A Markov Chain Model for an EMS System with Repositioning

Ramon Alanis, Armann Ingolfsson, and Bora Kolfal
School of Business, University of Alberta, Edmonton, AB T6G 2R6, Canada,
ramon@ualberta.ca, armann.ingolfsson@ualberta.ca, bora.kolfal@ualberta.ca

We propose and analyze a two-dimensional Markov chain model of an Emergency Medical Services system that repositions ambulances using a compliance table policy, which is commonly used in practice. The model is solved via a fixed-point iteration. We validate the model against a detailed simulation model. We demonstrate that the model provides accurate approximations to various system performance measures, such as the response time distribution and the distribution of the number of busy ambulances, and that it can be used to identify near-optimal compliance tables. Our numerical results show that performance depends strongly on the compliance table that is used, indicating the importance of choosing a well-designed compliance table.

Key words: Health care; Ambulance service; Repositioning; System status management; Fleet management

Date for this draft: August 20, 2010.

1. Introduction

Emergency medical services (EMS) systems are designed and operated with the aim of minimizing response times to emergency calls. In order to reduce the response time most emergency systems adapt to changing conditions using “system status management”—a set of strategies that include dynamic repositioning, by which dispatchers “move” ambulances to provide better coverage. In this paper, we propose, validate, and illustrate the use of a two-dimensional Markov chain model of an EMS system that uses a form of repositioning based on compliance tables.

Repositioning strategies are made possible by computer-aided dispatch (CAD) systems and global positioning system (GPS) technology, which make it possible for dispatchers to keep track of the location and status of every ambulance in the system. Also known as “flexing,” repositioning strategies stand in contrast to every ambulance returning to its “home station” at the conclusion of every call, which is how many EMS systems operated in the past, and some still do. Surveys of North American EMS operators in 2001 (Cady 2002) and 2009 (Williams 2009) showed the percentage of operators who used a static deployment strategy decreasing from 41% to 30%, those using a dynamic strategy increasing from 23% to 37%, and those using a combination of static and dynamic strategies changing from 36% to 33%.

Many operations research models of ambulance operations, for example the Hypercube Queueing Model (HQM, Larson 1974), assume a static deployment policy, either implicitly or explicitly, where ambulances are assigned to a fixed home station regardless of the system state; while in systems that use repositioning, dispatchers may “move” one or more ambulances when there is a change in system state (a call arrival or a call completion). A well-designed repositioning policy can improve performance through better dynamic matching of ambulance supply and call demand. For example, a simulation study for the Edmonton EMS system (Erkut et al. 2005), which was using a compliance table policy at the time, indicated that to achieve equal performance without repositioning would require the addition of eight new ambulances, around the clock.

Repositioning policies are controversial among EMS workers, because such policies increase the time workers need to spend in their vehicles, which may lead to increased back problems (Morneau and Stothart 1999). Bledsoe (2003) questions the effectiveness of system status management and summarizes the concerns of EMS staff, whose workloads are increased by repositioning. On the other hand, Stout (1989) argues in favor of system status management and details some misconceptions and common implementation problems which could potentially render it ineffective, or result in excessive stress for paramedics.

Response time for an EMS system is the time interval from the arrival of an emergency call until the ambulance reaches the scene of the incident. The performance goal for most EMS systems is to have the response times of at least a certain percentage of high priority calls below a specific time threshold. For example, having response times of at least 90% of urgent calls below 9 minutes is a common performance goal in North America (Fitch 2005). Our Markov chain model can be used

Available Ambulances	Stations							
1	42/RAH							
2	42	42/RAH						
3	42	42/RAH	32					
4	42	42/RAH	32	33				
5	42	42/RAH	28/32	33	34/27			
6	42	42/RAH	28/32	33	34/27	7/25/NEC		
7	42	42/RAH	28/32	33	34/27	7/25/NEC	9/19/GNH/MIS	
8	42	42/RAH	32/UAH	33	34/27	7/25/NEC	9/GNH	19/28/MIS

Table 1 Compliance table example.

to compute such performance measures as the response time distribution, the distribution of the number of busy ambulances, expected travel time, and average ambulance utilization.

The repositioning policy that we focus on uses a so-called "compliance table." Table 1 provides an example of a real compliance table that has been used in Edmonton, Alberta. Each row in a compliance table shows, for a given number of available ambulances, the desired ambulance locations. For example, based on Table 1, with one available ambulance, it should be at station 42 or at the Royal Alexandra Hospital (RAH); with two available ambulances, one should be at station 42 and the other at station 42 or at RAH, and so on. When the system state changes, the dispatchers decide which ambulances to move in order to reach compliance in the new system state. We view the choice of a compliance table as a tactical decision, while the choice of moves to reach compliance is a real-time operational decision. We focus on the tactical compliance table level but our model can accommodate different algorithms for the choice of moves to reach compliance. We also envision our model being used as part of a larger model to support such strategic decisions as when and where to build new ambulance stations or close existing ones.

In practice (Stout 1989), dispatchers usually choose the moves. Sometimes, the dispatchers use information that is not captured in the CAD system. For example, with two available ambulances located at stations 42 and 32, according to the compliance table, the ambulance at station 32 should be moved to either station 42 or RAH. However, if a dispatcher knows that an ambulance crew expects to finish its current call at RAH in the next five minutes, then he might decide that no moves are necessary. In some EMS systems that we are familiar with, dispatchers are evaluated, in part, based on the fraction of time that the system is "in compliance," with the system considered to be in compliance as soon as moves have been initiated that, when completed, will results in ambulance locations that are consistent with the compliance table.

Real-time move choice decisions could be supported by computer systems that suggests moves or the decision could even be completely automated. Commercial systems that provide such functionality have started to appear and we discuss related academic work later in this section.

Our Markov chain model has two state variables: The number $B(t)$ of busy ambulances and an indicator variable $C(t)$ for whether the system is in compliance or not. The model has three transition types: Call arrivals, call completions, and reaching compliance. We assume a constant exogenous arrival rate but completion rates are state-dependent and we determine them by analyzing the components of an EMS call: Evaluation and dispatch, chute, travel, on scene, transport, and hospital times. The rates at which the system reaches compliance are also state-dependent. We posit that the call completion rates depend on whether the last change to the number of busy ambulances was the result of a call arrival or a call completion. This results in transition rates that depend on state probabilities and therefore, to solve the model, we iterate between computing steady state probabilities and computing transition rates. After obtaining the steady state probabilities, we use convolution to approximate the response time distribution.

We focus our review of related work on models for performance evaluation and optimization of fleets of emergency vehicles, particularly ambulances. Green and Kolesar (2004) provide a recent review of the literature on the use of management science to support the design and operation of emergency service systems. Kolesar and Walker (1974) were one of the first to model repositioning, specifically of fire companies for New York City. This work helped to establish repositioning as an effective technique to improve performance, and the repositioning strategies proposed in Kolesar and Walker (1974) were still in use during the World Trade Center terrorist attacks of September 11, 2001 to re-balance the remaining units to maintain service for the rest of the city (Green and Kolesar 2004).

Three distinctions are important in relating our work to past research: (1) Whether repositioning is incorporated or not, (2) whether repositioning is done at planned times, in anticipation of predictable shifts in demand and travel speeds, or dynamically, based on the current system state, and (3) the time horizon for dynamic repositioning. We discuss each of these distinctions in turn.

Although HQM does not incorporate repositioning, it is an important point of departure for our model. HQM was initially introduced by Larson (1974) as a Markov chain model of a queueing system with *distinguishable* servers. In an EMS context, what makes a server (an ambulance) distinguishable from other servers is its home station, to which it is assumed to return at the end of every call. In a system with repositioning, home stations are much less important and it becomes reasonable to consider the servers to be indistinguishable. Viewing the servers as identical, we were able to construct a model that is more tractable than the HQM, even though the operations of a system that uses repositioning are more complicated than the operations of a system where ambulances always return to their home stations. Specifically, for a system with n ambulances, the cardinality of the HQM Markov chain's state space of 2^n grows exponentially with n, while our model's state space cardinality of $2n+1$ is linear in n. To address the computational and storage requirements of the exact HQM, researchers have developed a series of approximate versions (Larson 1975, Jarvis 1985, Budge et al. 2009) of the HQM, but such approximations are not necessary for our model.

We mention preplanned repositioning only to emphasize that it differs from the dynamic repositioning that we consider. See Rajagopalan et al. (2008) for recent work on preplanned repositioning.

Dynamic repositioning policies can be modeled using Markov decision processes or stochastic dynamic programming. Berman (1981a,b) took this approach and demonstrated how one could find optimal solutions for small systems. More recently, Maxwell et al. (2009, 2010) used approximate dynamic programming to find solutions for larger systems with more realistic assumptions. These approaches provide policies that prescribe optimal or near-optimal repositioning moves, taking into account the probability distribution of future consequences of the moves. In contrast, Gendreau et al. (2001) proposed a parallel tabu search heuristic for the real-time repositioning of ambulances, given the current locations of all available ambulances and demand rates throughout the region, which one can view as forecasts for the location of the next call to arrive. We view the compliance table policies that we analyze as being at the tactical level. A compliance table provides a high-level and easy-to-understand policy, which must be complemented with a method for real-time decisions about how to reach compliance. That method could, for example, be a heuristic similar to the one proposed by Gendreau et al. (2001). Even though compliance table policies could be suboptimal, they have the advantage of being much easier to use than optimal dynamic programming policies and already being accepted practice among many EMS operators. Compliance tables do have the drawback that, to our knowledge, no tractable analytical models have been available to predict their impact on system performance. Our aim in this paper is to provide such a model.

We make the following contributions:

1. We propose and analyze a tractable analytical model that has the same data requirements and can produce the same outputs as the HQM, but models repositioning policies, which HQM does not do.

2. We validate the model against a realistic simulation model and find that the Markov chain model provides a good approximation to several performance measures.

3. We demonstrate that the Markov chain model can be used to identify solutions that are near-optimal, as measured by a realistic simulation model.

4. Our numerical results show that different compliance tables lead to large variations in performance, which demonstrates the importance of using a well-designed compliance table.

In Section 2, we describe the operations of an EMS system and discuss the components of the service and response time. Section 3 presents the Markov chain model and an iterative algorithm to obtain the steady state probabilities. Section 4 discusses estimation of the input parameters for the Markov chain model. Section 5 describes how we approximate the response time distribution using convolution. In Section 6, we discuss how we validated the results of the Markov chain model by comparing them to simulation results. Section 7 presents a numerical study to show that our Markov chain model can be used by EMS system managers to choose the best or a near-best compliance table. Finally, in Section 8, we comment on the importance of our model and summarize its performance.

2. EMS System with Repositioning

The purpose of this section is to describe how a real EMS system operates. We support the discussion with 2008 empirical data from the Edmonton, Alberta, EMS system. EMS systems differ in their operational practices and our description will not apply in all respects to all such systems, but we believe that our description captures the main aspects of EMS operations for many EMS operators in medium to large urban areas in North America.

We have developed a discrete event simulation model of an EMS system, for validation purposes. The simulation model follows the description in this section. In the following section, we present our Markov chain model and explain the assumptions that it is based on.

One important assumption that we make is stationarity. Specifically, we assume that calls arrive to the system according to a homogeneous Poisson process with rate λ and the number of ambulances (servers) on duty is fixed at n. In reality, the arrival rate and the number of servers varies by time of day and day of the week. We envision our model being used separately for periods (say, hours) over which the arrival rate and number of servers remain at least approximately constant.

We aggregate city neighborhoods into demand nodes and we denote the set of demand nodes by D, with cardinality $|D|$. The probability that a particular call arrives to demand node d is f_d, with $\sum_{d\in D} f_d = 1$. We use S and H to denote the sets of stations and hospitals, with cardinalities $|S|$ and $|H|$. For calls that require transport, the probability of transport to hospital h is g_h, with $\sum_{h\in H} g_h = 1$.

We organize the discussion around the response to a typical emergency call, that is, the sequence of steps followed by the EMS system from the moment when a call arrives until the moment when the emergency crew finishes their duties for the call, as illustrated in Figure 1. If one views an ambulance as a server and the EMS system as a queueing system, then it is important to distinguish between the ambulance *service time* and the system *response time.* Figure 1 illustrates the difference between the two and we discuss it throughout this section.

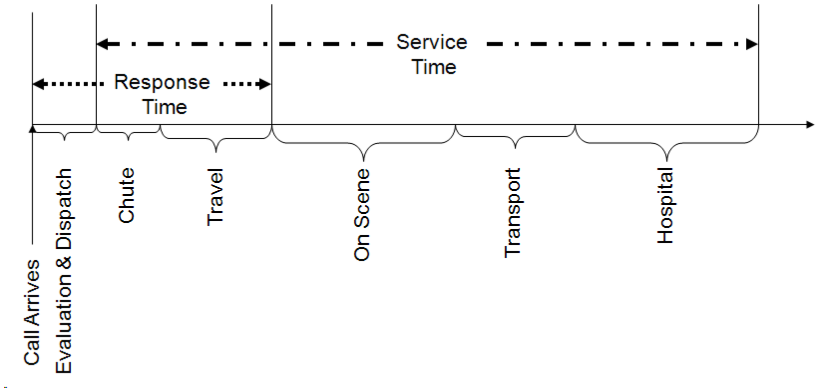

Figure 1 Service and Response Times

We define random variables to represent the various time intervals. These random variables will be conditional on the system state (b,c), where b is the number of busy ambulances and c equals one if the system is in compliance and zero otherwise. (We use these state variables in the Markov chain model described later but we recognize that the real system, with only these two state variables, is unlikely to satisfy the Markov property.) The sequence of time instants and corresponding time intervals for a typical call is:

- Call arrival: An medical emergency occurs somewhere in the EMS system's service area and the patient or a bystander calls 911.

- Dispatch: A dispatcher answers the call, evaluates the situation, and dispatches the ambulance closest to the scene of the incident. We define $R_{b,c}^{\text{Eval and Disp}}$ for the evaluation and dispatch time.

- Chute: The crew of the dispatched ambulance receives the notification and boards the ambulance, if they are not already on board, and start driving towards the scene of the incident. We define $S_{b,c}^{\text{Chute}}$ for the chute time.

- Travel: The dispatched ambulance moves from its original location to the scene of the incident. We define $S_{b,c}^{\text{Travel}}$ for the travel time.

- On-Scene: The crew locates the patient, provides emergency medical attention and, if required, transports the patient to the ambulance. In some cases the call ends at the scene of the incident, whereas in others, the patient requires transportation to a hospital (event T). We define $S_{b,c}^{\text{On-Scene}}$ for the on-scene time.

- Transport: If transportation is required, then the ambulance transports the patient to a hospital. We define $S_{b,c}^{\text{Transport}}$ for the transport time.

- Hospital: At the hospital, the crew transfers the patient to the care of the emergency room personnel. This is the end of the call. We define $S_{b,c}^{\text{Hospital}}$ for the hospital time.

- Relocation: After the call has been completed, either at a hospital or at the scene of the incident, the ambulance travels to a station (possibly different from its original station) to wait for the next call. (At the same time, the dispatcher may ask other ambulances to move, in order to reach compliance.) During this time the ambulance is free, and it can be assigned to a new call.

2.1. Service Time vs. Response Time

Service time, $S_{b,c}$, as illustrated in Figure 1, is the total time an ambulance and crew remain "busy" with a call, i.e., not available to take other calls. It is composed of the chute, travel, and on-scene time, as well as transport and hospital times for calls that require transport (event T):

$$S_{b,c} = \begin{cases} S_{b,c}^{\text{Chute}} + S_{b,c}^{\text{Travel}} + S_{b,c}^{\text{On-Scene}} + S_{b,c}^{\text{Transport}} + S_{b,c}^{\text{Hospital}} & \text{if } T, \\ S_{b,c}^{\text{Chute}} + S_{b,c}^{\text{Travel}} + S_{b,c}^{\text{On-Scene}} & \text{Otherwise.} \end{cases} \tag{1}$$

The response time $R_{b,c}$ is the basis for most EMS performance measures. It is the time from the moment when a call is received to the moment when the ambulance arrives to the call address, composed of the evaluation and dispatch, chute, and travel times:

$$R_{b,c} = R_{b,c}^{\text{Eval and Disp}} + S_{b,c}^{\text{Chute}} + S_{b,c}^{\text{Travel}}. \tag{2}$$

In the remainder of this section we will discuss the components of the service and response time in more detail.

2.2. Evaluation and Dispatch Time; Chute Time

Evaluation and dispatch is the process by which the emergency personnel answering the call question the caller to obtain basic information such as the nature of the emergency, and the location where the help is needed and based on this information, decide which unit should be called. Chute time is the time from the moment when an ambulance is dispatched, to the moment when the vehicle starts moving. The left panel of Figure 2 shows that the average evaluation and dispatch

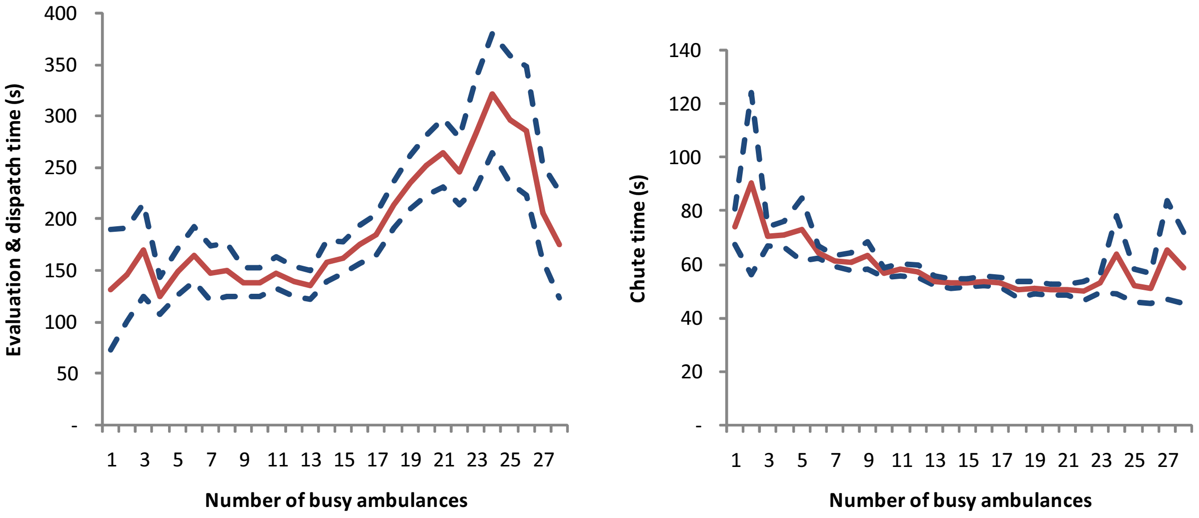

Figure 2 **Means and 95% confidence intervals for evaluation and dispatch time and for chute time .**

time increased with the number of busy ambulances, possibly because of increased workload on the control center. In contrast, the right panel shows that the average chute time decreased with the number of busy ambulances, likely because the probability that an ambulance responds while traveling (as opposed to being at a station) increases with the number of busy ambulances. (Figure 2 and all subsequent figures in this section are based on 2008 Edmonton data.)

2.3. Travel Time and Transport Time

Travel time is the largest component of the response time, and it is the component that is most influenced by repositioning. Accordingly, we need to model travel times carefully. Travel time depends not only on the number $n-b$ of available ambulances but also on the locations of those ambulances. The basis for our Markov chain model is the assumption that the *locations* of the available ambulances are determined largely by the *number* of available ambulances, because of the use of a compliance table. The compliance table does not provide the exact locations of the ambulances at all times, but the dispatchers continuously relocate ambulances attempting to match the locations specified by the compliance table.

We define a set Γ of possible origin-destination pairs as:

$$\Gamma = \{(o,d) : o \in S \cup H \cup D \text{ and } d \in D\}, \tag{3}$$

where the origin o is an ambulance location and the destination d is a demand node. Our travel time model for a given origin-destination pair is based on the work of Kolesar (1975) and Budge et al. (2010), with two exceptions, which we note in a moment. Let l be the distance from the origin o to the destination d. Parameters a (acceleration) and v_c (cruising speed) determine the median travel time $m(l)$ and parameters b_0, b_1, b_2 determine the travel time coefficient of variation $c(l)$ for the travel time $T(l)$ from o to d. The model is:

$$m(l) = \begin{cases} 2\sqrt{l/a} & l \le v_c^2/a \\ v_c/a + l/v_c & l > v_c^2/a \end{cases} \tag{4}$$

$$c(l) = \min\left\{ \frac{\sqrt{b_0(b_2+1) + b_1(b_2+1)m(l) + b_2 m(l)^2}}{m(l)}, c_{\max} \right\} \tag{5}$$

$$T(l) = m(l)\exp(c(l)\epsilon) \tag{6}$$

where ϵ follows a standard normal distribution.

Budge et al. (2010) found that a Student's t distribution for ϵ fit empirical data better than a standard normal distribution. The heavier tails of the t-distribution make the statistical estimation method used in Budge et al. (2010) more robust against errors in recorded travel times, but unfortunately, if ϵ follows a t distribution, then $T(l)$ has infinite expected value. We assume a standard normal distribution instead, to avoid having to solve a queueing model with a service time whose mean is infinite. The other difference between our travel time model and that of Budge et al. (2010) is that we set an upper bound $c_{\max}$ on the coefficient of variation, to avoid implementation difficulties when the distance l approaches zero.

By using (6) and the moment generating function for a normal distribution, we obtain the first two moments of T as follows:

$$E[T(l)] = m(l)\exp\left(c(l)^2/2\right), \tag{7}$$

$$E[(T(l))^2] = (m(l))^2 \exp\left(2c(l)^2\right). \tag{8}$$

We use the same equations ((4)-(6)) to model transport time from a given origin to a given destination, but with different parameters to reflect that transport to hospital is often less urgent

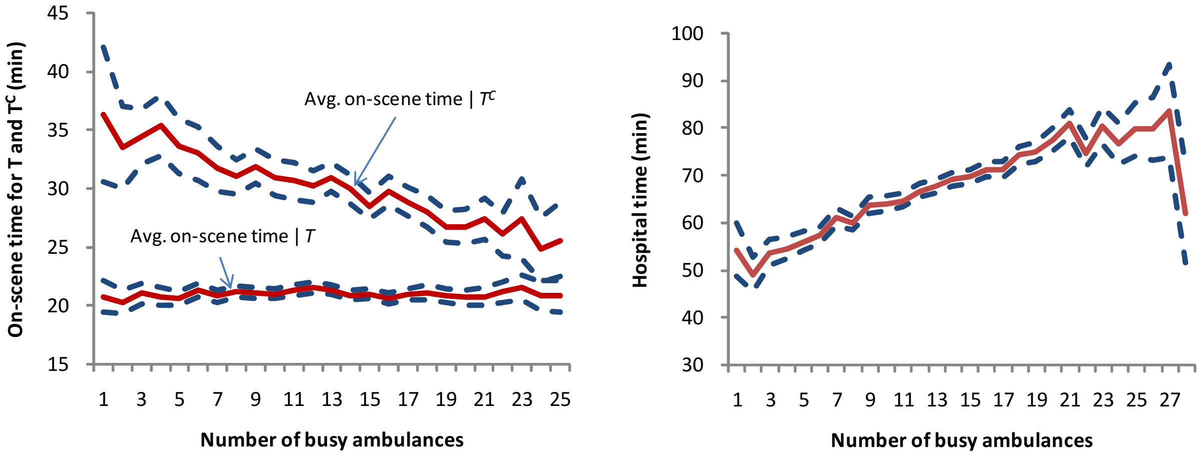

Figure 3 **Means and 95% confidence intervals for on-scene time and hospital time.**

than reaching the call address. The transport time is zero if the call does not require transport, which we denote as follows:

$$S_{b,c}^{\text{Transport}} = \begin{cases} S_{b,c}^{\text{Transport}|T} & \text{if } T, \\ 0 & \text{Otherwise.} \end{cases} \tag{9}$$

2.4. On-Scene Time and Hospital Time

On-scene time may vary between calls that require transportation to a hospital and those that do not. We represent this as follows:

$$S_{b,c}^{\text{On-Scene}} = \begin{cases} S_{b,c}^{\text{On-Scene}|T} & \text{if } T, \\ S_{b,c}^{\text{On-Scene}|T^C} & \text{Otherwise.} \end{cases} \tag{10}$$

The left panel of Figure 3 shows that the average on-scene time is approximately constant for calls that require transport but the average is higher and it decreases with the number of busy ambulances for calls that do not require transport. Possibly, ambulance staff provide some discretionary care (that could be performed by other health care providers later) to patients that do not require transport, especially if the EMS system is uncongested.

Hospital time is zero if the call does not require transport, and we represent this as follows:

$$S_{b,c}^{\text{Hospital}} = \begin{cases} S_{b,c}^{\text{Hospital}|T} & \text{if } T, \\ 0 & \text{Otherwise.} \end{cases} \tag{11}$$

In contrast to on-scene time, the right panel of Figure 3 shows that the average hospital time

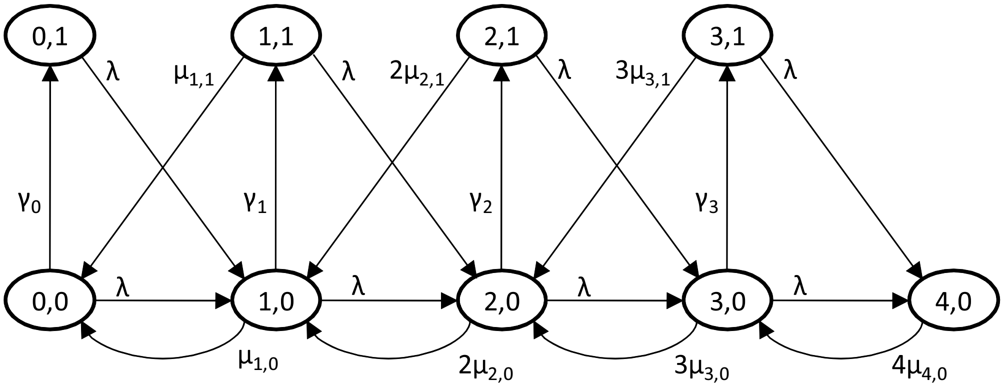

Figure 4 Transition rate diagram for $n = 4$

increases with the number of busy ambulances, which is what one would expect, because when the EMS system is busy, hospital emergency departments are likely to be busy as well.

With this, we have completed our discussion of the components of the service and response times. The discrete event simulation model that we developed for validation purposes follows the assumptions set out in this section. In addition, the simulation model includes a move choice module, based on a greedy heuristic (described in Section 4), which attempts to minimize the number of moves needed to reach compliance and uses minimization of total estimated time to reach compliance as a tie breaker. This module could be replaced with another move choice algorithm, such as the tabu search heuristic proposed by Gendreau et al. (2001)—both in our simulation model and in the Markov chain model, which we describe next.

3. Markov Chain Model

The model we use to approximate an EMS system with n ambulances that are repositioned according to a compliance table policy is a finite, continuous-time Markov chain, with state variables $B(t)$, the number of busy ambulances, and $C(t)$, an indicator variable for whether the system is in compliance at time t. The state space for this process is $\Omega \equiv \{(b,c) : b = 0,\ldots,n-1, c = 0,1\} \cup \{(n,0)\}$. In state $(n,0)$, all n ambulances are busy and therefore the term compliance is not meaningful. We arbitrarily set $C(t) = 0$ when $N(t) = n$. Figure 4 show a transition rate diagram for this process when $n = 4$.

We assume that whenever the number of busy ambulances increases or decreases, the system goes out of compliance. This assumption is made to limit the number of configurations that we need to consider when computing expected travel times.

The Markov chain has three types of transitions:

Call arrival: Call arrivals occur at rate λ. Arrivals are lost when $B(t) = n$. We assume that the system is out of compliance immediately after an arrival, and therefore an arrival to state $(b, 0)$ and an arrival to state $(b, 1)$ both result in a transition to $(b+1, 0)$, for $b = 0, \ldots, n-1$.

Call completion: Call completions occur at rate $b\mu_{b,c}$ when in state (b, c), for $b > 0$. Similar to arrivals, we assume that the system is out of compliance immediately after a call completion and therefore a call completion when in state $(b, 0)$ and a call completion when in state $(b, 1)$ both result in a transition to state $(b-1, 0)$, for $b > 0$. We interpret $\mu_{b,c}$ as the rate at which each individual busy unit completes its call, given that the system is in state (b, c).

Achieve compliance: When the system is out of compliance, in state $(b, 0)$, we assume it reaches compliance at rate γ_b, resulting in a transition to state $(b, 1)$, for $b < n$.

If $\mu_{b,c}$ is independent of b and c and if γ_b approaches infinity for all b, then this model approaches the Erlang B (or $M/M/n/n$) loss model. Given that the steady state probabilities of the Erlang B model are insensitive to the service time distribution beyond its mean (Gross and Harris 1998, p. 245), the similarity of our model to the Erlang B model suggests that the steady state probabilities for our model might be relatively insensitive to the shape of the service time distribution. We confirm, in Section 6, that the steady state probabilities for the Markov chain model are close to those of a simulation model that makes more realistic assumptions about the distributions of service time and time to reach compliance (see Figure 6).

If all transition rates are known, then it is straightforward to compute the steady state probabilities, $\pi_{b,c}$, for our model, as we describe next. For convenience in stating our equations, we set $\mu_{0,1} = \mu_{n,1} = \pi_{n,1} = 0$. The local balance equation for state $(b, 1)$ is

$$(\lambda + b\mu_{b,1})\pi_{b,1} = \gamma_b \pi_{b,0} \Rightarrow \pi_{b,1} = \frac{\gamma_b}{\lambda + b\mu_{b,1}} \pi_{b,0}, \text{ for } b = 0, \ldots, n-1, \tag{12}$$

and this equation allows us to compute $\pi_{b,1}$ when $\pi_{b,0}$ is known. Next, for a given b, if we consider transitions between states in $\{(b',c)\in\Omega : b'\leq b\}$ and states in $\{(b',c)\in\Omega : b'\geq b+1\}$, we obtain the following balance equation:

$$\lambda(\pi_{b,0}+\pi_{b,1})=(b+1)\mu_{b+1,0}\pi_{b+1,0}+(b+1)\mu_{b+1,1}\pi_{b+1,1} \text{ for } b=0,\ldots,n-1. \tag{13}$$

By combining equations (12) and (13), we obtain the following, which allows us to compute $\pi_{b,0}$ once $\pi_{b+1,0}$ and $\pi_{b+1,1}$ are known:

$$\pi_{b,0}=\frac{(b+1)(\lambda+b\mu_{b,1})(\mu_{b+1,0}\pi_{b+1,0}+\mu_{b+1,1}\pi_{b+1,1})}{\lambda(\lambda+\gamma_b+b\mu_{b,1})}, \quad \text{for } b=0,\ldots,n-1. \tag{14}$$

We compute the steady state probabilities recursively, starting with state $(n,0)$, as follows.

Algorithm 1:

1. Set $\pi_{n,0}\leftarrow 1$.

2. Decrement b from $n-1$ to 0 in steps of 1. For each value of b, do the following: Use (14) to compute $\pi_{b,0}$ and then use (12) to compute $\pi_{b,1}$.

3. Normalize: Set $\Pi\leftarrow\sum_{(b,c)\in\Omega}\pi_{b,c}$ and $\pi_{b,c}\leftarrow\pi_{b,c}/\Pi$.

The call completion rates, $\mu_{b,0}$, and the rates at which compliance is reached, γ_b, from the non-compliant states depend on exactly where the $n-b$ available ambulances are located. To clarify this dependence, we make the following approximation:

$$\mu_{b,c}=1/\mathrm{E}[S_{b,c}]. \tag{15}$$

Similarly, we define random variable L_b as the time it takes to achieve compliance from state $(b,0)$, by repositioning one or more ambulances, and we use the following approximation:

$$\gamma_b=1/\mathrm{E}[L_b]. \tag{16}$$

These are approximations because the service rates and the rates of reaching compliance are state-dependent. The non-compliant state $(b,0)$ must have been entered either as a result of a call arrival or a call completion. Call arrivals into state $(b,0)$ occur at rate

$$\alpha_b\equiv\lambda(\pi_{b-1,0}+\pi_{b-1,1}) \tag{17}$$

and call completions into state $(b,0)$ occur at rate

$$\beta_b \equiv (b+1)(\mu_{b+1,0}\pi_{b+1,0} + \mu_{b+1,1}\pi_{b+1,1}). \tag{18}$$

Therefore, if the process is in state $(b,0)$, the probability that it was entered via a call arrival is $\alpha_b/(\alpha_b+\beta_b)$ and the probability that the state was entered via a call completion is $\beta_b/(\alpha_b+\beta_b)$. We assume that the arrival rate, λ, the call completion rate $\mu_{b,1}$ in the compliant states, and the following parameters are available as input to the Markov chain model:

$$\begin{aligned}
\tau_{b,0,\text{arrival}} &= \mathrm{E}[S_{b,0}|(b,0) \text{ entered via a call arrival}],\\
\tau_{b,0,\text{completion}} &= \mathrm{E}[S_{b,0}|(b,0) \text{ entered via a call completion}],\\
\phi_{b,0,\text{arrival}} &= \mathrm{E}[L_b|(b,0) \text{ entered via a call arrival}],\\
\phi_{b,0,\text{completion}} &= \mathrm{E}[L_b|(b,0) \text{ entered via a call completion}].
\end{aligned}$$

(We discuss how to estimate the input parameters in Section 4.) Using these inputs and the expressions for the probability of entering state $(b,0)$ via a call arrival vs. a call completion, we compute the expected service time and the expected time to reach compliance when in state $(b,0)$:

$$\begin{aligned}
\mathrm{E}[S_{b,0}] &= \frac{\alpha_b\tau_{b,0,\text{arrival}} + \beta_b\tau_{b,0,\text{completion}}}{\alpha_b+\beta_b},\\
\mathrm{E}[L_b] &= \frac{\alpha_b\phi_{b,0,\text{arrival}} + \beta_b\phi_{b,0,\text{completion}}}{\alpha_b+\beta_b}.
\end{aligned}$$

The resulting expressions for the call completion rates and the rate of reaching compliance are:

$$\mu_{b,0} = \frac{\alpha_b+\beta_b}{\alpha_b\tau_{b,0,\text{arrival}} + \beta_b\tau_{b,0,\text{completion}}}, \tag{19}$$

$$\gamma_b = \frac{\alpha_b+\beta_b}{\alpha_b\phi_{b,0,\text{arrival}} + \beta_b\phi_{b,0,\text{completion}}}. \tag{20}$$

To determine the steady state probabilities $\pi_{b,c}$, we need to iterate between solving the balance equations and computing the call completion rates and the rates of reaching compliance. We use the following iterative algorithm. The superscript k is an iteration counter.

Algorithm 2:

1. Initialization: Set $\pi^0_{b,c} \leftarrow 1/(2n+1)$ for all $(b,c) \in \Omega$.

2. Use (17)-(20) to compute α_b, β_b, γ_b and $\mu_{b,0}$ for $b=0,\dots,n-1$ and set $\mu_{n,0} = 1/\tau_{n,0,\text{arrival}}$.

3. Use *Algorithm 1* to solve the balance equations to obtain $\pi^{k+1}_{b,c}$. Set $k \leftarrow k+1$.

4. Iterate steps 2 and 3 until $|\pi^{k+1}_{b,c} - \pi^k_{b,c}|/\pi^k_{b,c} < \epsilon$, $|\mu^{k+1}_{b,0} - \mu^k_{b,0}|/\mu^k_{b,0} < \epsilon$, and $|\gamma^{k+1}_b - \gamma^k_b|/\gamma^k_b < \epsilon$ for all $(b,c) \in \Omega$.

4. Parameter Estimation for the Markov Chain Model

The Markov chain model requires as input parameters the arrival rate λ, the average service times $\tau_{b,1} = 1/\mu_{b,1}$ in the compliant states, the average service times $\tau_{b,0,\text{arrival}}$ and $\tau_{b,0,\text{completion}}$ in the non-compliant states, and the average times to reach compliance $\phi_{b,0,\text{arrival}}$ and $\phi_{b,0,\text{completion}}$. The arrival rate is straightforward to estimate from empirical data. In this section, we describe how we estimate the state-dependent input parameters.

From (1) we obtain:

$$\begin{aligned}
\tau_{b,0,\text{arrival}} &= \mathrm{E}[S^{\text{Chute}}_{b,0}] + \mathrm{E}[S^{\text{On-Scene}}_{b,0}] + \Pr\{T\}(\mathrm{E}[S^{\text{Transport}}_{b,0}] + \mathrm{E}[S^{\text{Hospital}}_{b,0}]) \\
&\quad + \mathrm{E}[S^{\text{Travel}}_{b,0} | (b,0) \text{ entered via a call arrival}], \text{ for } b = 1,\dots,n, \\
\tau_{b,0,\text{completion}} &= \mathrm{E}[S^{\text{Chute}}_{b,0}] + \mathrm{E}[S^{\text{On-Scene}}_{b,0}] + \Pr\{T\}(\mathrm{E}[S^{\text{Transport}}_{b,0}] + \mathrm{E}[S^{\text{Hospital}}_{b,0}]) \\
&\quad + \mathrm{E}[S^{\text{Travel}}_{b,0} | (b,0) \text{ entered via a call completion}], \text{ for } b = 0,\dots,n-1.
\end{aligned}$$

We estimate the means of $S^{\text{Chute}}_{b,c}$, $S^{\text{On-Scene}}_{b,c}$, $S^{\text{Transport}}_{b,0}$, and $S^{\text{Hospital}}_{b,c}$ and the probability $\Pr\{T\}$ directly from empirical data. That leaves the estimation of $\mathrm{E}[S^{\text{Travel}}_{b,0} | (b,0)$ entered via a call arrival] and $\mathrm{E}[S^{\text{Travel}}_{b,0} | (b,0)$ entered via a call completion] and our focus in this section is on those two quantities.

For the purpose of estimating $\mathrm{E}[S^{\text{Travel}}_{b,0} | (b,0)$ entered via a call arrival], we assume that at the moment prior to the arrival transition all ambulances were located in compliance for $(b-1)$ busy ambulances. (Note that we make this assumption only for the purpose of estimating the mean travel time. The Markov chain model allows arrival transitions to $(b,0)$ not only from $(b-1,1)$ but also from $(b-1,0)$.) We begin with the ambulance locations specified in the compliance table for

$b-1$ busy ambulances and we construct different configurations by removing one of the available ambulances, corresponding to the location of the arriving call. We define the catchment area of an ambulance as the set of demand nodes for which the ambulance is the closest. We define $A_{b,c,j}$ as the j-th configuration of the available ambulances in state $(b,0)$ and we use negative values of $j \in J_{b,0,\text{arrival}} = \{-(n-b+1),\ldots,-1\}$ to index the possible configurations. We assign probabilities $\Pr\{A_{b,c,j}\}$ to each configuration, equal to the sum of the probabilities f_d for demand nodes d in the catchment area of the ambulance answering the call that generated the configuration.

Similarly, to estimate $\mathrm{E}[S_{b,0}^{\text{Travel}}|(b,0)$ entered via a call completion] we assume that at the moment prior to the call completion transition, all ambulances were located in compliance for $(b+1)$ busy ambulances, and we have therefore $(n-b-1)$ available ambulances with known locations. The ambulance that completed its call could have done so at a hospital $h \in H$, with probability $\Pr\{A_{b,c,h}\} = \Pr\{T\} \times g_h$ or it could have completed the call at a demand node $d \in D$, because no transport was required, with probability $\Pr\{A_{b,c,|H|+d}\} = (1-\Pr\{T\}) \times f_d$. Note that we index the $|H|$ configurations $A_{b,c,j}$ for calls completed at a hospital using $j = 1,\ldots,|H|$, we index the $|D|$ configurations for calls completed at a demand node using $j = |H|+1,\ldots,|H|+|D|$, and we define the index set $J_{b,0,\text{completion}} = \{1,\ldots,|H|+|D|\}$.

We reserve configuration $A_{b,c,0}$ for the case when the ambulances are located in compliance and therefore $A_{b,1,0}$ is the compliance configuration.

For each configuration $A_{b,c,j}$, we can enumerate $|D|$ possible trips for the next call arrival, with one trip for each demand node $d \in D$, with probability f_d, and a travel distance l from the closest available ambulance to d based on configuration $A_{b,c,j}$. We store distances in an $(|S|+|H|+|D|) \times |D|$ matrix L, where $L[o,d]$ is the distance from origin $o \in S \cup H \cup D$ (a station, hospital, or demand node) to destination $d \in D$ (a demand node). For each configuration $A_{b,c,j}$, we store trip probabilities in a matrix $P_{b,c,j}$ of the same dimension as L, where $P_{b,c,j}[o,d]$ is the probability that the next call arrival will be to demand node d and that the closest available ambulance, given by configuration $A_{b,c,j}$, is at origin o. Note that column d of $P_{b,c,j}$ will have only one non-zero entry, equal to $f_d \Pr\{A_{b,c,j}\}$.

In a real system, for each state $(b,0)$ there is an infinite set of configurations specifying the locations of each available ambulance, but we approximate this infinite set with the finite set indexed by $J_{b,0} = J_{b,0,\text{arrival}} \cup J_{b,0,\text{completion}} = \{-(n-b+1), \ldots, -1, +1, \ldots, |D|+|H|\}$. For a compliant state $(b,1)$, by definition there is only one configuration, $A_{b,1,0}$, and therefore $J_{b,1} = \{0\}$.

We add the matrices $P_{b,0,j}$ to obtain

$$P_{b,0,\text{arrival}} = \sum_{j \in J_{b,0,\text{arrival}}} P_{b,0,j}$$
$$P_{b,0,\text{completion}} = \sum_{j \in J_{b,0,\text{completion}}} P_{b,0,j},$$

where $P_{b,0,\text{arrival}}[o,d]$ and $P_{b,0,\text{completion}}[o,d]$ give the probability that the next call arrival will result in an ambulance traveling from ambulance location o to demand node d, when in state $(b,0)$, given that the last change to the number of busy ambulances was a call arrival or a call completion, respectively. Given parameters a, v_c, b_0, b_1, and b_2 and the distance $L[o,d]$, we can now use (4)-(5) and (7)-(8) to compute the first and second moment of the travel time from o to d. By aggregating over origin-destination pairs, we can compute the mean travel time to the next call, given that the system is in state $(b,0)$ and conditional on the previous state as:

$$\mathrm{E}[S^{\text{Travel}}_{b,0} | (b,0) \text{ entered via a call arrival}] = \sum_{(o,d)\in\Gamma} \mathrm{E}[T(L[o,d])] P_{b,0,\text{arrival}}[o,d] \tag{21}$$
$$\mathrm{E}[S^{\text{Travel}}_{b,0} | (b,0) \text{ entered via a call completion}] = \sum_{(o,d)\in\Gamma} \mathrm{E}[T(L[o,d])] P_{b,0,\text{completion}}[o,d]. \tag{22}$$

where $T(l)$ is the travel time random variable for a trip of distance l. For future reference, we note that we can obtain the second moment and variance of the travel time when in state $(b,0)$ using

$$\mathrm{E}\left[(S^{\text{Travel}}_{b,c})^2\right] = \frac{\alpha_b}{\alpha_b+\beta_b} \sum_{(o,d)\in\Gamma} \mathrm{E}[(T(L[o,d]))^2] P_{b,0,\text{arrival}}[o,d]$$
$$+ \frac{\beta_b}{\alpha_b+\beta_b} \sum_{(o,d)\in\Gamma} \mathrm{E}[(T(L[o,d]))^2] P_{b,0,\text{completion}}[o,d] \tag{23}$$
$$\mathrm{var}[S^{\text{Travel}}_{b,c}] = \mathrm{E}[(S^{\text{Travel}}_{b,c})^2] - (\mathrm{E}[S^{\text{Travel}}_{b,c}])^2. \tag{24}$$

We use the same assumption to estimate $\phi_{b,0,\text{arrival}}$ and $\phi_{b,0,\text{completion}}$: That just before the most recent call arrival or a call completion, the system was in compliance, and just after the call arrival or completion, the system is in one of the configurations that we have listed.

In order to estimate the average times to reach compliance, we compare the configuration $A_{b,0,j}$ just after an event (a call arrival or completion) to the corresponding compliant configuration $A_{b,1,0}$. We define two sets:

$$\text{Non-compliant ambulances: } O_{b,j} = \text{locations in } A_{b,0,j} \text{ but not in } A_{b,1,0}$$

$$\text{Non-compliant stations: } U_{b,j} = \text{locations in } A_{b,1,0} \text{ but not in } A_{b,0,j}$$

Both sets are conditional on having just arrived to state $(b,0)$ and the ambulance locations being specified by the configuration $A_{b,0,j}$. We need to move the non-compliant ambulances to the non-compliant stations. Our models (both the Markov chain model and the simulation model) can accommodate any method for matching non-compliant ambulances to non-compliant stations. In the computational experiments that we report in this paper, we used the following greedy heuristic:

1. Set $m = 1$ (an iteration counter).

2. Select the ambulance-station pair $(o_i \in O_{b,j}, u_k \in U_{b,j})$ with the shortest distance and record that distance as $r_{b,j,m}$.

3. Eliminate o_i from $O_{b,j}$ and u_k from $U_{b,j}$ and set $m = m + 1$.

4. Repeat Steps 2 and 3 until the sets $O_{b,j}$ and $U_{b,j}$ are empty (that is, until all non-compliant ambulances have been assigned to non-compliant stations).

5. Set $r_{b,j} = r_{b,j,m-1}$.

The relocation distance for the last assignment, $r_{b,j}$, is the maximum distance. It corresponds to the estimated time $\mathrm{E}[T(r_{b,j})]$ to reach compliance, starting from configuration $A_{b,0,j}$. Using the configuration probabilities, we obtain the following weighted average times to reach compliance:

$$\phi_{b,0,\text{arrival}} = \sum_{j \in J_{b,0,\text{arrival}}} \mathrm{E}[T(r_{b,j})] \Pr\{A_{b,0,j}\} \tag{25}$$

$$\phi_{b,0,\text{completion}} = \sum_{j \in J_{b,0,\text{completion}}} \mathrm{E}[T(r_{b,j})] \Pr\{A_{b,0,j}\}. \tag{26}$$

5. Approximating the Response Time Distribution

To compute system performance measures, we require the cumulative distribution function of the response time, $F_R(x)$. We compared three methods to approximate this distribution: (1) moment-matching to a lognormal distribution, (2) the moment generating function based method described in Wu et al. (2005) to approximate the sum of lognormal random variables, and (3) convolution. Our numerical experiments indicated that moment-matching was the least accurate method and convolution was the most accurate method. Therefore, we describe the convolution method in this section and we use that method in the numerical experiments that follow.

We begin by computing the mean, second moment, and variance of the three components of the response time—evaluation and dispatch; chute; and travel time. For example, we perform the following computations for the evaluation and dispatch time:

$$\begin{aligned}
\mathrm{E}[R^{\text{Eval and Disp}}] &= \sum_{(b,c)\in\Omega} \pi_{b,c}\mathrm{E}[R_{b,c}^{\text{Eval and Disp}}] \\
\mathrm{E}[(R^{\text{Eval and Disp}})^2] &= \sum_{(b,c)\in\Omega} \pi_{b,c}\mathrm{E}[(R_{b,c}^{\text{Eval and Disp}})^2] \\
\mathrm{Var}[R^{\text{Eval and Disp}}] &= \mathrm{E}[(R^{\text{Eval and Disp}})^2] - \mathrm{E}[R^{\text{Eval and Disp}}]^2.
\end{aligned}$$

For the travel time, we compute the first two moments of $S_{b,0}^{\text{Travel}}$ using (21)-(23) (and similar but simpler formulas for state $S_{b,1}^{\text{Travel}}$).

Next, we approximate $R^{\text{Eval and Disp}}$, S^{Chute}, and S^{Travel} as independent lognormally distributed random variables. The parameters of the lognormal distribution for the evaluation and chute time are determined as follows (and similarly for the chute and travel times):

$$\mu = \ln(\mathrm{E}[R^{\text{Eval and Disp}}]) - \frac{1}{2}\ln\left(1 + \frac{\mathrm{Var}[R^{\text{Eval and Disp}}]}{\mathrm{E}[R^{\text{Eval and Disp}}]^2}\right) \tag{27}$$

$$\sigma^2 = \ln\left(1 + \frac{\mathrm{Var}[R^{\text{Eval and Disp}}]}{\mathrm{E}[R^{\text{Eval and Disp}}]^2}\right) \tag{28}$$

Finally, we approximate the lognormal distributions with discrete distributions and use discrete convolution to compute the distribution of $R = R^{\text{Eval and Disp}} + S^{\text{Chute}} + S^{\text{Travel}}$.

6. Markov Chain Model Validation

We developed a discrete event simulation model using the Stochastic Simulation in Java (SSJ) simulation library (L'Ecuyer 2009) to validate the Markov chain model, which we programmed in Matlab. The simulation model uses realistic distributions for the response and service time components, as discussed in Section 2, and it models the details of the dispatching and repositioning of ambulances. In the convolution, we used a bin size of 0.01 minutes to discretize the lognormal distributions referred to in the preceding section. In this section, we compare average travel times, state probabilities, and the response time distributions from the two models.

We used the road network and current ambulance station locations for Edmonton, Alberta and we assumed a fleet of 19 ambulances. We chose an arrival rate of $\lambda = 6$ calls per hour (average hourly arrival rates in Edmonton varied from 4 to 10 per hour in 2008). We used a compliance table that was adapted from compliance tables that have been used in Edmonton. The simulation results in this section are based on 100 replications of approximately 20,000 calls each (excluding a warm-up period with approximately 2,000 calls), which took roughly 5 minutes of computation time. The average simulated service time was 77.8 minutes and the average simulated ambulance utilization was 40.9%, which is consistent with typical utilization levels in EMS systems that we are familiar with. The Markov chain calculations took 5.4 seconds and converged in five iterations.

Figure 5 shows that mean travel times increase with the number of busy ambulances, as expected. The mean travel time more than doubles, from four to nine minutes, when the number of busy ambulances increases from zero to 18 (corresponding to a single available ambulance.) The figure also shows that mean travel times are shorter when the system is in compliance. (Note that this is not guaranteed to happen—poorly designed compliance tables could *increase* mean travel times.) The mean travel times from the Markov chain model are, on average, within 2.2% of the simulated mean travel times. Figure 6 shows that the probability distributions for the number of busy ambulances (probability of compliance and out of compliance added) from simulation and from the Markov chain agree closely, differing by at most 0.0007. The probabilities of compliance from the two models also agree closely, with 0.332 from the Markov chain model and 0.330 from

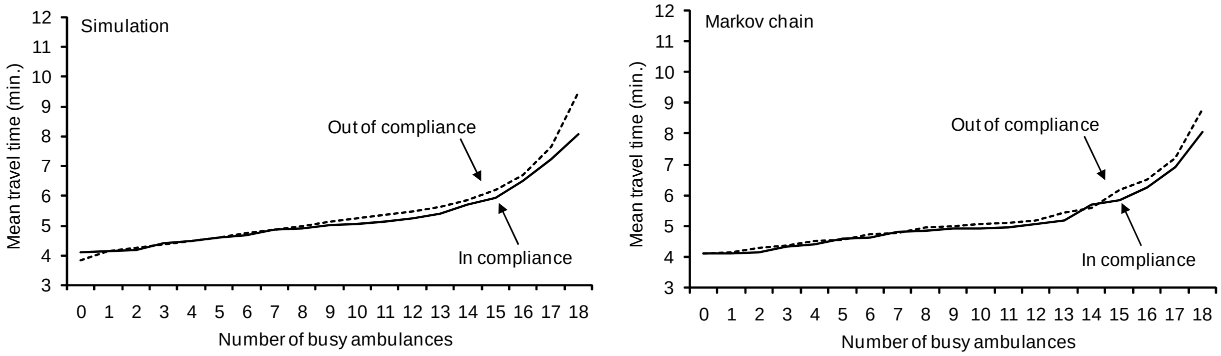

Figure 5 Mean travel times, based on the simulation and Markov chain models.

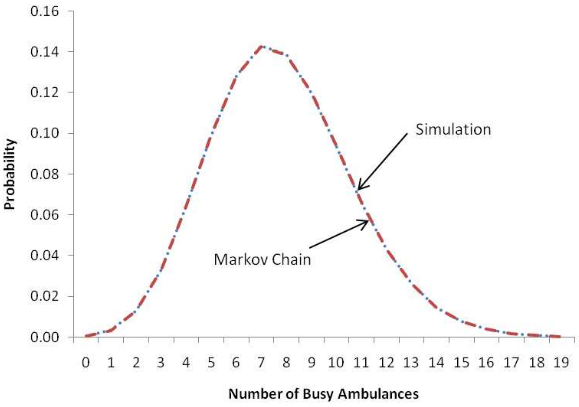

Figure 6 Probability distributions for number of busy ambulances

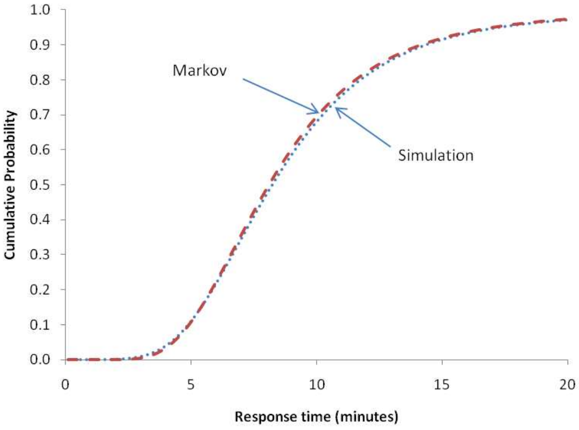

Figure 7 Response time distribution

the simulation. Figure 7 shows the response time distribution obtained from the simulation and Markov chain models.

Arrival rate (/hr.)	Utilization	Avg. service time (min.)	$\Pr\{R<9\}$	All ambulances available	All ambulances busy
Low: $\lambda = 2.28$	0.155/0.155	77.3/77.3	0.655/0.660	0.053/0.053	0.000/0.000
Medium: $\lambda = 6$	0.409/0.410	77.8/77.8	0.585/0.601	0.000/0.000	0.000/0.000
High: $\lambda = 10.74$	0.708/0.708	79.0/78.5	0.471/0.495	0.000/0.000	0.048/0.049

Table 2 Impact of load, with baseline compliance table. Simulation estimates are shown first, followed by Markov chain estimates.

In order to stress-test the Markov chain model, we repeated the computations for systems with *unrealistically high* and *unrealistically low* utilization. We chose the "high load" system by increasing the arrival rate until the probability that all ambulances are busy reached 0.05 and we chose the "low load" system by decreasing the arrival rate until the probability that all ambulances are available reached 0.05, keeping all other parameters constant. The resulting arrival rates of 10.74 per hour and 2.28 per hour are outside the range of average hourly arrival rates observed in Edmonton. In addition, even though we kept the number of ambulances fixed at 19 for testing purposes, in reality, the number of scheduled ambulances varies by time of day, making the high and the low load scenarios unrealistic. Table 2 compares these three scenarios, showing that the utilization increases roughly proportionally to the arrival rate, the average service time increases slightly with the load, and performance (measured by the response time probability) decreases with the load. The Markov chain estimates of the utilization, average service time, and the probability that all ambulances are busy or all are available are quite close to the simulation estimates. The Markov chain estimates of the response time probabilities are quite accurate as well, except for the high load scenario, where the estimate differs by 2.4% from the simulation estimate (recall that the high load scenario is likely to be less realistic in practice).

In the next section, we investigate the accuracy of the Markov chain model when it is used to determine the best from a set of possible compliance tables.

7. Impact of Changing the Compliance Table

One important use that we envision for our model is to screen a large number of potential compliance tables to identify promising ones, that could then be evaluated more carefully via simulation or other means. In order to assess the ability of the Markov chain model to identify the best compliance table, measured using simulation, we randomly generated 100 nested compliance tables, where "nested" means that the set of desired stations for a available ambulances was a subset of the desired stations for $a+1$ available ambulances, for $a=1,...,n-1$. We evaluated each compliance table using both the Markov chain model and the simulation model, using the probability of response times being within nine minutes as the performance measure. The results in this section are based on the same example for Edmonton, Alberta, as in Section 6, with medium load ($\lambda=6$ per hour).

The results are shown in Figure 8. The performance estimates from the simulation and Markov chain models are highly correlated (Spearman's rank order correlation coefficient is 0.97). The compliance table ranked highest by the Markov chain model is also the one ranked highest by the simulation. The Markov chain performance estimates are slightly biased, compared to the simulation estimates but this does not impact the model's ability to identify the best-performing table (point A in Figure 8). The difference in simulated performance between the best and the worst compliance table is over seven percentage points. Increasing performance even by one percentage point for an EMS system for a city of this size typically requires considerable resources—on the order of adding one ambulance to the system, 24 hours a day. The wide range in performance across compliance tables indicates the importance of choosing compliance tables carefully.

To quantify the differences between the performance estimates from the two models, we define two measures: *Estimation error* and *optimality error*. We define the estimation error, Δ_{est}, for each compliance table as

$$\Delta_{\text{est}} = \frac{|M_{\text{Model}} - S_{\text{Sim}}|}{S_{\text{Sim}}} \times 100\%, \tag{29}$$

where S_{Sim} is the response time probability obtained from the simulation and M_{Model} is the response time probability obtained from the Markov chain model. For the 100 compliance tables in the test suite, the average, minimum, and maximum percentage estimation errors were 1.29%, 0.08%, and 3.06%, respectively. (For comparison, the simulation estimates of the response time probabilities are accurate to about ±0.1% with 95% confidence.)

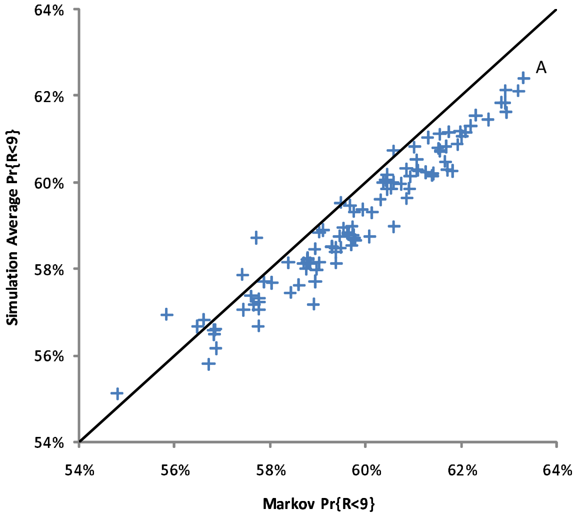

Figure 8 Response time probabilities with medium arrival rate ($\lambda = 6$ per hour).

The optimality error is a measure of the ability of the Markov chain model to select the optimal or a near-optimal compliance table from a given set of tables, as measured by simulation. We define the optimality error, Δ_{opt}, as the percentage difference between the best response time probability according to the simulation results and the response time probability obtained from simulation for the compliance table chosen as the best by the Markov chain model, that is:

$$\Delta_{\text{opt}} = \frac{|S_{\text{Model}} - S_{\text{Sim}}|}{S_{\text{Sim}}} \times 100\%, \tag{30}$$

where S_{Sim} is the best response time probability obtained from the simulation and S_{Model} is the response time probability obtained from the simulation for the compliance table chosen by the Markov chain model as the best. In this test suite with 100 compliance tables, as both the simulation and the Markov chain model picked the same compliance table as the best one (point A in Figure 8), $S_{\text{Sim}} = S_{\text{Model}} = 61.92\%$ is the vertical axis coordinate of point A and the overall optimality error of our Markov chain model is $\Delta_{\text{opt}} = 0\%$. The optimality error is our best estimate for the relative performance loss in the actual system, if the system planners decide to use the results of our Markov chain model, instead of the actual best compliance table in the test suite. In this test suite with 100 compliance tables, our Markov chain model picked the best compliance table and did not result in a performance loss.

We repeated the calculations described above for the high load and low load scenarios mentioned in Section 6. The graphical results are shown in Figure 9. Table 3 compares the low, medium, and high load results, using the Spearman correlation, the average estimation error, the optimality error, and the performance range, which we define as the difference between the maximum and minimum response time probability, as estimated with simulation. The results in the table suggest that the Spearman correlation between the simulation and Markov chain estimates decreases and the average estimation error and optimality error increase with the system load. From Figures 8-9, we see that for the low and medium load, the Markov chain model consistently overestimates performance but with the high load, the estimation errors are more dispersed—the Markov chain model underestimates performance for some tables and overestimates performance for others. The results for the performance range suggest that the potential for performance improvement using dynamic repositioning increases with the system load, as one would expect.

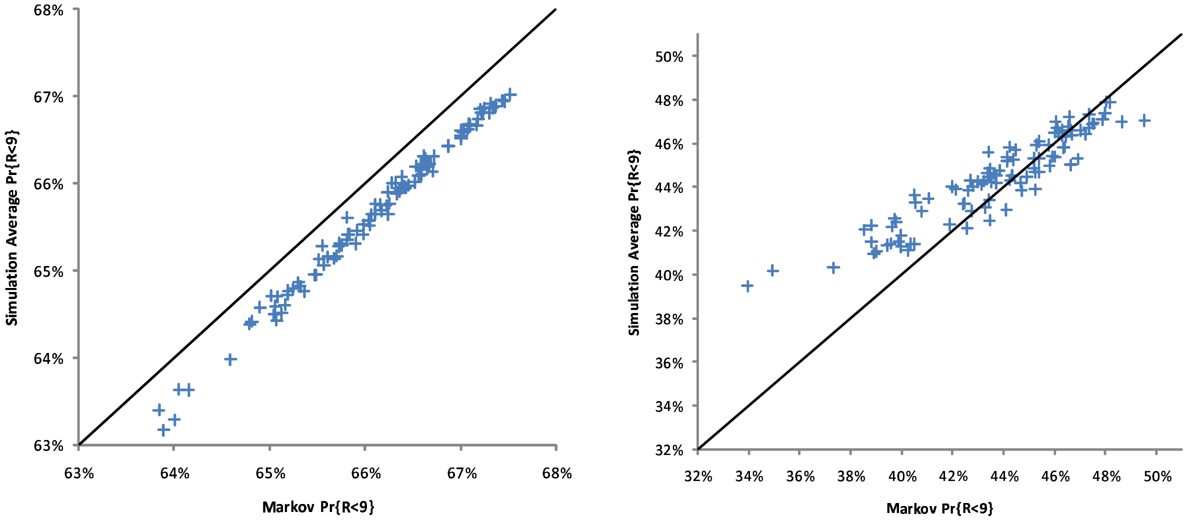

Figure 9 **Response time probabilities with low (left panel) and high (right panel) arrival rates.**

8. Conclusions

We have developed and validated a Markov chain model of an EMS system with repositioning. The model requires iteration between computing transition rates and computing steady state probabilities, but the computations converged in five iterations or less in all examples that we have tried

Arrival rate (/hr.)	Spearman correlation	Avg. estimation error	Optimality error	Performance range
Low: $\lambda = 2.28$	0.98	0.68%	0%	3.9%
Medium: $\lambda = 6$	0.97	1.29%	0%	7.3%
High: $\lambda = 10.74$	0.93	2.73%	1.8%	8.4%

Table 3 Summary of compliance table comparisons.

(more than 300 compliance tables). The cardinality of the model's state space scales linearly with the number of ambulances and solving the model for a realistic system takes only a few seconds. The model provides accurate estimates of the system response time distribution, the distribution of the number of busy ambulances, and various other performance measures.

We found that the Markov chain model provided estimates that were typically within 2% of estimates found using a more realistic simulation model. When we used the Markov chain model to choose the best from a set of 100 randomly compliance tables, under realistic conditions, it selected the one that performed best, as measured by simulation. This bodes well for using the Markov chain model as part of a heuristic to search for optimal compliance tables.

References

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

Berman, O. 1981b. Repositioning of distinguishable urban service units on networks. *Computers & Operations Research* **8**(2) 105–118.

Bledsoe, B. E. 2003. EMS mythology. EMS myth #7. System status management (SSM) lowers response times and enhances patient care. *Emergency Medical Services* **32**(7) 158–167.

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

Budge, S., A. Ingolfsson, D. Zerom. 2010. Empirical analysis of ambulance travel times: The case of Calgary Emergency Medical Services. *Management Science* **56**(4) 716–723.

Cady, G. 2002. 2001 JEMS 200 city survey, JEMS 2001 annual report on EMS operational & clinical trends in large, urban areas. *JEMS : A Journal of Emergency Medical Services* **27**(2) 46–71.

Erkut, E., A. Ingolfsson, S. Budge, D. Haight, J. Litchfield, O. Akyol, G. Holmes, J. Cheng. 2005. Final report: The impact of ambulance system status management. Unpublished report, prepared for the Emergency Response Department, City of Edmonton.

Fitch, J. 2005. Response times: Myths, measurement & management. *JEMS : A Journal of Emergency Medical Services* **30**(1) 46–56.

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

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

Gross, D., C. M. Harris. 1998. *Fundamentals of Queueing Theory*. Wiley, New York.

Jarvis, J. P. 1985. Approximating the equilibrium behavior of multi-server loss systems. *Management Science* **31**(2) 235–239.

Kolesar, P. 1975. Model for predicting average fire engine travel times. *Operations Research* **23**(4) 603–613.

Kolesar, P., W. E. Walker. 1974. An algorithm for the dynamic relocation of fire companies. *Operations Research* **22**(2) 249–274.

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

Larson, R. C. 1975. Approximating the performance of urban emergency service systems. *Operations Research* **23**(5) 845–868.

L'Ecuyer, P. 2009. *SSJ: Stochastic Simulation in Java*. Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Montréal, Québec. URL `http://www.iro.umontreal.ca/~simardr/ssj/indexe.html`. Last accessed 19 December 2009.

Maxwell, M. S., S. G. Henderson, H. Topaloglu. 2009. Ambulance redeployment: An approximate dynamic programming approach. M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, R. G. Ingalls, eds., *Proceedings of the 2009 Winter Simulation Conference*. 1850–1860.

Maxwell, M. S., M. Restrepo, S. G. Henderson, H. Topaloglu. 2010. Approximate dynamic programming for ambulance redeployment. *INFORMS Journal on Computing* **22**(2) 266–281.

Morneau, P. M., J. P. Stothart. 1999. My aching back. The effects of system status management & ambulance design on EMS personnel. *JEMS : A Journal of Emergency Medical Services* **24**(8) 36–50, 78–81.

Rajagopalan, H. J., C. Saydam, J. Xiao. 2008. A multiperiod set covering location model for dynamic redeployment of ambulances. *Computers & Operations Research* **35**(3) 814–826.

Stout, J. 1989. System status management. The fact is, its everywhere. *JEMS : A Journal of Emergency Medical Services* **14** 65–71.

Williams, D. M. 2009. JEMS 2008 200 city survey: The future is your choice. *JEMS : A Journal of Emergency Medical Services* **34**(2) 36–51.

Wu, Jingxian, N.B. Mehta, Jin Zhang. 2005. Flexible lognormal sum approximation method. *Global Telecommunications Conference, 2005. GLOBECOM '05. IEEE*, vol. 6. 3413 –3417.