Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

The Vehicle Mix Decision in Emergency Medical Service Systems

Kenneth C. Chong, Shane G. Henderson, Mark E. Lewis

To cite this article:
Kenneth C. Chong, Shane G. Henderson, Mark E. Lewis (2016) The Vehicle Mix Decision in Emergency Medical Service Systems. Manufacturing & Service Operations Management 18(3):347-360. http://dx.doi.org/10.1287/msom.2015.0555

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 18, No. 3, Summer 2016, pp. 347–360
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs

http://dx.doi.org/10.1287/msom.2015.0555

The Vehicle Mix Decision in Emergency Medical Service Systems

Kenneth C. Chong, Shane G. Henderson, Mark E. Lewis
School of Operations Research and Information Engineering, Cornell University, Ithaca, New York 14853
{kcc66@cornell.edu, sgh9@cornell.edu, mel47@cornell.edu}

We consider the problem of selecting the number of advanced life support (ALS) and basic life support (BLS) ambulances—the *vehicle mix*—to deploy in an emergency medical service (EMS) system, given a budget constraint. ALS ambulances can treat a wider range of emergencies, whereas BLS ambulances are less expensive to operate. To this end, we develop a framework under which the performance of a system operating under a given vehicle mix can be evaluated. Because the choice of vehicle mix affects how ambulances are dispatched to incoming calls, as well as how they are deployed to base locations, we adopt an optimization-based approach. We construct two models—one a Markov decision process, the other an integer program—to study the problems of dispatching and deployment in a tiered system, respectively. In each case, the objective function value attained by an optimal decision serves as our performance measure. Numerical experiments performed with both models suggest that, under reasonable choices of inputs, a wide range of tiered systems perform comparably to all-ALS fleets.

Keywords: ambulance dispatching; ambulance deployment; Markov decision processes; queues; integer programming; advanced life support; basic life support
History: Received: October 3, 2014; accepted: July 1, 2015. Published online in *Articles in Advance* October 12, 2015.

1. Introduction

Emergency medical service (EMS) systems operate in an increasingly challenging environment characterized by rising demand, worsening congestion, and unexpected delays (such as those caused by ambulance diversion). Achieving a desirable level of service requires the coordination of a wide range of medical personnel, as well as careful and effective management of system resources. To this end, EMS providers can make a number of strategic decisions, one of which is the composition of their ambulance fleets.

Ambulances can be differentiated by the medical personnel on board, and thus, by the types of treatment they can provide. These personnel can be roughly divided into two groups: emergency medical technicians (EMTs) and paramedics. EMTs use noninvasive procedures to maintain a patient's vital functions until more definitive medical care can be provided at a hospital, whereas paramedics are also trained to administer medicine and to perform more sophisticated medical procedures on scene. The latter may be necessary for stabilizing a patient prior to transport to a hospital, or for slowing the deterioration of the patient's condition en route. Thus, they may improve the likelihood of survival for patients suffering from certain life-threatening pathologies, such as cardiac arrest, myocardial infarction, and some forms of trauma. See, for instance, Bakalos et al. (2011), Gold (1987), Isenberg and Bissell (2005), Jacobs et al. (1984), McManus et al. (1977), and Ryynänen et al. (2010) for discussions of the effects of paramedic care on patient outcomes.

Ambulances staffed by at least one paramedic are referred to as advanced life support (ALS) units (we use the terms "ambulance" and "unit" interchangeably), whereas ambulances staffed solely by EMTs are known as basic life support (BLS) units. Although ALS ambulances are more expensive to operate (because of higher personnel and equipment costs), they are viewed as essential components of most EMS systems, as such systems are frequently evaluated by their ability to respond to life-threatening emergencies.

The selection of a *vehicle mix*—that is, a combination of ALS and BLS ambulances to deploy—has been a topic of some debate in the medical community, with the primary issue being whether BLS units should be included in a fleet at all. Proponents of all-ALS systems, such as Ornato et al. (1990) and Wilson et al. (1992), cite the risk of sending a BLS unit to a call requiring a paramedic, either because of errors in the dispatch process or system congestion. An ALS unit may also need to be brought on scene before transport can begin, thus diverting system resources, and allowing the patient's condition to deteriorate. Proponents of *tiered systems*, such as Braun et al. (1990), Clawson (1989), Slovis et al. (1985), and Stout et al. (2000),

argue that BLS ambulances enable EMS providers to operate larger fleets, leading to decreased response times. This may be more desirable, as they observe that a significant fraction of calls do not require an ALS response. Both sides of the debate allude to a trade-off inherent in the vehicle mix decision—that between improving a system's responsiveness and reducing the risk of inadequately responding to calls.

A number of secondary considerations may influence the decision-making process. For instance, dispatchers in a tiered system must also assess whether an ALS or BLS response is needed, complicating triage, and delaying the response to a call (Henderson 2011). As another example, EMS providers may have difficulty hiring and training the number of paramedics needed to operate an all-ALS system, as well as providing these paramedics with opportunities to hone and maintain their skills (Braun et al. 1990, Stout et al. 2000). Although many of these issues are quantifiable, they have received less attention in the literature, and we do not consider them here.

In this paper, we contribute to this debate by constructing models under which quantitative comparisons among vehicle mixes can be made. To do so, we consider a system in which the EMS provider has a fixed annual operating budget B, and for which annual operating costs for ALS and BLS ambulances are C_A and C_B, respectively (where $C_A \geq C_B$). We define a vehicle mix to be an integer pair (N_A, N_B)—corresponding to the number of ALS and BLS ambulances deployed, respectively—for which $C_A N_A + C_B N_B \leq B$. Let $f(N_A, N_B)$ denote a scalar measure of system performance that is attained when the EMS provider deploys the vehicle mix (N_A, N_B).

We formally describe our notion of performance in later sections, but comment that it is an aggregate measure of the system's ability to respond both to life-threatening and less urgent calls. Regardless of how we define it, performance depends on two closely-related decisions: *dispatching* decisions, the policy by which ambulances are assigned to emergency calls in real time, and *deployment* decisions, the base locations to which ambulances are stationed. These decisions, in turn, are affected by an EMS provider's choice of vehicle mix. Thus, it is reasonable for $f(N_A, N_B)$ to be the output of an optimization procedure. However, we do not intend for any "optimal solutions" we obtain to directly aid decision making in practical contexts. Doing so would require tailoring our model to a specific EMS system, and we are more interested in obtaining general insights.

Although we could construct a model that jointly optimizes with respect to both dispatching and deployment decisions, such a model may not be tractable. To obtain general insights, we require a procedure that can be quickly applied to large collections of problem instances. Thus, instead of studying one sophisticated model, we base our analysis upon two more stylized models. If we make the (admittedly large) assumption that the locations of arriving calls and ambulances can be ignored, we need only optimize with respect to dispatching decisions, and the resulting problem can be modeled as a Markov decision process (MDP). Alternatively, if we assume that the EMS provider operates under a fixed dispatching policy, the resulting problem of deployment can be modeled as an integer program (IP). Both models treat (N_A, N_B) as input, and we take $f(N_A, N_B)$ in each case to be the objective function value associated with an optimal solution.

We contend that there is value in using both models to study the vehicle mix problem. An analysis conducted using the MDP model alone would leave open the question of whether taking geographical factors into account would lead to qualitatively different results. Furthermore, an analysis conducted solely using the IP model largely omits the real-time decision-making aspects of the problem, and raises similar concerns. Thus, we view our two models as complementary to one another. Numerical experiments indicate that our models yield similar qualitative insights, but different quantitative results. However, we demonstrate in §6 that these discrepancies can be reconciled, allowing us to draw stronger conclusions than is possible with either model alone.

Our paper's primary finding is that there are rapidly diminishing marginal returns associated with biasing a fleet toward all-ALS, suggesting there is a fairly wide range of vehicle mixes that perform comparably to an all-ALS fleet. This can be interpreted in several ways. For systems that already operate all-ALS fleets, our analysis suggests that there is not a compelling reason to switch to a different vehicle mix. For systems that operate a mixture of ALS and BLS ambulances, the benefit that can be attained by a conversion to an all-ALS fleet may not justify the costs involved. With respect to the vehicle mix debate, we contend that the most appropriate fleet for a given system should depend more heavily on the secondary considerations previously described, or at least, that these considerations can be weighted more heavily in the decision-making process without significantly affecting performance.

The remainder of this paper is organized as follows. Following a literature review in §2, we describe and formally construct in §3 our MDP model for ambulance dispatching in a tiered EMS system. We perform a computational study on this model in §4, which is based upon a large-scale EMS system loosely modeled after Toronto EMS. In §5, we formulate our IP model for ambulance deployment, and conduct a numerical study in §6 similar to that in §4. We conclude and discuss future research directions in §7.

2. Literature Review

There is a sizable body of literature relating to the use of operations research models to guide decision making in EMS systems. We do not give a detailed overview here, and instead refer the reader to surveys, such as those by Brotcorne et al. (2003), Goldberg (2004), Green and Kolesar (2004), Henderson (2011), Ingolfsson (2013), Mason (2013), McLay (2010), and Swersey (1994). We draw primarily from two streams of literature.

The first stream of literature we consider pertains to dynamic models, which are used to analyze real-time decisions faced by an EMS provider. Such models typically assume that ambulances have already been deployed to bases, and instead consider how to dispatch these ambulances to incoming calls, or alternatively, how to redeploy idle ambulances to improve coverage of future demand. The first such model is due to Jarvis (1975), who constructs an MDP for dispatching in a small-scale EMS, whereas McLay and Mayorga (2012) consider a variant in which calls can be misclassified. Work relating to ambulance redeployment was initiated by Berman (1981a, b, c); Zhang (2012) conducts a more refined analysis for the case of a single-ambulance fleet. The aforementioned models succumb to the curse of dimensionality, and so to analyze large-scale systems, Maxwell et al. (2010) and Schmid (2012) develop redeployment policies using approximate dynamic programming. Although our model is less detailed than those previously cited, it captures the essential features of a tiered EMS system, and is amenable to sensitivity analysis, and can thus be used to obtain quick insights. Our model can also be extended to include a fairly wide range of system dynamics, such as call queueing, without significantly increasing the size of the state space; see Online Appendix A (available as supplemental material at http://www.dx.doi.org/10.1287/msom.2015.0555).

The second stream of literature we consider relates to integer programming models for ambulance deployment, canonical examples of which include Toregas et al. (1971), Church and ReVelle (1974), and Daskin (1983). These models base their objective functions upon some measure of the system's responsiveness to emergency calls. This can be quantified via the proportion of emergency calls that survive to hospital discharge, as in Erkut et al. (2008) and Mayorga et al. (2013), or more commonly, via the concept of coverage: the long-run average number of calls to which an ambulance can be dispatched within a given time threshold. These models have been extended to study the problem of deploying multiple types of emergency vehicles; see, for instance, Charnes and Storbeck (1980), Mandell (1998), and McLay (2009). Our IP model is most similar in spirit to that of Daskin (1983), in that we use a similar notion of coverage in our objective function, and is also closely related to that of McLay (2009), who considers calls of varying priority. However, our IP takes into account more of the nuances of dispatching in a tiered EMS system, by affording some flexibility in the dispatching policy. We also consider a more general notion of coverage that quantifies the system's ability to respond both to high-priority and low-priority calls.

Closely related to the previously mentioned body of work is a stream of literature pertaining to descriptive models of EMS systems, which aims to develop accurate and detailed performance measures for a system operating under a given set of deployment decisions. Larson's hypercube model (1974) and its variants, such as those by Larson (1975) and Jarvis (1985), are perhaps the most influential of this kind. Simulation has been widely used since Savas (1969); see, for instance, Henderson and Mason (2004). Although descriptive models allow for more thorough comparisons between candidate deployment decisions, they are not as amenable to optimization.

There is a related body of literature on the flexible design of manufacturing and service systems. The seminal work in this area is due to Jordan and Graves (1995), who observe that much of the benefit associated with a "fully flexible" system (which, in their case, represents the situation in which all plants in manufacturing system can produce every type of product) can be realized by a strategically configured system with limited flexibility. A similar principle has been shown to hold for call centers (Wallace and Whitt 2005), as well as for more general queueing systems (Gurumurthi and Benjaafar 2004, Tsitsiklis and Xu 2012). Theoretical justifications thereof are provided in Aksin and Karaesmen (2007) and Simchi-Levi and Wei (2012). In relating our work to this body of literature, we can define "flexibility" as the fraction of the budget that an EMS provider expends on ALS ambulances, but our paper's conclusions do not directly follow from this work. This is because servers in our model are geographically distributed, and their locations affect the system's ability to respond to incoming demand. The models we formulate provide a way to quantify these effects, and in turn, to study their effects on performance.

3. An MDP-Based Dispatching Model

3.1. Setup

Consider an EMS system operating N_A ALS and N_B BLS units. Incoming emergency calls are divided into two classes: urgent, *high-priority* calls for which the patient's life is potentially at risk, and less urgent *low-priority* calls. We assume that high-priority and low-priority calls arrive according to independent Poisson

processes with rates λ_H and λ_L, respectively, and that they only require single-ambulance responses.

Service times are exponentially distributed with rate μ, independent of the priority of the call and of the type of ambulance dispatched. Although we use the exponential distribution for tractability, our assumption of a single μ may be reasonable in large-scale systems, as on-scene treatment times are typically small relative to those for other components of the emergency response, such as travel and hospital drop-off times. Arrivals occurring when all ambulances are busy do not queue, but instead leave the system without receiving service. This is consistent with what occurs in practice, as EMS providers may redirect calls to external services, such as a neighboring EMS or the fire department, during periods of severe congestion, but we revisit this assumption in §3.3.

Dispatches to high-priority calls must be performed whenever an ambulance is available, but a BLS unit can be sent in the event that every ALS unit is busy, so as to provide the patient with some level of medical care. In this case, we assume that the BLS unit can adequately treat the high-priority call, but that such a dispatch is undesirable, in a way that we clarify in §3.1.2; we also revisit this assumption in §3.3. Similarly, dispatches to low-priority calls must be made if a BLS ambulance is available, but if this is not the case, the dispatcher may either respond with an ALS unit (if one is available), or redirect the call to an external service (to reserve system resources for potential future high-priority calls).

3.1.1. State and Action Spaces. We define the state space to be $\mathcal{S} = \{0, 1, \ldots, N_A\} \times \{0, 1, \ldots, N_B\}$, where $(i, j) \in \mathcal{S}$ denotes the state in which i ALS and j BLS units in the system are busy. We define the action space to be $\mathcal{A} = \bigcup_{(i,j)\in\mathcal{S}} A(i, j)$, where

$$A(i, j) = \begin{cases} \{0, 1\} & \text{if } i < N_A \text{ and } j = N_B, \\ \{0\} & \text{otherwise.} \end{cases}$$

In states where both actions are available, action 1 dispatches an ALS unit to the next arriving low-priority call, whereas action 0 redirects the call instead. In all other states, the dispatcher does not have a decision to make, and action 0 represents a dummy action.

3.1.2. Rewards. Let R_{HA} and R_{HB} be the rewards associated with dispatching an ALS unit or a BLS unit to a high-priority call, respectively, and R_L be the reward for dispatching an ambulance (of either type) to a low-priority call. We assume $R_{HA} = 1$ (without loss of generality) and $R_{HA} \geq \max\{R_{HB}, R_L\}$, but make no assumptions about the relative ordering of R_{HB} and R_L, as this may depend, for instance, on the skill gap between EMTs and paramedics, or on the incentives of the EMS in question. This is an unconventional modeling choice, but it results in an objective function that takes into account the system's responsiveness to both high-priority and low-priority calls. Typically, these two goals conflict, and we can adjust how they are weighted in the objective function by changing the reward structure. Our framework is also flexible. By letting R_{HA}, for instance, be the probability of patient survival when an ALS ambulance responds to a high-priority call (and defining R_{HB} and R_L similarly), we can mimic the reward structure adopted by McLay and Mayorga (2012). If an EMS provider is concerned solely with high-priority calls, we can let $R_{HA} = 1$, $R_{HB} = 0$, and $R_L = 0$. More generally, these rewards can represent a measure of the utility that the EMS provider derives from a successful dispatch. Nevertheless, identifying suitable choices for R_{HA}, R_{HB}, and R_L may be difficult, and we discuss this issue further in §4.

3.1.3. Uniformization. Because the times between state transitions in our model are exponentially distributed with a rate that is bounded above by $\Lambda = \lambda_H + \lambda_L + (N_A + N_B)\mu$, our MDP is uniformizable in the spirit of Lippman (1975), and we can consider an equivalent process in discrete time. Suppose without loss of generality (by rescaling time, if necessary) that $\Lambda = 1$. The discrete-time process is such that at most one event can occur during a single (uniformized) time period, and given the system is in state $(i, j) \in \mathcal{S}$, that event can be

- with probability λ_H, the arrival of a high-priority call;
- with probability λ_L, the arrival of a low-priority call;
- with probability $i\mu$, an ALS unit service completion;
- with probability $j\mu$, a BLS unit service completion;
- with probability $(N_A - i)\mu + (N_B - j)\mu$, a dummy transition.

3.1.4. Rewards and Transition Probabilities. Let $R((i, j), a)$ denote the expected reward collected over a single uniformized time period, given that the system begins the period in state (i, j), and the dispatcher takes action $a \in A(i, j)$. We have

$$R((i, j), a) = \begin{cases} \lambda_H R_{HA} + \lambda_L R_L & \text{if } i < N_A,\ j < N_B,\ a = 0, \\ \lambda_H R_{HB} + \lambda_L R_L & \text{if } i = N_A,\ j < N_B,\ a = 0, \\ \lambda_H R_{HA} & \text{if } i < N_A,\ j = N_B,\ a = 0, \\ \lambda_H R_{HA} + \lambda_L R_L & \text{if } i < N_A,\ j = N_B,\ a = 1, \\ 0 & \text{if } i = N_A,\ j = N_B,\ a = 0. \end{cases} \quad (1)$$

Let $P((i', j') \mid (i, j), a)$ denote the one-stage transition probabilities from state (i, j) to state (i', j') under

action $a \in A(i,j)$. There are several cases to consider, as the system dynamics change slightly at the boundary of the state space. For brevity, we consider only the case when $0 < i < N_A$ and $j = N_B$, in which case, we have

$$P((i',j') \mid (i,j),a) = \begin{cases} \lambda_H + \mathbf{I}(a=1)\lambda_L & \text{if } (i',j') = (i+1,j), \\ i\mu & \text{if } (i',j') = (i-1,j), \\ j\mu & \text{if } (i',j') = (i,j-1), \\ 1-\lambda_H - \mathbf{I}(a=1)\lambda_L - (i+j)\mu & \text{if } (i',j') = (i,j). \end{cases}$$

The first transition corresponds to an arrival of a high-priority call (or a low-priority call, if the dispatcher performs action 1), the second and third to service completions by ALS and BLS units, respectively, and the fourth to dummy transitions because of uniformization.

3.2. Optimality Equations

We seek a policy that maximizes the long-run average reward collected by the system, where we define a policy to be a stationary, deterministic mapping $\pi\colon \mathcal{S} \to \{0,1\}$ that assigns an action to every system state. Because state and action spaces are finite, by Theorem 8.4.5 of Puterman (2005), we can restrict attention to this class of policies Π without loss of optimality.

For a fixed policy $\pi \in \Pi$, let S_n^π be the state of the system at time n, and A_n^π be the action selected by π at this time. We define the long-run average reward collected attained by policy π as

$$J^\pi = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}\left[\sum_{n=1}^{N} R^\pi(S_n^\pi, A_n^\pi)\right],$$

By Theorem 8.3.2 of Puterman (2005), J^π is well-defined and independent of the system's initial state, as the Markov chain induced by π is irreducible. Indeed, suppose (i,j) and (i',j') are distinct states in $\mathcal{S}$. Then state (i',j') can be reached from state (i,j) under any policy via $i+j$ consecutive service completions, followed by i' high-priority and j' low-priority call arrivals. We wish to find $J^* := \max_{\pi\in\Pi} J^\pi$, the long-run average reward attained by an optimal policy. This quantity is well-defined because Π is finite, and can be found by solving the optimality equations

$$J^* + h(i,j) = \max_{a\in A(i,j)} \left\{ R((i,j),a) + \sum_{(i',j')\in\mathcal{S}} P\big((i,j),(i',j'),a\big) h(i',j') \right\} \quad \forall (i,j) \in \mathcal{S} \qquad (2)$$

for $h(\cdot)$ and J^*. We do this using the policy iteration algorithm in Section 8.6.1 of Puterman (2005).

3.3. Extensions to the MDP

Our MDP model can be modified to relax some of the previously mentioned assumptions without dramatically increasing the size of the state space. In Online Appendix A, we formulate an extended MDP in which low-priority calls can be placed in queue during periods of congestion, and in which ALS units may need to be brought on scene to assist BLS first responses to high-priority calls. For our computational work in §4, we focus primarily on the base model we formulated in §3.1, as our extended model yields very similar numerical results; see Online Appendix A.7.

4. Computational Study of the MDP

In this section, we consider a hypothetical system loosely modelled after that operated in Toronto, Canada. We use the term "loosely" because we select inputs to our model using a data set obtained from Toronto EMS, but we assume that the interarrival and service time distributions do not vary with time. Thus, the results we obtain in this section may be indicative of—but not necessarily predict—how the vehicle mixes we consider would perform in practice.

4.1. Experimental Setup

Our data set contains records of all ambulance dispatches occurring within the Greater Toronto Area between January 1, 2007, and December 31, 2008; we restrict our attention to calls originating from the City of Toronto. Emergency calls are divided into eight priority levels, two of which require a "lights and sirens" response. We treat calls belonging to these two priority levels as high-priority, and all other emergency calls as low-priority. Estimating arrival rates by taking long-run averages over the two-year period for calls originating within the City of Toronto, we obtain $\lambda_H = 8.1$ and $\lambda_L = 13.1$ calls per hour. We define service time as the length of the interval beginning with an ambulance dispatch and ending with the call being cleared (either on scene or following drop-off at a hospital). Because the mean service times for low-priority and high-priority calls do not differ substantially in the data set (by less than 5%), our assumption of a single service rate for all calls is reasonable. We set $\mu = 3/4$ per hour, corresponding to a mean service time of 80 minutes.

As a starting point for our analysis, we let $R_{HA} = 1$, $R_{HB} = 0.5$, and $R_L = 0.6$; we investigate in §4.3 the sensitivity of our findings to the latter two quantities. To estimate C_A and C_B, we assume that an ambulance requires three crews to operate 24 hours per day, that an ALS crew consists of two paramedics, that a BLS crew consists of two EMTs, and that ALS and BLS vehicles cost \$110,000 and \$100,000 to equip and operate annually, respectively. Assuming that annual salaries for paramedics and EMTs are \$90,000 and \$70,000 per year, respectively, we obtain $C_A = 650{,}000$

and $C_B = 520{,}000$, which we normalize to $C_A = 1.25$ and $C_B = 1$.

To determine our system's operating budget B, we assume that the average utilization of ambulances in our system is 0.4. Since $\lambda_H + \lambda_L = 21.2$ and $\mu = 0.75$, we would need approximately 70 ambulances to achieve the desired utilization. Assuming our system operates an all-ALS fleet (as in Toronto), we obtain $N_A = 70$ and $B = 87.5$. Restricting our attention to fleets that use as much of the budget as possible, we evaluate vehicle mixes in the set

$$\Gamma = \{(N_A, N_B)\colon N_A \le 70 \text{ and } N_B = \lfloor 87.5 - 1.25N_A \rfloor\}. \quad (3)$$

4.2. Findings

Using the previously specified inputs, we construct an MDP instance for each vehicle mix (N_A, N_B) in the set Γ specified in (3). We solve each instance numerically using policy iteration, and record the long-run average reward attained under the corresponding optimal policy. Plotting the resulting values with respect to N_A, we obtain Figure 1.

The curve in Figure 1 increases fairly steeply when N_A is small, indicating there is significant benefit associated with including ALS ambulances in a fleet. However, for larger values of N_A, the curve plateaus. This suggests that the marginal benefit associated with continuing to increase N_A diminishes rapidly. One might suspect that the marginal benefits decrease too rapidly. Indeed, the long-run average reward attainable under any dispatching policy is upper bounded by $\lambda_H R_{HA} + \lambda_L R_L = 15.96$, but every vehicle mix for which $N_A \ge 20$ performs within 0.1% of this upper bound. We would not expect to find systems attaining this level of service in practice.

This behavior can be attributed to resource pooling. In formulating our MDP, we implicitly assumed that any ambulance can respond to any incoming call. In practice, most ambulances will be too far from a particular call to respond in time, and so only a small number of ambulances may be effectively pooled. Thus, the MDP-based system can more readily respond to calls during periods of congestion; see, for instance, Whitt (1992) for a formal discussion of this phenomenon.

To offset the effects of resource pooling, two alternatives include accelerating call arrivals, or reducing the number of ambulances in the system by decreasing the budget B; we adopt the latter approach here. Shrinking the fleet allows us to more easily see the effects of the vehicle mix decision under an optimal dispatching strategy, and in turn, draw insights from our MDP. This is because we subject the system to periods of congestion, which occur in practice, and represent the situations in which vehicle mix may have the greatest impact on system performance.

It is not at all obvious to what extent the fleet should be shrunk. We proceed by first selecting a service level, which we define as the long-run fraction of time in which at least one ambulance is available. If we assume that the EMS provider operates an all-ALS fleet, and adopt a dispatching policy in which low-priority calls are not redirected unless all ambulances are busy, then we can model the system as an $M/M/N_A/N_A$ queue, and find the blocking probability under a given budget using the Erlang loss formula. We consider five different budgets: 48.75, 46.25, 43.75, 41.25, and 37.50, which allow for all-ALS fleets of size 39, 37, 35, 33, and 30, respectively. These fleets can provide service levels of 0.990, 0.980, 0.965, 0.943, and 0.898, respectively, and operate under utilizations of 0.846, 0.808, 0.779, 0.749, and 0.717, respectively. Solving the corresponding MDP instances, and plotting the resulting five curves, we obtain Figure 2.

Figure 1 (Color online) Long-Run Average Reward Attained by the Optimal Dispatching Policy, as a Function of N_A

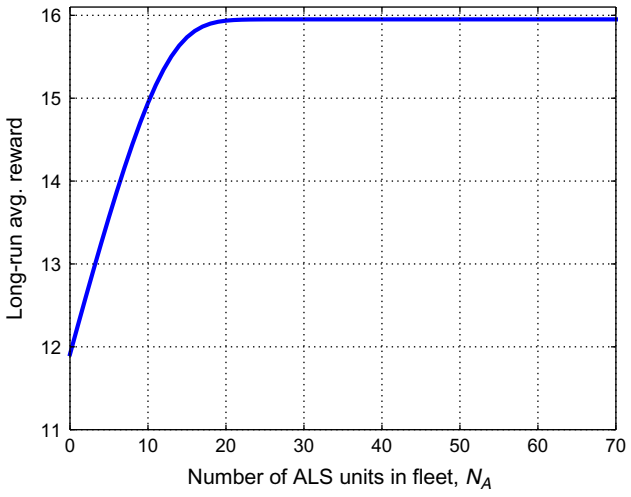


Figure 2 (Color online) Long-Run Average Reward as a Function of Vehicle Mix, for Several Reduced Values of the Budget B

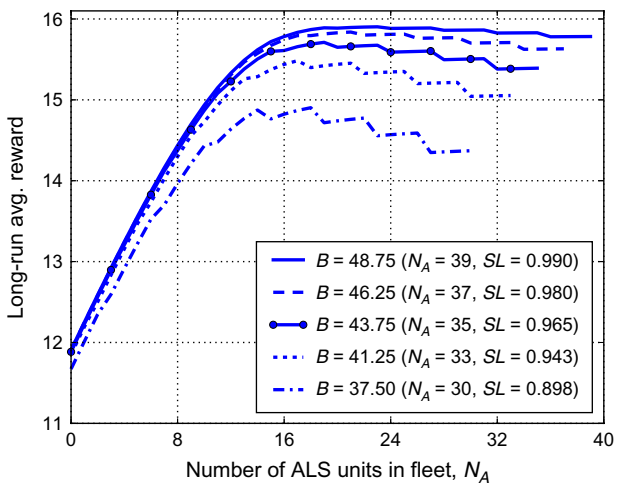


Figure 3 (Color online) Long-Run Proportion of Calls Receiving an Appropriate Dispatch as a Function of N_A, When $B = 41.25$

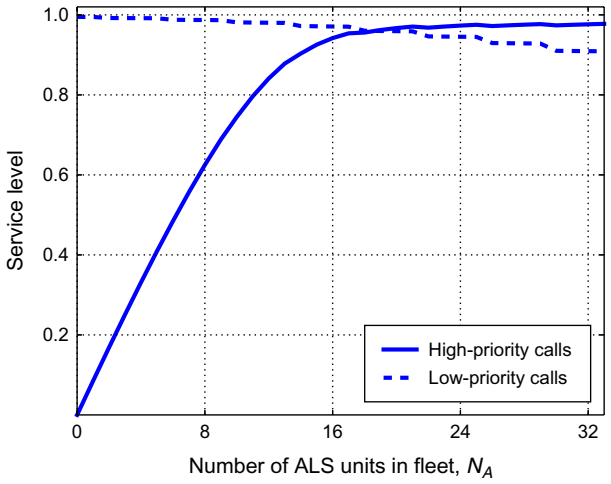
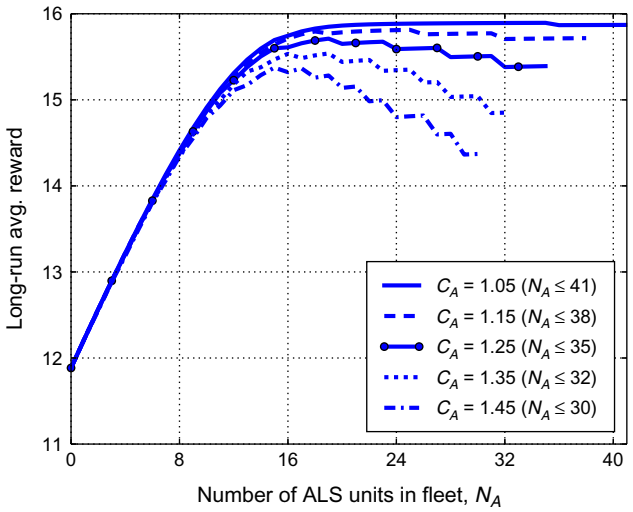


Figure 4 (Color online) Long-Run Average Reward as a Function of Vehicle Mix, for Various Choices of C_A



Although the curves in Figure 2 maintain the same basic structure observed in Figure 1, they taper off for larger values of N_A, suggesting that in certain situations, an all-ALS fleet may be detrimental. This is intuitive, as all-ALS systems tend to operate smaller fleets than their tiered counterparts, and a larger fleet may be preferable in heavily congested systems. To explore this idea further, we examine two related performance measures: the level of service provided to high-priority and to low-priority calls. Figure 3 plots these performance measures for the case $B = 43.75$.

As we expect, increasing N_A improves the system's responsiveness to high-priority calls, but worsens the system's responsiveness to low-priority calls. This results in a trade-off that is influenced by the relative importance of responding to high-priority and low-priority calls, as captured by the rewards R_{HA}, R_{HB}, and R_L. In this case, the marginal improvement attained by increasing N_A is eventually offset by the loss in ability to respond to low-priority calls. Because Figures 1 and 2 more succinctly describe this trade-off, we restrict our attention to plots of long-run average reward for the remainder of the paper.

4.3. Sensitivity Analysis

We next study the robustness of our findings by constructing curves analogous to those in Figures 1 and 2, but for problem instances in which we vary our model's input parameters. We also perform numerical experiments on the extended MDP model briefly described in §3.3. In the experiments that follow, $B = 43.75$.

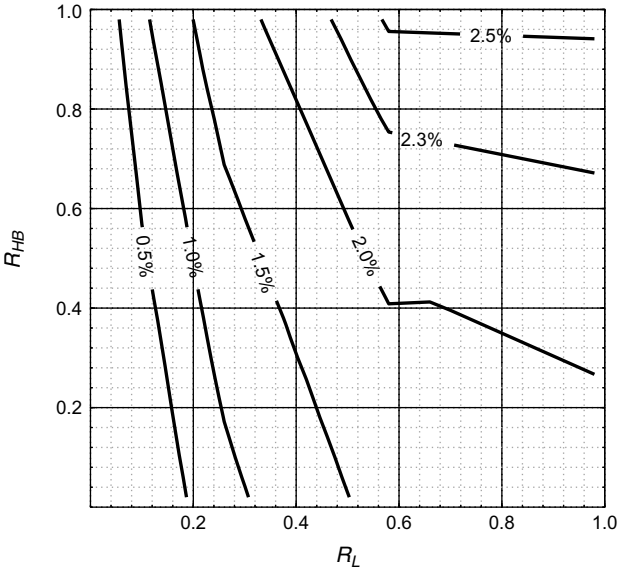
We begin with a sensitivity analysis with respect to C_A, the annual cost of deploying an ALS unit. Figure 4 depicts five curves, each similar in spirit to that in Figure 1, but for values of C_A ranging from 1.05 to 1.45. Although we do not expect ALS units to cost 45% more than BLS units to operate in practice, we choose a wide range of values for illustrative purposes.

Not surprisingly, operating an all-ALS fleet can be suboptimal when C_A is large. Perhaps the more interesting observation is that when C_A is small, an all-ALS fleet is not necessarily the obvious choice. As we observed in Figure 3, an EMS provider would consider a fleet with more ALS ambulances if there is a strong incentive to provide ALS responses to high-priority calls. However, a high level of service can be achieved without an all-ALS fleet, and the marginal benefit associated with increasing N_A shrinks rapidly—even when ALS ambulances are inexpensive.

Next, we consider the robustness of our findings to our rewards. Because we assume $R_{HA} = 1$, we need only perform a sensitivity analysis with respect to R_{HB} and R_L. We restrict our attention to the all-ALS fleet $(35, 0)$ and the tiered system $(19, 20)$: the optimal vehicle mix from our "base case" analysis in Figure 2. We consider a collection of 625 MDP instances where R_{HB} and R_L take on one of 25 values in the set $\{0.02, 0.06, \ldots, 0.94, 0.98\}$. Plotting the relative difference in long-run average reward collected by the two systems for each instance, we obtain Figure 5.

In each problem instance, the tiered system outperforms the all-ALS fleet. This is unsurprising for instances where R_{HB} and R_L are close to 1 (as emergency calls become effectively indistinguishable, and the tiered system deploys eight more ambulances), but is counterintuitive for instances when R_{HB} and R_L are small. Even when performance is determined almost entirely by the system's responsiveness to high-priority calls, the all-ALS fleet does not outperform the tiered system. This suggests that the tiered system can adequately respond to these calls, which is

Figure 5 **Contour Plot of the Relative Difference Between the Long-Run Average Reward Collected by the Tiered System (19, 20) and That by the All-ALS Fleet (35, 0), for Various Values of R_{HB} and R_L**



consistent with Figure 3. We also observe that the two systems perform comparably in all of our problem instances; the largest observed difference was roughly 2.53%. We obtain similar results when comparing the all-ALS fleet to other tiered systems, although we omit the corresponding plots of these results in the interest of brevity. This suggests our findings are insensitive to our choice of rewards, which is encouraging, as selecting appropriate values of R_{HB} and R_L is difficult.

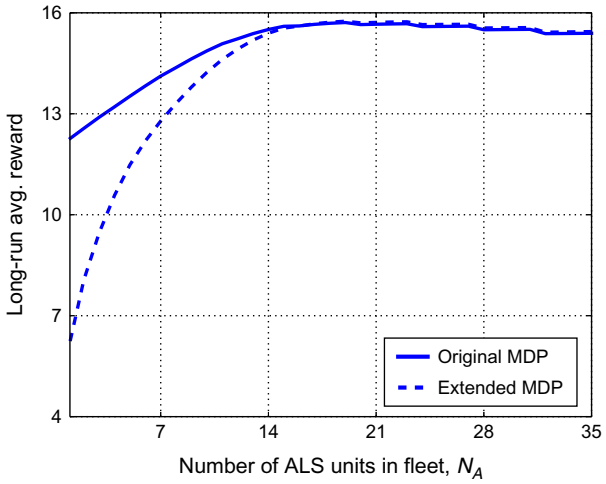
We conclude this section with a brief examination of the extended MDP model alluded to in §3.3. This model allows low-priority calls to be placed in a finite buffer, and considers the possibility that high-priority calls receiving a BLS first response may later receive treatment from an ALS ambulance. Plotting the "base case" curves for both the original and extended models, we obtain Figure 6. We truncate the curves at $N_A = 1$, as an all-BLS fleet cannot adequately treat high-priority calls in the extended model.

Fleets operating too few ALS ambulances perform noticeably worse in the extended model. This is because high-priority calls will primarily be assigned to BLS ambulances, and many of these ambulances will be forced to idle as they wait for ALS units to become free. However, for vehicle mixes under which high-priority demand can be adequately met, our two models are roughly in agreement. We conduct a more detailed analysis of the extended model in Online Appendix A.7.

4.4. Discussion

Taken together, our numerical experiments suggest that a wide range of tiered systems perform

Figure 6 **(Color online) Long-Run Average Reward as a Function of Vehicle Mix, for the Original and Extended MDP Models (see §3.3 and Online Appendix A)**



comparably to all-ALS fleets. The relatively small gap in long-run average reward that we observe between the two types of systems appears to be robust to changes in operating costs, arrival patterns, reward structure, and changes to system dynamics.

In §6, we perform a similar set of numerical experiments, to determine whether we draw similar conclusions when we study our IP model, in which we mitigate the effects of resource pooling by geographically dispersing the fleet.

5. An IP-Based Deployment Model

5.1. Formulation

Consider an EMS system whose service area is represented by a connected graph $G = (N, E)$, where N is a set of demand nodes and E is a set of edges. High-priority and low-priority calls originate from node $i \in N$ at rates λ_i^H and λ_i^L, respectively. An EMS provider can respond to these calls with a fleet of N_A ALS and N_B BLS ambulances, which can be deployed at a set of base locations $\bar{N} \subseteq N$. For convenience we set $\bar{N} = N$, but this assumption can easily be relaxed.

We define t_{ij} as the travel time along the shortest path between nodes i and j. A call originating from node i can only be treated by an ambulance based at node j if $t_{ij} \leq T$, where T is a prespecified response time threshold. This gives rise to the *neighborhoods* $C_i = \{j \in N\colon t_{ij} \leq T\}$—the set of bases from which an ambulance can promptly respond to a call originating from node i. If a ALS and b BLS units are deployed within C_i, we say that node i is *covered* by a ALS and b BLS units.

Let p_A denote the *busy probability* associated with each ALS ambulance: the long-run fraction of time that a given ALS unit is not available for dispatch; we define p_B similarly for BLS ambulances. We treat p_A

and p_B as model inputs, and discuss our procedure for approximating these quantities in §5.2. We assume that ambulances of the same type operate under the same utilization, and that all ambulances are busy independently of one another. Thus, if node i is covered by a ALS units and b BLS units, then $(p_A)^a(p_B)^b$ is the long-run proportion of time that the system cannot respond to calls originating from that node. Calls to which an ambulance cannot be immediately dispatched are redirected to an external service. We revisit these assumptions in §5.3.

As in the MDP model, we allow BLS units to be dispatched to high-priority calls, and ALS units to be dispatched to low-priority calls, but we do not require that ALS ambulances respond to every low-priority call that arrives when all BLS ambulances are busy. Let ϕ denote the long-run proportion of low-priority calls receiving an ALS response in this situation. This quantity does not specify how decisions are made in real time, but provides a succinct measure of the system's willingness, in the long run, to dispatch ALS ambulances to low-priority calls. As with p_A and p_B, we assume ϕ to be given, and discuss in §5.2 how it can be approximated. Finally, we define rewards R_{HA}, R_{HB}, and R_L as before.

We construct our objective function as follows. Suppose that node $i \in N$ is covered by a ALS and b BLS ambulances, and consider the level of coverage provided to low-priority calls at that node. With probability $1-(p_B)^b$, a BLS unit can be dispatched. Conditional on this not being the case, an ALS unit is available with probability $1-(p_A)^a$, but a dispatch only occurs with probability ϕ. Thus, the expected reward collected by the system from a single low-priority call is

$$R_L(a,b) = R_L[1-(p_B)^b + \phi(p_B)^b(1-(p_A)^a)]. \quad (4)$$

Similar reasoning yields that the system collects, in expectation, a reward

$$R_H(a,b) := R_{HA}(1-(p_A)^a) + R_{HB}(p_A)^a(1-(p_B)^b) \quad (5)$$

from a single high-priority call. This implies that the system obtains reward from node i at a rate $\lambda_i^H R_H(a,b) + \lambda_i^L R_L(a,b)$ per unit time. We want to deploy ambulances such that the sum of this quantity over all nodes in N is maximized. Let x_i^A and x_i^B be the number of ALS units and BLS units stationed at node $i \in N$, respectively, and let y_{iab} take on the value 1 if node $i \in N$ is covered by exactly a ALS units and b BLS units, and 0 otherwise. We thus obtain the following formulation:

$$\max \sum_{i\in N} \lambda_i^H \sum_{a=0}^{N_A}\sum_{b=0}^{N_B} y_{iab} R_H(a,b) + \sum_{i\in N} \lambda_i^L \sum_{a=0}^{N_A}\sum_{b=0}^{N_B} y_{iab} R_L(a,b) \quad \text{(IP)}$$

$$\text{s.t.} \sum_{i\in N} x_i^A \le N_A, \quad (6)$$

$$\sum_{i\in N} x_i^B \le N_B, \quad (7)$$

$$\sum_{a=0}^{N_A} a \sum_{b=0}^{N_B} y_{iab} \le \sum_{j\in C_i} x_j^A \quad \forall\, i \in N, \quad (8)$$

$$\sum_{b=0}^{N_B} b \sum_{a=0}^{N_A} y_{iab} \le \sum_{j\in C_i} x_j^B \quad \forall\, i \in N, \quad (9)$$

$$\sum_{a=0}^{N_A}\sum_{b=0}^{N_B} y_{iab} \le 1 \quad \forall\, i \in N, \quad (10)$$

$$x_i^A \in \{0,1,\ldots,N_A\} \quad \forall\, i \in N, \quad (11)$$

$$x_i^B \in \{0,1,\ldots,N_B\} \quad \forall\, i \in N, \quad (12)$$

$$y_{iab} \in \{0,1\} \quad \forall\, i \in N, a, b. \quad (13)$$

Constraints (6) and (7) state that at most N_A ALS units and N_B BLS units can be deployed. Constraints (8)–(10) link the x variables to the y variables, by ensuring for each node i that if $\sum_{j\in C_i} x_j^A = a$ and $\sum_{j\in C_i} x_j^B = b$, then $y_{iab} = 1$ if and only if $a = \bar{a}$ and $b = \bar{b}$. This holds because the coefficients multiplying the y variables are strictly increasing in a and b. Finally, constraints (11)–(13) restrict the x variables and the y variables to integer values.

5.2. Approximating p_A, p_B, and ϕ

To approximate the quantities p_A, p_B, and ϕ, we use the inputs of our integer program to construct an instance of our MDP from §3.1, and examine the stationary distribution induced by an optimal dispatching policy. The MDP model takes as input the arrival rates λ_H and λ_L, as well as the service rate μ. We set $\lambda_H = \sum_i \lambda_i^H$ and $\lambda_L = \sum_i \lambda_i^L$, and for the computational work we perform in §6, we again use $\mu = 0.75$. Let ν be the stationary distribution of the Markov chain induced by an optimal policy. We approximate the busy probabilities p_A and p_B using the average utilizations of ALS and BLS ambulances, respectively:

$$p_A \approx \frac{1}{N_A} \sum_{(i,j)\in\mathcal{S}} i\nu(i,j) \quad \text{and} \quad p_B \approx \frac{1}{N_B} \sum_{(i,j)\in\mathcal{S}} j\nu(i,j). \quad (14)$$

To approximate ϕ, we could similarly use the quantity

$$\phi \approx \frac{\sum_{i=0}^{N_A-1} \nu(i,N_B)\cdot \mathrm{I}(A(i,N_B)=1 \text{ under the optimal policy})}{\sum_{i=0}^{N_A-1} \nu(i,N_B)}. \quad (15)$$

The denominator of (15) corresponds to the long-run fraction of time that all BLS ambulances are busy and at least one ALS ambulance is available. The numerator denotes the long-run fraction of time in which the dispatcher would send an ALS ambulance

to a low-priority call. Equations (14) and (15) are approximations because p_A, p_B, and ϕ depend on how ambulances are located within the system. Nevertheless, these approximations allow us to capture, to an extent, the dependencies of these parameters on vehicle mix and on dispatching decisions.

5.3. Extensions to the IP

Perhaps the three most significant assumptions we make in formulating our integer program (IP) are that calls do not queue, that ambulances are busy independently of one another, and that these probabilities do not depend on location. The former assumption can be relaxed by estimating p_A, p_B, and ϕ from the output of an MDP that includes call queueing, such as that formulated in Online Appendix A, and by modifying the objective function to take queued calls into account.

The independence assumption can be relaxed using correction factors, which adjust the probabilities obtained under this assumption by a multiplicative constant to account for dependence. This idea is due to Larson (1975), and has been used, for instance, in Ingolfsson et al. (2008) and McLay (2009) to formulate IP-based models for ambulance deployment. In Online Appendix B, we formulate an extended IP model that incorporates correction factors to our objective function.

Relaxing the assumption of location-independent busy probabilities is difficult, as the utilization of a single ambulance depends upon how all other ambulances are deployed, resulting in nonlinear interactions. Budge et al. (2009) develop an iterative procedure for the case when ambulances have already been deployed. Ingolfsson et al. (2008) alternate between solving an integer program for a given set of utilizations, and using the resulting optimal solution to compute updated values, also in an iterative fashion. However, these approaches are computationally intensive.

For our computational work in §6, we study the IP model we formulated in §5.1, as the extended IP model yields qualitatively similar results; see Online Appendix B.3.

6. Computational Study of the IP

6.1. Experimental Setup

We base our computational experiments in this section upon the same hypothetical EMS considered in §4. To construct our graph G, we bound the service area within a rectangular region. Using latitude and longitude information included with call records in the data set, we find that a 26×19 mile region suffices. We divide this region into a 52×38 grid of 0.5×0.5 mile cells, which we treat as demand nodes.

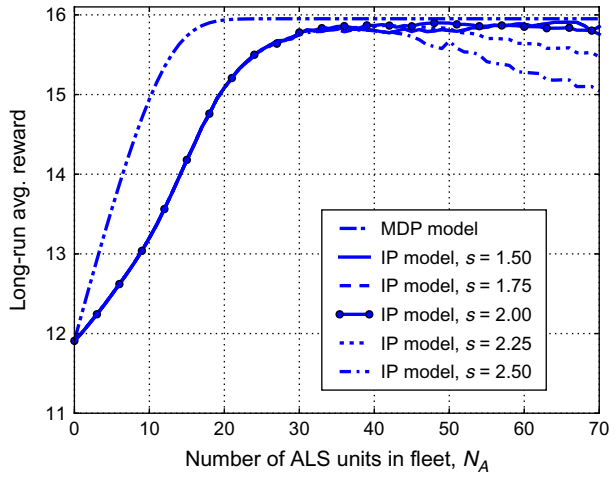
To compute call arrival rates associated with each node, we map each call to a cell in the grid, and take a long-run average over the two-year period for which we have data. We define the distance between two nodes as the Manhattan distance between the centers of their corresponding cells. For each node i, we define the neighborhood C_i as the set of bases from which an ambulance can be brought on scene within 9 minutes. This response interval includes the time taken by the dispatcher to assign an ambulance to a call, and by the corresponding crew to prepare for travel to the scene. Assuming this process takes two minutes, and that ambulances travel at 30 miles per hour, C_i contains all nodes lying no more than 3.5 miles away from node i. As before, we set $R_{HA} = 1$, $R_{HB} = 0.5$, $R_L = 0.6$, $C_A = 1.25$, and $C_B = 1$, and again evaluate vehicle mixes in the set $\Gamma = \{(N_A, N_B)\colon N_A \leq 70 \text{ and } N_B = \lfloor 87.5 - 1.25N_A \rfloor\}$.

6.2. Findings

For each vehicle mix in the set Γ, we solve the corresponding instance of (IP) to within 1% of optimality, and store the resulting objective function value. Although we introduce some error by not finding the optimal integer solution, the impact on our overall findings is negligible. To decrease computation times, we remove any decision variables y_{iab} for which either $a \geq 30$ or $b \geq 30$. Thus, we consider any demand node that is covered by more than 30 ALS or BLS ambulances to be covered by exactly 30 ambulances of the corresponding type instead. In doing so, we do not render infeasible any solutions that cover a node with more than 30 units, but we disregard the contributions of these excess units to the objective function. We thereby underestimate the coverage provided by a given deployment decision, but not to a significant degree, as p_A^{30} and p_B^{30} are very small for reasonable choices of p_A and p_B. Since a and b can be as large as 70 and 87, respectively, this dramatically reduces the number of decision variables in the model.

We use the procedure specified in §5.2 to approximate p_A, p_B, and ϕ, but find that in each of the resulting problem instances that $\phi = 1.0$. That is, when we solve the MDP instances corresponding to our IP instances, the optimal policy always dispatches ALS ambulances to low-priority calls when BLS ambulances are busy. These anomalous results can again be attributed to resource pooling. Although a dispatcher may sometimes want to reserve ALS ambulances for future high-priority calls, this is not the case in the MDP model, as the system rarely becomes congested enough for the policy to have detrimental effects. To identify more suitable values for ϕ, we instead solve modified MDP instances in which we accelerate arrivals using a scaling factor s. The resulting optimal policies may more closely reflect decisions made in practice, as they are derived from more heavily congested systems. Applying (15) to these policies, we obtain new values for ϕ, which we then use

Figure 7 **Long-Run Average Reward as a Function of Vehicle Mix, Attained Under a Near-Optimal Deployment Policy for Various Arrival Scaling Factors s, Overlaid with the Analogous Curve from the MDP Model**



to construct modified IP instances in which all other inputs (including arrival rates and busy probabilities) are kept fixed.

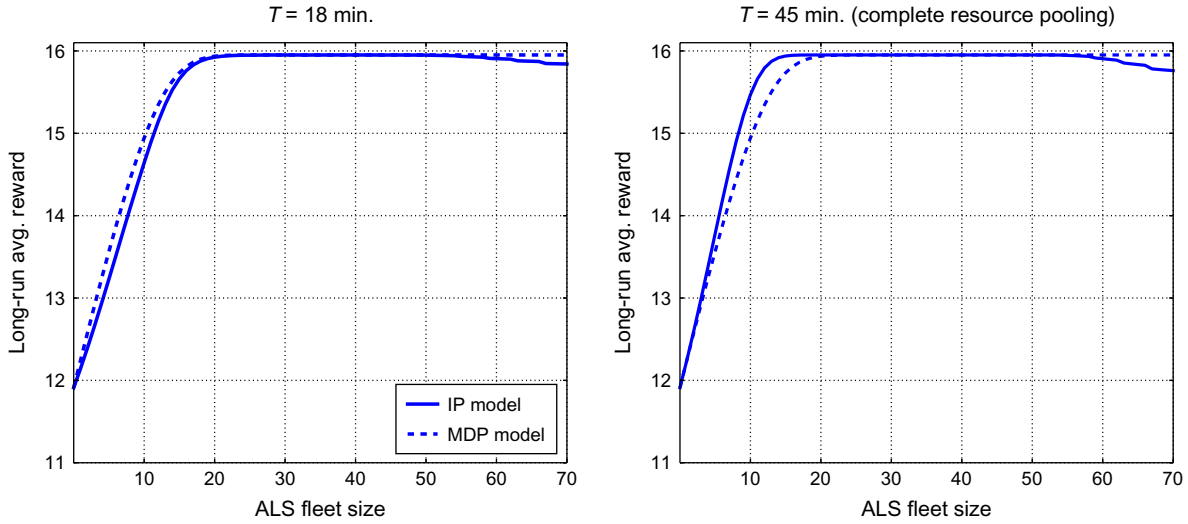
It is not at all clear how arrivals should be scaled, and so we begin our numerical study by examining the sensitivity of our results to s. Figure 7 plots long-run average reward with respect to vehicle mix for values of s ranging from 1.50 to 2.50. The resulting curves are analogous to that in Figure 1 of §4, which we also include in Figure 7.

All of the curves in Figure 7 exhibit the same trend: a relatively sharp increase for small values of N_A, followed by rapidly diminishing marginal returns. The curves obtained from the IP model almost completely overlap when N_A is less than about 30, as ALS ambulances in the corresponding systems are overwhelmed by high-priority calls. Scaling factors have very little effect on the optimal policies of the corresponding MDP's, and so we obtain very similar values for ϕ. As we move toward an all-ALS fleet, we observe a decline in performance for more extreme values of s. This is because in a heavily loaded system, the dispatcher may prefer to reserve ALS ambulances for high-priority calls (which are more likely to occur when s is large), resulting in smaller values of ϕ. However, this translates into an overly conservative dispatching policy, and thus lower performance, within the context of the IP model. Nonetheless, it is encouraging that our findings are not particularly sensitive to the dispatching policy that we employ in the IP model, as captured by the parameter ϕ. In the experiments that follow, we restrict our attention to the case when $s = 2$.

Another observation we draw from Figure 7 is that the IP and MDP models yield very different numerical results, particularly for smaller values of N_A. This may again be due to the effects of resource pooling; in the MDP model, any ambulance can respond to any call, whereas in the IP model, this is not the case. To test this hypothesis, we consider a collection of IP instances in which we artificially magnify the effects of resource pooling, to see whether we obtain quantitative results that are more consistent with those from the MDP.

In our IP model, the degree of resource pooling is captured by a single input parameter: the response time threshold T. Increasing this threshold increases the number of ambulances that can cover a given demand node. By letting T grow sufficiently large (in this case, to 45 minutes), we obtain a system with complete resource pooling, as in the MDP model. Thus, we proceed by considering modified problem instances in which T is increased, but all other parameters are unchanged. Figure 8 illustrates the curves we obtain by setting T to 18 and 45 minutes, respectively.

Figure 8 **Long-Run Average Reward Attained Under the IP and MDP Models, for Two Choices of the Response Time Threshold T**



As we increase T, the gap between the two curves narrows, but interestingly, for larger values of T, the IP model yields objective values larger than those obtained by the MDP model, particularly when N_A is small. This can partly be attributed to our assumption that ambulances are busy independently of one another. In particular, during periods of congestion, this assumption could lead to optimistic estimates of ambulance availability. Indeed, our extended IP model, which employs correction factors, improves the fit slightly; see Online Appendix B.3.

Our experiments suggest that our two models yield similar qualitative results, and that much of the observed discrepancy in our quantitative results can be accounted for by resource pooling. Thus, we contend that our models are generally in agreement, and that both support our claim of rapidly diminishing marginal returns associated with biasing a fleet toward all-ALS.

6.3. Sensitivity Analysis

We begin with a sensitivity analysis with respect to C_A that similar to that in §4; Figure 9 is analogous to Figure 4. We again observe the same general trends. This theme recurs if we perform sensitivity analyses with respect to rewards or arrival rates, suggesting that the agreement between our two models is also robust to changes to our input parameters.

We conclude this section with a sensitivity analysis with respect to ambulance travel speeds. Because ambulances must arrive on scene within a response time threshold T, changes to these speeds can affect the distance an ambulance can travel to cover a call. Although we examined this to an extent in Figure 8, we consider a more realistic range of values here. Figure 10 illustrates the curves we obtain for speeds ranging from 21.43 to 38.57 mph. We choose these values so that the resulting ambulance coverage radii, in miles, are integer multiples of 0.5. (Recall that we discretized the service area into 0.5×0.5 mile squares.) We observe that the profile of the curves does not change dramatically with the speed at which ambulances travel. This is encouraging, as travel speed can depend on the time of day, as well as on geographical factors. Although we assume away these effects in our IP, Figure 10 suggests that doing so does not substantially change our results.

Figure 9 **Long-Run Average Reward as a Function of Vehicle Mix for Several Choices of C_A**

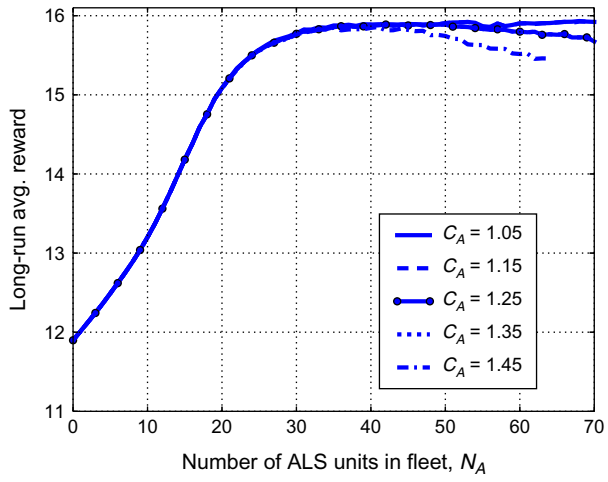
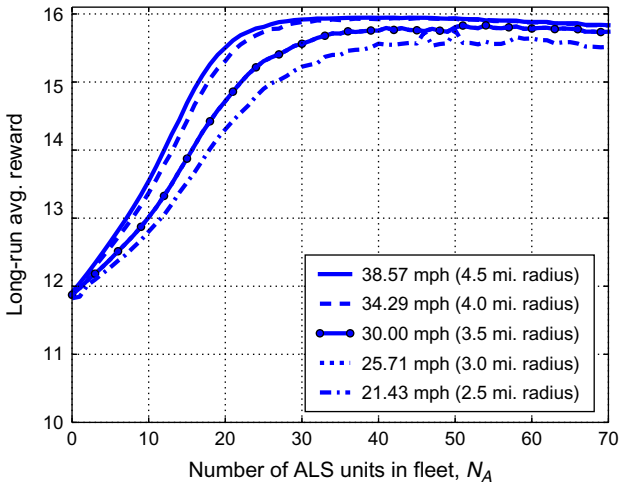


Figure 10 **Long-Run Average Reward as a Function of Vehicle Mix for Several Choices of Ambulance Travel Speeds (or Alternatively, for Several Choices of Response radii)**



7. Conclusion

In this paper, we studied the effects of the vehicle mix decision on the performance of an EMS system. Inherent in this decision is a trade-off between improving the quality of service provided to high-priority calls, and increasing the size of the fleet. We analyzed this trade-off via two complementary optimization models of decision making in a tiered EMS system. Specifically, we formulated a Markov decision process that examined the operational problem of ambulance dispatching, as well as an integer program that modeled the tactical problem of deploying ambulances within a geographical region. To aid decision making, we assigned rewards for individual responses to emergency calls, which formed the basis of a performance measure allowing quantitative comparisons between vehicle mixes to be made. Numerical experiments suggest that although ALS ambulances are essential components of EMS fleets, a wide range of mixed fleets can perform comparably to (or occasionally, outperform) all-ALS fleets. This was corroborated by both of our models, and appears to be robust to reasonable changes to the values of our input parameters. As a consequence, when constructing an ambulance fleet, secondary considerations, such as those described in the introduction, can

be weighed into the decision-making process without significantly decreasing performance. Mathematically modeling these considerations, and their effects on the performance of a given vehicle mix, may be a direction for future research.

Although our focus in this paper was to construct models that can be used to quickly obtain basic insights, a natural question to ask is what additional insights can be drawn from a more sophisticated model. Another possible direction for future research would be to consider the problem of dispatching in a tiered EMS system, when geographical locations of ambulances factor into decision making. The resulting decision problem would have a considerably larger state space, but may be approachable using Approximate Dynamic Programming (ADP), as in Maxwell et al. (2010) or in Schmid (2012). This model could incorporate a wider range of system dynamics, such as time-varying call arrival rates, multiple call priority classes, and patient transport to a hospital.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/msom.2015.0555.

Acknowledgments

This work was partially supported by the National Science Foundation [Grant CMMI-1200315].

References

Aksin OZ, Karaesmen F (2007) Characterizing the performance of process flexibility structures. *Oper. Res. Lett.* 35(4):477–484.

Bakalos G, Mamali M, Komninos C, Koukou E, Tsantilas A, Tzima S, Rosenberg T (2011) Advanced life support versus basic life support in the pre-hospital setting: A meta-analysis. *Resuscitation* 82(9):1130–1137.

Berman O (1981a) Dynamic repositioning of indistinguishable service units on transportation networks. *Transportation Sci.* 15(2):115–136.

Berman O (1981b) Repositioning of distinguishable urban service units on networks. *Comput. Oper. Res.* 8(2):105–118.

Berman O (1981c) Repositioning of two distinguishable service vehicles on networks. *IEEE Trans. Systems, Man, and Cybernetics* 11(3):187–193.

Braun O, McCallion R, Fazackerley J (1990) Characteristics of midsized urban EMS systems. *Ann. Emergency Medicine* 19(5): 536–546.

Brotcorne L, Laporte G, Semet F (2003) Ambulance location and relocation models. *Eur. J. Oper. Res.* 147(3):451–463.

Budge S, Ingolfsson A, Erkut E (2009) Approximating vehicle dispatch probabilities for emergency service systems with location-specific service times and mulitple units per location. *Oper. Res.* 57(1):251–255.

Charnes A, Storbeck J (1980) A goal programming model for the siting of multilevel EMS systems. *Socio-Economic Planning Sci.* 14(4):155–161.

Church R, ReVelle C (1974) The maximal covering location problem. *Papers of the Regional Sci. Assocation* 32(1):101–118.

Clawson JJ (1989) Emergency medical dispatching. Roush WR, ed. *Principles of EMS Systems: A Comprehensive Text for Physicians* (American College of Emergency Physicians, Dallas), 119–133.

Daskin MS (1983) A maximum expected covering location model: Formulation, properties and heuristic solution. *Management Sci.* 17(1):48–70.

Erkut E, Ingolfsson A, Erdoğan G (2008) Ambulance location for maximum survival. *Naval Res. Logist.* 55(1):42–58.

Gold CR (1987) Prehospital advanced life support vs. "scoop and run" in trauma management. *Ann. Emergency Medicine* 16(7):797–801.

Goldberg JB (2004) Operations research models for the deployment of emergency medical services vehicles. *EMS Management J.* 1(1):20–39.

Green LJ, Kolesar PJ (2004) Improving emergency responsiveness with management science. *Management Sci.* 50(8):1001–1014.

Gurumurthi S, Benjaafar S (2004) Modeling and analysis of flexible queueing systems. *Naval Res. Logist.* 51(5):755–782.

Henderson SG (2011) Operations research tools for addressing current challenges in emergency medical services. *Wiley Encyclopedia Oper. Res. Management Sci.*

Henderson SG, Mason AJ (2004) Ambulance service planning: Simulation and data visualisation. Brandeau ML, Sainfort F, Pierskalla WP, eds. *Operations Research and Health Care: A Handbook of Methods and Applications* (Kluwer Academic Publishers, Boston), 77–102.

Ingolfsson A (2013) EMS planning and management. Zaric GS, ed. *Operations Research and Health Care Policy* (Springer Science and Business Media, New York), 105–127.

Ingolfsson A, Budge S, Erkut E (2008) Optimal ambulance location with random delays and travel times. *Health Care Management Sci.* 11(3):262–274.

Isenberg DL, Bissell R (2005) Does advanced life support provide benefits to patients? A literature review. *Prehospital and Disaster Medicine* 20(4):266–270.

Jacobs LM, Sinclair A, Beiser A, D'agostino RB (1984) Prehospital advanced life support: Benefits in trauma. *J. Trauma* 24(1):8–12.

Jarvis JP (1975) Optimization in stochastic service systems with distinguishable servers. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Jarvis JP (1985) Approximating the equilibrium behavior of multiserver loss systems. *Management Sci.* 31(2):235–239.

Jordan WC, Graves SC (1995) Principles on the benefits of manufacturing process flexibility. *Management Sci.* 41(4):577–594.

Larson RC (1974) A hypercube queueing model for facility location and redistricting in urban emergency services. *Comput. Oper. Res.* 1(1):67–95.

Larson RC (1975) Approximating the performance of urban emergency service systems. *Oper. Res.* 23(5):845–868.

Lippman SA (1975) Applying a new device in the optimization of exponential queueing systems. *Oper. Res.* 23(4):687–710.

Mandell MB (1998) Covering models for two-tiered emergency medical services systems. *Location Sci.* 6(1):355–368.

Mason AJ (2013) Simulation and real-time optimised relocation for improving ambulance operations. Denton BT, ed. *Handbook of Healthcare Operations Management: Methods and Applications* (Springer Science and Business Media, New York), 289–317.

Maxwell MS, Restrepo M, Henderson SG, Topaloglu H (2010) Approximate dynamic programming for ambulance redeployment. *INFORMS J. Comput.* 22(2):266–281.

Mayorga ME, Bandara D, McLay LA (2013) Districting and dispatching policies for emergency medical service systems to improve patient survival. *IIE Trans. Healthcare Systems Engrg.* 3(1):39–56.

McLay LA (2009) A maximum expected covering location model with two types of servers. *IIE Trans.* 41(8):730–741.

McLay LA (2010) Emergency medical service systems that improve patient survivability. *Wiley Encyclopedia Oper. Res. Management Sci.*, ePub ahead of print June 15, http://onlinelibrary.wiley.com/doi/10.1002/9780470400531.eorms0296/abstract.

McLay LA, Mayorga ME (2012) An optimal dispatching model for server-to-customer systems with classification errors. *IIE Trans.* 45(1):1–24.

McManus WF, Tresch DD, Darin JC (1977) An effective prehospital emergency system. *J. Trauma* 17(4):304–310.

Ornato JP, Racht EM, Fitch JJ, Berry JF (1990) The need for ALS in urban and suburban EMS systems. *Ann. Emergency Medicine* 19(12):1469–1470.

Puterman ML (2005) *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley Series in Probability and Statistics (John Wiley & Sons, Hoboken, NJ).

Ryynänen O, Iirola T, Reitala J, Pälve H, Malmivaara A (2010) Is advanced life support better than basic life support in prehospital care? A systematic review. *Scandinavian J. Trauma, Resuscitation, Emergency Medicine* 18(62), doi: 10.1186/1757-7241-18-62.

Savas ES (1969) Simulation and cost-effectiveness analysis of New York's emergency ambulance service. *Management Sci.* 15(12):608–627.

Schmid V (2012) Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. *Eur. J. Oper. Res.* 219(3):611–621.

Simchi-Levi D, Wei Y (2012) Understanding the performance of the long chain and sparse designs in process flexibility. *Oper. Res.* 60(5):1125–1141.

Slovis CM, Carruth TB, Seitz WJ, Thomas CM, Elsea WR (1985) A priority dispatch system for emergency medical services. *Ann. Emergency Systems* 14(11):1055–1060.

Stout J, Pepe PE, Mosesso VN Jr (2000) All-advanced life support vs. tiered-response ambulance systems. *Prehospital Emergency Care* 4(1):1–6.

Swersey AJ (1994) The deployment of police, fire, and emergency medical units. Pollock SM, Rothkopf MH, Barnett A, eds. *Operations Research and the Public Sector* (Elsevier, North Holland, Amsterdam), 151–200.

Toregas C, Swain R, ReVelle C, Bergman L (1971) The location of emergency service facilities. *Oper. Res.* 19(6):1363–1373.

Tsitsiklis JN, Xu K (2012) On the power of (even a little) resource pooling. *Stochastic Systems* 2(1):1–66.

Wallace RB, Whitt W (2005) A staffing algorithm for call centers with skill-based routing. *Manufacturing Service Oper. Management* 7(4):276–294.

Whitt W (1992) Understanding the efficiency of multi-server service systems. *Management Sci.* 38(5):708–723.

Wilson B, Gratton MC, Overton J, Watson WA (1992) Unexpected ALS procedures on non-emergency ambulance calls: The value of a single-tier system. *Prehospital and Disaster Medicine* 7(4):380–382.

Zhang L (2012) Simulation optimisation and Markov models for dynamic ambulance redeployment. Ph.D. thesis, University of Auckland, New Zealand.