OPERATIONS RESEARCH
Vol. 62, No. 5, September–October 2014, pp. 1014–1027
ISSN 0030-364X (print) | ISSN 1526-5463 (online)

http://dx.doi.org/10.1287/opre.2014.1302

A Bound on the Performance of an Optimal Ambulance Redeployment Policy

Matthew S. Maxwell, Eric Cao Ni, Chaoxu Tong, Shane G. Henderson, Huseyin Topaloglu
School of Operations Research and Information Engineering, Cornell University, Ithaca, New York 14853
{msm57@cornell.edu, cn254@cornell.edu, ct423@cornell.edu, sgh9@cornell.edu, ht88@cornell.edu}

Susan R. Hunter
School of Industrial Engineering, Purdue University, West Lafayette, Indiana 47907, susanhunter@purdue.edu

Ambulance redeployment is the practice of repositioning ambulance fleets in real time in an attempt to reduce response times to future calls. When redeployment decisions are based on real-time information on the status and location of ambulances, the process is called system-status management. An important performance measure is the long-run fraction of calls with response times over some time threshold. We construct a lower bound on this performance measure that holds for nearly any ambulance redeployment policy through comparison methods for queues. The computation of the bound involves solving a number of integer programs and then simulating a multiserver queue. This work originated when one of the authors was asked to analyze a response to a request-for-proposals (RFP) for ambulance services in a county in North America.

Subject classifications: system-status management; ambulance location; ambulance relocation; ambulance deployment; move-up; coupling.
Area of review: Policy modeling and public sector OR.
History: Received July 2013; revision received March 2014; accepted May 2014. Published online in *Articles in Advance* July 16, 2014.

1. Introduction

Emergency medical service (EMS) providers respond to calls for assistance, providing medical services at the scene of the call, and transporting the sick or injured to a hospital. The local municipality contracts these services from an EMS provider for a particular geographic region. Municipalities evaluate potential EMS providers by how quickly the provider can respond to calls. In particular, contracts typically require that the percentage of *late* calls in a set time interval, for example, one month, be at most some threshold level, for example, 10%. A call is considered late if the *response time*, i.e., the time from when the call is first received to the arrival of an ambulance on the scene, is greater than a threshold, usually on the order of nine minutes. See Ingolfsson (2012) for an excellent introduction to, and overview of, EMS management from an operations perspective.

To win the EMS contract in a municipality, EMS providers must propose a *shift schedule* that gives the start and finish times of all ambulance shifts, thereby determining the number of ambulances deployed as a function of time. The key question we consider is: How can a municipality know if a proposed shift schedule can meet response-time targets without knowing how ambulances will be deployed around a city, or how ambulances will be dispatched to calls?

EMS providers have traditionally used *static* policies for ambulance deployment. In a static policy, ambulances are assigned to fixed bases for an entire shift and return to the assigned base on call completion. Usually each base is staffed with two or three ambulances. More recently, increasing roadway congestion has motivated many EMS providers to spread ambulances out to meet response time performance targets. In these policies, ambulances are no longer located in groups, so a dispatched ambulance leaves a hole in coverage. A location is covered at time t if an available ambulance at time t can reach that location within the response-time threshold. The term coverage (at time t) refers to the set of locations in a city that are covered at time t.

To reduce the impact of coverage holes on the expected response times for future calls, many EMS providers now use *ambulance redeployment*, also known as system-status management. Ambulance redeployment is the practice of repositioning idle ambulances in real-time to better respond to future calls. Typically such policies have an apparently modest impact on performance, reducing the percentage of late calls by up to about 5%. So, for example, 16% of calls might be late in a system operated with static policies, while 11% of calls might be late in a system using redeployment (Maxwell 2011). While these reductions may sound modest, they can allow an EMS provider to meet contractual performance targets without adding ambulances. This can be costly.

Ambulance redeployment policies may be constructed in myriad ways. The diversity of methods available perhaps reflects the mathematical intractability of finding a truly

optimal policy. Redeployment policies may be constructed by solving integer programs in real time (Gendreau et al. 2001, Brotcorne et al. 2003, Richards 2007, Nair and Miller-Hooks 2009, Zhang 2012, Naoum-Sawaya and Elhedhli 2013), by presolving certain integer programs to construct look-up tables (Gendreau et al. 2006, Alanis et al. 2013), by solving or approximately solving stochastic dynamic-programming formulations (Berman 1981a, c, b; Zhang et al. 2010; Zhang 2010; Maxwell et al. 2010, 2013), and by screening potential policies with approximate models and subsequently comparing the top candidate policies through simulation (Alanis et al. 2013), through online simulation (Yue et al. 2012) or through heuristics (Andersson 2005, Andersson and Vaerband 2007). See Mason (2013) and Henderson (2009) for further discussion of redeployment methods.

Returning to our original question: How can a municipality know whether a proposed ambulance schedule can meet contractual response time targets, without knowing the deployment/dispatch policy? Straightforward techniques, such as simulation, are not applicable, as one must know the deployment/dispatch policy to simulate the system. To answer this question, which arose when the fourth author of this paper was asked to analyze the proposed ambulance services for a county in North America, we derive a lower bound on the fraction of late calls achievable by nearly *any* ambulance redeployment policy (mild restrictions on the policies we consider are discussed in §2). The bound also applies to static policies, which are special cases of redeployment policies. The bound we derive can be used to determine when a proposed ambulance schedule is not sufficient to meet contractual targets. Indeed, a version of the bound was used in the fourth author's testimony in the case of an EMS provider who believed another company's proposal was unrealistic. The bounds may also be used to determine when to terminate searches for effective redeployment policies and to determine when measures other than ambulance redeployment are necessary to improve response times. We know of no other practical bounds in the open literature for ambulance deployment, including for static policies.

Constructing useful bounds for ambulance redeployment is nontrival, as exemplified by the following two proposals, neither of which work well in our context. First, one might expect that a lower bound on the performance of any ambulance-redeployment policy is given by the proportion of unreachable calls, that is, the proportion of call arrivals that arise in locations that are so far from any ambulance base that they cannot be reached within the time threshold. If all calls are attended by ambulances responding from a base, then this bound is valid. However, this was shown to have little practical value for two example cities in Maxwell (2011). In one city, this unreachable bound returned values on the order of 1% when the fraction of late calls obtained by our best redeployment policy was on the order of 17%. The proportion of unreachable calls does not provide a lower bound when ambulances may respond to calls from the road, e.g., when returning to a base after completing a call, thereby enabling a response within the time threshold.

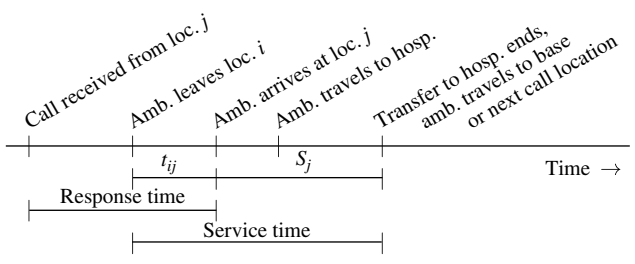
The second type of bounds one might consider is generic, i.e., when the ambulance redeployment problem is modeled as a stochastic dynamic program. Bounds on the optimal value of generic stochastic dynamic programs have recently been devised (Brown et al. 2010) and are obtained by solving a version of the dynamic program that has more information at the time of decisions than the original formulation. Unfortunately we have been unsuccessful in our attempts to apply this general methodology to obtain bounds in ambulance redeployment, owing to the complexity of solving the relaxed dynamic program where more information is assumed to be known.

Nevertheless, we develop a computationally tractable lower bound on the expected number of late calls over a finite time horizon achievable by nearly any policy. To our knowledge, this paper gives the first such lower bound, which we call the *cover* bound. It arises by modeling the ambulance system as a multiserver queue where servers represent ambulances, and customers represent calls for ambulance service.

To analyze the ambulance system as a multiserver queue, consider the timeline of events when a call is received (Figure 1). Recall that our performance metric is the percentage of late calls, which is a function of response times. In the queueing model, response time corresponds to the amount of time in the queue plus t_{ij}, the time for an available ambulance to travel from its location i to the call location j, which we assume is deterministic. The *service time* corresponds to the total amount of time the ambulance spends responding to a call, including traveling to the call location, assisting at the scene, and providing transport and transfer to a hospital if needed. Assume that when a call is received, it is assigned to the nearest available ambulance. If no ambulance is available, the calls are served in first in, first out fashion as ambulances become free.

Now consider two such ambulance queueing systems, identical in every way except for their service-time distributions. The first is the *real* system; the second is the *bounding*

Figure 1. Timeline of events from when a call is received at location j and assigned to an ambulance that will start traveling from location i, to the completion of the call.



Notes. The quantity t_{ij} is the deterministic time to travel between locations i and j. Some events may not occur or may require zero time, e.g., no hospital travel may be required, or there is no queueing time.

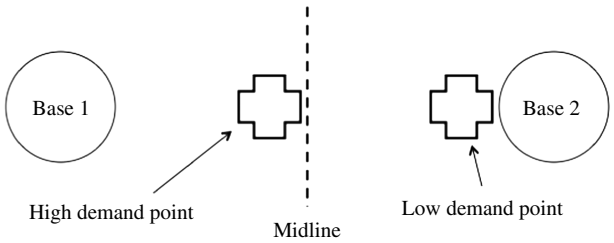
system. If the service times for calls in the bounding system are always smaller than in the real system, then the bounding system always has at least as many ambulances available to serve calls as the real system. In the real system, the service-time distribution depends on the time and location of the call, the number of ambulances available, and the geographic configuration of ambulances. The geographic configuration of ambulances at the time of the call in the real system, in particular, depends on the redeployment policy π. To create an analogous service-time distribution for the bounding system, which does not depend on π, we generate a stochastic lower bound on the service-time distribution of the real queue, that is, a distribution that is stochastically smaller than the service-time distribution obtained from any configuration of available ambulances. Computing this bounding service-time distribution requires solving a set of integer programs, several for each potential number of available ambulances.

It may be intuitive that one can now run the bounding system as a multiserver queue using the bounding service-time distribution and record the estimated fraction of late calls as an estimated lower bound on the fraction of late calls for the real system. However, the bounding service-time distribution collapses information about t_{ij} into the service time, which makes recovering information about the response times, and hence late calls, from the service times impossible. Therefore an additional step is required. Just before each call is received in the bounding system, we instantaneously reposition all the available ambulances to maximize *coverage*, thereby creating a lower-bounding probability that the call is late. We show that the expected value of the sum of these lower-bounding probabilities is a lower bound on the performance of any ambulance-redeployment policy that implements mild call-queueing assumptions. Because the expected value can be accurately estimated using discrete-event simulation, the bounding system yields the desired lower bound.

We call our bound the *cover* bound because idle ambulances are instantaneously repositioned into coverage-optimal locations that minimize the probability that the call will be late, thus randomizing over the call location. These ambulance locations can be identified by solving a set of standard integer programs (Church and ReVelle 1974) that maximize coverage subject to a constraint on the number of available ambulances. This observation allows us to recover sufficient information from simulating the bounding system to bound the response time, even though we do not keep track of specific ambulance locations.

Interestingly, simply assuming that available ambulances are located so as to maximize coverage just before each call does not produce a bound in and of itself, although in experiments reported in Maxwell (2011), the difference between the supposed bound obtained through this assumption alone and the value of a true bound is very small. If ambulances were to actually respond from the coverage-optimal locations, then the service time in the queueing system can be inflated compared to some other set of locations. This policy could lead to higher use of the ambulances and poorer response-time performance; see Figure 2 for an example.

Figure 2. We wish to locate a single ambulance at one of two bases.



Notes. Demand can arise in only two locations, at the high and low demand points; no hospital transport is required. The high demand point is reachable within the response-time threshold only from Base 1. Thus the coverage-optimal location is at Base 1, but the service times include driving a longer distance to the low demand point. Therefore service times may be larger on average than if the ambulance were positioned at Base 2.

The cover bound is computed assuming that ambulances can only be at a finite number of locations. Of course, ambulances can be anywhere on the road network, so the bound is based on an approximation to the true dynamics. Mathematically, we derive the cover bound with notation that assumes that calls and ambulances can be at, for example, any intersection in the road network. However in practice, when we compute the bound assuming ambulances can be at any intersection in the road network, the resulting bound is too loose to be useful. (Maxwell 2011 reports that the lower bound ranges between 3% and 4% while the ambulance redeployment policy in use results in 18% to 22% late calls.) Consequently, in our numerical experiments, we use a stylized model of ambulance redeployment where ambulances always respond from ambulance bases. Even though ambulances only respond from bases, this assumption still allows for dynamic policies because ambulances may respond from *different* bases throughout the day, the identity of which will usually depend on the overall state of the system. This stylized model is justified by empirical computations indicating that the fraction of calls for which ambulance response originates from a base is quite large (on the order of 80%) for realistic simulation models (Maxwell 2011).

Simulation results (Maxwell 2011) suggest that the performance of ambulance redeployment policies on the stylized model where all ambulances respond from bases is often very similar to performance on more realistic models where ambulance responses can originate anywhere. Because the lower bound is quite tight for the stylized model, it seems reasonable to use the value of the bound as computed on the stylized model as a heuristic bound on the potential performance of ambulance redeployment polices. Indeed, this is how we recommend using the results of this paper in practice.

For the stylized model, simulation results reported here show that the cover bound is very tight for one realistic

example, but not as tight in another. To better understand these results, we design an experiment that involves artificial cities. These artificial cities differ in terms of the number of modes of the distribution of call locations, the dispersion of the demand around these modes, and the degree to which ambulance bases cluster around demand modes. For each artificial city we compute both the cover bound and the performance of a redeployment policy. The results shed light on what features of a city are such that we can expect the gap between the cover bound and the performance of the specific redeployment policy to be small (or large).

The rest of the paper is organized as follows. In §2 we state our assumptions on the ambulance system that, taken together, define a model of ambulance dynamics. Section 3 constructs the cover bound and proves its validity. We also formulate the well known integer program that maximizes coverage for a given number of available ambulances. In §4, we show in detail how to construct the stochastic lower bound on the service-time distribution that the bound uses. Section 5 provides simulation results demonstrating the performance of the bound on two realistic examples. Section 6 presents the results of the designed experiment that sheds light on when we can expect the bound to be effective. Section 7 provides closing remarks and suggestions for future research..

This paper is an outgrowth of Chapter 5 of Maxwell (2011), where a version of the cover bound was derived, and Ni et al. (2012), where the results of the designed experiment were first reported.

2. Model and Assumptions

Consider a model of ambulance operations over a finite deterministic time horizon $[0, \bar{t}\,]$. Calls arrive in time according to an arbitrary call-arrival process that is independent of the state of the system. Usually the arrival process is assumed to be a nonhomogeneous Poisson process, but for now we only assume that the expected number of calls received in $[0, \bar{t}\,]$ is finite. Call locations, ambulance positions, and hospital locations are restricted to a finite set $\{1, 2, \ldots, J\}$ of locations. There are H hospitals at locations $s(1), s(2), \ldots, s(H) \in \{1, 2, \ldots, J\}$ and A ambulances. Successive call locations are independent and identically distributed (iid) and the probability that a call originates at location j is d_j, $j = 1, 2, \ldots, J$. The time t_{ij} required to travel from Location i to Location j is deterministic and does not vary with time. This is perhaps the most restrictive of our assumptions, as travel times on road networks can depend heavily on the time of day.

We highlight the following assumption about the deployment policies we consider.

Assumption 1. *When a call is received, it is immediately assigned to an available ambulance. If no ambulance is available, calls are queued and served in a first in, first out fashion as ambulances become free.*

Ambulances that are traveling but not currently assisting a patient (e.g., an ambulance has finished transferring a patient to a hospital and is returning to a base) are considered available. To understand the restrictions imposed by Assumption 1, consider a call arriving when only a single ambulance, 30 minutes away in travel time, is available, but an ambulance that is 5 minutes away in travel time is expected to complete its job in 5 more minutes. We do not wait for the closer but currently occupied ambulance to become available before assigning the job.

One artificial and inappropriate way to reduce response times is to ensure that ambulances are never at an ambulance base. This works because ambulance crews at a base require a brief activation delay or turn-out time to finish what they are doing, get to their ambulance and depart; this time is either eliminated or at least substantially reduced when the crew is on the road or parked at some other location. Naturally, ambulance crews prefer waiting at ambulance bases for calls rather than sitting in their vehicles for entire shifts, and so it is inappropriate for a policy to induce reductions in response times by simply requiring that the vehicles be away from a base at almost all times. Hence, in the results given in this paper we assume that all turn-out times are zero, thereby eliminating this artificial advantage of roving ambulances.

Upon being assigned to a call at Location i, an ambulance travels to Location i (with no turn-out time) and spends a random amount of time at the scene. (Unless otherwise stated all random quantities are independent of one another.) Then, with probability h_{i0}, the call at Location i is completed at the scene and the ambulance is free to answer other calls and/or be directed elsewhere. Otherwise the ambulance transports the patient to hospital j with probability h_{ij}, $j = 1, 2, \ldots, H$, so that $\sum_{j=0}^{H} h_{ij} = 1$ for all $i = 1, 2, \ldots, J$. Our experience is that this assumption of a location-dependent hospital-choice distribution is reasonable in many cities. However, the assumption does not hold universally. For example, if hospital diversion is used, then the hospital-choice distribution may depend on the state of the system, and our argument establishing the bound does not hold. Our bounds might still be of practical use in such a setting, even if they are not strictly valid, if the service-time distribution (i.e., the distribution of the time required for an ambulance to handle the call, from first being notified, to the time that the ambulance clears the call, either at the scene or at the hospital where the patient has been dropped off) is not dramatically altered.

The distribution of the (random) time required to transfer a patient at hospital j can be hospital specific, although we use a common distribution in our work. After the patient transfer is complete the ambulance is free to answer other calls and/or be directed elsewhere.

The ambulance redeployment policy may arbitrarily direct available ambulances to any location at any time, and there is no limit to the number of vehicles that can simultaneously occupy a location or the number of times a vehicle may be asked to move. (In practice, frequent redeployments are avoided to minimize crew frustration, but our bound applies irrespective of the frequency of redeployments.)

3. Cover Bound

Let N be the random number of calls received over the time horizon $[0, \bar{t}\,]$. Fix an arbitrary redeployment policy π and let L denote the random number of late (i.e., response time exceeds the time threshold) calls that result. We will compute a lower bound on $E(L)$ that does not depend on the policy π. This bound can then be used to bound the long-run fraction of late calls by dividing by $E(N)$. To see why, consider an infinite sequence of i.i.d instances of the interval $[0, \bar{t}\,]$, in the ith of which the number of calls (late calls) is $N(i)$ $(L(i))$. Then the fraction of late calls over n such instances is

$$\frac{L(1)+L(2)+\cdots+L(n)}{N(1)+N(2)+\cdots+N(n)}.$$

If we divide both the numerator and denominator by n and apply the strong law of large numbers to both, we see that the long-run fraction of late calls is $E(L)/E(N)$.

So how do we bound $E(L)$? Let L_k be the indicator taking value 1 if the kth call received is late, and 0 otherwise, and let $I(\cdot)$ denote a generic indicator function. Let T_k be the time of arrival of the kth call, and let A_k and C_k denote the number and configuration (i.e., set of locations) of the available ambulances at time T_k- (just before the kth call arrival). Then the expected number of late calls is given by

$$\begin{aligned}
E(L) &= E\left(\sum_{k=1}^{N} L_k\right)\\
&= E\left(\sum_{k=1}^{\infty} L_k I(k \leqslant N)\right)\\
&= \sum_{k=1}^{\infty} E[L_k I(k \leqslant N)]\\
&= \sum_{k=1}^{\infty} E(E[L_k I(k \leqslant N) \mid T_k, A_k, C_k])\\
&= \sum_{k=1}^{\infty} E(I(k \leqslant N) E[L_k \mid T_k, A_k, C_k]).
\end{aligned}$$

Now suppose the existence of a decreasing function v: $\{0, 1, 2, \ldots, A\} \to [0, 1]$ such that

$$E[L_k \mid T_k, A_k, C_k] \geqslant v(A_k) \tag{1}$$

for all k almost surely. The quantity on the left-hand side of (1) is the conditional probability that a call is late, conditional on the time of the call, and the number and configuration of available ambulances just before the call is received. Our lower bound on this quantity depends only on the number of available ambulances at the time the call is received. Shortly, we will show how to compute $v(\cdot)$ by solving integer programs. The condition that $v(\cdot)$ be decreasing is natural in that we expect the lateness probability to be decreasing in the number of available ambulances.

Equipped with the function v we then have

$$E(L) \geqslant \sum_{k=1}^{\infty} E(I(k \leqslant N) v(A_k)) = E\left(\sum_{k=1}^{N} v(A_k)\right).$$

We complete the construction of the bound and the proof that it is valid using a comparison of queues, i.e., a coupling (i.e., joint construction) of the ambulance dynamics under the policy π and the ambulance dynamics of the bounding system. In this construction the call arrival process is identical in both systems, but the number of available ambulances $\tilde{A}_k$ in the bounding system at the time of the kth arrival satisfies $\tilde{A}_k \geqslant A_k$ for all k on all sample paths. Since $v(\cdot)$ is decreasing,

$$\sum_{k=1}^{N} v(A_k) \geqslant \sum_{k=1}^{N} v(\tilde{A}_k),$$

and taking expectations we conclude that

$$E(L) \geqslant E\left(\sum_{k=1}^{N} v(\tilde{A}_k)\right). \tag{2}$$

The right-hand side of (2) is our desired bound. It can be estimated to arbitrary accuracy via simulation provided that we can simulate the bounding system and compute the function v. We defer the construction of the bounding system to §4.

The quantity $v(m)$ gives the fraction of demand that cannot be covered by m ambulances. (Recall that the demand at a location is said to be covered if an available ambulance can reach that location within the time threshold.) For $1 \leqslant m \leqslant A$, let $v(m)$ be the optimal objective value of the integer program (Church and ReVelle 1974)

$$\begin{aligned}
\min\ & \sum_{j=1}^{J} d_j(1-w_j)\\
\text{s.t.}\ & \sum_{i=1}^{J} x_i \leqslant m,\\
& w_j \leqslant \sum_{i=1}^{J} \delta(i,j) x_i \quad \forall\, j = 1, 2, \ldots, J,\\
& x_i \in \{0,1\} \quad \forall\, i = 1, 2, \ldots, J,\\
& w_j \in \{0,1\} \quad \forall\, j = 1, 2, \ldots, J.
\end{aligned} \tag{3}$$

Here the parameter $\delta(i, j)$ equals 1 if the travel time from i to j is no larger than the time threshold, and 0 otherwise. The decision variable x_i indicates whether an ambulance is stationed at Location i, and the decision variable w_j indicates whether Location j is covered. The first constraint limits the number of ambulances to m, and the second allows w_j to be 1 only if Location j is covered. The integer program thus seeks a set of locations for the m ambulances that minimizes the fraction of uncovered demand, as required. Furthermore, as m increases, the feasible region increases, so that $v(\cdot)$ is decreasing in m. Finally, we set $v(0) = v(1)$. The intuition here is that calls arriving when no ambulance is free must wait until at least one becomes free; the chance that the call

is not covered at that stage is at least $v(1)$ because of the queueing delay.

4. Bounding System

We now give the joint construction that ensures that the number of available ambulances at the time of the kth call in the bounding system and the true system satisfy $\tilde{A}_k \geqslant A_k$ for all k (pathwise).

Recall that the service time for a call is the sum of the time spent traveling to the scene, at the scene, and if necessary traveling to a hospital and transferring the patient to the hospital. We exploit a set of distribution functions $(\tilde{G}_m\colon 0 \leqslant m \leqslant A)$, the mth of which gives a stochastic lower bound on the conditional distribution of the service time conditional on the locations of the m available ambulances at the time the call is received. The stochastic lower bound holds uniformly for all possible locations of the m ambulances. We further require that the distribution functions are stochastically decreasing in m, i.e., that $\tilde{G}_0(x) \leqslant \tilde{G}_1(x) \leqslant \cdots \leqslant \tilde{G}_A(x)$ for all x. We will construct such a set of distribution functions shortly, but for now assume that they exist.

Let the number of calls received N and the times of arrivals of the calls $T_1 \leqslant T_2 \leqslant \cdots \leqslant T_N$ on the interval $[0, \bar{t}\,]$ be given. These quantities are common to both the original and bounding systems. Let $(U_j\colon j \geqslant 1)$ be an i.i.d sequence of $U(0, 1)$ random variables that is independent of N and $(T_1, T_2, \ldots, T_N)$. We will use these uniform random variables to generate service times in both the original system and in the bounding system.

We use the concept of virtual workload (Kiefer and Wolfowitz 1955) to establish our results. The virtual workload (henceforth workload) of an ambulance is the time required for that ambulance to complete its currently assigned call (if any) and also any calls that are queued and will eventually be served by that ambulance. To track the workloads as calls arrive, we assume that the service times of calls are realized at the time at which the call is received. This means that we generate the time spent traveling to the scene, time at the scene, time spent traveling to a hospital and transferring the patient to the hospital for each call at the time at which the call arrives. This allows us to track the workloads; the resulting system dynamics are statistically identical to the case where these service-time components are generated as they unfold. This assumption is a purely theoretical device for establishing that the bound is valid.

The workload dynamics are as follows. For $k \geqslant 1$, let W_k be the virtual workload vector in increasing order (Kiefer and Wolfowitz 1955) at time T_k+, i.e., just after the arrival of the kth call, and set W_0 to be the initial vector of workloads (at time 0). The ith component $W_k(i)$ of the vector W_k gives the ith smallest workload of any of the A ambulances in the system controlled by the fixed redeployment policy π just after the kth call has arrived and the work associated with that call has been assigned to one of the ambulances under Assumption 1. Let $(\tilde{W}_k\colon k \geqslant 0)$ be the corresponding vector process in the bounding system, and set $\tilde{W}_0 = W_0$.

We show inductively that $\tilde{W}_k \leqslant W_k$ for all k (pathwise). This immediately implies that the number of free ambulances just before the time of the kth call, $\tilde{A}_k \geqslant A_k$, for all k, since all components of the workload vectors decrease linearly at rate 1 between call arrival times, at least until they hit 0. The number of available ambulances just before the time of the kth arrival is the number of components of the workload vector that are equal to 0 at time T_k-, just before the kth call.

To show inductively that $\tilde{W}_k \leqslant W_k$, suppose that for some $k \geqslant 1$, $\tilde{W}_{k-1} \leqslant W_{k-1}$. Recall that $\tilde{A}_k$ is the number of available ambulances in the bounding system just before the time of the kth call arrival. Generate the service time for the kth call in the bounding system via $\tilde{G}^{-1}_{\tilde{A}_k}(U_k)$. The service time of the kth call under policy π depends on the number, A_k, and location, C_k, of the available ambulances at time T_k-. Denote the distribution function of this service time by $G(\cdot \mid A_k, C_k)$. Generate the service time of the kth call in the system controlled by π via $G^{-1}(U_k \mid A_k, C_k)$. Now, $\tilde{A}_k \geqslant A_k$ since $\tilde{W}_{k-1} \leqslant W_{k-1}$ and all components of the workload vectors decreased at the same unit rate (until they hit zero) between the call arrival times T_{k-1} and T_k. Hence,

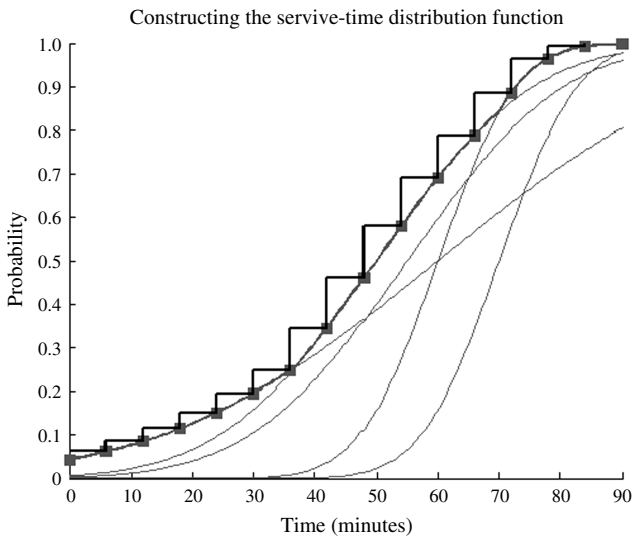
$$\tilde{G}_{\tilde{A}_k}(\cdot) \geqslant \tilde{G}_{A_k}(\cdot) \geqslant G(\cdot \mid A_k, C_k),$$

and it follows that $\tilde{G}^{-1}_{\tilde{A}_k}(U_k) \leqslant G^{-1}(U_k \mid A_k, C_k)$. In other words, the kth service time in the bounding system is no more than that in the system controlled by π. We therefore have $\tilde{W}_k \leqslant W_k$, exactly as in the proof of Theorem 6.2.1 of Stoyan (1983). Now generate the travel time, scene time, hospital selection (if any) and hospital transfer time (if any) in the system controlled by π from their appropriate conditional distributions given the service time. These conditional distributions may be complicated, but they exist and that is all that we require to establish the validity of the bound. This then allows us to advance the ambulance dynamics in the system controlled by π, and the inductive step is complete. We have proved the following result.

THEOREM 1. *Consider an arbitrary ambulance redeployment policy operating on an ambulance system that is modeled as in §2. Under the construction of* $(\tilde{A}_k\colon k \geqslant 1)$ *given above, the expected number of late calls over the interval* $[0, \bar{t}\,]$ *is bounded below by* (2).

We use integer programming to compute the distribution functions $(\tilde{G}_m\colon 1 \leqslant m \leqslant A)$. We then set $\tilde{G}_0 = \tilde{G}_1$. The idea behind this choice of distribution function for $m = 0$ is that when no ambulances are available at the time at which a call arrives, the call will be queued and served when at least one ambulance becomes available. The service-time distribution bound is then taken to be that associated with one ambulance, since the time the call spends in the queue does not count towards service time.

Figure 3. The construction of $\tilde{G}_m$ for a fixed m.



Notes. The service-time distribution function (including travel time to the scene) depends on the configuration of available ambulances at the time the call is received. There is therefore one such distribution function for each set of locations as illustrated by the 5 hypothetical fine curves. There are an enormous number of such curves but we do not attempt to enumerate them. Instead we solve integer programs to obtain the points indicated by squares. The heights of these squares are the values $\eta_m(r)$ (r takes values on the horizontal axis). The squares lie on the pointwise maximum of the fine curves, indicated by the heavier line. We then construct the right-continuous increasing function $\tilde{G}_m(\cdot)$ by connecting the dots with a staircase in such a way that the staircase lies above the pointwise upper bound.

Fix $m \geqslant 1$. We use an integer program to compute temporary values $\eta_m(r)$ for each of a finite selection of values $r_1 < r_2 < \cdots < r_k$ say. We then define $\tilde{G}_m(r) = 0$ for $r < 0$ and for $r \geqslant 0$ we use the construction depicted in Figure 3. More precisely, given the values $(\eta_m(r_i)\colon 1 \leqslant i \leqslant k)$, for $r \geqslant 0$ we set $\tilde{G}_m(r) = \eta_m(r_{i+1})$ if $r_i \leqslant r < r_{i+1}$ (taking $r_0 = 0$, $r_{k+1} = \infty$ and $\eta_m(\infty) = 1$).

The distribution functions that result from this construction are piecewise constant, and have jumps only at the times r_i, $i = 0, 1, \ldots, k$. In other words, they are the distribution functions of discrete random variables taking at most $k+1$ values. As such they are straightforward distributions from which to simulate.

It remains to specify how we compute the points $\eta_m(r)$ for each fixed m and r. To explore this computation in detail, let $F_j(\cdot)$ denote the distribution function of the service time of a call that originates at location j, not including the travel time to the scene for an ambulance (that is, $F_j(\cdot)$ is the distribution function of S_j in Figure 1). This distribution function therefore captures only the time at scene, the time to transport to hospital, and the time spent at the hospital. We need to compute F_j and we henceforth assume it is given. For a call originating at Location j served by an ambulance at Location i, the probability that the service time (including travel time to the scene) is at most r is equal to $F_j(r - t_{ij})$.

The integer program below selects locations for the m ambulances and assignments of every Demand Location j to an ambulance to maximize the probability that the service time will be completed within r time units. The quantity $\eta_m(r)$ is the optimal value of the integer program

$$
\begin{aligned}
\max\ & \sum_{j=1}^{J} d_j p_j \\
\text{s.t.}\ & \sum_{i=1}^{J} x_i \leqslant m, \\
& y_{ij} \leqslant x_i, \quad \forall\, i, j = 1, 2, \ldots, J, \\
& \sum_{i=1}^{J} y_{ij} = 1, \quad \forall\, j = 1, 2, \ldots, J, \\
& p_j = \sum_{i=1}^{J} F_j(r - t_{ij}) y_{ij}, \quad \forall\, j = 1, 2, \ldots, J, \\
& x_i \in \{0, 1\}, \quad \forall\, i = 1, 2, \ldots, J, \\
& y_{ij} \in \{0, 1\}, \quad \forall\, i, j = 1, 2, \ldots, J, \\
& p_j \in [0, 1], \quad \forall\, j = 1, 2, \ldots, J.
\end{aligned}
$$

The decision variables in this formulation are x_i, y_{ij}, p_j with i, j taking values in $\{1, 2, \ldots, J\}$. As with the integer program (3), x_i is the indicator that an ambulance is placed at Location i; the first constraint limits the available ambulances to m. The variables y_{ij} are indicators as to whether demand at Location j is served from Location i; the second and third constraints state that demand can be served from Location i only if an ambulance is positioned there, and each location j must be served from exactly one location. The variable p_j gives the probability that the service time will be at most r given that the call originates at Location j, and the final constraint assigns this probability in accordance with the Location i that serves demand at Location j.

Recall that we required the distribution functions to satisfy the condition $\tilde{G}_0 \leqslant \tilde{G}_1 \leqslant \cdots \leqslant \tilde{G}_A$. As m increases, the feasible region increases, so $\eta_m(r)$ is increasing in m for fixed r, which then implies that $\tilde{G}_m(r)$ is increasing in m for each fixed r, provided that the points $r_1, r_2, \ldots, r_k$ used to define the distribution functions are the same for each m.

This formulation is a p-median problem, which is difficult to solve to optimality for large instances. If we relax the integer constraints and solve the resulting linear program, we obtain an upper bound on the optimal solution. A check of the construction in Figure 3 shows that we can replace the values $\eta_m(r)$ with upper bounds, and the resulting cdf will still be a stochastic lower bound as desired. Hence, our bounds remain valid if we solve the relaxation instead of the full integer program. We found in numerical experiments that solving the linear programming relaxation yields no more than a percent of integrality gap, and so in the next section we solve the linear programming relaxation instead of the tighter integer program.

In summary, the computation of the bound proceeds as follows.

1. Select a set of points $r_1, r_2, \ldots, r_k$ that cover the range of reasonable call service times (e.g., from 0 to 3 hours) and are sufficiently fine to ensure that the staircase function in Figure 3 is a reasonable approximation to the smooth curve (e.g., 30 second increments).

2. For each point r_i, $i = 1, \ldots, k$, and for each number of available ambulances $m = 1, 2, \ldots, A$, solve the p-median problem (or its relaxation for a weaker bound) to obtain $\eta_m(r_i)$. This also yields the distribution functions $\tilde{G}_0, \tilde{G}_1, \ldots, \tilde{G}_A$.

3. Simulate a queueing system with the same arrival process of calls as in the real system, but with the kth service time obtained as $\tilde{G}^{-1}_{\tilde{A}_k}(U_k)$, where $\tilde{A}_k$ indicates the number of available ambulances in the simulation just before the time of arrival of the kth call, and $(U_k\colon k \geqslant 1)$ is a sequence of i.i.d. $U(0, 1)$ random variables that is independent of the arrival process. Use the simulation to estimate the expectation (2) yielding the desired bound.

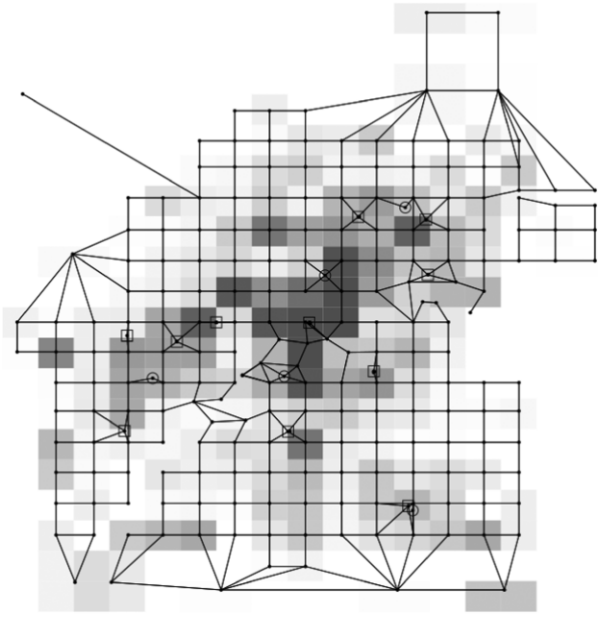
5. Computational Results on Realistic Models

In this section we provide computational results for realistic but not real models of ambulance operations in Edmonton, Canada and Melbourne, Australia. The models are realistic in the sense that calls arise according to a Poisson process in space and time, ambulances respond on road networks, and patients are delivered to a variety of hospitals when transport is required. The models are not real in the sense that we use a constant (in time) arrival rate that we select to be representative, we use a simplified road network with travel times that do not depend on the time of day, and our ambulances do not work shifts but instead are simply fixed in number throughout the simulation. In other respects, the models are identical to those used in Maxwell et al. (2010). We emphasize that the results reported here are meant to be *representative*, and should not be construed as reflecting actual on-time performance in Edmonton and Melbourne.

The Edmonton model has 4,725 locations where calls can arise, 11 bases, 16 ambulances, and 5 hospitals. See Figure 4 for the road network (modeled at the avenue level), locations of bases and hospitals, and a depiction of the demand distribution. The choice of hospital to transport to corresponds to that seen in historical data. The time spent at the scene is exponentially distributed with mean 12 minutes. The probability that transport to hospital is required is 0.75, and the time spent at the hospital is Weibull distributed with shape parameter 2.5 and mean 30.4 minutes. These distributions and parameters for the various components of a call are similar to what is seen in general EMS systems (Ingolfsson 2012, Maxwell 2011), and not necessarily what is seen in Edmonton. Travel times are deterministic and are not time dependent.

The Melbourne model has 1,413 locations where calls can arise, 87 bases, 97 ambulances, and 22 hospitals. See Figure 5 for the road network, locations of bases, and a depiction of the demand distribution. The hospital-choice distribution depends on the location of the call and is inferred from historical data. The distributions for time at scene and transfer to hospital care are the same as in Edmonton, as is the probability that transport to hospital is required.

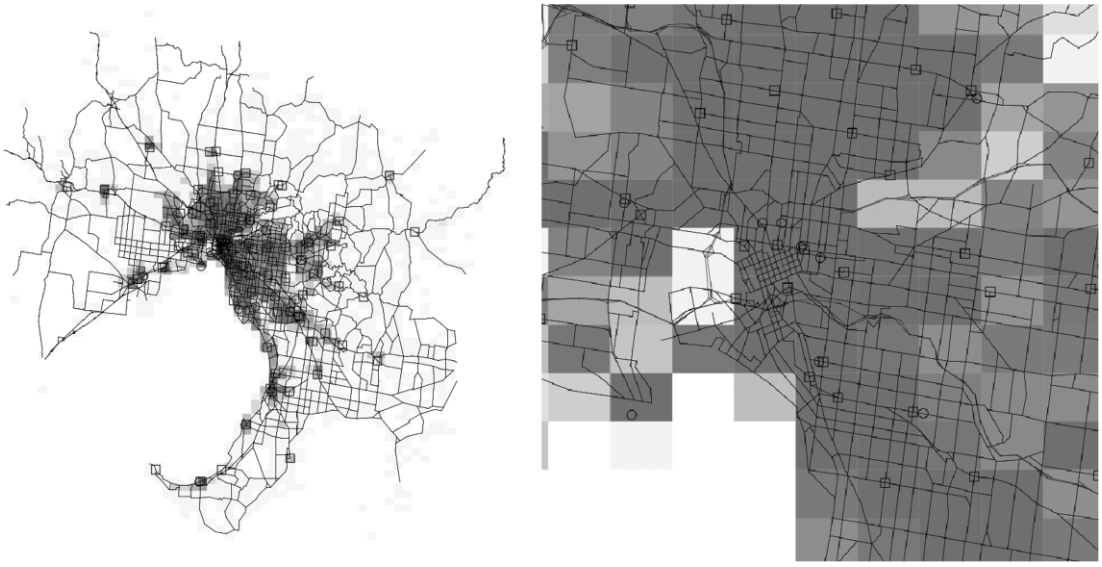
Figure 4. Road network, distribution of demand (darker regions correspond to greater demand), bases (squares), and hospitals (circles) in Edmonton.



For both Edmonton and Melbourne models, the ambulance redeployment policy we use is based on formulating the ambulance redeployment problem as a dynamic program and building approximations of the value function. This approximate dynamic programming (ADP) policy takes the current state of the ambulance fleet into consideration when deciding where to reposition an available ambulance, unlike static policies that always reposition an available ambulance back to its fixed home base. The ADP policy is built on Maxwell (2011), where detailed computational experiments show that this policy can provide practically significant improvements over static policies.

A complete version of the lower bound calculations allows ambulances to respond from any node in the road network. This allows for ambulances being redirected to a new call while returning to base, and for ambulances being parked at arbitrary locations throughout a city, and not just bases. Unfortunately, the resulting bound is too loose to be useful. We therefore restrict candidate ambulance locations to bases only. This means the results presented here are not true bounds in the strict sense, but they can certainly be interpreted as bounds from a practical perspective. Indeed, simulation results for representative redeployment policies in Maxwell (2011) show that ambulances respond from bases more than 80% of the time. Furthermore, being stationed

Figure 5. Road network, distribution of demand (darker regions correspond to greater demand), bases (squares), and hospitals (circles) in Melbourne.



Notes. Left panel: The entire region. Right panel: The city center.

away from the amenities of a base is not desirable from the crew perspective.

For both Edmonton and Melbourne models, the values r_i were evenly spaced from 0 to 200 minutes in steps of 0.4 minutes. An interval of width 200 minutes was sufficient to ensure that calls were served within this time window. To compute the location-dependent distribution functions $F_j(\cdot)$ we used numerical convolution with step size 0.2 minutes. The resulting integer programs take approximately 2 days of computation on a system consisting of a 3.4 GHz CPU with 8 GB of memory running under Arch Linux.

For an arrival rate in Edmonton of 4 calls per hour, a 95% confidence interval on the cover bound is $(14.1 \pm 0.1)\%$. The ADP policy obtained using Maxwell et al. (2013) methods achieves $(16.9 \pm 0.4)\%$. When the arrival rate is 6 calls per hour representing a very busy scenario, the cover bound is $(14.7 \pm 0.4)\%$, with the ADP policy achieving $19.7\% \pm 0.4\%$.

For the Melbourne case, we take the average call arrival rate over day and night as our homogenous arrival rate, which is 23 calls per hour. In this case, the cover bound is $(11.2 \pm 0.1)\%$ and our best ADP policy is $(18.3 \pm 0.4)\%$.

6. Computational Results on Artificial Models

The results in the previous section suggest that the Edmonton policy is very close to optimal, while in Melbourne we have less information. Indeed, for Melbourne we do not know whether the bound is weak, the policies are substantially suboptimal, or if both the bounds and the policies need improvement. In this section, we conduct experiments on artificially-constructed cities to gain insight into why we see such differences in performance between Edmonton and Melbourne.

To explore the difference in performance of the bounds on a set of artificial cities, we vary three defining characteristics across a set of 18 total cities:

1. the number of modes in the two-dimensional probability distribution of demand,

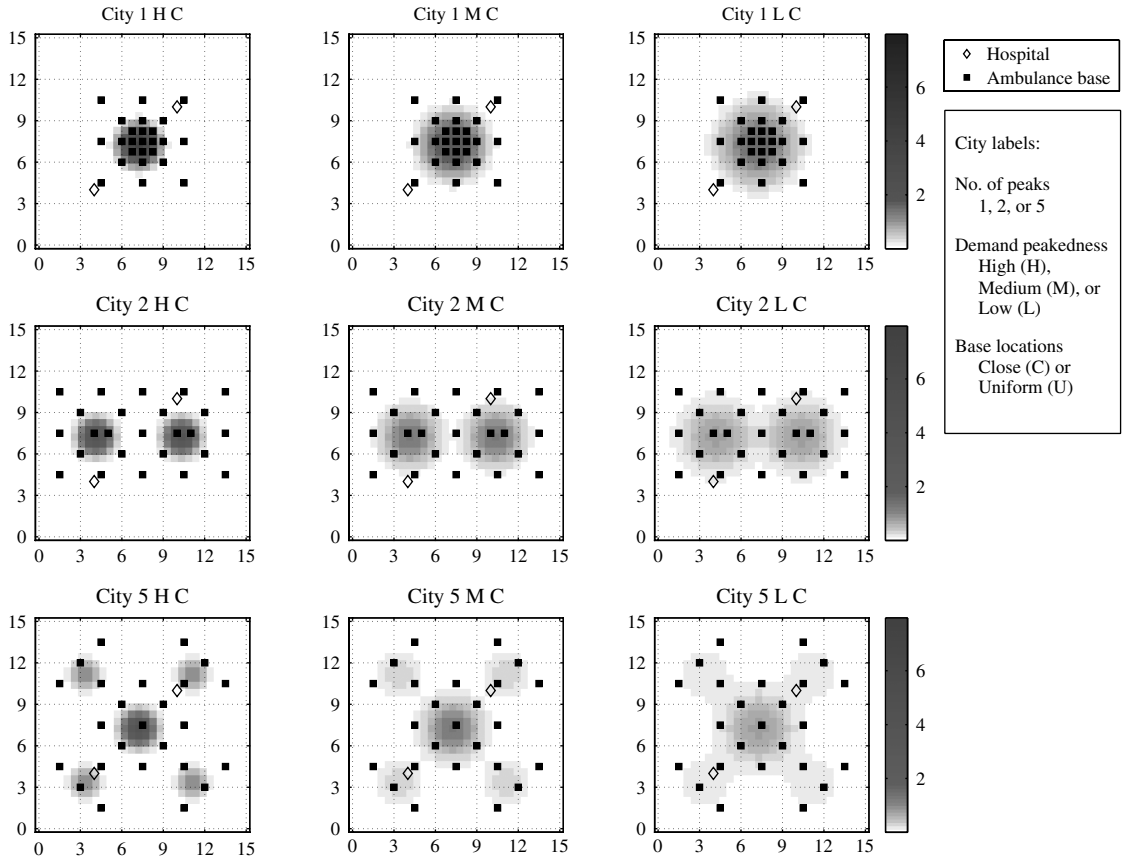
2. the concentration of the demand distribution around its modes, and

3. the degree to which the bases are concentrated under the modes of the demand distribution.

The first factor, the number of modes in the demand density, represents different city configurations. We construct cities with 1, 2, or 5 peaks in demand density, representing one city center, a twin city, or one city center with four suburban areas, respectively. We also vary the second factor, the concentration of the demand distribution, over three different levels. The demand at each peak can be highly concentrated (H), moderately concentrated (M), or lightly concentrated (L). Different levels of demand concentration around the peaks indicate the amount of city sprawl. Finally, we consider two possible configurations of bases in the city. In the first, the ambulance bases are concentrated under the demand (C), and in the second, the ambulance bases are uniformly distributed throughout the city (U). Considering each level of these factors results in 18 cities from a $3^2 \times 2$ full-factorial design. For each city, we compute the performance of the ADP policy and the cover bound, and explore the performance results.

Our fictional cities, nine of which are shown in Figure 6, are 15 miles by 15 miles, with 7 ambulances each traveling at 24 miles per hour. Each city has 25 ambulance bases and 2 hospitals. The word base means a location where ambulances might be asked to wait for their next call, and is often referred to as a post in the industry. We

Figure 6. Distribution of demand, bases, and hospitals in artificial cities.



Notes. Only cities with bases clustered under demand are shown. Cities with uniform bases have a single base in the center of each city square.

use the Manhattan metric to compute the distances from point to point. The call arrival process is Poisson with a constant rate of 3 calls per hour, and call locations are chosen independently from a two-dimensional probability distribution. The density of the location distribution is constant within each of the 25 cells on a 5×5 grid on the city, but the value varies from cell to cell. Figure 6 shows the maps of nine cities, with each panel depicting the demand density (shaded so that regions with high demand intensities are darker), the hospital locations, and the base locations in the clustered-base cities. The base locations for the uniform-base cities are at the centers of the 3 mile-by-3 mile squares of the grid. As shown in Figure 6, each fictional city has two hospitals, where the hospital locations do not vary by city. The name of a fictional city is given by XYZ, where X is the number of peaks, 1, 2, or 5; Y is the peakedness, H, M, or L; and Z is the configuration of the ambulance bases, C or U.

We assume zero turnout time for ambulances responding to a new call. All distributions and parameters are the same as used in the Edmonton and Melbourne models except that when hospital transport is required the choice between the two hospitals is random with probabilities 0.4 and 0.6, respectively. As with those models, after completing a call, an ambulance may be redeployed to one of the bases, unless calls have queued up, in which case the ambulance is deployed to the first call received. We then compute the cover bound and compare it to the performance of the ADP policy. For each of the 18 configurations, we use common random numbers to simulate the bounding system in parallel with the ADP policy for 10,000 iterations.

Figures 7–9 summarize our computational results. Figure 7 shows box plots of the estimated percent of late calls as output from the simulation for each fictional city. An estimated lower bound on the percent of late calls is also shown for each city. We are particularly interested in the gap between the simulation performance and the lower bound; box plots of this gap are plotted for each city in Figure 8. Finally, to better assess trends in the data across the main effects, we show main-effects box plots for the gap between the simulation performance and the lower bound in Figure 9.

From Figure 7, it is clear that as the number of modes in the demand distribution increases, the lower bound and the percent of late calls from the simulated policy both increase. That is, unimodal cases are easier than multimodal cases. We suspect this increase in the percent of late calls occurs because multimodal demand distributions imply less cooperation among the ambulances since each peak is covered somewhat separately from the others. Thus the economies

Figure 7. Box plots of the simulation results for each city and lines for the mean simulation performances and estimated lower bounds.

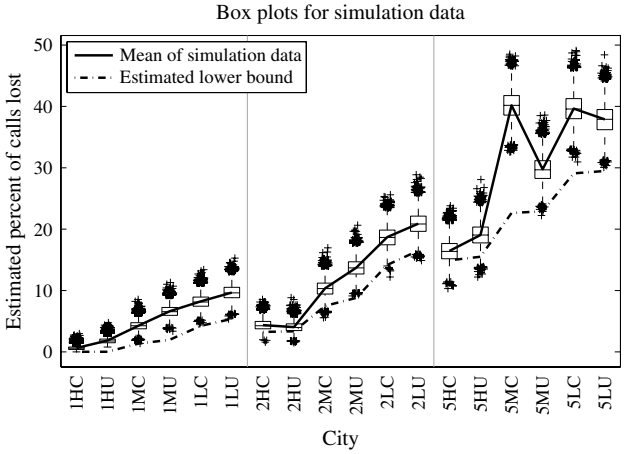
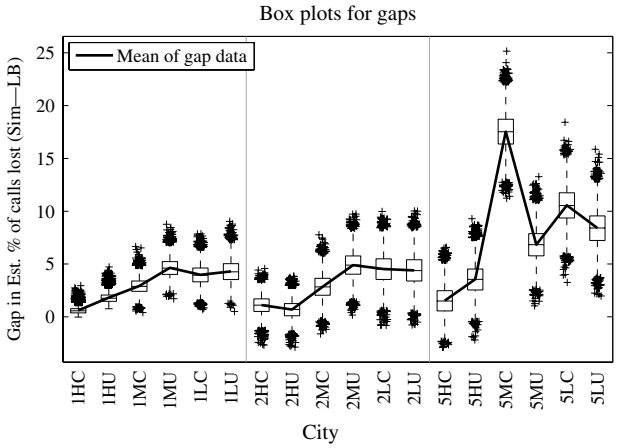


Figure 8. Box plots of the gap between the simulation and lower bound results for each city.



of scale associated with larger queueing systems are lost. Furthermore, the lower bound is only moderately affected by base clustering, which makes sense as the lower bound is based on *covering* demand. As long as base locations are positioned so that large fractions of the demand can be covered with available ambulances, small changes in the location of the ambulance are not important.

Figure 7 also shows anomalous simulation results for cities 5MC and 5LC. While the overall trend for the effect of base locations is that concentrated base locations perform better than uniform base locations, this trend is reversed in the simulation data for two city pairs: 5MC and 5MU, and 5LC and 5LU. While the lower bound is unaffected by this anomaly, the simulation results for cities 5MC and 5LC are unexpectedly poor. This poor performance also reverses the mean trend, but not the median trend, for base locations in Figure 9. To investigate this performance, we plot each of the 5-mode cities in Figure 10, where the size of the base square is proportional to the average percent, over 10,000 simulation runs, of the time an ambulance was reassigned to that base.

Figure 10 broadly shows that clustering the bases under the demand results in more advantageous ambulance assignments, except for cities 5MC and 5LC. In city 5HC, each of the satellite modes is strong enough to warrant ambulance assignments to bases under each of the demand modes,

Figure 9. Box plots of the main effects for the gap between the simulation and lower bound results for each city.

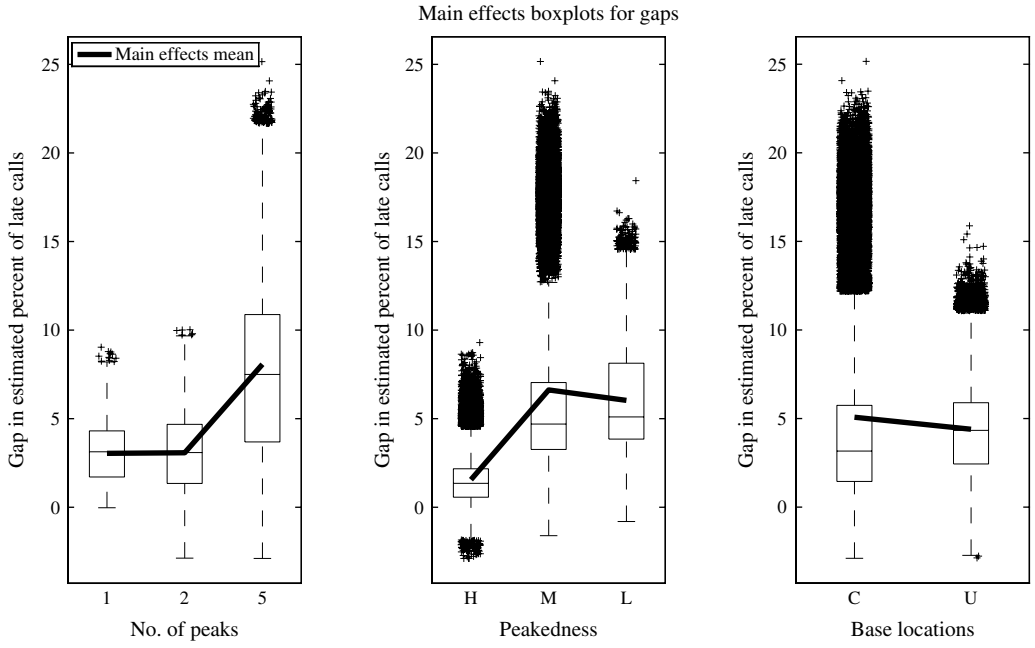
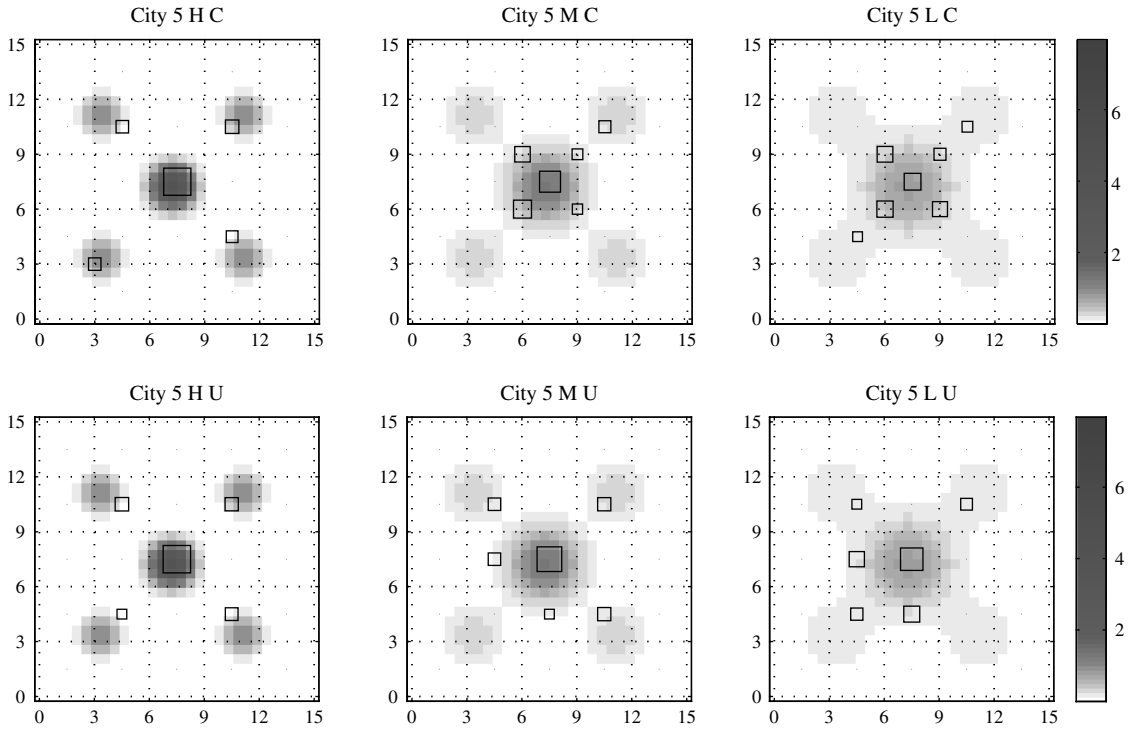


Figure 10. Maps of bases and demands for the artificial cities.



Note. The size of the bases (squares) are proportional to the average percent of ambulance reassignments to that base across 10,000 simulation replications.

as in City 5HU. However in cities 5MC and 5LC, this satellite demand is not strong enough to force assignment of ambulances to bases under the demand, which hurts overall performance. In contrast, because the bases are not clustered in cities 5MU and 5LU, the ambulances are placed nearer the demand modes, resulting in improved performance. Indeed, removing all bases from cities 5MC and 5LC except the five used by city 5HC (the top-left frame of Figure 10), significantly reduced the fractions of missed calls in those cities to levels lower than those in cities 5MU and 5LU.

With these anomalous results explained, we now revisit the gap data in Figures 8 and 9. As the number of demand modes increases, the gap between the lower bound and the simulation data increases, and as the demand distribution becomes less peaked, the gap increases, except for the previously discussed special case of city pairs 5MC, 5MU and 5LC, 5LU. Our results suggest that unimodal cases are easier than multimodal cases, and that as the number of modes increases, the location distribution becomes less concentrated. The cover bound assumes that as long as there is one ambulance near a location, that location is covered. In contrast, quickly reaching calls is expected to become more difficult when location distributions are less concentrated.

As an aside, in all of our cases the demand can be covered with a full complement of ambulances. This is not typically true in practice because city outskirts tend to contain low-density populations. If we were to include such areas in our cities, then we believe that both the lower bound and the ADP policy results would be shifted upwards, thereby disguising the effects we wish to uncover.

The effects above do not appear to be due to different ambulance utilizations in the different cases. We estimated utilizations in all cases to approximately one decimal place, and all estimated utilizations fell between 34% and 39.8%; all estimated utilizations in the 5-mode cases fell between 37.8% and 39.9%. These utilizations levels are typical of EMS systems in practice; we chose our experimental design parameters to ensure that this was the case.

Melbourne has satellite demand points that each represent a nontrivial fraction of overall demand so, as with our artificial cities, it is possible that the policy in Melbourne can be improved by increased ambulance assignments to those locations. To check this, we performed an experiment in which we try to force the ADP policy to use bases at the centers of demand, rather than bases that lie between demand peaks. We first determined the minimal amount of uncovered demand, $\tilde{v}(97)$ for example, with a full complement of 97 available ambulances, given by the optimal objective value of the integer program (3), with the added restriction that ambulances can only be located within the 87 bases. We then found the smallest value m such that the optimal objective function value, $\tilde{v}(m)$, of (3), again with the restriction that ambulances can only be located at bases, is still equal to $\tilde{v}(97)$. The minimal value of m is 60, i.e., we can cover the same fraction of demand with as few as 60 ambulances (stationed only in bases) as we can with 97 ambulances. We then simulate the ADP policy from Maxwell et al. (2013)

as before, but restricting ambulance redeployment to the 60 bases identified as the optimal locations in obtaining $\tilde{v}(60)$. The fraction of late calls for the policy working with these 60 locations was $(17.5 \pm 0.3)\%$, representing a practically significant reduction over the value of $(18.3 \pm 0.4)\%$ obtained when using all 87 bases.

We believe the observations in this section will lead to additional improvements in deployment policies. For example, in the redeployment policies of Maxwell et al. (2013) the value function approximation relies on partitioning demand between ambulance bases; cooperation between bases is ignored. When the system becomes busy, the partition is not an accurate model of operations. We now conjecture that the partition should not be relative to the bases, but rather to the locations of the available ambulances. It is not yet clear how to implement this change in an ADP framework; that is a topic of future research.

7. Conclusions

We developed a bound on the long-run fraction of calls with response times over some threshold for ambulance redeployment policies. This cover bound is built by optimally positioning available ambulances at the time calls are received, and using a comparison-of-queues (i.e., a coupling) to ensure that a certain bounding queueing system always has more ambulances available than would be the case for any redeployment policy. By simulating the bounding system, we obtain the cover bound.

Our results on realistic but not real cases based on the cities of Edmonton and Melbourne indicate that the bound is very tight in Edmonton, but for the much larger city of Melbourne, there is a nontrivial gap between the bound and the performance of our best redeployment policy to date. Simulation results for artificial cities suggest that the larger gap in Melbourne may be primarily due to a less effective policy, rather than a poor bound. Based on these results we subsequently improved the policy in Melbourne by a practically significant amount, although an appreciable gap remains between the performance of our best policy to date (approximately 17.5% of calls are late) and our bound (approximately 11.2% of calls are late). Further improvements to the redeployment policy in Melbourne are the subject of future research.

Our assumption that travel times are deterministic and do not depend on time is restrictive. We would like to explore methods for avoiding this assumption, perhaps through a more sophisticated calculation that allows time dependence or by applying our bounding approach separately to time periods over which the travel-time assumption is approximately satisfied.

Acknowledgments

The authors thank Andrew Mason, Armann Ingolfsson, and Ambulance Victoria for helpful discussions and data, and the editorial team for comments that improved the paper. This work was partially supported by National Science Foundation [Grants CMMI-0758441 and CMMI-1200315] and by the Extreme Science and Engineering Discovery Environment (XSEDE) [Grant DDM-120004]. XSEDE is supported by NSF [Grant OCI-1053575].

References

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

Andersson T (2005) Decision support tools for dynamic fleet management. Unpublished doctoral dissertation, Department of Science and Technology, Linkoepings Universitet, Norrkoeping, Sweden.

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

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.* 8:105–118.

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

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

Brown DB, Smith JE, Sun P (2010) Information relaxations and duality in stochastic dynamic programs. *Oper. Res.* 58(4, Part 1 of 2):785–801.

Church R, ReVelle C (1974) The maximal covering location problem. *Papers of the Regional Sci. Assoc.* 32:101–108.

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

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

Henderson SG (2009) Operations research tools for addressing current challenges in emergency medical services. Cochran JJ, ed. *Wiley Encyclopedia of Operations Research and Management Science* (John Wiley & Sons, Hoboken, NJ).

Ingolfsson A (2012) EMS planning and management. Zaric G, ed. *Operations Research and Health Care Policy*, International Series in Operations Research and Management Science, Vol. 190 (Springer, New York), 105–128.

Kiefer J, Wolfowitz J (1955) On the theory of queues with many servers. *Trans. Amer. Math. Soc.* 78(1):1–18.

Mason AJ (2013) Simulation and real-time optimised relocation for improving ambulance operations. Denton B, ed. *Handbook on Healthcare Operations Management*, International Series in Operations Research and Management Science, Vol. 184 (Springer, New York), 298–318.

Maxwell MS (2011) Approximate dynamic programming policies and performance bounds for ambulance redeployment. Unpublished doctoral dissertation, School of Operations Research and Information Engineering, Cornell University, Ithaca, NY.

Maxwell MS, Henderson SG, Topaloglu H (2013) Tuning approximate dynamic programming policies for ambulance redeployment via direct search. *Stochastic Systems* 3(2):322–361.

Maxwell MS, Restrepo M, Henderson SG, Topaloglu H (2010) Approximate dynamic programming for ambulance redeployment. *INFORMS J. Comput.* 22(2):266–281. http://joc.journal.informs.org/cgi/content/abstract/ijoc.1090.0345v1.

Nair R, Miller-Hooks E (2009) Evaluation of relocation strategies for emergency medical service vehicles. *Transportation Res. Record* 2137:63–73.

Naoum-Sawaya J, Elhedhli S (2013) A stochastic optimization model for real-time ambulance redeployment. *Comput. Oper. Res.* 40(8): 1972–1978.

Ni EC, Hunter SR, Henderson SG, Topaloglu H (2012) Exploring bounds on ambulance deployment policy performance. Laroque C, Himmelspach J, Pasupathy R, Rose O, Uhrmacher AM, eds. *Proc. 2012 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 497–508.

Richards DP (2007) Optimised ambulance redeployment strategies. Unpublished thesis, Department of Engineering Science, University of Auckland, Auckland, New Zealand.

Stoyan D (1983) *Comparison Methods for Queues and Other Stochastic Models* (John Wiley & Sons, New York).

Yue Y, Marla L, Krishnan R (2012) An efficient simulation-based approach to ambulance fleet allocation and dynamic redployment. *AAAI Conf. Artificial Intelligence (AAAI), Special Track Comput. Sustainability and Artificial Intelligence*.

Zhang L (2010) Optimisation of small-scale ambulance move-up. Ehrgott M, Mason A, eds. *Proc. 45th Annual Conf. Oper. Res. Soc. New Zealand* (Operations Research Society of New Zealand, Auckland, New Zealand), 150–159.

Zhang L (2012) Simulation optimisation and Markov models for dynamic ambulance redeployment. Unpublished doctoral dissertation, The University of Auckland, Auckland, New Zealand.

Zhang L, Mason A, Philpott A (2010) Optimization of a single ambulance move up. Technical report, University of Auckland Faculty of Engineering. Accessed July 8, 2014, https://researchspace.auckland.ac.nz/handle/2292/6031.

Matthew S. Maxwell is a Senior Operations Research Specialist in the Revenue Management and Pricing Optimization division at SAS Institute, Inc. His research interests include approximate dynamic programming, stochastic simulation, and revenue management. His current focus is network revenue management and customer choice revenue management models for hospitality industries.

Eric Cao Ni is a doctoral student in the School of Operations Research and Information Engineering at Cornell University. His research interests include simulation optimization, emergency services, and queuing theory.

Chaoxu Tong is a doctoral student in the School of Operations Research and Information Engineering at Cornell University. His research interests include design of approximation algorithms for NP-hard problems and proving inapproximability results for such problems. His current focus is logistics problems and problems with online structure with applications in timely delivery services.

Shane G. Henderson is a professor in the School of Operations Research and Information Engineering at Cornell University. His primary research interests lie in simulation and simulation optimization, and in emergency-services applications.

Susan R. Hunter is an assistant professor in the School of Industrial Engineering at Purdue University. Her research interests include Monte Carlo methods and simulation optimization.

Huseyin Topaloglu is a professor in the School of Operations Research and Information Engineering at Cornell University. His research interests include stochastic programming and approximate dynamic programming with applications in revenue management, inventory control, and healthcare systems.