Real-Time Ambulance Dispatching and Relocation

Amir Ali Nasrollahzadeh,[a] **Amin Khademi,**[a] **Maria E. Mayorga**[b]

Received: May 24, 2016
Revised: November 26, 2016; March 9, 2017
Accepted: April 30, 2017
Published Online in Articles in Advance: April 11, 2018

https://doi.org/10.1287/msom.2017.0649

Abstract. In this study, we develop a flexible optimization framework for real-time ambulance dispatching and relocation. In addition to ambulance redeployment, we consider a general dispatching and relocation strategy by which the decision maker has the option to (i) select any available ambulance to dispatch to a call or to queue the call and (ii) send an idle ambulance to cover the location of an ambulance just dispatched to a call. We formulate the problem as a stochastic dynamic program, and, because the state space is unbounded, an approximate dynamic programming (ADP) framework is developed to generate high-quality solutions. We assess the quality of our solutions by developing a lower bound on the expected response time and computing a lower bound on the expected fraction of late calls of any relocation policy. We test the performance of our policies and available benchmarks on an emergency medical services system in Mecklenburg County, North Carolina. The results show that our policies are near optimal and significantly outperform available benchmarks. In particular, our ADP policy reduces the expected response time and fraction of high-priority late calls by 12% and 30.6%, respectively, over the best available static benchmarks in the case study. Moreover, the results provide insights on the contribution of each dispatching, redeployment, and reallocation strategy.

Keywords: ambulance operations management • dispatching • redeployment • reallocation • approximate dynamic programming

1. Introduction

1.1. Motivation

Emergency medical services (EMS) provide out-of-hospital acute medical care and transport the sick or injured to hospitals for definitive care. Typically, EMS providers' performance is evaluated based on their response time (National Association of State EMS Officials 2009), the amount of time that an ambulance takes to arrive to the scene of a call once the call is received, as reducing the response time is an essential factor in lowering patient mortality rates (Wilde 2013). In particular, a target for the proportion of urgent calls whose response time is less than a threshold is a common measure of performance. For example, the U.S. National Fire Protection Association suggests a target that 90% of emergency medical calls be reached by a first responder within four minutes, followed by an advanced life support response within eight minutes (NFPA 2010). Also, in North America, a common target is reaching 90% of urgent urban calls within nine minutes (Fitch 2005).

Factors such as increased nonemergency calls, which (by law) require that an ambulance be dispatched, and insufficient funding have increased pressure on EMS providers to "do more with less," or, at best, to use the same level of resources to achieve response time targets set by municipalities or contracts (Ward 2014). This has spurred EMS providers to better manage their ambulances by using more complex dispatching and location policies. Studies of realistic settings show that the performance of static policies, those that send the closest ambulance and preassign a location to each ambulance, can be quite poor (Maxwell et al. 2010). Recently, the availability of real-time information to dispatchers via geographical information systems and the affordability of computing power has facilitated using real-time ambulance management, which provides a platform that enables EMS providers to consider more sophisticated operational strategies to improve the performance of ambulance deployment policies. One potential strategy is ambulance relocation, which refers to repositioning idle ambulances in real time to better respond to future calls. It is possible for some locations to be covered by more than one ambulance; therefore, some ambulances might be idle at their locations while providing no additional coverage value. Note that an area is covered if an idle ambulance can reach it in a specific time threshold. Repositioning these ambulances to improve the coverage level is a strategy we call "ambulance reallocation." This strategy can improve the performance of

EMS systems because idle ambulances at other locations can compensate for a "coverage hole" caused by dispatching the only ambulance covering a region. A second type of strategy is to send an ambulance that just finished service to a new location rather than sending it to a preassigned base, which we call "ambulance redeployment." A third potential strategy is to decide which ambulance should serve a call (if immediately), which we call "ambulance dispatching," as sending the closest ambulance for every incident may be suboptimal (Swersey 1994). This strategy can significantly improve the performance of the system, as supported by our numerical study. For example, suppose a high-priority call arrives and the closest ambulance is 20 minutes away from the call location; however, another ambulance is currently in service just two minutes away from the call location and will be available in three minutes. Sending the closest ambulance immediately will result in a late call in this example and is shown to be suboptimal in realistic settings. The aim of this work is to develop a flexible mathematical framework to explore a variety of strategies for real-time ambulance operations management. Pursuant to this goal, we formulate the problem as a stochastic dynamic program and use an approximate dynamic programming (ADP) approach to produce efficient real-time dispatching and relocation policies.

1.2. Main Contributions and Results

In this study, we make the following contributions:

1. We develop a flexible optimization framework by simultaneously considering general dispatching, redeployment, and reallocation strategies for real-time stochastic dynamic ambulance operations management. We consider a general dispatching rule upon receiving a call in that the decision maker can send any available ambulance in addition to not serving the call immediately. Therefore, we let the model decide which ambulance should immediately be dispatched to a received call or the call has to wait for an ambulance in the near future. EMS providers are also motivated to spread out ambulances on the roads to meet the performance standards. To improve coverage level, we consider relocating an available ambulance to the location of an ambulance just dispatched to serve a call, and we name this "ambulance reallocation."

2. To assess the quality of solutions produced by our ADP approach, we develop a novel lower bound on the expected response time of any relocation policy. To create this bound, we consider a lower bounding system as in Maxwell et al. (2014). However, instead of solving a maximum covering location problem (MCLP), upon receiving a call in the original system, we reposition the available ambulances to minimize the expected response time by solving a different p-median integer program.

3. We develop new basis functions that estimate the expected response time of the calls in the system and modify some of the available basis functions in the literature to enhance the performance of the ADP policies. In particular, we introduce new basis functions that estimate a future state of the system in which a busy ambulance becomes available, thus enabling the ADP algorithm to react to the future coverage level. These basis functions serve an important purpose in that they allow the algorithm to delay or alter a dispatching or relocation decision in response to a situation by considering future costs of an appropriate response when a new ambulance configuration has emerged.

4. We measure the contribution of each strategy in terms of a variety of objective functions such as the expected discounted priority-adjusted response time and the expected discounted priority-adjusted fraction of late calls. We discover insights regarding the relative contribution of each strategy, as well as available benchmarks.

We test the performance of six static benchmarks in the literature on our data set to find the best static benchmark in terms of expected response time and fraction of late calls. Our analysis shows that the maximum expected covering location problem (MEXCLP) and maximum covering location problem (MCLP) outperform other static benchmarks when the objective is to minimize the expected fraction of late calls and the expected response time, respectively. Thus, the static policy, hereafter, refers to the MEXCLP (MCLP) when the objective is to minimize the fraction of late calls (response time).

In addition, we consider five dynamic benchmarks, including a heuristic that has been reported to be efficient in the literature. To analyze the contribution of each dispatching, redeployment, and reallocation strategy on performance improvement, we design three dynamic benchmarks by adding each strategy to the static policy one at a time; that is, Benchmark 1 builds on the static policy by considering a general dispatching rule instead of sending the closest available ambulance; Benchmark 2 builds on the static policy by considering a redeployment strategy after an ambulance has finished serving a call; Benchmark 3 builds on the static policy by sending an available ambulance to the location of an ambulance just dispatched to serve a call; Benchmark 4, which consists of redeployment and reallocation strategies, is used to compute the optimality gap as both lower bounds assume the closest ambulance is dispatched; and Benchmark 5 uses the dynamic heuristic relocation policy proposed by Jagtenberg et al. (2015) to evaluate its performance with respect to our relocation strategies.

Our results show that when the ADP objective function is to minimize the expected response time, ADP

policies generated by Benchmarks 1, 2, and 3 outperform the MCLP static benchmark by 2.7%, 6.8%, and 1.3%, respectively. However, when the ADP objective function is to minimize the expected fraction of late calls, the ADP policies produced in Benchmarks 1, 2, and 3 outperform the MEXCLP static benchmark by 13.5%, 21.3%, and 9.8%, respectively. This shows that each strategy can significantly improve the static benchmarks. Note that the contribution of a redeployment strategy is significantly greater than those of dispatching and reallocation strategies. Also, this observation is consistent in both ADP objective functions, i.e., minimizing the response time and fraction of late calls. Furthermore, the expected frequency with which Benchmark 2 deviates from the static benchmarks is significantly greater than that of the other dynamic benchmarks. The ADP approach that simultaneously considers all three strategies outperforms both the static and dynamic benchmarks. In particular, when the ADP objective is to minimize the expected fraction of late calls, our ADP approach outperforms Benchmark 2 in expected response time and fraction of late calls by 4.3% and 14.5%, respectively. Benchmark 2 is similar to the setting studied in Maxwell et al. (2010), where only ambulance redeployment is considered. Our results suggest that expanding the action space beyond redeployment can significantly improve the performance of the system, e.g., 14.5% improvement in the fraction of late calls when the ADP objective is to minimize the fraction of late calls. Furthermore, Benchmark 1, which uses a general dispatching rule, provides novel insights to the discussion around the optimality of policies that deviate from sending the closest available ambulance. Our results show that Benchmark 1 simultaneously reduces the fraction of late calls and response time, as our ADP policy shifts the entire response time distribution toward shorter times (see Figure 1). This result is different from that in Jagtenberg et al. (2016), where deviating from sending the closest ambulance resulted in an improvement on fraction of late calls, but significant increase in response time.

2. Related Work

The literature on ambulance operations management is quite rich. Therefore, we briefly discuss previous works related to our study and refer the reader to the following survey papers and the references therein for a comprehensive review. Swersey (1994) and Brotcorne et al. (2003) reviewed deterministic and probabilistic ambulance location and relocation models. Also, Ingolfsson (2013) provided a survey on the analytical stochastic models focusing on ambulance station selection and ambulance allocation to stations with respect to performance measures such as response time.

Early models of the ambulance location problem sought to minimize the number of ambulances required to respond to future calls for a determined time threshold or to maximize the demand covered using a fixed fleet size; see Church and ReVelle (1974) and the references therein. These approaches did not consider the fact that when an ambulance is dispatched, the coverage level might fall below a minimum threshold. One possibility to address the unavailability of dispatched ambulances over time includes considering multiple coverage, i.e., demand points that are supposed to be covered by more than one vehicle. Gendreau et al. (1997) introduced the double standard model (DSM) by including multiple coverage. Doerner et al. (2005) extended their work with respect to capacity constraints and different demand density in each location. Gendreau et al. (2001, 2006) developed dynamic models to formulate the ambulance repositioning problem, where the objective function is to maximize the total covered demand. Because these approaches require solving an integer program every time the dispatcher makes a decision, they are computationally very intensive. Also, these models are deterministic and do not capture the effect of randomness in the system.

Berman (1981a, b) used Markov decision theory to minimize the long-run cost of repositioning ambulances. They provided an exact dynamic programming approach to find available ambulances to compensate for coverage level drop induced by dispatched ambulances. However, these exact formulations are tractable only in oversimplified settings for a small number of ambulances over a small network of routes. Restrepo et al. (2009) used an Erlang loss function to compute the fraction of late calls, those not responded to within a time threshold, and embedded it into an optimization model to minimize the percentage of late calls by static deployment of ambulances. McLay and Mayorga (2013) formulated the ambulance dispatching problem as a Markov decision process to optimally dispatch ambulances to prioritized patients. Alanis et al. (2013) developed a two-dimensional Markov chain model to analyze ambulance repositioning according to a compliance table, which suggested where to reposition an ambulance based on the number of available ambulances. (In this setting, the closest ambulance is dispatched to serve the call. After the ambulance finishes its service, the decision maker seeks to reposition ambulances in such a way that maintains a configuration of the available ambulances similar to the one suggested by the compliance table.) Andersson and Varbrand (2007) measured the ability of an ambulance to cover a future call by introducing a "preparedness function," which approximates the value function in a dynamic program. However, to apply it to real-time applications even with small sets of available ambulances and relocation destinations, the dynamic relocation

problem must be solved heuristically. van Barneveld et al. (2016) designed a heuristic dynamic repositioning policy by minimizing the "unpreparedness function," which returns the expected penalty that the next request generates. In that setting, the best relocation policy is found in terms of a "motion" from an origin base to a destination base. To prevent long transition times, a linear bottleneck assignment problem is solved to determine how available ambulances should move to reach the new configuration.

Mason (2013) developed a dynamic repositioning policy that relocates ambulances in demand zones when coverage levels drop below a threshold. However, repositioning idle ambulances every time coverage levels fall below a certain level may result in shortage of available ambulances at times of dispatch. Therefore, to limit the repositioning time, a neighborhood search strategy is developed to solve the base allocation problem, which leads to solutions that differ only slightly from the initial base locations. Jagtenberg et al. (2015) developed a heuristic approach to real-time ambulance relocation by maximizing the expected marginal contribution of each available ambulance to the coverage level. van Barneveld (2016) extended the MEXCLP to incorporate a nonnegative, nondecreasing function of response time into the objective function to calculate performance measures related to response time instead of coverage level. By solving the extended formulation for different levels of available ambulances, compliance tables are obtained offline, and when the number of available ambulances changes, an assignment problem is carried out to reconfigure the system, i.e., move the ambulances to new positions according to the compliance tables. Sudtachat et al. (2016) modified the steady-state probabilities calculated by Alanis et al. (2013) and incorporated them into an integer program to maximize the coverage level in a single type ambulance and call priority system with a zero-length queue. The resulting nested compliance policy, when out of compliance, requires at most one vehicle movement at a time to reconfigure accordingly. Bélanger et al. (2016) modified the double standard model to consider multiperiod and dynamic settings with and without relocations. In the multiperiod settings, ambulances are relocated only between periods and returned to the same base throughout the period. In the dynamic settings, the double standard model is solved whenever an ambulance is dispatched if a certain amount of time has passed since the last relocation and a secondary coverage level falls below a threshold.

To make ambulance redeployment decisions in an uncertain dynamic setting, Maxwell et al. (2010) developed an ADP approach based on approximate policy iteration. They formulated the ambulance redeployment problem as a dynamic program and approximated the value function by an affine combination of basis functions. They used an iterative simulation-based procedure to estimate tunable parameters of the approximation. The objective was to minimize the fraction of late calls only through redeployment. Their model, however, does not consider ambulance reallocation and uses a myopic dispatching rule; i.e., the closest ambulance is sent to a call, calls are served in decreasing order of priority, and a first-come, first-served strategy is considered for each priority. Schmid (2012) also used ADP to model real-time ambulance dispatching and relocation. Our study is different in both the problem scope and methodology used. In particular, Schmid (2012) did not consider the relocation of idle ambulances and assumed that an ambulance must be immediately dispatched to a call. In terms of ADP, Schmid (2012) used a general ADP framework based on aggregation and postdecision states. However, we develop an ADP approach specific to ambulance operations management by exploiting novel basis functions, as well as developing a lower bound on expected response time.

There is another stream of research related to our work: the literature on approximate dynamic programming. Many researchers have used ADP to come up with high-quality solutions for a variety of applications, e.g., allocating resources in service systems (Adelman 2007), resource allocation in healthcare (Bertsimas et al. 2013, Khademi et al. 2015), and supply chain management (Lai et al. 2010, Van Roy et al. 1997).

3. Problem Formulation

This section presents an infinite-horizon Markov decision process formulation of the problem. Let $\mathscr{L} := \{0, 1, 2, \ldots, L\}$ be the set of call locations, and let $\mathscr{B} := \{0, 1, 2, \ldots, B\}$ be the set of all ambulance bases. We assume a total of N ambulances are available and at most J calls are tracked. This assumption is not restrictive because one may consider a large J.

3.1. State Space

An ambulance i is represented by $m_i = (f_i, o_i, d_i, t_i)$, where f_i is the status of the ambulance, o_i is the original location of the ambulance, d_i is the destination of the ambulance, and t_i is the start time of the latest movement of the ambulance. For the purposes of this work, it is sufficient to consider six possibilities for the status of an ambulance, i.e., $f_i \in \{0, 1, 2, 3, 4, 5\}$, where 0 shows that the ambulance is available at base, 1 shows that the ambulance is going to a call location, 2 shows that the ambulance is serving a call on the scene, 3 shows that the ambulance is going to the hospital, 4 shows that the ambulance has finished serving a call, and 5 shows that the ambulance is being reallocated and going to another base or the ambulance is going to a base after finishing service. Note that if ambulance i

is idle in a location, the original location is set to the current location and the destination to null. Similarly, when an ambulance is serving a call on scene, we set the original location to the call location and destination to null. We let vector $m = (m_1, m_2, \ldots, m_N) \in M$ represent the state of all ambulances. A call j is represented by $c_j = (g_j, l_j, p_j, q_j)$, where g_j is the status, l_j is the location, p_j is the priority, and q_j is the arrival time of the call. In particular, $g_j \in \{0, 1\}$, where 0 shows that the call is waiting for service, and 1 shows that the call is assigned to an ambulance. When an ambulance reaches the call scene, the call is removed from the list. Aligned with literature, we consider two priority levels for a call, $p_j \in \{0, 1\}$, where 0 shows that the priority of a call is low, and 1 shows that the priority of the call is high (Maxwell et al. 2010). Extending the framework of this study to consider more priority levels is straightforward. We let vector $c = (c_1, c_2, \ldots, c_J) \in C$ represent the state of all calls.

Without loss of generality, we assume that decisions are made at transition times. In our model, transition times are associated with the following events: call j arrives, ambulance i is in transit to call j, ambulance i arrives at the location of call j, ambulance i is finished serving call j at scene, ambulance i is finished serving call j at hospital, and ambulance i arrives at a base. Let E be the set of all possible events. Therefore, the state space of the system is represented by $S := \{s = (\tau, e, m, c)\colon e \in E,\ m \in M,\ c \in C\}$, where τ corresponds to the current time.

3.2. Action Space

The action space is described in four cases. We assume that dispatching, reallocating, and redeploying the ambulances are nonpreemptive. One can relax this assumption by defining an event "consider preemption," which occurs with a certain frequency, and upon occurrence, one may preempt any of the ambulance services and reconsider actions. However, Maxwell et al. (2010) showed that considering only the service preemption of ambulances that are returning to base significantly increases the computational effort, while its benefit may be marginal.

Case 1. If call j arrives, the decision maker has two types of decision: (i) which ambulance should be immediately dispatched to serve the call (if any) and (ii) which ambulances should be reallocated to other bases (if any). Note that in this case, an ambulance is not necessarily dispatched upon receiving a call immediately. If this happens, the call will join the queue and will be served later. The rationale for considering reallocation decisions is that by spreading out the ambulances over the area, it is likely that a location is covered by only one ambulance; thus, sending the ambulance to a call may cause a coverage hole. Since coverage level does not decrease unless an idle ambulance becomes unavailable, reallocation decisions are considered when an ambulance is dispatched. Because multiple reallocations in short intervals are expensive and could become a burden on the ambulance crew (van Barneveld et al. 2016, Jagtenberg et al. 2015), we assume that reallocations are limited to at most one ambulance upon dispatching an ambulance. Let $\mathcal{M}(s)$ be the set of available ambulances, i.e., $\mathcal{M}(s) := \{i\colon f_i = 0\}$; let $\mathcal{B}_1(s)$ represent the location of the ambulance just dispatched to a call; and let $\mathcal{M}_1(s)$ represent the set of all available ambulances right after dispatching ambulance i when the state is s. If no ambulance is dispatched upon receiving a call, we set $\mathcal{B}_1(s) = \varnothing$ and do not consider ambulance reallocation. Note that when ambulances are in transit toward a base ($f_i = 5$), they are not considered available because of the nonpreemption assumption. It is possible to use the event "consider preemption" to preempt an ambulance that is moving toward a base and dispatch it to a call. However, the benefit of such preemptions may be marginal (Maxwell et al. 2010).

Define $X_{i,j} = 1$ if ambulance i is assigned to call j and $X_{i,j} = 0$ otherwise. Also, define $Y_{i,b} = 1$ if ambulance i is reallocated to location b and $Y_{i,b} = 0$ otherwise. Therefore, if event e is of the type "a call arrives," and $\mathcal{J}_1(s)$ denotes a set that points to the index of the call, the action space is given by

$$A_1(s) := \Bigg\{(X_{i,j}, Y_{i,b})\colon \sum_{i \in \mathcal{M}(s)} X_{i,j} \le 1,\ j \in \mathcal{J}_1(s);\ \sum_{i \in \mathcal{M}_1(s)} Y_{i,b} \le 1,\ b \in \mathcal{B}_1(s)\Bigg\},$$

where the first constraint ensures that at most one ambulance is assigned to a received call j, and the second constraint ensures that at most one ambulance is reallocated to the location of the dispatched ambulance.

Case 2. Let $\mathcal{Q}(s)$ denote the set of all calls waiting in the queue for ambulance assignment; i.e., $\mathcal{Q}(s) := \{j\colon g_j = 0\}$. If $\mathcal{Q}(s) = \varnothing$ and event e is of the type "ambulance i has finished serving a call at scene" or "ambulance i has finished serving a call at hospital," the decision is where to redeploy the ambulance. Let $\mathcal{M}_2(s)$ denote the set of ambulances available for redeployment in state s, and $Z_{i,b} = 1$ if ambulance i is redeployed to location b, and $Z_{i,b} = 0$ otherwise. We set $\mathcal{M}_2(s) = \{i\}$ in this case. The action space is presented by

$$A_2(s) := \Bigg\{(Z_{i,b})\colon \sum_{b \in \mathcal{B}} Z_{i,b} = 1,\ i \in \mathcal{M}_2(s)\Bigg\},$$

where the constraint ensures that ambulance i is redeployed to only one location.

Case 3. If $\mathcal{Q}(s) \neq \varnothing$ and event e is of type "ambulance i has finished serving a call at scene," "ambulance i has finished serving a call at hospital," or "ambulance i has arrived at a base," the decision is to dispatch an available ambulance to a call in the queue or to redeploy it

to a location. If there is more than one call in the queue, the calls are served in decreasing order of priority, and within a given priority level, they are served based on a first-come, first-served rule. Let $\mathcal{J}_3(s)$ denote the set that points to the highest-priority call with the longest waiting time in the queue, and let $\mathcal{M}_3(s)$ denote the set of available ambulances for redeployment when the state is s. Set $\mathcal{M}_3(s) = \{i\}$ in this case. The action space is given by

$$A_3(s) := \Big\{(X_{i,j}, Z_{i,b}): \sum_{i\in\mathcal{M}(s)} X_{i,j} \le 1,\ j \in \mathcal{J}_3(s);\ \sum_{b\in\mathcal{B}} Z_{i,b} = 1 - X_{i,j},\ i \in \mathcal{M}_3(s),\ j \in \mathcal{J}_3(s)\Big\},$$

where the first constraint considers dispatching an ambulance to call j in the queue, and the second constraint ensures that the ambulance that has just become available will be redeployed to a location if it is not already assigned to a call.

Case 4. If an event is of type "ambulance i is in transit to call j" or "ambulance i arrives at the location of call j," we set $A(s) = \varnothing$.

3.3. Transitions

We assume that call arrivals in location l follow a nonhomogeneous Poisson process with rate λ_l^τ at time τ. If an ambulance arrives at call j scene, it completes the service at the scene with probability ρ_j, and it will transfer the patient to a hospital with probability $1 - \rho_j$. We assume that travel times are deterministic, and the time required to serve a call at scene or taking a patient to hospital follows an arbitrary distribution with a finite mean, independent of call location. Note that if the destination is a hospital, in addition to travel time, our historical data also consider both the service time on scene before going to hospital and the time that it takes to hand over the patient to hospital personnel. We estimate all of the distributions using historical data from Mecklenburg County, North Carolina. Let s_κ be the state of the system when the κth event happens. The evolution of state s_κ can then be characterized by action a_κ, a random element $\omega(s_\kappa, a_\kappa)$, and a function F, i.e., $s_{\kappa+1} = F(s_\kappa, a_\kappa, \omega(s_\kappa, a_\kappa))$.

3.4. Objective Function

We consider minimizing the expected discounted priority-adjusted total response time and the expected discounted priority-adjusted fraction of late calls as the primary ADP objective functions for the optimization framework. We also report other performance measures such as response time of late calls and fraction of late high-priority calls in our case study. Let $h(s_\kappa, a_\kappa, s_{\kappa+1})$ denote the cost of a transition from s_κ to $s_{\kappa+1}$, when action a_κ is taken. To minimize the expected discounted priority-adjusted response time, define

$$h(s_\kappa, a_\kappa, s_{\kappa+1}) = \begin{cases} w_1(\tau(s_{\kappa+1}) - q_j) \\ \quad \text{if a high-priority call } j \text{ arrives and} \\ \quad \text{the event } e(s_{\kappa+1}) \text{ is of the form} \\ \quad \text{“ambulance } i \text{ arrives at the scene of call } j\text{,”} \\ w_2(\tau(s_{\kappa+1}) - q_j) \\ \quad \text{if a low-priority call } j \text{ arrives and} \\ \quad \text{the event } e(s_{\kappa+1}) \text{ is of the form} \\ \quad \text{“ambulance } i \text{ arrives at the scene of call } j\text{,”} \\ 0 \quad \text{otherwise,} \end{cases}$$

where $(\tau(s_{\kappa+1}) - q_j)$ measures the response time of call j, and w_1 and w_2 are priority adjustment weights. This cost structure is flexible in that w_1 and w_2 can be tuned to capture the relative importance of high-priority versus low-priority calls.

Similarly, to minimize the long-run priority-adjusted fraction of late calls, define

$$h(s_\kappa, a_\kappa, s_{\kappa+1}) = \begin{cases} w_3(\mathbb{1}_{\{\tau(s_{\kappa+1})-q_j\ge\triangle\}}) \\ \quad \text{if a high-priority call } j \text{ arrives and} \\ \quad \text{the event } e(s_{\kappa+1}) \text{ is of the form} \\ \quad \text{“ambulance } i \text{ arrives at the scene of call } j\text{,”} \\ w_4(\mathbb{1}_{\{\tau(s_{\kappa+1})-q_j\ge\triangle\}}) \\ \quad \text{if a low-priority call } j \text{ arrives and} \\ \quad \text{the event } e(s_{\kappa+1}) \text{ is of the form} \\ \quad \text{“ambulance } i \text{ arrives at the scene of call } j\text{,”} \\ 0 \quad \text{otherwise,} \end{cases}$$

where $\triangle$ denotes the given time threshold, and $\mathbb{1}_{\{\tau(s_{\kappa+1})-q_j\ge\triangle\}}$ is an indicator function that takes a value of one if the call is not responded to within the time threshold. The cost structure can capture the relative importance of call priorities by tuning w_3 and w_4. Note that one might use different time thresholds for different priorities.

3.5. Optimality Equation

Let $J_\pi(s)$ denote the expected total discounted cost when $s_0 = s$ under policy $\pi \in \mathcal{P}$, where $\mathcal{P}$ denotes the set of all stationary nonanticipative policies; that is,

$$J_\pi(s) = \mathbb{E}\Big\{\sum_{\kappa=1}^{\infty} \gamma^{\tau(s_\kappa)} h(s_\kappa, \pi(s_\kappa), s_{\kappa+1}) \,\Big|\, s_0 = s\Big\}, \quad s \in S,\ \pi \in \mathcal{P},$$

where $\pi(s_\kappa)$ denotes the action selected by policy π in state s_κ at time $\tau(s_\kappa)$, and $0 \le \gamma < 1$ is a discount factor. The decision maker solves for $v(s) = \inf_{\pi\in\Pi}\{J_\pi(s)\}$, where $\Pi \subseteq \mathcal{P}$ denotes the set of admissible policies under consideration and $v(s)$ satisfies the Bellman optimality equation

$$v(s) = \min_{a\in A(s)} \big\{\mathbb{E}_a(h(s, a, s') + \gamma^{(\tau(s')-\tau(s))} v(s') \mid s)\big\}, \quad \forall\, s \in S, \qquad (1)$$

where the expectation is taken with respect to action a and $s' = F(s, a, \omega(s, a))$ (Puterman 2005). Moreover, a stationary optimal policy exists, which is myopic relative to the optimal value function.

4. Approximate Solutions and Performance Guarantee

Solving formulation (1) to optimality is impractical because of the curse of dimensionality. The state space of the system, S, is unbounded, and the traditional methods do not apply. Section 4.1 adapts approximate policy iteration to produce high-quality solutions, which provide an upper bound on the optimal value function, and Section 4.3 computes lower bounds on the long-run fraction of late calls and response time under any relocation strategy to assess the quality of the solutions.

4.1. Upper Bound

The standard policy iteration algorithm starts with an arbitrary policy π^0. At iteration n, it evaluates π^n by calculating $v^n(s)$ for all $s \in S$ by solving $v^n(s) = L_{\pi^n} v^n(s)$, where $L_{\pi^n} v^n(s) = \mathbb{E}\{h(s, a, s') + \gamma^{\tau(s')-\tau(s)} v(s')\}$. Next, it improves the policy by choosing a myopic policy relative to v^n, i.e., $\pi^{n+1}(s) \in \arg\min_{d \in D^{MD}} \{\mathbb{E}(h(s, a, s') + \gamma^{\tau(s')-\tau(s)} v^n(s'))\}$, where $d \in D^{MD}$ denotes a decision rule in the set of stationary Markovian deterministic policies (D^{MD}). This iterative procedure is continued until $\pi^{n+1} = \pi^n$ (Puterman 2005). Because the state space is unbounded, the policy evaluation and improvement steps are intractable in this problem. To overcome this issue, the value function is approximated by an affine combination of basis functions, i.e., $v(s) \approx \hat{v}(s) = \alpha_0 + \sum_{k=1}^{K} \alpha_k \phi_k(s)$, where each $\phi_k(s)$ is a basis function, and α_k is its associated weight in the approximation. The quality of the approximation depends on the choice of basis functions, which should be able to characterize the optimal value function (Powell 2007). Section 4.2 discusses our choice of basis functions in detail. Therefore, by replacing the value function with its approximation, the policy improvement step at iteration n of the approximate policy iteration involves solving

$$\pi^n(s) \in \arg\min_{a \in A(s)} \left\{ \mathbb{E}_a \left(h(s, a, s') + \gamma^{(\tau(s')-\tau(s))} \hat{v}(s') \mid s \right) \right\}, \quad \forall s \in S, \qquad (2)$$

where $\mathbb{E}_a(\cdot)$ denotes the expectation with respect to action a. We use Monte Carlo simulation to approximate $\mathbb{E}_a(\cdot)$ via a sample average. Starting from state s and taking action a, we simulate the system for one step and use $\hat{v}(s)$ as the cost-to-go estimate. Because the simulation is evaluated only until the next event, enumerating all trajectories is manageable. The next event could be a call arrival or a busy ambulance completing one stage of its transition, which is either reaching a call scene, finishing service (at scene or hospital), or arriving at a location after finishing service.

Solving formulation (2) involves enumerating all actions for a given state. In our setting, this is manageable because if the event is "call j arrives," the decision maker has to determine which ambulance should be immediately dispatched to the call (if any) and which ambulance should be reallocated to the location of the ambulance just dispatched (if any). Let $|\mathcal{M}(s)|$ denote the number of available ambulances. The size of the action set will be $1 + |\mathcal{M}(s)|(|\mathcal{M}(s)| - 1)$. If the event is of the type "ambulance becomes available after serving a call" and no calls are in the queue, then the decision maker determines which location the ambulance should be redeployed to. This is equal to the number of locations, denoted by $|\mathcal{B}|$, which in our case study is 40. If there are calls in the queue and the event is of type "ambulance becomes available after serving a call" or "ambulance has just arrived at its location," then the decision maker determines which ambulance to dispatch to the call based on a first-come, first-served rule, and if the decision is not to dispatch, which location the ambulance should be redeployed to, which is at most $|\mathcal{M}(s)| + |\mathcal{B}|$. Once the expectation is estimated for all actions, the decision that yields the smallest value is chosen by the policy. Formulation (2) provides $\hat{v}$-improving decision rules for a fixed state s. However, solving it for each state is not possible because the state space is unbounded. Therefore, to evaluate a policy, we use formulation (2) upon visiting a state on the fly in the Monte Carlo simulation; that is, it is solved only for states observed in simulation. In the settings of interest, our computational experiments demonstrate that solving formulation (2) for a state is instantaneous.

Algorithm 1 (Approximate policy iteration)
Set $n = 0$, $\epsilon > 0$ and $\alpha = \alpha^0$.
while $|\hat{v}^n(s) - \hat{v}^{n-1}(s)| > \epsilon$ or $n \neq 0$ **do**
- Policy improvement: Find a myopic policy induced by $\hat{v}^n(s)$ by solving formulation (2).
- Policy evaluation: Use Monte Carlo to simulate the system; find actions for each state visited by the simulation via solving formulation (2); calculate $C^r(s)$, the total discounted cost for each initial state s and replication r.
- Projection: Use $C^r(s)$ from the Monte Carlo simulation and solve formulation (3) to estimate α^{n+1} for the next iteration.
- Set $n \leftarrow n + 1$.

Next, we develop an algorithmic approach to estimate $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_k)$ and consequently derive high-quality solutions. Consider an appropriately large finite horizon, initialize $\alpha = \alpha^0$, and evaluate the policy associated with it for states in $\hat{S}$, where $\hat{S}$ is a subset of S. We construct $\hat{S}$ by sampling states that are more likely to be visited by the optimal policy and

update $\hat{v}^n(s)$ at iteration n of the algorithm (de Farias and Van Roy 2004). For the policy evaluation step, we propose the following procedure. Start from an initial state s, use Monte Carlo simulation to simulate the system, and upon observing a state, find actions by solving formulation (2), and calculate the total discounted cost for that realization of the system. Let $C^r(s)$ denote the total discounted cost of the realization of the system, starting from state s in replication r, i.e., the simulated value function for state s in replication r. Let R_s be the total number of replications of the Monte Carlo simulation for state s. To estimate α, solve the following optimization problem:

$$\min_{\alpha} \sum_{s\in\hat{S}} \sum_{r=1}^{R_s} \left(C^r(s) - \alpha_0 - \sum_{k=1}^{K} \alpha_k \phi(s) \right)^2, \tag{3}$$

which minimizes the squared error between the approximate value function and the simulated value function. Note that formulation (3) is a regression model, and computational experiments in our case study show that solving it is instantaneous. This procedure continues until convergence in some norm is achieved. Algorithm 1 formalizes this approach.

4.2. Basis Functions

This section describes the basis functions $\{\phi_k(\cdot)\colon k = 1,\dots,K\}$ used for the value function approximation. Jagtenberg et al. (2015) noted that basis functions in the ambulance relocation literature may not produce high-quality solutions in general, but our computational results show that our ADP approach based on the following basis functions produces near-optimal solutions.

4.2.1. Response Time. This novel basis function estimates the expected response time of a call when the state of the system is s. To that end, let $r_l(s)$ denote the expected response time of a call in region l in state s, and set $\phi_1(s) = \sum_{l\in\mathcal{L}} \lambda_l^\tau r_l(s)$. Response time is comprised of travel time of an ambulance to reach a call plus potential waiting time of a call in queue for ambulance assignment. To estimate the expected waiting time of a call in a region, we develop an $M/G/c$ queueing system for each region and estimate the expected waiting time in the queue. Let $N_l(s)$ denote the number of available ambulances (the number of servers in the $M/G/c$ queuing system) that cover location l when the state of the system is s. We consider a region covered by ambulance i, if the time that it takes for an available ambulance to reach to the center of the region, $\bar{l}$, is less than a threshold $\triangle$. Therefore, $N_l(s) = \sum_{i\in\mathcal{M}(s)} \mathbb{1}_{\{t(o_i,\bar{l})\le\triangle\}}$, where $t(x,y)$ denotes the travel time between locations x and y. We estimate the service rate of an ambulance in region l, $\mu_l(s)$, by considering the average travel time in region l plus the average time that an ambulance spends on scene plus the average time of handing over a patient to hospital personnel. Because ambulances in a region may also serve other regions, we adjust the arrival rate of calls in region l by summing over call arrival rates of regions covered by an available ambulance in region l, using $\lambda'_{l,\tau}(s) = \sum_{u\in\mathcal{L}} \sum_{i\in\mathcal{M}_l(s)} \lambda_u^\tau \mathbb{1}_{\{t(o_i,\bar{u})\le\triangle\}}$, where $\mathcal{M}_l(s)$ denotes the set of available ambulances that cover region l; i.e., $\mathcal{M}_l(s) := \{i\in\mathcal{M}(s)\colon t(o_i,\bar{l})\le\triangle\}$, and λ_l^τ is the call arrival rate in region l at time τ. We use $\mu_l(s)$ and $\lambda'_{l,\tau}(s)$ to compute the expected waiting time in queue in an $M/M/c$ queuing system in region l, i.e., $W^{q,l}_{(M/M/c)}(s)$. The expected waiting time in the queue in an $M/G/c$ queuing system is then approximated by

$$W^{q,l}_{(M/G/c)} \approx \frac{1+cv_l^2}{2} W^{q,l}_{(M/M/c)},$$

where cv_l denotes the coefficient of variation of the service time in region l (Allen 1980). Let $\bar{t}^l$ denote the average travel time within region l; then,

$$r_l(s) = \mathbb{1}_{\{\mathcal{M}_l(s)=\varnothing\}} \left[W^{q,l}_{M/G/c}(s) + \bar{t}^l \right] + \mathbb{1}_{\{\mathcal{M}_l(s)\neq\varnothing\}} \left[\min_{i\in\mathcal{M}_l(s)} t(o_i,\bar{l}) \right].$$

Note that the queuing theory approach may not accurately estimate arrival rates, service rates, and the number of the servers for a region. To overcome this issue, we calibrate the model by scaling the arrival rates and find the scaling factor through experimentation.

4.2.2. Future Response Time. Ambulances in transit can serve a (currently in queue or future) call after their service is finished. Therefore, the destinations of the busy ambulances are as important as their current locations. This is the underlying motivation for the second and the fourth basis functions. Let $\vec{s}$ denote the state that corresponds to the earliest time when one of the following events occur: an ambulance finishes serving a call (at scene or hospital) or an ambulance arrives at a base. The future response time in state $\vec{s}$ is important because it evaluates the trade-off between immediate and future cost. Given that the current state of the system is $s = (\tau, e, m, c)$, we construct a new state $\vec{s}(s) = (\vec{\tau}(s), \vec{e}(s), \vec{m}(s), \vec{c}(s))$, where $\vec{\tau}(s)$ denotes the time that the future state $\vec{s}$ will be visited, and $(\vec{e}(s), \vec{m}(s), \vec{c}(s))$ denotes predicted future event, ambulance status, and call status at time $\vec{\tau}(s)$. Also, $\vec{s}(\cdot)$ is determined by searching the earliest time that a busy ambulance becomes available. Predicting future events, ambulance statuses, call statuses, and the earliest time that a busy ambulance becomes available is possible by searching the future event list in the simulation. We then set $\phi_2(s) = \phi_1(\vec{s})$. This basis function is novel in that $\vec{s}$ computes the state that corresponds to the earliest time that an ambulance becomes available, compared to Maxwell et al. (2010), where the future state is computed by replacing the locations of all busy ambulances with their destinations.

4.2.3. Uncovered Call Rate. The third basis function computes the rate of uncovered calls. Recall that $N_l(s)$ is the number of available ambulances in region l in state s, and calls arrive with rate λ_l^τ from location l at time τ. If no ambulance covers region l, then the call may be late (Restrepo et al. 2009). We define the uncovered call rate by $\phi_3(s) = \sum_{l\in\mathscr{L}} \lambda_l^\tau \mathbb{1}_{\{N_l(s)=0\}}$.

4.2.4. Future Uncovered Call Rate. The fourth basis function calculates the uncovered call rate for a future state $\vec{s}$, which is constructed in the same way discussed in the second basis function, i.e., $\phi_4(s) = \phi_3(\vec{s})$.

4.2.5. Unreachable Calls. The fifth basis function computes the number of calls for which an ambulance is assigned but cannot reach the scene within the time threshold $\triangle$, i.e.,

$$\phi_5(s) = \sum_{j=1}^{J} \mathbb{1}_{\{g_j=1\}} \sum_{i=1}^{N} \mathbb{1}_{\{f_i=\text{“ambulance } i \text{ is going to call scene } j\text{”}\}} \cdot \mathbb{1}_{\{t_i+t(o_i,\bar{l})-g_j\geq\triangle\}},$$

where t_i is the time that ambulance i started to move to the scene of call j. The above expression first checks whether the call is assigned to an ambulance and then checks whether the ambulance will fail to reach the call location within the time threshold (Maxwell et al. 2010).

4.2.6. Aggregated Delay Time. This novel basis function computes the aggregated delay time for calls in the queue for which an ambulance is assigned but is not going to reach to the call scene within the time threshold $\triangle$, i.e.,

$$\phi_6(s) = \sum_{j=1}^{J} \mathbb{1}_{\{g_j=1\}} \sum_{i=1}^{N} \mathbb{1}_{\{f_i=\text{“ambulance } i \text{ is going to call scene } j\text{”}\}} \cdot \mathbb{1}_{\{t_i+t(o_i,\bar{l})-g_j\geq\triangle\}}(t_i + t(o_i,\bar{l}) - g_j \geq \triangle).$$

The indicator function $\mathbb{1}_{\{f_i=\text{“ambulance } i \text{ is going to call scene } j\text{”}\}} \cdot \mathbb{1}_{\{t_i+t(o_i,\bar{l})-g_j\geq\triangle\}}$ ensures that only late calls are counted.

4.3. Lower Bound

This section provides a lower bound on the expected total response time for a broad class of relocation policies over a finite time horizon. The bound is based on a lower bounding system in which the call arrival process is exactly the same as in the original system, and the number of available ambulances just before the arrival of a call is greater than or equal to that in the original system under policy π. This is achieved by computing a stochastic lower bound on the service time distribution of ambulances in the original system, which is independent of the ambulance configuration in an EMS system, and thus the relocation policies (Maxwell et al. 2014). Therefore, we simulate a multiserver queuing system (ambulances resemble servers and calls resemble customers) with the bounding service time distribution, where calls arrive according to the same process as the original system. However, just before the arrival of a call, the available ambulances are repositioned to minimize the expected response time by solving an integer program. Because we use the same bounding system as Maxwell et al. (2014), the same set of assumptions hold true.

Let D be the (random) number of calls over a horizon, and let T denote the (random) total response time over the same horizon. The goal is to compute a lower bound on $\mathbb{E}(T)$ independent of relocation policy π, which is given by

$$\mathbb{E}(T) = \mathbb{E}\left(\sum_{j=1}^{\infty} T_j \mathbb{1}_{\{j\leq D\}}\right) = \sum_{j=1}^{\infty} \mathbb{E}(\mathbb{1}_{\{j\leq D\}} \mathbb{E}[T_j \mid \mathscr{A}_j, \tau_j, \mathscr{C}_j]) \geq \sum_{j=1}^{\infty} \mathbb{E}(\mathbb{1}_{\{j\leq D\}} \nu(\mathscr{A}_j)) = \mathbb{E}\left(\sum_{j=1}^{D} \nu(\mathscr{A}_j)\right),$$

where T_j is the response time of the jth call, τ_j is the arrival time of the jth call, $\mathscr{C}_j$ is the configuration of ambulances at time τ_j, $\mathscr{A}_j$ is the number of available ambulance at time τ_j, and ν: $\{0,1,\ldots,\mathscr{A}\} \to [0,\infty]$ is a decreasing function such that $\mathbb{E}(T_j \mid \mathscr{A}_j, \tau_j, \mathscr{C}_j) \geq \nu(\mathscr{A}_j)$. Maxwell et al. (2014) constructed a bounding system by a coupling of the ambulance dynamics such that the number of available ambulances in the bounding system, $\tilde{\mathscr{A}}_j$, at the arrival time of the jth call satisfies $\tilde{\mathscr{A}}_j \geq \mathscr{A}_j$ for all j almost surely. Therefore, $\mathbb{E}(T) \geq \mathbb{E}(\sum_{j=1}^{D} \nu(\mathscr{A}_j)) \geq \mathbb{E}(\sum_{j=1}^{D} \nu(\tilde{\mathscr{A}}_j))$.

Having $\nu(\cdot)$ allows us to approximate the above expectation by simulating the bounding system. Let $\nu(\mathscr{A}_j)$ denote the minimum response time when $\mathscr{A}_j$ ambulances are available at the arrival time of the jth call. For $1 \leq \mathscr{A}_j \leq \mathscr{A}$, $\nu(\mathscr{A}_j)$ is the optimal objective function of the following integer program:

$$\begin{aligned}
\nu(\mathscr{A}_j) = \min \quad & \sum_{l=1}^{\mathscr{L}} d_l \sum_{k=1}^{|\mathscr{A}_j|} \sum_{b=1}^{\mathscr{L}} y_{kbl}\, t(b,l) && (4)\\
\text{s.t.} \quad & \sum_{k=1}^{|\mathscr{A}_j|} \sum_{b=1}^{\mathscr{L}} y_{kbl} = 1, \quad \forall\, l,\\
& y_{kbl} \leq x_{kb}, \quad \forall\, b,l,k,\\
& \sum_{b=1}^{\mathscr{L}} x_{kb} = 1, \quad \forall\, k,\\
& x_{kb} \in \{0,1\},\ y_{kbl} \in \{0,1\},\\
& \quad \forall\, b,l = 1,2,\ldots,\mathscr{L} \text{ and } \forall\, k = 1,2,\ldots,|\mathscr{A}_j|,
\end{aligned}$$

where y_{kbl} is an indicator taking a value of 1 if ambulance k is stationed at base b and is assigned to serve location l, x_{kb} takes a value of 1 if ambulance k is stationed at base b, $t(b,l)$ denotes the travel time between base b and location l, and d_l denotes the proportional call arrival rate in location l. The first constraint ensures that each location is served by exactly one ambulance, and the third constraint prevents an ambulance

from being located at different bases at the same time. Thus, formulation (4) seeks to minimize the expected response time to the demand. We set $\nu(0) = \nu(1)$. We also use the "cover bound" developed in Maxwell et al. (2014) to assess the quality of our solutions when the ADP objective function is minimizing the expected fraction of late calls.

5. Case Study: Mecklenburg County, North Carolina

This section presents the result of implementing our ADP framework using data from the EMS provider in Mecklenburg County, which contains the city of Charlotte and is the most populated and densely populated county in the state of North Carolina, with a population of over a million as of 2014 estimates (U.S. Census Bureau 2016). The EMS system in the county has, on average, 17 ambulances and three hospitals, and we consider 40 potential ambulance locations. We divide the county into 168 regions, where each region is a 2×2 mile square rectangle. As many EMS providers distribute ambulances along the road to meet the performance targets, we could consider all regions as potential locations for ambulances. However, to keep computations tractable, we limit the number of possible ambulance locations to 40 bases (see Figure 2 in the online companion). Section 5.1 provides further details on the choice of base locations, which serve as the main contributing factor in designing static benchmarks. We assume that all ambulances are the same and that turnout time (the activation delay needed for the ambulance crew to get ready and depart the base) is 45 seconds. We also assume that travel times are deterministic and estimate them based on historical data from more than 40,000 incidents. We divide a day into four time intervals (12:00 A.M.–06:00 A.M., 06:00 A.M.–12:00 P.M., 12:00 P.M.–06:00 P.M., and 06:00 P.M.–12:00 A.M.) and estimate the rates λ_i^τ using historical data. A few regions on the borders of the county had too few data points to fit a distribution and are excluded from the study. The amount of time that an ambulance spends serving a call at a scene has a normal distribution, with mean and standard deviation of 54.18 and 15.18 minutes, respectively. Historical data show that 77% of calls are transferred to a hospital. The service times of calls that are transferred to a hospital have a normal distribution, with a mean and standard deviation of 56.7 and 13.6 minutes, respectively. Note that this time includes the amount of time that an ambulance spends on the scene, travel time to hospital, and the time that it takes to hand over the patient to the hospital. A call not reached within eight minutes is considered to be late. The simulation horizon is two weeks, and we set $\gamma = 0.99$ per day. A sample of the 100 most visited states is used in formulation (3). Increasing the sample size to 200 and 500 states had minor effects on the results. We initialize the approximate policy iteration algorithm by setting $\alpha = (1, 1, 1, 1, 1, 1, 1)$ and $R_s = 5$ in each iteration of ADP. (Recall that R_s is the total number of replications of the Monte Carlo simulation for state s.) After a warm-up period under the best static benchmark, we begin collecting the statistics at the extant state when the warm-up period ends. Priority adjustment weights (w_1, w_2) and (w_3, w_4) are such that high-priority calls are 10 times more important than low-priority calls. Each iteration of ADP takes about two days of central processing unit time on an Intel Core i7 3.4 GHz processor with 16 GB of random access memory. However, this procedure is carried out offline, and after estimating an appropriate α, solving formulation (2) is instantaneous, which is what an EMS system needs for real-time ambulance management.

5.1. Choice of Static Benchmark

This section investigates the performance of several static benchmarks and considers the best in terms of response time and the best in terms of fraction of late calls, as benchmarks to dynamic policies. A static policy sends the closest available ambulance to a call and returns an ambulance after finishing service to its predetermined base if no call is in the queue. If no ambulances are available, the call will join a queue and will be served according to a first-come, first-served rule in a decreasing order of priority. In the absence of repositioning policies, identifying the base for each ambulance is the key to design efficient static benchmarks to ensure that a certain fraction of demand is reached within a specified response time target. Some models seek to maximize the fraction of demand covered by available ambulances, and others focus on minimizing the response time. We consider six frequently used models and refer the reader to van den Berg et al. (2016) for a complete description and formulation of each model. The MCLP maximizes the weighted number of demand locations covered by at least one ambulance. The DSM focuses on covering a demand location with two ambulances to prevent a coverage drop if an ambulance becomes busy. The DSM guarantees a certain level of coverage within the target response time for at least a fraction of demand and defines a second type of coverage with higher target response time that must be maintained for all demand locations. The MEXCLP maximizes the expected coverage of all demand locations by calculating the marginal contribution of each ambulance to coverage while considering that the ambulance might not be available with a certain probability, called the "busy fraction," which is calculated by dividing the priority-adjusted total workload of the system in minutes by total ambulance capacity in minutes. The maximum availability location problem (MALP) calculates the minimum number of available ambulances to guarantee a specific coverage level prior to formulating an instance of the model

and uses it to maximize the covered demand. The average response time model (ARTM) is equivalent to the p-median model applied to ambulance location problem and minimizes the average response time from the closest base. The expected response time model (ERTM) is similar to MEXCLP in that it minimizes the expected response time by incorporating the probability of a demand location being served by the pth nearest ambulance.

The initial bases of ambulances in the static benchmarks are determined by solving each model to optimality. The static benchmarks are simulated for two weeks, and their performance is measured with respect to fraction of late calls and average response time. Table 1 shows that the MEXCLP and MCLP outperform other models in the fraction of late calls and expected response time, respectively. Therefore, the performance of static benchmarks based on the MEXCLP and MCLP is used to compare with that of dynamic benchmarks and the ADP policy in terms of fraction of late calls and average response time, respectively.

5.2. Dynamic Benchmark Policies

We design three dynamic benchmarks to assess the contribution of each strategy: dispatching, redeployment, and reallocation. The fourth dynamic benchmark is used to test the quality of our solutions, and the fifth is a relocation heuristic designed by Jagtenberg et al. (2015). Benchmark 1 (dispatching only) allows the dispatcher to assign any available ambulance when a call is received. However, redeployment and reallocation decisions are not considered; that is, every ambulance in the EMS system is preassigned to a base and returns to that base after serving a call, and the repositioning of idle ambulances to the base of an ambulance that was just dispatched to a call is not considered. Note that if the dispatcher does not immediately send an available ambulance to a received call in this benchmark, the call will join a queue. After an ambulance finishes serving a call, the dispatcher decides whether the ambulance serves a call in the queue or returns to its preassigned base. Calls in the queue are served based on a first-come, first-served rule in a decreasing order of priority. Benchmark 2 (redeployment only) sends immediately the closest available ambulance to a call and does not consider the possibility of ambulance reallocation after dispatching an ambulance. However, Benchmark 2 determines the redeployment policy; i.e., after an ambulance has finished serving a call, the dispatcher decides whether the ambulance serves a call in the queue or is redeployed to a base. Benchmark 2 is similar to the settings studied by Maxwell et al. (2010). Benchmark 3 (reallocation only) sends the closest available ambulance to serve a call, and after an ambulance has finished serving a call, decides whether the ambulance serves a call in the queue or returns to its preassigned base. However, upon dispatching an ambulance to a call, Benchmark 3 considers the possibility of reallocating an available ambulance to the base of the ambulance that was just dispatched. The performance of Benchmark 4, which considers both redeployment and reallocation, is used to test the quality of the solutions by calculating the optimality gap with respect to the lower bounding system. The ADP approach, presented in Section 3, considers all of the dispatching, redeployment, and reallocation strategies simultaneously.

5.2.1. Relocation Heuristic. Jagtenberg et al. (2015) developed a simple relocation heuristic, which is easy to implement and showed strong performance in some data sets. We use this heuristic as another benchmark. The dispatching policy in this benchmark is to send the closest ambulance to a received call; however, the relocation policy can reallocate an available ambulance to a base, or redeploy an ambulance that just finished its service to a base that results in the largest marginal contribution to coverage according to the MEXCLP model. The marginal contribution of adding a kth ambulance to cover demand in region l is given by $E_k - E_{k-1} = \lambda_l^\tau(1-\alpha)\alpha^k$, where α denotes a "busy fraction" similar to the MEXCLP, and λ_l^τ is the demand (call arrival) rate in region l at time τ. The base that gives the largest marginal contribution over all demand is chosen as the destination for relocation.

5.3. Results and Managerial Insights

We compare the performance of ADP and benchmark policies with the static benchmarks with respect to four major measures: (i) average response time, (ii) fraction of late calls, (iii) average response time of late calls, and (iv) fraction of late high-priority calls. Table 2 reports the performance of each policy when the ADP objective is to minimize the expected discounted priority-adjusted total response time. The average response time for the ADP policy is 6.5 ± 0.2 minutes (95% confidence interval), while the MCLP static

Table 1. Performance of Static Benchmarks

	Ambulance location models					
	MCLP	DSM	MEXCLP	MALP	ARTM	ERTM
Fraction of late calls (%)	25.3	25.6	24.4	27.1	48.3	24.7
Average response time (min.)	7.3	7.9	7.5	7.4	9.7	7.8

Table 2. Performance of the ADP Policy and Benchmarks

	Baseline performance (95% confidence interval)			
	Average response time (min.)	Fraction of late calls (%)	Average response time of late calls (min.)	Fraction of late high-priority calls (%)
MCLP	7.3 ± 0.2	25.3 ± 0.2	11.7 ± 0.1	4.9 ± 0.2
Benchmark 1: Dispatching only	7.1 ± 0.1	22.5 ± 0.1	10.5 ± 0.1	4.2 ± 0.1
Benchmark 2: Redeployment only	6.8 ± 0.1	20.2 ± 0.1	9.1 ± 0.1	3.8 ± 0.1
Benchmark 3: Reallocation only	7.2 ± 0.1	22.9 ± 0.1	10.4 ± 0.1	4.5 ± 0.1
Benchmark 4: Redeployment + reallocation	6.7 ± 0.1	19.2 ± 0.2	9.0 ± 0.2	3.5 ± 0.2
Relocation heuristic	6.8 ± 0.1	19.7 ± 0.1	10.1 ± 0.1	3.7 ± 0.2
ADP	6.5 ± 0.2	18.3 ± 0.1	8.9 ± 0.2	3.4 ± 0.2

Note. The ADP objective function in this table is to minimize the expected total discounted priority-adjusted response time.

benchmark is estimated to have an average response time of 7.3 ± 0.2 in 30 replications. Table 2 shows that the fraction of late calls for the ADP policy is significantly less than that of the MCLP static and other benchmarks. In particular, the fraction of late calls is 18.3 ± 0.1% for the ADP policy and 25.3 ± 0.2% for the MCLP static benchmark. The performance of the ADP policies for the average response time of late calls and the fraction of high-priority late calls is also significantly better than that of the benchmarks. Similarly, Table 3 reports the performance of the ADP policy and benchmarks when the ADP objective is to minimize the expected discounted priority-adjusted fraction of late calls. Tables 2 and 3 show that the ADP policy improves various performance measures compared to other benchmarks. Our results indicate that the contribution of the redeployment-only strategy is significantly greater than that of the dispatching-only and reallocation-only strategies in improving the performance over static benchmarks in all measures. Our analysis shows that one reason for this observation may be that the expected proportion of time that the dispatching-only ADP strategy deviates from the best static benchmark is much less than the proportion of time that the redeployment-only ADP strategy deviates from it. Specifically, the redeployment-only ADP strategy sends the ambulance to its previous base after finishing service only 19% of times, while the dispatching-only ADP strategy immediately sends the closest ambulance to a received call nearly 70% of times. Further analysis of the dispatching-only ADP strategy shows that, conditioned on not immediately sending the closest ambulance, a nonclosest ambulance is dispatched in nearly 87% of times, while calls are delayed in 13% of times. Moreover, the performance of the dispatching-only ADP strategy does not significantly change if the dispatcher is not allowed to queue a call when an ambulance is available. Although both high- and low-priority calls can be queued in our framework, our numerical analysis shows that only 1% of high-priority calls are delayed. Our results also show that the reallocation-only ADP strategy relocates an idle ambulance to the base that just emptied in nearly 10% of times, and the reallocation flows are toward empty bases in high demand zones. Comparing the results for the relocation heuristic and Benchmark 4 shows that the relocation heuristic is an efficient policy when only relocation strategies are considered.

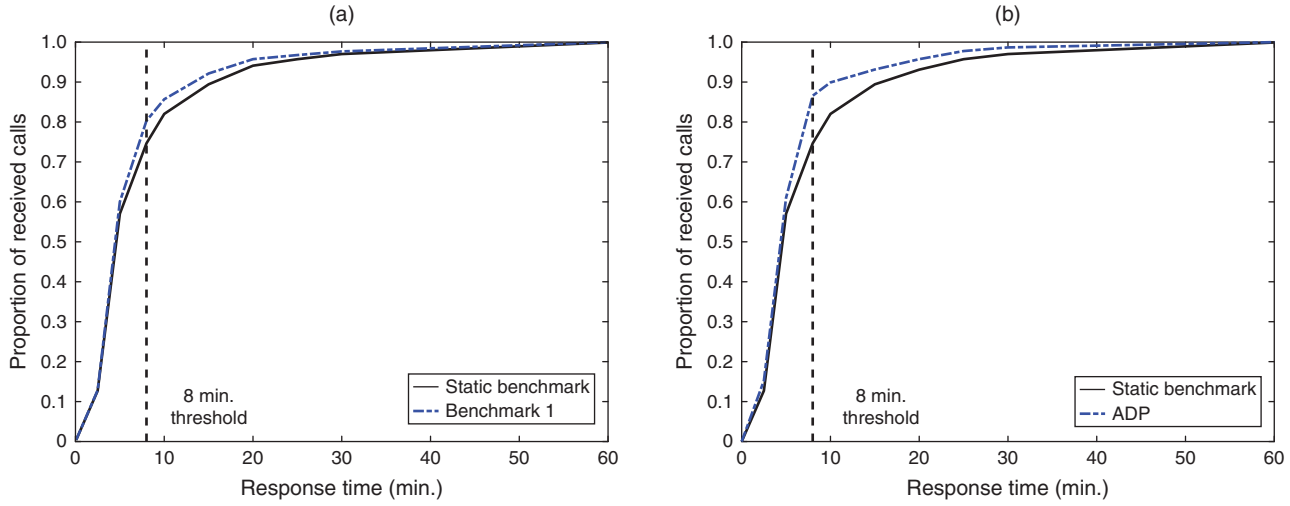
Figure 1 shows the empirical cumulative distribution function of the response times for the MCLP static benchmark to (a) the dispatching-only strategy and to (b) the ADP policy. One could think that minimizing the expected discounted priority-adjusted total response time might involve the risk of losing some of the closer calls by trying to concentrate the optimization on calls with larger response times. Figure 1 suggests that the ADP policies do not abandon a few calls

Table 3. Performance of the ADP Policy and Benchmarks

	Baseline performance (95% confidence interval)			
	Average response time (min.)	Fraction of late calls (%)	Average response time of late calls (min.)	Fraction of late high-priority calls (%)
MEXCLP	7.5 ± 0.2	24.4 ± 0.2	12.7 ± 0.2	4.6 ± 0.2
Benchmark 1: Dispatching only	7.1 ± 0.1	21.1 ± 0.1	11.2 ± 0.1	3.9 ± 0.1
Benchmark 2: Redeployment only	6.9 ± 0.1	19.2 ± 0.1	10.1 ± 0.1	3.5 ± 0.1
Benchmark 3: Reallocation only	7.3 ± 0.1	22.0 ± 0.1	10.7 ± 0.1	4.2 ± 0.1
Benchmark 4: Redeployment + reallocation	6.8 ± 0.1	18.3 ± 0.2	9.9 ± 0.2	3.4 ± 0.2
Relocation heuristic	6.8 ± 0.1	19.7 ± 0.1	10.1 ± 0.1	3.7 ± 0.2
ADP	6.6 ± 0.2	16.4 ± 0.1	9.2 ± 0.1	3.1 ± 0.1

Note. The ADP objective function in this table is to minimize the expected total discounted priority-adjusted fraction of late calls.

Figure 1. (Color online) Empirical Cumulative Distributions of the Response Time



to wait for a long time; instead, they shift the entire distribution of response times to the left.

To illustrate the quality of solutions produced by the ADP framework, we compute a lower bound on the expected response time and fraction of late calls. Recall that we assumed a nonhomogeneous Poisson process for call arrivals in Section 3. However, in reporting the results for comparing our lower bounds with Benchmark 4, an assumption of a constant call arrival rate for each location is forced in both the lower bounding system and Benchmark 4. Our results show that in the lower bounding system, the expected response time and fraction of late calls are 5.1 minutes and 11.7%, respectively. We use Benchmark 4 to assess the quality of solutions with respect to the lower bounding system, because both Benchmark 4 and the lower bounding system use a myopic dispatching rule, i.e., immediately sending the closest ambulance and relying only on relocating available ambulances to improve performance, which in the case of Benchmark 4 consists of redeployment and reallocation strategies. Our results show that the absolute difference between Benchmark 4 and the lower bound on average response time (fraction of late calls) is 1.6 minutes (6.6%) when the objective function is to minimize the response time (fraction of late calls). The results of the sensitivity analysis are reported in the online companion.

6. Conclusion

In this study we formulated a real-time ambulance dispatching and relocation problem as a stochastic dynamic program and solved it via approximate dynamic programming. We extended the literature on real-time ambulance management via ADP, which considers only ambulance redeployment, in two dimensions. First, we considered a general dispatching strategy in which the decision maker can send any available ambulance to a received call in addition to having the option of not dispatching an ambulance immediately, but rather waiting for an ambulance that may become available soon. Second, we introduced an ambulance reallocation strategy in which the decision maker may send an available ambulance to the location of an ambulance just dispatched to a call. The ambulance reallocation strategy can improve performance by reducing the expected time that a region is uncovered, which is caused by dispatching the only ambulance that covers it. We tested the performance of policies generated by our ADP framework on an EMS system in Mecklenburg County, North Carolina, and our results show that our policies significantly improve static benchmarks. In particular, our near-optimal policies reduce the response time and fraction of high-priority late calls by 12% and 30.6%, respectively, compared to the best static benchmarks.

We designed three benchmarks to analyze the contribution of each strategy, general dispatching, redeployment, and reallocation, by adding strategies to the static policy one at a time. Our results show that each strategy significantly improves the static benchmarks, and considering all three strategies simultaneously is significantly better than each strategy alone. We also showed that the redeployment strategy is the best when only one strategy could be added to the static policy. This observation, our analysis shows, is due to the fact that the expected frequency with which the redeployment-only ADP policy deviates from the static benchmarks is significantly greater than the frequency with which the dispatching-only ADP policy deviates from the static benchmarks. Allowing the dispatcher to queue a received call when an ambulance is available did not significantly improve the performance. Considering a general dispatching rule, redeployment and

reallocation can significantly improve the performance of an EMS system.

Acknowledgments

The authors are thankful to Morris Cohen, the associate editor, and two anonymous referees for their constructive feedback. They also thank Dr. Huseyin Topaloglu for his comments.

References

Adelman D (2007) Dynamic bid prices in revenue management. *Oper. Res.* 55(4):647–661.

Alanis R, Ingolfsson A, Kolfal B (2013) A Markov chain model for an EMS system with repositioning. *Production Oper. Management* 22(1):216–231.

Allen AO (1980) Queueing models of computer systems. *IEEE Comput. Soc.* (4):13–24.

Andersson T, Varbrand P (2007) Decision support tools for ambulance dispatch and relocation. *J. Oper. Res. Soc.* 58(2):195–201.

Bélanger V, Kergosien Y, Ruiz A, Soriano P (2016) An empirical comparison of relocation strategies in real-time ambulance fleet management. *Comput. Indust. Engrg.* 94(1):216–229.

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

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

Bertsimas D, Farias VF, Trichakis N (2013) Fairness, efficiency, and flexibility in organ allocation for kidney transplantation. *Oper. Res.* 61(1):73–87.

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

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

de Farias D, Van Roy B (2004) On constraint sampling in the linear programming approach to approximate dynamic programming. *Math. Oper. Res.* 29(3):462–478.

Doerner KF, Gutjahr WJ, Hartl RF, Karall M, Reimann M (2005) Heuristic solution of an extended double-coverage ambulance location problem for Austria. *Central Eur. J. Oper. Res.* 13(4): 325–340.

Fitch J (2005) Response times: Myths, measurement, and management. *J. Emergency Medical Services* 30(9):47–56.

Gendreau M, Laporte G, Semet F (1997) Solving an ambulance location model by tabu search. *Location Sci.* 5(2):75–88.

Gendreau M, Laporte G, Semet F (2001) A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Comput.* 27(12):1641–1653.

Gendreau M, Laporte G, Semet F (2006) The maximal expected coverage relocation problem for emergency vehicles. *J. Oper. Res. Soc.* 57(1):22–28.

Ingolfsson A (2013) EMS planning and management. Zaric GS, ed. *Operations Research and Health Care Policy*, International Series in Operations Research and Management Science, Vol. 190 (Springer, New York), 105–128.

Jagtenberg CJ, Bhulai S, van der Mei RD (2015) An efficient heuristic for real-time ambulance redeployment. *Oper. Res. Health Care* 4(March):27–35.

Jagtenberg CJ, Bhulai S, van der Mei RD (2016) Dynamic ambulance dispatching: Is the closest-idle policy always optimal? *Health Care Management Sci.* 20(4):1–15.

Khademi A, Saure DR, Schaefer AJ, Braithwaite RS, Roberts MS (2015) The price of nonabandonment: HIV in resource-limited settings. *Manufacturing Service Oper. Management* 17(4):554–570.

Lai G, Margot F, Secomandi N (2010) An approximate dynamic programming approach to benchmark practice-based heuristics for natural gas storage valuation. *Oper. Res.* 58(3):564–582.

Mason AJ (2013) Simulation and real-time optimised relocation for improving ambulance operations. Denton BT, ed. *Handbook of Healthcare Operations Management*, International Series in Operations Research and Management Science, Vol. 184 (Springer, 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.

Maxwell MS, Cao Ni E, Tong C, Henderson SG, Topaloglu H, Hunter SR (2014) A bound on the performance of an optimal ambulance redeployment policy. *Oper. Res.* 62(5):1014–1027.

McLay LA, Mayorga ME (2013) A model for optimally dispatching ambulances to emergency calls with classification errors in patient priorities. *IIE Trans.* 45(1):1–24.

National Association of State EMS Officials (2009) EMS performance measures: Recommended attributes and indicators for system and service performance. Accessed July 2017, http://www.nasemso.org/Projects/PerformanceMeasures/documents/EMSPerformanceMeasuresDec2009.pdf.

National Fire Protection Association (NFPA) (2010) NFPA 1710: Standard for the organization and deployment of fire suppression operations, emergency medical operations, and special operations to the public by career fire departments. Report, National Fire Protection Association, Quincy, MA.

Powell WB (2007) *Approximate Dynamic Programming: Solving the Curses of Dimensionality* (Wiley-Interscience, Hoboken, NJ).

Puterman ML (2005) *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (Wiley-Interscience, New York).

Restrepo M, Henderson SG, Topaloglu H (2009) Erlang loss models for the static deployment of ambulances. *Health Care Management Sci.* 12(1):67–79.

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

Sudtachat K, Mayorga ME, Mclay LA (2016) A nested-compliance table policy for emergency medical service systems under relocation. *Omega* 58(January):154–168.

Swersey AJ (1994) The deployment of police, fire, and emergency medical units. *Handbooks Oper. Res. Management Sci.* 6(December): 151–200.

U.S. Census Bureau (2016) Mecklenburg County, North Carolina quick facts. Accessed July 2017, https://www.census.gov/quickfacts/fact/table/mecklenburgcountynorthcarolina/PST045216.

van Barneveld T (2016) The minimum expected penalty relocation problem for the computation of compliance tables for ambulance vehicles. *INFORMS J. Comput.* 28(2):370–384.

van Barneveld TC, Bhulai S, van der Mei RD (2016) The effect of ambulance relocations on the performance of ambulance service providers. *Eur. J. Oper. Res.* 252(1):257–269.

van den Berg PL, van Essen JT, van Harderwijk EJ (2016) Comparison of static ambulance location models. El Hilali Alaoui A, Benadada Y, Boukachour J, eds. *Proc. 3th Internat. IEEE Conf. Logist. Oper. Management* (Institute of Electrical and Electronics Engineers, New York), 1–10.

Van Roy B, Bertsekas DP, Lee Y, Tsitsiklis JN (1997) A neuro-dynamic programming approach to retailer inventory management. Peshkin M, ed. *Proc. 36th IEEE Conf. Decision Control*, Vol. 4 (Institute of Electrical and Electronics Engineers, New York), 4052–4057.

Ward MJ (2014) Saving more lives? *J. Emergency Medical Services* 39(2): 46–53.

Wilde ET (2013) Do emergency medical system response times matter for health outcomes? *Health Econom.* 22(7):790–806.