing Models Involving Busy Probabilities: MEXCLP vs. MEXSLP, and MEXCLP+PR vs. MEXSLP+PR

For the models involving busy probabilities (namely the MEX*** family), we used a two-dimensional experimental design, the first parameter being r (number of EMS units) and the second being p (system-wide average busy probability of EMS units). We estimate the average system-wide busy probability for MEXCLP+PR and MEXSLP+PR as $p = \lambda\tau / q$ where $\lambda = \sum_{i=1}^{m} d_i$ is the total arrival rate of calls to the system, and τ is the average time that an EMS unit is tied up with a call. We used the same formula in reverse by inserting the number of EMS units and the targeted system-wide busy probability, taking the total arrival rate of calls as output, and scaling the demand data used for busy probability estimation accordingly. The purpose of including p in the experimental design is to control for the overall level of congestion in the system. This does not mean that we force the busy probabilities $\hat{p}_j$ to be the same for all stations—we still allow them to vary, as indicated in the formulations for MEXCLP+PR and MEXLSP+PR. We caution that for these models, our solutions may not be optimal since the iterative algorithm is not guaranteed to find an optimal solution.

Table 3 summarizes the comparison of MEXCLP with MEXSLP. The expected number of survivors was computed using the approximate hypercube model for the solutions to both models. Table 3 shows the percent improvement in the expected number of survivors achieved by MEXSLP over MEXCLP. The average improvement is 0.6%, the maximum improvement is 3.7%, and the maximum degradation is 1.8%. The MEXSLP solution is superior to that of MEXCLP by at least 1% in 32 instances, and the opposite is true in only 4 instances. It seems that the improvements are most significant for smaller number of EMS vehicles and lower levels of system congestion (the value of p). This makes sense because it is more challenging to locate 9 EMS vehicles in 16 stations than it is to locate 24 EMS

vehicles in 16 stations. With higher congestion, the locations of the vehicles become less important. (To see why, consider the limiting case when the congestion is so high that most of the time, only one vehicle is available. In that case, the closest vehicle to the call will usually be busy, and the vehicle that responds will simply be the one that happens to be available.) Based on our experiment, it is fair to say that the survival function improves the performance of MEXCLP in most instances.

Table 3: The percent improvement in the expected number of survivors achieved by using MEXSLP instead of MEXCLP. The number of EMS vehicles is varied from 5 to 25 (with 16 stations), and the busy probability for the EMS vehicles is varied from 0.1 to 0.6.

***r*\p**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**
5	3.2%	3.2%	1.7%	1.2%	1.9%	1.9%
6	0.0%	0.0%	0.0%	0.0%	3.1%	1.9%
7	-0.1%	0.8%	1.6%	2.4%	0.0%	1.5%
8	0.2%	0.8%	1.1%	1.7%	2.2%	0.4%
9	3.5%	3.4%	0.9%	3.7%	0.0%	0.0%
10	0.0%	0.0%	0.0%	3.6%	0.6%	1.4%
11	2.7%	0.9%	2.7%	2.8%	0.2%	1.6%
12	2.7%	0.8%	2.5%	2.5%	0.2%	0.2%
13	2.6%	0.9%	0.0%	0.0%	0.6%	0.0%
14	3.2%	-1.4%	-0.7%	0.0%	0.1%	-0.4%
15	0.1%	1.0%	0.7%	-1.0%	1.3%	-1.0%
16	-0.8%	0.6%	0.2%	0.0%	-1.3%	-1.8%
17	0.9%	0.8%	0.6%	-0.3%	0.0%	-0.8%
18	0.7%	0.4%	0.1%	1.1%	1.2%	-0.7%
19	0.0%	-0.1%	-0.8%	-0.1%	1.2%	-1.2%
20	-0.2%	-0.9%	0.7%	0.2%	0.6%	-0.6%
21	-0.3%	-0.9%	0.4%	0.3%	0.0%	0.2%
22	-0.4%	-0.1%	-0.3%	-0.5%	-0.4%	0.8%
23	-0.2%	0.0%	0.0%	0.0%	0.0%	0.6%
24	0.1%	0.2%	0.2%	0.0%	-0.3%	-0.5%
25	0.1%	0.2%	0.1%	-0.1%	-0.4%	-0.5%

Table 4 summarizes the comparison of MEXCLP+PR with MEXSLP+PR. These results are mixed. The expected number of survivors is the same in 51 cases (40%), higher with MEXSLP+PR in 43 cases (34%), and higher with MEXCLP+PR in 32 cases (25%). Overall, the two models appear to perform about equally well. MEXSLP+PR appears to find solutions that perform a little better when congestion is low and MEXCLP+PR appears to find solutions that perform a little better when congestion is high. Our earlier observation (when discussing the results in Table 2) that once the probabilistic response times are included in the models the survival function makes less of a difference is probably relevant here as well. Note that Table 4 summarizes the results for the most refined pair of models which include probabilistic response times as well as busy probabilities. MEXCLP+PR is a sophisticated model and the inclusion of the survival function does not add much to the performance of its solutions.

Table 4: The percent improvement in the expected number of survivors achieved by using MEXSLP+PR instead of MEXCLP+PR. The number of EMS vehicles is varied from 5 to 25 (with 16 stations), and the busy probability for the EMS vehicles is varied from 0.1 to 0.6.

r\p	0.1	0.2	0.3	0.4	0.5	0.6
5	0.0%	0.0%	-0.2%	0.9%	1.2%	0.6%
6	0.0%	0.0%	0.0%	-0.5%	0.6%	0.0%
7	0.0%	0.9%	0.8%	1.8%	-1.1%	0.0%
8	0.2%	0.3%	0.0%	0.4%	0.1%	-0.9%
9	0.0%	0.0%	0.0%	0.2%	0.0%	-0.7%
10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
11	0.3%	0.0%	0.0%	-0.5%	0.0%	-1.0%
12	0.3%	0.2%	0.0%	0.2%	0.0%	0.0%
13	0.6%	0.2%	0.0%	-0.1%	-0.8%	-0.7%
14	0.5%	0.7%	0.0%	0.0%	0.0%	0.0%
15	1.1%	1.4%	-0.2%	-0.7%	0.0%	-0.7%
16	0.4%	0.8%	0.9%	0.0%	-1.3%	-1.8%
17	0.0%	0.1%	-0.2%	-1.2%	0.0%	-1.3%
18	0.0%	0.0%	0.0%	0.8%	0.0%	-1.1%
19	0.0%	0.0%	0.0%	0.0%	-0.5%	-0.6%
20	0.1%	0.1%	0.1%	0.2%	-0.5%	-0.6%
21	0.1%	0.2%	0.0%	0.0%	-0.5%	-0.5%
22	0.0%	0.0%	0.0%	0.0%	-0.4%	0.6%
23	0.0%	0.0%	0.0%	0.0%	-0.3%	0.5%
24	0.0%	0.0%	0.0%	0.0%	-0.6%	0.0%
25	0.0%	0.0%	0.0%	0.0%	-0.4%	0.0%

6.3 Comparing MEXSLP+PR with Solutions from the Jarvis (1975) Heuristic

Table 5 compares MEXSLP+PR solutions to those obtained with Jarvis's (1975) average response time minimization heuristic. Somewhat surprisingly, this heuristic produces solutions that perform quite poorly with respect to the expected number of survivors. MEXSLP+PR improves these solutions by 6.5% on average (and by as much as 21.6%). The differences are higher when the number of vehicles is smaller and (surprisingly) when the system congestion is higher. These results may indicate either that average response time is a poor proxy for the expected number of survivors, or that Jarvis's heuristic fails to globally minimize the average response time.

6.4 Comparing All Models

As a final comparison of the ten optimization models presented, we fixed the number of EMS vehicles r at 6 and the average busy probability p at 0.3, solved all ten models, and then used the approximate hypercube model to evaluate the expected number of survivors for each of the ten solutions. The results are shown in Table 6.

Table 5: The percent improvement in the expected number of survivors when using MEXSLP+PR compared to minimizing average response time using Jarvis's heuristic. The number of EMS vehicles is varied from 5 to 25 (with 16 stations), and the busy probability for the EMS vehicles is varied from 0.1 to 0.6.

r\p	0.1	0.2	0.3	0.4	0.5	0.6
5	21.6%	16.3%	4.4%	5.4%	5.1%	7.2%
6	9.8%	12.0%	7.9%	12.2%	6.1%	7.0%
7	11.3%	13.9%	16.1%	10.8%	9.3%	13.2%
8	8.8%	10.2%	12.7%	15.7%	9.6%	12.0%
9	9.9%	12.0%	13.8%	13.5%	11.6%	8.2%
10	8.8%	10.1%	13.1%	8.4%	13.0%	12.9%
11	6.7%	8.5%	10.2%	11.6%	14.9%	14.9%
12	6.3%	7.5%	8.8%	11.0%	13.4%	8.5%
13	1.6%	3.0%	4.6%	6.7%	8.4%	12.4%
14	1.6%	3.2%	4.8%	6.9%	8.5%	10.3%
15	1.5%	3.0%	4.8%	6.3%	8.4%	7.9%
16	1.3%	2.4%	4.4%	5.8%	6.9%	8.3%
17	0.0%	0.0%	1.4%	2.2%	4.6%	5.4%
18	2.2%	3.9%	4.7%	5.5%	7.4%	8.2%
19	2.5%	4.4%	5.4%	5.2%	7.2%	8.1%
20	2.4%	4.0%	4.8%	5.1%	5.8%	7.3%
21	2.4%	4.0%	4.8%	5.1%	4.7%	6.9%
22	0.5%	1.5%	2.3%	2.6%	2.3%	4.4%
23	0.4%	1.2%	1.7%	1.8%	1.9%	3.9%
24	0.0%	0.0%	0.0%	0.0%	0.0%	1.4%
25	0.6%	1.4%	1.6%	1.6%	1.5%	1.9%

Table 6: The comparison of the output of all 10 models, solved for $r = 6$, $p = 0.3$, and using the approximate hypercube to evaluate the expected number of survivors.

	Incorporation of uncertainty			
Type of model	None	Response times	Server availability	Both
Avg. response time	697.3			745.0
Coverage	761.6	809.1	809.1	809.1
Survival	809.1	809.1	809.1	809.1

The results summarized in Table 6 are quite striking. We see that if average response time is used as a proxy for expected number of survivors (by solving the q-median problem), then the inclusion of response time uncertainty and server availability (using the Jarvis (1975) heuristic) improves the number of survivors, but it still falls far short of the best possible. In contrast, if expected coverage is used as a proxy, then the inclusion of either response time uncertainty or server availability in the model brings the expected number of survivors up to its best known value. More importantly, just changing the objective function to expected number of survivors achieves all of the benefits, even without incorporating the two elements of uncertainty in the model.

7. Concluding remarks

This paper points to a weakness of covering models for locating emergency vehicles. We discuss research from the medical literature that allows for accurate modeling of consequences of different response times for cardiac arrest patients. We then show how a survival function that maps response time to survival probability can be incorporated into the deterministic maximum covering model so that the objective becomes one of maximizing the expected number of survivors. We proceed to modify three more sophisticated covering models by including the survival function in each. The incorporation of the survival function does not complicate the optimization problems much.

We highlight several weaknesses of the deterministic maximum covering model, including its inability to recommend additional beneficial facilities once the whole region has been covered and its use of the abstract concept of coverage. Our computational experiment indicates that incorporating survival functions can result in EMS unit locations that save more lives. The standard covering approach is a blunt tool for emergency facility location and it should be used with great caution. In terms of computational effort, optimization models that maximize expected number of survivors are only slightly less tractable than covering models. In terms of data requirements, survival models are more data-intensive, but some EMS agencies are already collecting the necessary information. We have illustrated how this data can be incorporated in the optimization models.

One obvious shortcoming of our approach is that we only have survival functions for one type of emergency call. The commonly used standard of responding to 90% of all high priority calls within 9 minutes shares this shortcoming, because it is also a cardiac arrest-driven standard (Eisenberg, 1979). The EMS world seems to be paying considerable attention to cardiac arrests given their relative frequency and their possible consequences. However, EMS practitioners and medical researchers recognize that quantifying the impact of response time for other call types is important (Pons and Markovchick, 2002). If future research leads to quantifiable survival functions for other call types, then they can be incorporated in the models we have presented, by combining survival functions for different call types using weights corresponding to the frequency of different call types.

In addition to the deployment of ambulances, the framework we have used also permits study of broader policy issues, such as the impact of actions to increase rates of bystander CPR. We hope that this paper will help encourage further research on survival functions and other more direct and realistic models of EMS operations.

References

Anonymous (2006). *Emergency Medical Services: At the Crossroads*. National Academies Press, Washington, DC.

Berman, O., D. Krass, Z. Drezner (2003). The gradual covering decay location problem on a network. *European Journal of Operational Research*, 151, 474-480.

Blackwell, T. H., J. S. Kaufman (2002). Response time effectiveness: comparison of response time and survival in an urban emergency medical services system. *Academic Emergency Medicine*, 9(4), 288-295.

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

Budge, S. (2004). Emergency Medical Service Systems: Modelling Uncertainty in Response Time. Ph.D. thesis, University of Alberta.

Budge, S., A. Ingolfsson, E. Erkut (2007a). Approximating vehicle dispatch probabilities for emergency service systems with location-specific service times and multiple units per location. *Operations Research*, forthcoming.

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

Campbell, J. F. (1992). Selecting routes to minimize urban travel time. *Transportation Research B*, 26B(4), 261-274.

Carson, Y. L., R. Batta (1990). Locating an ambulance on the Amherst campus of the State University of New York at Buffalo. *Interfaces* 20(5), 43-49.

Chaiken, J., R. C. Larson (1972). Methods for allocating urban emergency units: a survey. *Management Science* 19(4),110-132.

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

Church, R. L., Roberts, K.L (1983). Generalized coverage models and public facility location. *Papers of the Regional Science Association*, 53, 117-135.

Cretin, S., T. R. Willemain (1979). A model of preshospital death from ventricular fibrillation following myocardial infarction. *Health Services Research*, 14(3), 221-234.

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

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

Davis, R. (2003a). Many lives are lost across USA because emergency services fail; turf wards between ambulance, fire crews cause deadly delays. *USA Today*, 28 July 2003, pp. A-01.

Davis, R. (2003b). Special report: sluggish responses to emergencies let patients die; precise measures of EMS response times can save lives. *USA Today*, 29 July 2003, pp. A-01.

De Maio, V.J., I.G. Stiell, G.A. Wells, D.W. Spaite (2003). Optimal defibrillation for maximum out-of-hospital cardiac arrest survival rates. *Annals of Emergency Medicine*, 42(2), 242-250.

Eisenberg, M.S., L. Bergner, A. Hallstrom (1979). Cardiac resuscitation in the community: importance of rapid provisioning and implications for program planning. *Journal of the American Medical Association*, 241(18): 1905-1907.

Eisenberg, M.S., B.T. Horwood, R.O. Cummins, R. Reynolds-Haertle, T.R. Hearne (1990). Cardiac arrest and resuscitation: a tale of 29 cities. *Annals of Emergency Medicine*, 19(2): 179-186.

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

Fitch, J (2005). Response times: myths, measurement and management. *Journal of Emergency Medical Services*, 30(9), 46–56.

Jarvis, J. (1975). Optimization in Stochastic Service Systems with Distinguishable Servers. Ph.D. thesis, Massachsuetts Institute of Technology.

Jia, H., F. Ordonez, M. Dessouky (2007). A modeling framework for facility location of medical service for large-scale emergencies. *IIE Transactions*, 39(1), 41-55.

Karasakal, E.K., O. Karasakal (2004). A maximal covering location model in the presence of partial coverage. *Computers and Operations Research*, 31(9), 1515–1526.

Kolesar, P., W. Walker, J. Hausner (1975). Determining the relation between fire engine travel times and travel distances in New York City. *Operations Research*, 23(4), 614-627

Larsen, M.P., M.S. Eisenberg, R.O. Cummins, A.P. Hallstrom (1993). Predicting survival from out-of-hospital cardiac-arrest—a graphic model. *Annals of Emergency Medicine*. 22(11), 1652-1658.

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

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

Larson, R. C. (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, K. B. Haley (ed), North-Holland Publishing Co., Amsterdam, Holland.

Marianov, V., ReVelle C. (1995). Siting emergency services. in *Facility Location: A Survey of Applications and Methods*, Drezner, Z., ed. Springer Series in Operations Research, 199-222.

Moeller, B. J. 2004. Obstacles to measuring emergency medical services performance. *EMS Management Journal* 1(2): 8 – 15.

Pedersen, R. (2002). Insurance hikes loom as response times lag. *Edmonton Journal*, June 28, B1.

Pons, P.T., V. J. Markovchick (2002). Eight minutes or less: does the ambulance response time guideline impact trauma patient outcome? *The Journal of Emergency Medicine* 23, 43-48.

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

Saydam, C., M. McKnew (1985). A separable programming approach to expected coverage: an application to ambulance location. *Decision Sciences*, 16, 381-398.

Swersey, A.J. (1994). The deployment of police, fire, and emergency medical units. In *Handbooks in Operations Research and Management Science, Vol. 6: Operations Research and the Public Sector*, Barnett, A., S.M. Pollock, and M.H. Rothkopf (eds.). North-Holland.

Toregas, C., C. ReVelle, L. Bergman (1971). The location of emergency service facilities. *Operations Research*, 19:1363-1373.

Valenzuela, T.D., D.J. Roe, S. Cretin, D.W. Spaite, M.P. Larsen (1997). Estimating effectiveness of cardiac arrest intervention–A logistic regression survival model. *Circulation*, 96(10), 3308-3313.

Valenzuela, T.D., D. J. Roe, G. Nichol, L. L. Clark, D. W. Spaite, R. G. Hardman (2000). Outcomes of rapid defibrillation by security officers after cardiac arrest in casions. *The New England Journal of Medicine*, 343(17), 1206-1209.

Waaelwijn, R.A., R. de Vos, J.G.P. Tijssen, R.W. Koster (2001). Survival models for out-of-hospital cardiopulmonary resuscitation from the perspectives of the bystander, the first responder, and the paramedic. *Resuscitation*, 51(2), 113-122.

Willemain, T.R. (1975). The Status of Performance Measures for Emergency Medical Services. *Journal of the American College of Emergency Physicians*, 4, 143-151.

Williams, D. M. (2007). 2006 JEMS 200-City Survey: EMS from all angles. *Journal of Emergency Medical Services*, 32(2), 38-54.

Appendix A: Maximum Coverage Formulations

Define m : the number of demand nodes,
n :the number of candidate locations,
q :the maximum number of stations,
d_i :the demand of node i,
t_c :the coverage radius of a station in time units,
t_{ji} :the travel time from candidate location j to demand node i,
t_d: the pre-travel delay,

$$x_j = \begin{cases} 1, & \text{if candidate location } j \text{ is selected} \\ 0, & \text{otherwise} \end{cases},$$

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

$$a_{ij} = \begin{cases} 1, & \text{if if demand node } i \text{ is covered by candidate location } j, \text{i.e.} t_{ji} + t_d \le t_c \\ 0, & \text{otherwise} \end{cases}.$$

The Maximum Coverage Location Problem (MCLP):

MCLP:

$$\max \quad \sum_{i=1}^{m} d_i y_i \qquad \text{(A1)}$$

$$\text{s.t.} \quad \sum_{j=1}^{n} a_{ij} x_j \ge y_i \,, \; i = 1, \ldots, m \qquad \text{(A2)}$$

$$\sum_{j=1}^{n} x_j \le q \quad \text{(A3)}$$

$$x_j \in \{0,1\},\ j = 1, \ldots, n \quad \text{(A4)}$$

$$y_i \in \{0,1\},\ i = 1,\ldots,m \quad \text{(A5)}$$

The objective function (A1) maximizes total demand covered. Constraints (A2) state that demand node *i* can only be covered if at least one candidate location that covers *i* is selected. Constraint (A3) limits the number of facilities to *q*. In this model, each station houses one EMS vehicle.

The Maximal Expected Covering Location Problem (MEXCLP)

There are two formulations for the MEXCLP in the literature. The first formulation by Daskin (1983) is an integer program and the second is a non-linear integer program by Saydam and McKnew (1985). Both models account for the probability that an EMS unit may be busy. We provide only the linear model for the sake of brevity.

Let r denote the maximum number of EMS units,
p denote the average fraction of time an EMS unit is busy,
c_j be the maximum number of EMS units that can be stationed at candidate location j,
z_j be the number of EMS units allocated to station j, and

$$\hat{y}_{ik} = \begin{cases} 1, & \text{if demand node } i \text{ is covered by at least } k \text{ units} \\ 0, & \text{otherwise} \end{cases}.$$

The linear programming model for the MEXCLP follows:

MEXCLP:

$$\max \quad \sum_{i=1}^{m} d_i \sum_{k=1}^{r} (1-p) p^{k-1} \hat{y}_{ik} \quad \text{(A6)}$$

$$\text{s.t.} \quad \sum_{k=1}^{r} \hat{y}_{ik} \le \sum_{j=1}^{n} a_{ij} z_j,\ i = 1,\ldots, m \quad \text{(A7)}$$

$$\sum_{j=1}^{n} z_j \le r \quad \text{(A8)}$$

$$\hat{y}_{ik} \in \{0,1\},\ i = 1,\ldots, m;\ k = 1,\ldots, r \quad \text{(A9)}$$

$$z_j \in \{0,1,\ldots,c_j\},\ j = 1,\ldots, n \quad \text{(A10)}$$

The inner summation of objective function (A6) calculates the probability that there will be an EMS unit available to service demand node *i*. Therefore, objective function (A6) maximizes the expected coverage of demand nodes. Constraints (A7) state that the actual number of EMS units covering node *i* (LHS of constraint) cannot exceed the total number of EMS units that can cover node *i* (RHS). Constraint (A8) limits the total number of EMS units to be allocated to all open candidate locations.

The Maximal Covering Location Problem with Probabilistic Response Time (MCLP+PR)

Let y_{ij} equal 1 if demand node i is closest to candidate location j and let P_{ij} be the probability that an ambulance at station j can reach demand node i within the coverage time standard.

Daskin (1987) provides the following formulation for the Maximal Covering Location Problem with Probabilistic Response Time:

MCLP+PR:

$$\max \quad \sum_{i=1}^{m} d_i \sum_{j=1}^{n} P_{ij} y_{ij} \tag{A11}$$

$$\text{s.t.} \quad \sum_{i=1}^{m} y_{ij} \le m x_j \,, j = 1, \ldots, n \tag{A12}$$

$$\sum_{j=1}^{n} y_{ij} = 1, i = 1, \ldots, m \tag{A13}$$

$$\sum_{j=1}^{n} x_j \le q \tag{A14}$$

$$x_j \in \{0,1\}, j = 1, \ldots, n \tag{A15}$$

$$y_{ij} \in \{0,1\}, i = 1, \ldots, m, \; j = 1, \ldots, n \tag{A16}$$

Objective function (A11) maximizes the total demand covered account for the coverage probabilities. Constraints (A12) and (A13) ensure that a demand node is assigned to only one EMS facility. Constraint (A14) requires that at most q candidate locations be chosen. As in MCLP, each candidate location houses at most one vehicle. We note that the MSLP formulation is structurally identical to the MCLP+PR formulation, with $s(t_{ji} + t_d)$ in MSLP replacing P_{ij} in MCLP+PR (i.e. {(10) – (14)} = {(A12) – (A16)}).

The Maximal Expected Covering Location Problem with Probabilistic Response Time (MEXCLP+PR)

Let $k(i, j)$ denote the j^{th} preferred station for demand node i, and $\hat{p}_j$ denote the average fraction of time an EMS unit at station j is busy.

Budge et al. (2007b) formulate the Maximal Expected Covering Location Problem with Probabilistic Response Time as follows:

MEXCLP+PR:

$$\max \quad \sum_{i=1}^{m} d_i \sum_{j=1}^{n} P_{i,k(i,j)} \prod_{u=1}^{j-1} \hat{p}_{k(i,u)}^{z_{k(i,u)}} \left(1 - \hat{p}_{k(i,j)}^{z_{k(i,j)}}\right) \tag{A17}$$

$$\text{s.t.} \quad \sum_{j=1}^{n} z_j \le r \tag{A18}$$

$$z_j \in \{0, 1, \ldots, c_j\}, \; j = 1, \ldots, n \tag{A19}$$

Objective function (A17) maximizes the total expected demand covered accounting for the coverage probabilities $P_{i,k(i,j)}$. Constraint (A18) ensures that at most r EMS units are assigned to open candidate locations, with at most c_j units in location j.

The authors propose the following iterative heuristic:

1) Initialize the vector $\hat{p}_1$ of busy probabilities to an estimated system-wide busy probability ($\hat{p}_{1j} = p, \forall j$).
2) Solve MEXCLP+PR for $\hat{p}_1$, and record the solution as z^*.
3) Compute a new vector of busy probabilities, $\hat{p}_2$, based on z^*.
4) If some convergence criterion is satisfied, stop. Else, replace $\hat{p}_1$ with $\hat{p}_2$ and go to Step 2.

To compute the busy probabilities, the authors generalize (Budge at al, 2007a) and employ an approximation scheme based on the well known hypercube queuing model of Larson (1974, 1975). A detail about the busy probabilities requires attention: The busy probability associated with an EMS station with no allocated EMS units is 1 (and not 0).

Appendix B: Minimizing Average Response Time

The *q*-Median Problem

The *q*-median problem is structurally identical to the MCLP+PR formulation. Using the notation for that formulation, the *q*-median problem can be formulated as:

$$\max \quad \sum_{i=1}^{m} d_i \sum_{j=1}^{n} E[R_{ij}]\, y_{ij} \qquad \text{(B1)}$$

s.t. (A12)-(A16)

When incorporating ambulance unavailability, we use a *locate-allocate* heuristic developed by Jarvis (1975) and further discussed by Larson (1979) to minimize average response time. As described by Jarvis and Larson, the heuristic uses the exact hypercube model to evaluate the average response time. The exact hypercube model is computationally expensive but has the advantage of permitting multiple vehicles per station. The approximate hypercube model developed by Larson (1974) assumes a single vehicle at each station. We used a version of the approximate hypercube model that allows multiple vehicles per station (see Budget et al, 2007b). As presented by Jarvis and Larson, the heuristic can be described as follows:

Initialization: Find an initial solution.

Allocation step: Evaluate the current solution using the hypercube model. This includes computing $f(i, j)$, the fraction of all demand that comes from node j and is served by vehicle i, or in other words, the fraction of patients that come from node j and are *allocated* to vehicle i.

Location step: For each vehicle, pretend that it is possible to move it to any station without changing the $f(i, j)$'s. Move the vehicle to the station that minimizes the average response

time to calls that that vehicle responds to. Note that in this step, multiple vehicles could be moved to the same station.

Convergence check: if the new solution equals the previous solution, stop, otherwise return to the allocation step.

Jarvis and Larson do not specify how to generate an initial solution. We used the following approach to allocate r vehicles to q stations. First, we solve an r-median problem. If $r > q$, then the solution will simply be to place one vehicle at each station. In this case, we reduce each station's capacity by one, and reduce the number of vehicles to be allocated by to $r - q$. Then we solve an $r - q$ median problem to allocate the remaining ambulances. If any ambulances remain, then we repeat the procedure, until all ambulances have been allocated.

We note the following property of the heuristic. Suppose that at some point in the execution of the heuristic, the current solution has more than one vehicle at a particular station. In the location step, all of these vehicles will be seen as having the same $f(i, j)$'s, and therefore, they will either remain at the current station or they will all be moved together to another station. As a consequence, the maximum number of vehicles at a station will never decrease during the algorithm. Possibly as a result of this property, we found that the heuristic almost always cycled, sometimes between 5 solutions. To reduce the impact of such cycling, we used at least 50 iterations when the heuristic did not converge. Furthermore, we kept track of the solution with the lowest average response time throughout the algorithm, and reported this solution at the end, even when the heuristic converged to a different solution.