Operations Research Tools for Addressing Current Challenges in Emergency Medical Services

Shane G. Henderson

March 26, 2009

Abstract

Emergency Medical Service (EMS) providers face a host of significant challenges, including traffic congestion, increasing call volumes, hospital diversion (i.e., emergency departments at hospitals not allowing ambulances to deliver patients to them), and increasing transfer times at emergency departments. Many of these challenges can be attacked through an application of operations research techniques in conjunction with EMS expertise. The purpose of this article is to outline some of the key challenges and potential operations research based remedies, and to discuss in detail one such remedy that has various names including "system-status management" and "move up."

The job of an Emergency Medical Service (EMS) provider is to respond to calls for assistance, render urgent medical care at the scene of a call and then, if necessary, transport the patient to an appropriate hospital. An EMS provider must coordinate the actions of many ambulances, the staff in which may have different levels of training, to address calls of many different types. Call volumes vary significantly according to cyclical patterns on a daily, weekly and annual basis, and calls are certainly not evenly spread over geographic regions.

EMS problems are similar to those faced by other emergency services including fire and police, although there are important differences. For a superb overview of work in these various fields to the early 1990s see [40], and for some recent commentary see [23]. The survey [20] is a valuable reference for both operations research specialists and EMS professionals, although it is written primarily for EMS professionals. EMS problems also have some similarities with the taxi and courier industries, although the latter industries can benefit from a-priori scheduling of a nontrivial fraction of their calls.

The focus in this article is on *operational* issues rather than *clinical* issues. Clinical issues are those issues that are essentially medical in nature, such as the difference that prompt intubation can make to patient survival rates. The clinical side is extremely active. See, for example, the Annals of Emergency Medicine, the Journal of Emergency Medical Services, or Academic Emergency Medicine. These journals also publish occasional articles that focus on operational aspects.

Operations research tools have been used to assist in decision making in EMS at least since the 1960s, e.g., [38]. Typical applications include the siting of new bases and staff scheduling. But there are still a host of challenges faced by EMS providers, some of which are variants of these well-known problems, while others have arisen only recently. The goal of this article is to describe some of these challenges and suggest possible avenues of (operations) research that might help address them. I will particularly emphasize system-status management as a tool for addressing several problems at once. The article is not meant to be a state of the art survey, but rather a "call to arms." To absorb the entire article the reader should have a background in the tools of operations research, say at the undergraduate or masters level, but much of the discussion should be accessible without this background.

This is one of two articles that deal with EMS. In the other article [33] the goal is to provide an overview of next-generation EMS design that helps address the problems faced today. In contrast, I start with the problems of today, indicate potential solution approaches, and go into depth on one approach (system-status management) that is already in use.

The good news for an operations researcher looking at EMS problems is that the area is data rich. Computer-Aided Dispatch (CAD) systems have been in place in many locations for over a decade and so in large centers it is invariably the case that data is available on every call. This is not necessarily the case in small centers, however. The amount of data varies, but there is usually at least three years available as of 2008, and often much more—even decades in some places. Time stamps are logged for each call, including the time the call was received, the time the call was assigned to an ambulance, the time the ambulance was en-route to the call, and so forth. Some of these time stamps are logged manually by ambulance officers, while others are logged automatically by the CAD system, so the quality of the time stamps varies (people occasionally forget). A key component to planning systems is travel-time information, including time-dependent travel times or speeds on road networks. This sort of data is often available from town-planning departments, but is usually intended for long-term planning questions, so often needs editing before it is suitable for EMS applications. A potential source of travel-time data is automatic vehicle location (AVL) data, which is becoming much more widely available as Global Positioning System (GPS) units on ambulances become standard. AVL data gives the location of ambulances as recorded by a GPS unit at regular intervals of time. GPS readings vary in quality depending on how accessible the GPS satellites are. For example, GPS readings often degrade when ambulances are in tunnels or in the urban canyons that are typical of the centers of large cities, so even this data source varies in quality.

The pressing need for efficiencies in EMS systems, together with the availability of large amounts of data, suggests that quantitative operations research methods, together with appropriate statistical methodology, should be well positioned to strengthen an already well-established presence in EMS.

1 An Overview of Emergency Medical Service Challenges

1.1 Measuring Performance

EMS systems are designed to help reduce morbidity and mortality by quickly and safely responding to calls, providing emergency medical assistance at the scene and transporting those patients that require further treatment to an appropriate hospital. As with many decision-making questions in healthcare it is difficult to quantify the medical impact of changes. The traditional approach is to measure performance through response times. The response time for a call is the elapsed time from when the call is received to when an ambulance arrives at the scene. Performance is then measured as the fraction of calls received that have response times of x minutes or less, where x is usually 8 or so; see [17] for historical notes and discussion from a practical perspective. Let us call this performance measure the Response Time Threshold Fraction (RTTF). RTTFs are often reported separately for different regions and different time intervals, so that a picture of performance in space and time can be obtained.

The RTTF is easily explained, easily measured and unambiguous. It can also be computed using existing methodologies including simulation and hypercube methods. (See [25] for an overview of simulation in EMS and [29, 30, 12] for how to do this using the hypercube method.) These are important advantages that help to explain why RTTF is so prevalent in the industry. RTTF has its origins in cardiac survival studies, where survival rates are linked to response times. But it is also somewhat arbitrary. The threshold of 8 minutes is chosen from the survival probability

versus response time curve, and is often adjusted depending on whether one is in a rural setting (larger values are used) or a metropolitan area. It is also highly city dependent, with values 8, 9, 10 and 12 being common choices. Different thresholds are used for different types of calls, since some calls have lower priorities than others, so naturally receive slower response times.

For all of these reasons, there is great interest in finding better performance measures than RTTF. One approach is to use a utility function u so that a call with response time t receives a utility or "score" $u(t)$, and the goal is to maximize the average utility over all calls. RTTF is actually a special case where u only takes values 0 and 1. The utility approach is a central theme in both [33] and [16], where the utility function adopted is again based on cardiac studies. Of course there are other types of calls including, e.g., trauma calls, so interpretation of the utility can be difficult, even when it is specialized to different types of calls. As another example of the difficulty in coming to agreement on utilities, many cardiac patients wait for some time before calling EMS so a few minutes difference in response time is somewhat immaterial when there has already been a delay of up to a few hours. Nevertheless, the idea of attaching utilities to calls and measuring average utility is important. For example, a key observation in [16] is that solutions to ambulance deployment problems obtained using RTTF can perform very poorly when measured in terms of average utility.

Perhaps we should continue to err on the side of simplicity and simply measure performance in terms of *average* response times instead of RTTFs or more complex measures? See [33] for some discussion of this question.

Some EMS providers additionally measure performance in other ways. For example, one approach to ambulance deployment used in Edmonton, Alberta is to specify, for each number of available ambulances, a set of locations that they should occupy. When ambulances are appropriately positioned, the service is said to be "in compliance," and performance is measured by the fraction of time the system is in compliance. One would expect that RTTFs are also tracked.

1.2 Delays at, and Diversion from, Emergency Departments

The time required to transfer a patient from an ambulance to an accepting emergency department varies greatly depending on the medical needs and priority of a patient and how busy the emergency department is. Typical desired transfer times are on the order of 15 minutes, but in many cases can be as large as 1 hour or more. This places tremendous pressure on EMS organizations because it dramatically increases—even doubles—the time required to handle each call. Another serious problem is known as "ambulance diversion," which occurs when an emergency department notifies dispatchers that it will temporarily not accept patients from ambulances. The resulting patient transport times are increased because ambulances are forced to go elsewhere, again impacting ambulance utilization.

Both transfer delays and diversion can have serious medical consequences. A statistical study [41] links ambulance diversion to increased death rates from heart attacks. The problem may be even more serious than this paper suggests, since transfer delays can be much larger than the additional driving time associated with diversion.

What can be done to reduce the impact of delays and diversion?

One approach that has been tried is to place an EMS paramedic at each receiving hospital. The paramedic receives patients from arriving ambulances and subsequently transfers those patients to the emergency department when possible. When not busy with this function the paramedic can assist in the emergency department. One can think of the paramedic as a buffer (in the sense of manufacturing) between arriving ambulances and the emergency department. In one case, this was observed to work well for about 1 week, until the hospital realized it could use the EMS paramedic for its own ends, and patient transfer delays returned.

This points to a key problem in that hospitals and EMS systems are usually operated as separate organizations. The incentives that are in place for both organizations are not geared towards inducing cooperative behaviour in the emergency department where these organizations interface.

There are at least three possible solutions to this problem.

The first, and most obvious, potential solution is to immediately conclude that hospitals need more resources to cope with the flows of patients entering their emergency departments. But as noted in several places, e.g., [21, 22], any problem may lie downstream from the emergency department, in hospital wards that need to free capacity to allow emergency-department patients to be transferred. It may also be the case that the systems in place in hospitals can be streamlined to make it easier for emergency-department staff to keep pace with patient flows and reduce the need for ambulance delays and diversion. This is a key place where operations research can help.

The second potential solution is to adopt some form of "regionalization," as is currently under way in a number of provinces in Canada to varying degrees. Here the hospital and pre-hospital organizations are placed under a single umbrella organization, with the goal that the pre-hospital and hospital organization are now a single organization and therefore will "pull in the same direction." It is not yet clear whether this step will help though, because the size of the pre-hospital and hospital organizations makes it difficult to maintain a "systems view" of operations, and it is very likely that a hierarchical organization will result, with EMS on one branch and emergency departments on another, perhaps leading to the same situation we have now. This is not a foregone conclusion of course, but it is one possible outcome.

The third potential solution is to redesign the incentive schemes for both EMS and emergency departments, i.e., change the performance measures that are used to report their performance, along with how those performance measures are written into contracts. Perhaps by aligning the incentive schemes using the game-theoretic field of mechanism design, one can design contracts that lead to outcomes that are simultaneously of high quality for patients, EMS providers and emergency departments.

1.3 Increasing Traffic Congestion

The high-profile aspect of traffic congestion is large-scale traffic jams. But traffic congestion can also mean large traffic flows on a daily basis, and therefore slow travel on routes that were never intended to handle the volumes that they receive. Even when ambulances travel with lights and sirens, it can take time for drivers to react and clear a path for an ambulance. Furthermore, ambulances spend significantly more time traveling at regular traffic speeds than at lights and sirens speeds, in which case they are slowed by traffic congestion, thereby increasing the load on ambulances for the same call volume. This, in turn, translates into increased response times due to unavailability of the closest appropriate ambulance.

So what can EMS providers do about this problem?

Rather than maintain ambulance bases with 3 or more ambulances at a base, many EMS providers now spread their available ambulances around the city. This is sometimes known as "parking on street corners." Crews are asked to sit in their ambulances in parking areas, waiting for calls. The particular locations are chosen depending on time-dependent demand and travel speeds, and various methods are available for such planning, e.g., [35]. This has the effect of reducing the driving distance to calls, and therefore can improve response times. It also creates the need for real-time ambulance location and relocation. To see why, picture a city with evenly spread ambulances. When a call comes in, one of those ambulances "disappears" from the picture, leaving a potentially large "gap" in coverage. It may then be advantageous to move one or more nearby ambulances to reduce the size of this gap. This process of using real-time

information to relocate available ambulances is known as "system-status management," and we will discuss it in depth later in this article. It is worth noting that under a system where multiple ambulances are stationed at a single base, only occasionally are all ambulances at the base busy with calls, so gaps in coverage do not open up as frequently. However, due to the clumping of ambulances at bases, the distance between available ambulances is typically large, and this translates into large response times.

The use of system-status management methods makes having accurate travel-time or speed information more important than ever. As mentioned earlier, currently available speed information on road networks is usually not very accurate, so one would like to "tune" the speeds, probably using AVL data. But how should this be done, given that the data consists of noisy observations of ambulance locations at discrete points in time, so one is uncertain even about the exact route that an ambulance has taken? The field of map matching (see, e.g., [31, 28]) is undoubtedly an important tool that can perhaps be combined with carefully designed statistical methods to update modeled travel speeds. See [13] for a recent statistical study of travel times that models travel times as random, and [24, 11] for the consequences of travel times being random.

1.4 Vehicle Mix Questions

Should one's ambulance fleet consist of advanced life support (ALS) vehicles only, or a mix of ALS and basic life support (BLS) vehicles? To understand the distinction, note that ambulance staff can have a variety of levels of training. *Ambulance officers* (also called Emergency Medical Technicians) typically have on the order of several weeks of training and are highly adept first responders. However, they cannot deliver the higher level of care needed on a significant fraction of calls that a *paramedic* can provide (e.g., administer drugs) due to a much higher level of training. ALS vehicles have a paramedic on board, while BLS vehicles do not.

Most cities have a mix of ALS and BLS vehicles, although some authors strongly advocate an ALS-only system, e.g., [3]. It is cheaper to run a BLS ambulance than an ALS ambulance, so for a given budget one can operate more vehicles in a mixed fleet than in an ALS-only fleet. With a mixed fleet one sends BLS vehicles to calls where paramedic-level attention is not needed. However, this requires that dispatchers make a determination at the time of receiving a call about which type of vehicle to send, thereby inflating the time required before dispatching a vehicle, and creating the possibility of sending the wrong vehicle, leading to additional delays before a high level of care can be provided. The tradeoffs between ALS-only fleets or a mixed fleet are therefore complex, and it is not clear which approach is better in a given situation. Operations research tools can help quantify the tradeoff to assist in making a determination. For example, simulation was used in [25] to model the tradeoffs between sending a vehicle to a call based on minimal information from a caller, as opposed to obtaining more information from the caller before dispatching (ProQA). Using ProQA ensures a better match of vehicle type to the call, but slows down dispatch times, thereby inflating response times.

The ALS/BLS mix is just one example of a number of vehicle mix questions. For example, some EMS organizations use vehicles staffed by a single paramedic that can rapidly respond to calls but cannot transport patients. How many such vehicles should be employed? Should scheduled patient transports be performed by emergency vehicles or by a fleet dedicated to the task? And to what extent should EMS organizations employ helicopters?

1.5 Increasing Call Volumes

Virtually all EMS providers are seeing their call volumes increase over time. This increase is probably a consequence of many factors, including population increase, demographic changes (which are strong predictors of call volumes; see, e.g., [1, 39]), and what the general public perceives as the threshold for calling an ambulance. This obviously poses challenges for an EMS provider whose budget is often constrained.

Using careful dispatch strategies can help. For example, it has been known for some time [14] that sending the closest available vehicle is not necessarily optimal, because it can lead to imbalances in workload and therefore inflated response times. Nevertheless, most dispatchers follow either exactly this policy or one very close to it, partly because of the danger of litigation in the event of an unfortunate outcome, but also partly because there is a lack of decision support systems that can help with this decision. What do optimal dispatching policies look like and how can they be deployed in a dispatcher-friendly manner?

Another question we must ask ourselves is whether all of the incoming calls actually *need* EMS response. During the September 11th attacks in New York city, some emergency departments noted a considerable *drop* in patient arrivals that was not due to a lack of ambulances. It appears that the general public felt that their problems were outweighed by the needs of potential casualties from the World Trade Center and did not call for help. In many cases, no doubt, they should still have called for help, but it does make one wonder how much EMS and emergency room demand could be mitigated using other strategies. Is there a way to shape demand, perhaps through public education schemes? Would such schemes lead to better outcomes in the sense of overall public health, or would such schemes reduce EMS call volumes at the expense of public health?

Yet another approach to dealing with increasing call volumes is to attempt to obtain better performance from the resources at one's disposal. One such method that can help immediately, and that is already in use, is system status management, and this is the subject of the remainder of this article.

2 Existing Approaches to System-Status Management

Many ambulance organizations are moving away from the model where ambulances are stationed in twos and threes at bases, instead preferring to spread the vehicles out over the city in an attempt to reduce the travel distance to calls. When a call is received and an ambulance is dispatched to the call, a "hole" is left in the coverage of the city. Should the nearby vehicles be rearranged? If so, how? Any method that gives a strategy for performing such rearrangements is known as system-status management (SSM), or by one of the synonyms move up, redeployment, dynamic relocation, or dynamic positioning.

In order to implement SSM, dispatchers need to be aware of the location and status of the ambulance fleet. This is not difficult for EMS organizations that have installed GPS units on all vehicles and where crews send updates on their status to the dispatch center, and indeed, virtually all large EMS organizations have adopted these practices.

Another key issue is ambulance crew receptiveness to SSM. Many implementations encounter some resistance from crews and perhaps with good reason. Ambulance crews have a stressful job that also requires them to work in shifts. On top of that, a poorly implemented SSM approach may lead to crews spending almost the entirety of their shift driving between locations to fill holes in coverage, or parked in some undesirable location that is far less comfortable than any ambulance base. The crews do not see the "birds-eye view" that leads to these redeployments,

and so the trips can seem pointless to them. However, if implemented with care, SSM can obtain material improvements in response time with only modest disruption to crews.

We now describe specific methods for implementing SSM.

2.1 Manual Search and Selection

Suppose a dispatcher is considering redeploying a specific ambulance to one of several potential locations. Experience often guides the selection of a specific location. What-if tools can also help. For example, one approach is to assume that all other vehicles will not change status or location in the short term, and then consider the candidate locations, one at a time. As long as the number of candidate locations is small, it is typically a simple matter to recompute some measure of quality (usually "coverage," where coverage means the fraction of future demand that can be reached within some time threshold from the candidate locations) at each candidate location and select the one that is best. Graphical tools that display coverage, e.g., by colouring locations in a map according to projected response times, can also help. This is the basic idea behind simple implementations of SSM. Such implementations are easy to understand by dispatchers, but they also have some important shortcomings.

1. It can be difficult to gain a picture of both call density by location and coverage in graphical displays. Therefore, dispatchers need experience to help in determining which vehicle relocations might prove effective.

2. It is difficult to consider all possibilities when simultaneously considering moving two or more vehicles because of the combinatorial increase in the number of candidate locations that need to be considered.

3. The status and location of ambulances is constantly in flux, so decisions based on such static models can miss the fact that the situation may look very different by the time any relocations have been completed.

4. In busy systems, it is quite possible that vehicles on their way to a new location may be dispatched to a call before they reach their intended destinations, calling into question the basis for the relocation decision.

5. It is difficult to get a sense of which relocations are the most important ones in terms of reducing response times, which is important when one wishes to limit relocations to avoid frustrating crews.

2.2 Real-Time Optimization

Several of the disadvantages of the real-time manual approach discussed above motivate an automated approach. All such searches rate candidate solutions using some measure of quality that is usually coverage, as defined above, although other measures have been used, e.g., [2]. Exhaustive searches can work quite well when the number of potential relocations is small, but they can quickly become overwhelmed when multiple ambulance moves are simultaneously considered. It is then that optimization-based methods become the preferred approach.

The first optimization-based method was developed by Kolesar and Walker in the context of fire-fighting operations [27]. It involved solving, in real time, three integer programs using heuristics. Unfortunately, one cannot just "cut and paste" this approach into the EMS setting, because there are some important differences in the way fire services and EMS services operate. The key differences are that (1) fire companies have much lower utilization than EMS (to ensure

that they can respond rapidly to major incidents), and (2) fire companies can be engaged at a structural fire for 8 hours or more, whereas it is unusual for an ambulance to be engaged in a single call for more than 2 hours. These differences mean that it is reasonable to assume that the dominant system conditions will not change during a fire company deployment (call) while, as we argue below, in the EMS setting such an assumption is problematic.

A real-time optimization approach was developed in [18]. The integer programming formulation ensures appropriate coverage of various locations, that vehicles are not relocated too often, and that vehicles do not take too many "round trips." A tabu-search heuristic on a parallel computing platform was used to solve the model. Similar formulations have been adopted in some commercial systems.

2.3 Offline Optimization

An alternative to the methods above is to develop a lookup table which gives, for each number of available ambulances, the set of locations where available ambulances should be positioned. Dispatchers then attempt to direct the ambulances to occupy those locations with as little disruption to crews as possible. Lookup tables can be time-dependent to reflect time-dependent travel speeds and demand patterns, and are compact and easily understood. But how can they be constructed?

Here one can exploit the large amount of available research on identifying optimal base locations in the *static* version of ambulance deployment where ambulances always return to their pre-assigned bases when free. Integer-programming methods (see [10] for a recent survey) are available for selecting ambulance locations for a given number of ambulances. One can solve one such optimization problem for each number of available ambulances. A difficulty here is that there is no attempt to ensure consistency between the set of locations chosen for n available ambulances, and those chosen for $n+1$ available ambulances. This can lead to many redeployments that should be avoidable. A partial solution is to require that the redeployment locations be *nested* in the sense that the set of locations chosen for n ambulances be a subset of those chosen for $n+1$ ambulances, for each $n = 1, 2, N-1$, where N is the total number of ambulances in the fleet [19]. These nesting constraints can be satisfied by solving a single large integer program that simultaneously finds the optimal locations for all $n = 1, 2, \ldots, N-1$, and includes explicit constraints that enforce the nesting.

This method is a highly appealing and important option for SSM implementation, but it does possess some disadvantages. In particular, Issues 3, 4 and 5 identified as difficulties for manual selection are also important difficulties for this approach.

2.4 Stochastic Dynamic Programming

The most important issues with the methods we have discussed thus far are that (a) they are based on the assumption that vehicles are exactly at the desired locations at all times, and (b) they cannot give a clear sense of the importance of particular redeployments. Both of these issues can be addressed using stochastic dynamic programming (DP) models.

DP models were first introduced for this problem in work by Berman and co-authors in [4, 5, 6, 7] and [8]. DP is certainly a very natural framework to employ, since it is the study of problems where decisions must be made over time in the presence of uncertainty, which exactly describes our problem. To apply DP, one tracks the locations and status of all ambulances over time. At any decision point, one defines a set of actions (potential redeployment decisions) that might be taken. The actions are of the form "send ambulance x to location y." The DP is designed to discover a *value function* V that gives the quality of a particular configuration of ambulances at

a particular time. The function V maps the system states (ambulance locations and status, and time) to a measure of the quality of the configuration, which can be interpreted as the expected performance of the system (e.g., number of calls answered on time) starting from that particular state. Once one obtains the value function V^* associated with an optimal policy, it can be used to determine the optimal decision at any stage by implementing the associated greedy policy: At each decision point, one selects the action that maximizes the expected value of the immediate reward plus "downstream" rewards. See, e.g., [34] for a comprehensive introduction to DP and full discussion of these concepts.

Recently, Zhang, Mason and Philpott [42] have revisited the use of DP for ambulance relocation, establishing a number of results that help build intuition for what optimal policies look like. Their results show that one should send vehicles to one of several distinguished locations, where the set of locations depends on the arrival rate of calls. Furthermore, they show numerically that the heavier the load of calls on the system, the less effective SSM can be relative to a static policy.

In the papers mentioned above, the approach is to obtain the optimal value function V^* by storing a value for *every possible state* and applying standard DP techniques. This necessitates keeping the number of states small, which means that only very small examples can be treated. Therefore the models addressed in this work are somewhat stylized. The results are important in that they help build our understanding of when SSM can be effective and what effective policies look like, but the approach cannot be applied directly to realistic-sized systems. To enable the leap to realistic-sized systems a different approach to representing the value function is needed, which brings us to the subject of approximate dynamic programming.

3 Approximate Dynamic Programming for System-Status Management

Approximate dynamic programming (ADP) is a general-purpose suite of tools for solving problems where successive decisions must be made over time under uncertainty. Essentially, ADP involves an implementation of dynamic programming where one stores an *approximation* for the value function, instead of the value function itself. This allows one to tackle problems that would be impossible with exact DP owing to the huge (often infinite) state space involved. The catch is that one cannot, in general, expect ADP to find an *optimal* policy. The goal, instead, is to find a *high-quality* policy in a computationally tractable manner.

I will sketch the key ideas behind an ADP approach to SSM developed in [36] and further explored and improved in [32]. I will refer to this approach as *our* approach because I was one of the authors of this work. The goal is to minimize the fraction of calls received that have a response time that exceeds a given threshold. We refer to such calls as *lost* or *missed*. The specific version of ADP we use is an adaptation of a generic approach for ADP that was initially described in [9] and involves a mix of simulation and regression.

Let V denote an approximation to the optimal value function. I will explain where the approximation comes from shortly, but for now assume that it is given. Associated with V is a policy, known as the greedy policy with respect to V, that works as follows. At each time point t at which a decision is required, let s denote the current state of the real system, which is essentially all of the information that dispatchers have at their disposal at time t. Therefore, s includes information on the location and status of all ambulances, including the time that each ambulance has been at its location at time t. It also includes information on any queued calls. It does not include information on future events such as might be obtained from an event list in a simulation, e.g., the time remaining until an ambulance completes a hospital transfer in

progress. One then considers each possible *action* in turn, where each action corresponds to a request for one or more ambulances to redeploy to specific locations. As a consequence of each action, the specified ambulances will begin to deploy to the new locations, but do not necessarily reach those locations before the situation changes. For each action a, we let $S(a)$ denote the random system state corresponding to the new situation at the time of the next system event, which could be an ambulance arriving at the scene of a call, a new call arriving, etc. Also, let $C(a)$ be the immediate cost incurred by this action, which in our case is either 0 (no missed calls) or 1 (one missed call). We then compute or approximate $E[C(a) + V(S(a))]$, and choose the action that minimizes this quantity as our decision at time t. (Notice that the expectation is computed *conditional* on the system state at time t.) We then continue the simulation to the next decision point, at which time the process above is repeated.

We cannot compute the required conditional expectation analytically, and instead use a small number of simulations to obtain a Monte Carlo estimate of it for each action that is considered. We refer to this calculation as a *micro simulation*, because each replication of a micro simulation involves simulating from the current system state forward in time for at most a few minutes of simulated time, and this can be accomplished with a very small amount of computation. In our implementation, the micro simulations for all potential actions take a fraction of a second. Furthermore, these micro simulations are only necessary when we make a relocation decision, and then only from the current state of the system, so they represent a minor computational overhead. The Monte Carlo error we incur means that we will occasionally take an action that does not minimize the true expectation. However, this will typically occur when the suboptimal action we choose has a value that is close to that of the optimal action, so we do not expect that this effect will cause practical problems, provided that the number of micro simulations is reasonable.

To summarize thus far, for a given function V there is an associated greedy policy that we can approximate using micro simulations, and this can be implemented in real time provided that one has a simulation model of the system along with the ability to observe the state of the system in real time.

So where does the function V come from? We use a linear combination of pre-specified basis functions $V_1, V_2, \ldots, V_n$, i.e., for each state s,

$$V(s) = a_1 V_1(s) + a_2 V_2(s) + \cdots + a_n V_n(s),$$

for some coefficients $a_1, a_2, \ldots, a_n$. The basis functions are chosen heuristically to reflect characteristics or features of the state s that are thought to be important in determining a policy. In the present case, after experimentation with potentially useful functions we settled on $n = 6$ functions as follows.

V_1 A constant.

V_2 The number of queued calls that will definitely not be reached within the time threshold.

V_3 The arrival rate of calls to areas that cannot be reached by any available ambulance.

V_4 The arrival rate of calls to areas that can be reached by an available ambulance but will likely be missed because of congestion. Here we measure the loss rate of calls as the product of two terms, the first being the arrival rate, and the second being the loss probability as measured by the Erlang loss function.

V_5 A version of V_3 where the ambulances are assumed to have reached their destinations if enroute somewhere, and where presently stationary ambulances remain where they are.

V_6 A version of V_4 that is modified as in V_5 above.

Why did we choose these functions? The Erlang loss function was helpful in choosing static locations for ambulances [37] so it seems reasonable to include it, and SSM is designed to help in the short run—say over the next hour or so—so using versions of the basis functions associated with predicted ambulance locations on that time scale seems reasonable. But there is plenty of room for argument and art in this selection, and we do not claim that this set of functions is the best possible. Much more remains to be done in basis function selection.

Given a set of basis functions, the problem then reduces to determining the coefficients a_i, $i = 1, 2, \ldots, n$. There are a variety of methods for doing this. We use regression as follows. The state s also includes the time t at which the state is observed. The function value $V(s)$ then represents the expected cost (number of missed calls) observed starting from state s at time t through to the end of the simulation horizon. (We use two weeks for the simulation horizon.) In implementing this algorithm we have found it necessary to *discount* the costs in time. This is purely a computational device that helps ensure convergence; see [32] for more discussion on why this is needed and why it works. (A similar trick was employed in a different setting in [26].) So the function value $V(s)$ actually represents the expected *discounted* number of missed calls from the time t until the end of the simulation horizon. The simulation provides *observations* of the discounted number of calls till the end of the horizon. So we can use linear regression to try to align $V(s)$ with these observations.

The result of the linear regression is a set of coefficients $a_1, a_2, \ldots, a_n$ which determine a new value function $\tilde{V}$. We can then iterate this process, with the next iteration simulating the greedy policy with respect to $\tilde{V}$ instead of V. After a modest number of iterations (usually less than 20 or so) we terminate the process, and select what appears to be the best policy (i.e., set of coefficients) seen thus far.

Figure 1 gives an example of the results of such a training session for data from Edmonton, Alberta. The data were modified prior to use to protect the confidentiality of Edmonton EMS performance. Here the vertical axis plots the estimated *undiscounted* percentage of missed calls. The apparent-best policy arises after three iterations. There are two reasons why this is the *apparent* rather than necessarily the *actual* best policy. First, at each point we run a simulation to estimate the performance of the given policy. Therefore, there is simulation error in the estimates of performance. Second, even though each point gives an unbiased estimate of the performance of the corresponding policy, when we focus in on the point with the lowest function value in the plot, we are biased towards observing a point where the estimated cost is lower than the true cost, i.e., the estimated cost associated with the best observed policy is biased low. If we perform an independent simulation of the estimated-best policy we obtain an unbiased estimate of its performance, which in the case of the observed best policy in Figure 1 gave an estimated performance of 25.4 ± 0.1 percent. By way of comparison, the static policy (where ambulances return to pre-assigned bases after completing each call) misses 29.5 ± 0.1 percent of calls.

The results of Figure 1 are based on policies that only consider an ambulance for redeployment when it is completing a call either at scene or at a hospital. This is appealing because it ensures that crews are not overly disrupted by excessive redeployments. However, it is possible to obtain even greater improvements in performance by considering additional redeployments. In [32] approximately an additional 3% of calls are not missed over and above the completing-calls-only redeployment policy when additional redeployments are considered at regular intervals.

One might reasonably ask whether ADP identifies the optimal choice of coefficients $a_1, a_2, \ldots, a_n$. As mentioned earlier, the algorithm discussed above is not designed to converge to any particular set of coefficients. This is because the computational burden in training is great, so we cannot simulate a large number of policies, which is essential if we wish to design a converging algorithm.

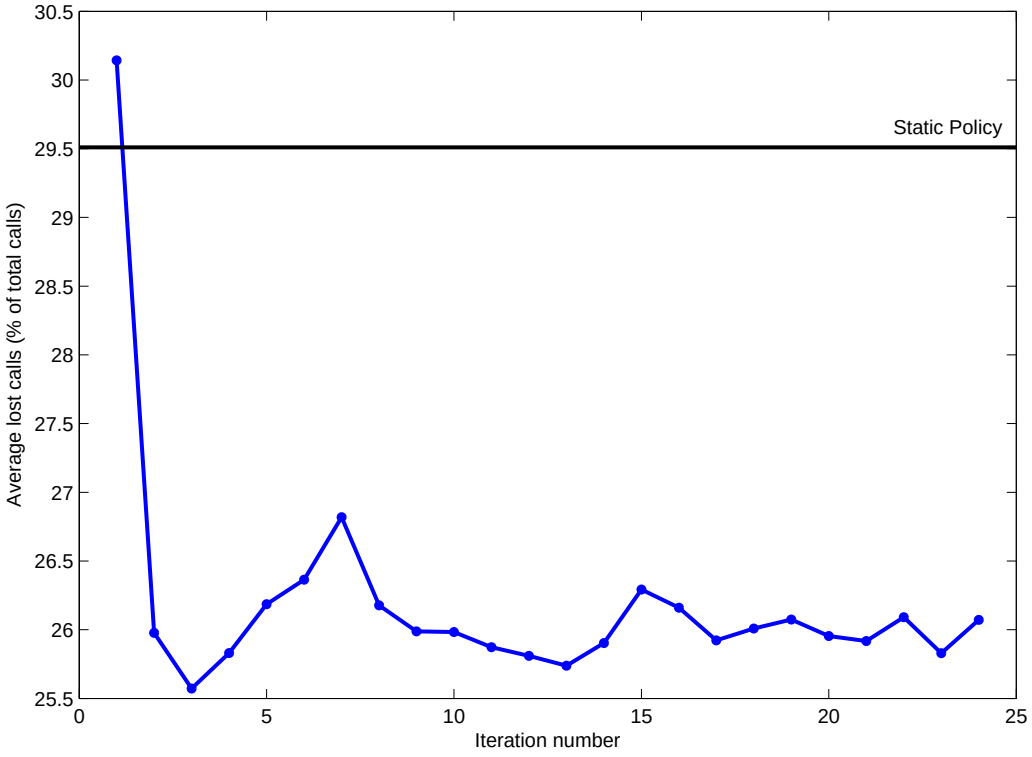

Figure 1: An example of the training process of ADP using representative data from Edmonton, Alberta.

Instead we focus on exploring the coefficient space for a highly effective policy.

The natural question to ask then is how the policy identified by ADP ranks amongst the potential policies. In Figure 2 we give a histogram of the performances of the top-performing policies amongst 1000 policies. These 1000 policies were selected uniformly at random from a cube surrounding the final recommended coefficient vector a with side lengths chosen to be commensurate with the variation in the coefficients seen during the course of the ADP iterations. We see that the performance of the optimal ADP policy is very close to the best policy seen. It may even be optimal, because we again obtain bias when we focus in on the policies appearing at the extreme left of the histogram. There are no theoretical guarantees that the ADP policy would do so well, and so it may be necessary to apply a separate optimization procedure after the ADP search to refine the coefficients. We are considering this step in current research.

4 Conclusion

EMS providers face significant challenges that, in the long run, could be addressed by a variety of remedies that involve operations research methodologies. We have suggested some such remedies, but they are a long way from being developed, much less implemented. System-status management can help essentially immediately with many of the difficulties faced by EMS providers, but needs to be implemented carefully to fully realize the potential benefits in response times while not unduly inconveniencing crews. Existing implementations vary in quality, which has led to the technique receiving a bad name with crews.

Offline and online optimization implementations of system-status management probably rep-

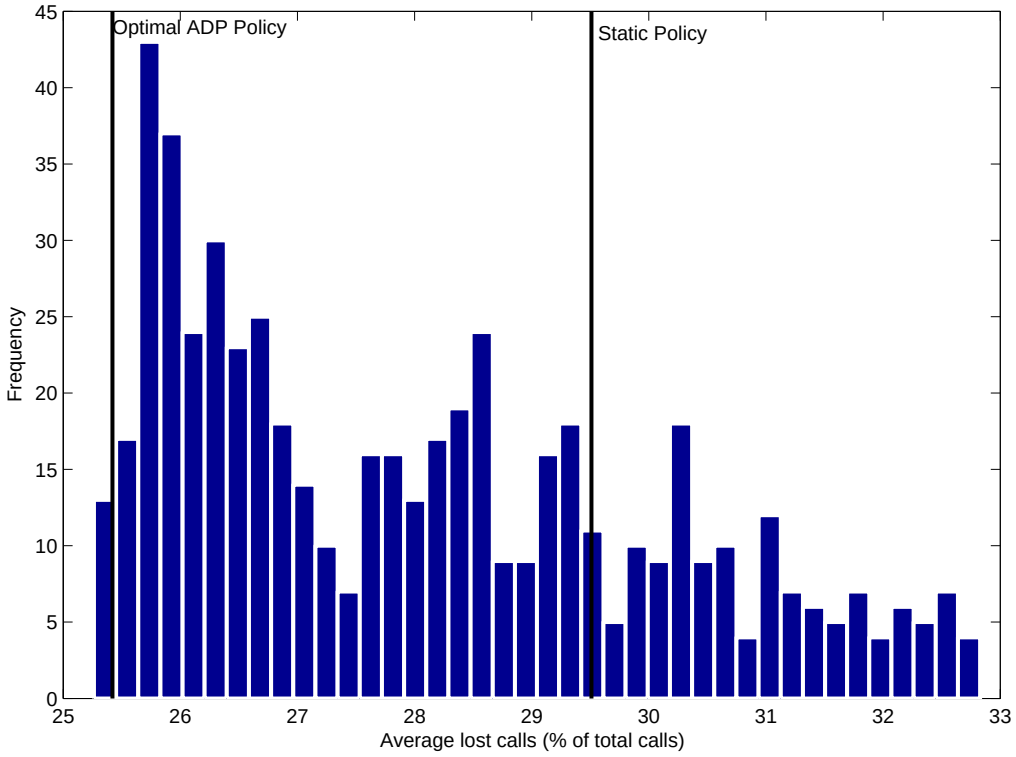

Figure 2: A histogram showing the estimated performance of a number of policies using representative data from Edmonton, Alberta.

resent the current "best practice" in the area, with offline optimization perhaps being more transparent and accepted by dispatchers than the online approach. We have explained in broad terms how these methods work. We have also sketched the main ideas behind an approach to system-status management based on Approximate Dynamic Programming. This new approach is already competitive with the existing optimization-based methods, and has a number of advantages over the existing methods. It is not quite ready for commercial implementation, but is certainly close to that point.

All of the methods for system-status management rely on quality estimates of call arrival rates broken down by time and location. Our ability to obtain quality estimates of these quantities lags our ability to exploit those estimates in operations research models. So while there are many potential research directions described in this article, perhaps the most pressing one is the need to expand existing efforts, e.g., [15], in statistical analysis of EMS data.

Acknowledgments

I thank the editors and referees for very helpful comments that improved this article. This article is based upon work supported by the National Science Foundation under the grants DMI 0400287 and CMMI 0758441. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

[1] C. A. Aldrich, J. C. Hisserich, and L. B. Lave. An analysis of the demand for emergency ambulance service in an urban area. *American Journal of Public Health*, 61(6):1156–1169, 1971.

[2] T. Andersson and P. Värband. Decision support tools for ambulance dispatch and relocation. *Journal of the Operational Research Society*, 58:195–201, 2007.

[3] T. Balaker and A. B. Summers. Emergency medical services privatization: Frequently asked questions. Available online at http://www.reason.org/policystudiesbysubject.shtml#priv_govreform, 2003.

[4] O. Berman. Dynamic repositioning of indistinguishable service units on transportation networks. *Transportation Science*, 15:115–136, 1981.

[5] O. Berman. Repositioning of 2 distinguishable service vehicles on networks. *IEEE Transactions on Systems Man and Cybernetics*, 11:187–193, 1981.

[6] O. Berman. Repositioning of distinguishable urban service units on networks. *Computers & Operations Research*, 8:105–118, 1981.

[7] O. Berman and B. LeBlanc. Location-relocation of mobile facilities on a stochastic network. *Transportation Science*, 18:315–330, 1984.

[8] O. Berman and M. R. Rahnama. Optimal location-relocation decisions on stochastic networks. *Transportation Science*, 19:203–221, 1985.

[9] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1997.

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

[11] S. Budge, A. Ingolfsson, and E. Erkut. Optimal ambulance location with random delays and travel times. *Health Care Management Science*, 11:262–274, 2008.

[12] S. Budge, A. Ingolfsson, and E. Erkut. Approximating vehicle dispatch probabilities for emergency service systems with location-specific service times and multiple units per location. *Operations Research*, 57(1):251–255, 2009.

[13] S. Budge, A. Ingolfsson, and D. Zerom. Empirical analysis of ambulance travel times: the case of Calgary emergency medical services. Submitted for publication, 2008.

[14] G. M. Carter, J. M. Chaiken, and E. Ignall. Response areas for two emergency units. *Operations Research*, 20(3):571–594, 1972.

[15] N. Channouf, P. L'Ecuyer, A. Ingolfsson, and A. N. Avramidis. The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta. *Health Care Management Science*, 10:25–45, 2007.

[16] E. Erkut, A. Ingolfsson, and G. Erdoğan. Ambulance deployment for maximum survival. *Naval Research Logistics*, 55:42–58, 2008.

[17] J. Fitch. Response times: Myths, measurement & management. *Journal of Emergency Medical Services*, 30(9):46–56, 2005.

[18] M. Gendreau, G. Laporte, and F. Semet. A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Computing*, 27:1641–1653, 2001.

[19] M. Gendreau, G. Laporte, and F. Semet. The maximal expected coverage relocation problem for emergency vehicles. *Journal of the Operational Research Society*, 57:22–28, 2006.

[20] J. B. Goldberg. Operations research models for the deployment of emergency services vehicles. *EMS Management Journal*, 1:20–39, 2004.

[21] L. V. Green. Capacity planning and management in hospitals. In M. L. Brandeau, F. Sainfort, and W. P. Pierskalla, editors, *Operations Research and Health Care: A Handbook of Methods and Applications*, pages 15–41. Kluwer, Boston, 2004.

[22] L. V. Green. Using Operations Research to reduce delays for healthcare. In Z.-L. Chen and S. Raghavan, editors, *Tutorials in Operations Research*, P. Gray, Series Editor. INFORMS, 2008.

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

[24] S. G. Henderson. Should we model dependence and nonstationarity, and if so how? In M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, editors, *Proceedings of the 2005 Winter Simulation Conference*, pages 120–129, Piscataway NJ, 2005. IEEE.

[25] S. G. Henderson and A. J. Mason. Ambulance service planning: simulation and data visualization. In M. L. Brandeau, F. Sainfort, and W. P. Pierskalla, editors, *Operations Research and Health Care: A Handbook of Methods and Applications*, pages 77–102. Kluwer Academic, Boston, 2004.

[26] S. G. Henderson, S. P. Meyn, and V. Tadić. Performance evaluation and policy selection in multiclass networks. *Discrete Event Dynamic Systems*, 13:149–189, 2003. Special issue on learning and optimization methods.

[27] P. Kolesar and W. E. Walker. An algorithm for the dynamic relocation of fire companies. *Operations Research*, 22:249–274, 1974.

[28] J. Krumm, J. Letchner, and E. Horvitz. Map matching with travel time constraints. In *Society of Automotive Engineers (SAE) 2007 World Congress*. SAE International, April 2007. Paper 2007-01-1102.

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

[30] R. C. Larson. Approximating the performance of urban emergency service systems. *Operations Research*, 23:845–868, 1975.

[31] A. J. Mason and S. G. Henderson. Analysing ambulance travel: A dynamic program for map matching with sparse GPS data. Working Paper, 2007.

[32] M. S. Maxwell, M. Restrepo, S. G. Henderson, and H. Topaloglu. Approximate dynamic programming-based ambulance redeployment. Submitted for publication, 2009.

[33] L. A. McLay. Emergency medical service systems that improve patient survivability. In J. J. Cochran, editor, *Wiley Encyclopedia of Operations Research and Management Science*. Wiley, 2009.

[34] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, New York, NY, 1994.

[35] H. K. Rajagopalan, C. Saydam, and J. Xiao. A multiperiod set covering location model for dynamic redeployment of ambulances. *Computers & Operations Research*, 35:814–826, 2008.

[36] M. Restrepo. *Computational methods for static allocation and real-time redeployment of ambulances*. PhD thesis, Cornell University, Ithaca NY, USA, 2008.

[37] M. Restrepo, S. G. Henderson, and H. Topaloglu. Erlang loss models for the static deployment of ambulances. *Health Care Management Science*, 2008. To appear.

[38] E. S. Savas. Simulation and cost-effectiveness analysis of New York's emergency ambulance service. *Management Science*, 15:B608–B627, 1969.

[39] H. Setzler, C. Saydam, and S. Park. EMS call volume predictions: A comparative study. *Computers & Operations Research*, 36(6):1843–1851, 2009.

[40] A. J. Swersey. The deployment of police, fire, and emergency medical units. In S. M. Pollock, M. H. Rothkopf, and A. Barnett, editors, *Operations Research and the Public Sector*. North Holland, Amsterdam, 1994.

[41] N. Yankovic, S. Glied, M. Grams, and L. V. Green. Ambulance diversion and myocardial infarction mortality. Submitted for publication, 2009.

[42] O. Zhang, A. J. Mason, and A. B. Philpott. Simulation and optimisation for ambulance logistics and relocation. Presentation at the INFORMS 2008 Conference, 2008.