Modeling Yellow and Red Alert durations for ambulance systems

Amir Rastpour*
Faculty of Business and Information Technology, University of Ontario Institute of Technology, ON L1H 7K4,
amir.rastpour@uoit.ca,
Phone: 905 721 8668 ext. 3625, Fax: 905 721 3167

Armann Ingolfsson
Alberta School of Business, University of Alberta, AB T6G 2R6,
armann.ingolfsson@ualberta.ca,
Phone: 780 492 7982, Fax: 780 492 3325

Bora Kolfal
Alberta School of Business, University of Alberta, AB T6G 2R6,
bora.kolfal@ualberta.ca,
Phone: 780 492 8466, Fax: 780 492 3325

Emergency systems are designed to almost always have enough capacity to respond to emergencies. However, capacity shortage periods do occur and these systems need to recover quickly. In this paper, we apply queueing models and study whether it is better for an emergency system to add or to expedite servers, in order to quickly recover from a capacity shortage period. We focus on emergency medical service (EMS) systems and use Erlang loss models to study Red Alerts (when all ambulances are busy) and Yellow Alerts (when the number of available ambulances falls below a threshold). We analyze two loss models: one with Markovian state-dependent service rates and one with generally and independently distributed service times. We validate the two models against EMS data sets from two cities. Despite the fact that the distribution of ambulance service times is a mixture of lognormal distributions, which is far from being exponential, we find that the loss model with Markovian state-dependent service rates provides a better representation of empirical Yellow and Red alert statistics. We build on the model with state-dependent rates and use the theory of absorbing Markov chains to quantify the impact of adding or expediting ambulances, with respect to two performance measures: (1) the duration of alert periods and (2) the number of lost calls. This quantification helps EMS staff (dispatchers and supervisors) to make better decisions to avoid, and to recover from, alert periods. For example, staff should not wait until a Red Alert before adding ambulances, which is a common practice, because the expected number of lost calls rapidly increases as the number of available ambulances at the action epoch decreases.

Key words: Service management; Ambulance service; Busy period analysis; Erlang loss model; Stochastic model applications

History: Received: May 2018; Accepted: March 2020 by Sergei Savin, after 2 revisions

1. Introduction

Capacity shortage in a mission-critical system like fire, police, and emergency medical service (EMS) can lead to a disaster if no contingency plans have been made. Although these systems

are designed to almost always have enough capacity to respond to emergency calls in a timely fashion, contingency plans are needed to minimize the frequency with which the capacity is highly utilized, and to shorten the duration of capacity shortage periods if they do happen. Motivated by the EMS in Calgary and Edmonton, Alberta, Canada, that have been experiencing relatively frequent capacity (ambulance) shortage periods for several years, we focus on the specific context of EMS, and use queueing theory to provide insights on managing ambulance shortage periods.

Ambulance shortage periods are not limited to Alberta—they occur in EMS systems all over the world. For media and organizational reports from EMS in Australia, the US, and Canada (outside Alberta), see ABC News (2015), Zekman (2014), and Brown (2018), respectively.

EMS practitioners distinguish between high, medium, and low levels of resource utilization (Fitch et al. 1993). The terms used to describe these utilization levels vary among countries and regions. Calgary and Edmonton EMS refer to a High utilization as a ''Red Alert,'' which corresponds to a period when no ambulances are available to respond to new medical emergencies. They refer to a Medium utilization as a ''Yellow Alert,'' which corresponds to a period when the number of available ambulances is below a threshold[1] θ ($\theta = 12$ in Calgary; $\theta = 8$ in Edmonton). In this paper, we use Red Alert and Yellow Alert to refer to High and Medium utilization levels, respectively.

Yellow Alert periods are important from two perspectives: (1) the onset of a Yellow Alert is a signal to EMS staff to take actions to prevent the situation from deteriorating into a Red Alert, and (2) a smaller number of available ambulances in the system increases the average distance between the call and the closest available ambulance, which results in longer response times. Practitioners view the threshold θ as the minimum number of ambulances needed to adequately cover the city's geographical area. As Figure 1 shows, average response time depends primarily on the number of available ambulances and is not highly sensitive to the number of scheduled ambulances (which varies with the call arrival rate). For these reasons, we keep θ fixed, independent of time, the number of scheduled ambulances, and the rate of call arrivals.

EMS staff manage Yellow and Red Alerts by taking three types of actions: *expediting*, *adding*, and *repositioning* ambulances. Expediting shortens ''hospital time'', during which EMS crews wait to transfer patient care to emergency department (ED) staff. Expediting can be accomplished by: (1) expediting the admission of a patient occupying an ED bed into a hospital ward to free up the bed for an EMS patient, and (2) consolidating the care of several waiting EMS patients under a single paramedic crew, allowing the other crews to leave the ED. Approach 1 requires collaboration and coordination between EMS and ED staff but Approach 2 can be carried out within

[1] Protocols that define when a Yellow Alert is triggered sometimes include additional considerations besides the number of available units, such as ''7 or fewer units ... sustained for 15 minutes.'' We use alert period definitions that are solely based on the number of available units, for simplicity.

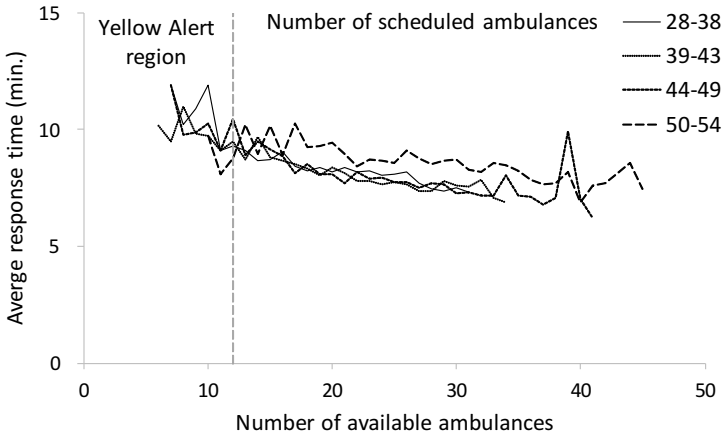


Figure 1 **Average response time vs. the number of available ambulances. Each curve corresponds to one quartile for the number of scheduled ambulances (Calgary 2009 data). Only data points with a sample size of** 10 **or more are shown.**

the EMS. Adding ambulances could take the form of supervisors or managers asking for ambulances from neighboring municipalities, from another service (an interfacility-transfer ambulance fleet, for example), or asking new ambulance crews to come on duty. Adding ambulances usually requires collaboration and coordination between different EMS systems. Repositioning entails relocating available ambulances to improve the coverage of arriving calls. A simple and commonly used repositioning policy involves the use of a compliance table, which specifies target locations for all ambulances, as a function of the number of available ambulances.

Discussions with Calgary and Edmonton EMS staff indicate that they decide on actions based on a combination of judgment and rules that are implemented in a decision support system. A Red Alert is a simple rule that often triggers the adding of ambulances (Rumbolt 2017). A Yellow Alert is another trigger that indicates the staff should consider taking action but does not specify precisely which actions should be taken. The dynamics of EMS systems are sufficiently complex that it is difficult, even for highly experienced practitioners, to reliably predict the consequences of different actions using unaided human judgment. For example, the impact of adding ambulances depends on the remaining duration of the current alert period, which is difficult to predict, because alert period durations are highly variable, with squared coefficients of variation larger than one in most cases (Table 2). If the current alert period ends before the new ambulance(s) arrive, then cost will have been incurred and dispatchers, supervisors, and ambulance crews will have experienced added stress, all to no avail.

In this study, we do not attempt to estimate the cost of actions. Instead, we mathematically model alert periods and analyze the impacts of the expedite and add actions on these periods. We

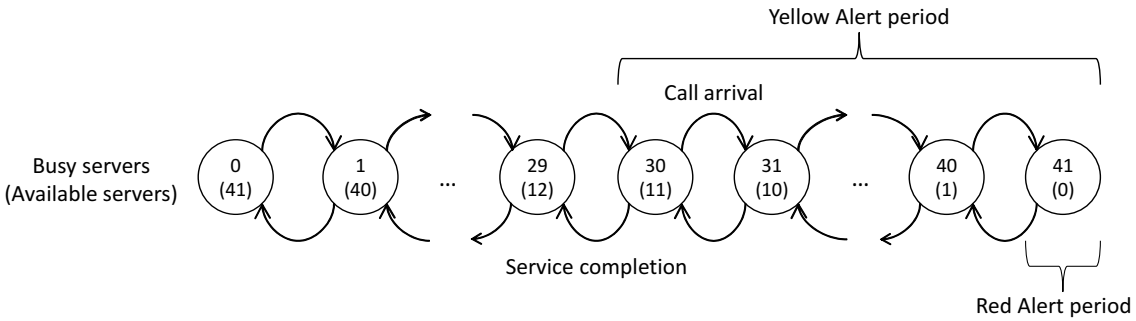


Figure 2 System states that correspond to Yellow and Red Alerts. $c = 41$, $\theta = 12$.

do not study repositioning because its impact on EMS operations has been investigated extensively by several researchers (Alanis et al. 2013, Maxwell et al. 2010, Schmid 2012).

We model an EMS system as a loss (as opposed to delay) system, similar to other researchers (Maxwell et al. 2010, Restrepo et al. 2009, for example), because the loss model is tractable, and calls that arrive during a Red Alert (which we refer to as "lost calls") are typically served by other resources, such as the fire department or backup ambulance units, rather than waiting in a queue (Chong et al. 2015). We model Yellow and Red Alerts as special cases of "k-partial busy periods": time intervals during which k or more of the c scheduled ambulances are busy. Red Alerts are c-partial busy periods and Yellow Alerts are $(c-\theta+1)$-partial busy periods. Figure 2 illustrates Yellow and Red Alerts when $\theta = 12$ and $c = 41$. Every Red Alert is contained within a Yellow Alert, and a Yellow Alert can contain multiple Red Alerts.

We view an ambulance and its crew as a server, and we define the period from when a patient is assigned to a server, until the server becomes available again, as the service time. We thoroughly analyze the loss model with Markovian state-dependent service rates, $M/M(k)/c/c$, and provide some results for the loss model with general service time, $M/G/c/c$. The former permits us to capture load-based speedup or slowdown effects, and to indirectly capture some of the impacts of the spatial distribution of available ambulances, especially for systems in which compliance table policies are used to reposition ambulances. The latter permits us to model service times as a mixture of two distributions, for calls that are transported to hospital and those that are not.

We extend $M/M(k)/c/c$ loss models to capture the impacts of add and expedite actions on alert periods through two performance measures: (1) the expected residual duration of a Yellow Alert, and (2) the expected number of lost calls. We provide insights on how the expedite and add actions compare with respect to these performance measures.

Our managerial contribution is that we show how to apply queueing models to quantify the impact of add and expedite actions on the expected remaining duration of Yellow Alert, and on the expected number of lost calls. For each performance measure, we use real EMS data and

provide a threshold policy for comparing add and expedite actions as a function of the expected value of the time until actions are realized. We show that staff should not wait until a Red Alert occurs before taking action (which is a common practice), especially if the performance measure of interest is the expected number of lost calls, because that measure escalates rapidly as the number of available ambulances decreases. Another drawback of waiting too long before taking action is that the improvement in comparison with taking no action decreases, as the number of available ambulances at the action epoch decreases.

Our technical contribution is that we provide recursions to calculate the first and second moments of k-partial busy periods for $M/M(k)/c/c$. We prove an insensitivity result for $M/G/c/c$ that the first moments (but not higher moments) of k-partial busy period durations depend on the service time distribution only through its mean. Using real EMS data, we show that although ambulance service rates depend strongly on the number of busy ambulances, and the service time distribution is a mixture of lognormal distributions, which is far from being exponential, the two loss systems perform similarly ($M/M(k)/c/c$ is slightly better), with respect to predicting the mean of alert periods.

The remainder of the paper is organized as follows. We review related literature in Section 2; we define and analyze k-partial busy period durations in Section 3; we validate our models in Section 4; we analyze the impacts of two actions, add and expedite, on two performance measures, expected remaining Yellow Alert duration and the expected number of lost calls, in Section 5; we provide managerial insights on taking actions in Section 6; and we conclude in Section 7. Appendices A-H contain proofs, additional computational results, and a list of notation.

2. Literature Review

We survey four streams of related literature: (1) modeling of EMS systems, (2) insensitivity results for loss systems, (3) modeling of partial busy periods, and (4) strategies to mitigate capacity or inventory shortages in various contexts.

EMS system models: Ingolfsson (2013) provides a recent general survey of research on planning and management for EMS systems. Modeling EMS systems as loss systems is common in this literature—either as a standard $M/G/c/c$ system (Restrepo et al. 2009) or as a more general loss system (Maxwell et al. 2010, Almehdawe et al. 2013, Alanis et al. 2013, Li and Whitt 2014, Chong et al. 2015). We adopt the Erlang loss model for simplicity and to make progress on modeling the duration of partial busy periods and on modeling the impact of actions to mitigate capacity shortages. We assess the impact of some of the simplifications that are inherent in the Erlang loss model in Section 4.

The impact of adding or expediting ambulances appears not to have been investigated before. Repositioning—another action that can be taken to mitigate capacity shortages—has been investigated by Alanis et al. (2013), Maxwell et al. (2010), Schmid (2012) and others. Our work complements theirs.

The Erlang loss model ignores two key aspects of EMS systems: Ambulances may not have the same service distribution, because of their geographic locations, and parameters (arrival rates and number of ambulances) vary with time or system state. Larson's (1974, 1975) exact and approximate hypercube queueing models (HQM) address the geographic heterogeneity of ambulances. Many researchers have used variants of HQM to study EMS systems. Fewer researchers have explicitly incorporated time-varying parameters in an analytical EMS system model; Ignall and Walker (1977) did this for an EMS system and Kolesar et al. (1975) for police patrol cars. Simulation models of EMS systems typically do incorporate time-varying parameters (Henderson and Mason 2004, Mason 2013).

Evidence in Alanis et al. (2013) suggests that ambulance service rates depend on the number of busy ambulances. Erlang loss models with state-dependent service rates have applications in traffic flow modeling (Jain and Smith 1997), and in designing evacuation networks (Weiss et al. 2012).

Insensitivity results for loss systems: Taylor (2011) defines an insensitive stochastic model as one whose ''stationary distribution depends on one or more of its constituent lifetime distributions only through the mean,'' and provides an extensive literature review. The best known insensitive stochastic models are $M/G/c/c$ and $M/G/\infty$.

Although the steady state probabilities of the $M/G/c/c$ system are insensitive to the service time distribution beyond its mean, the same is not true for the transient occupancy probabilities. We show that the first moments of the k-partial busy period durations, although they are measures of transient behavior, are insensitive to the service time distribution beyond its mean.

Partial busy periods: Busy periods are unambiguously defined and well studied for single-server queues; they begin when a customer arrives to an empty system and last until the server becomes idle again for the first time. For analytical results, see Gross and Harris (1998, p. 102), for example. For multi-server queues, however, the terminology for busy periods varies. Omahen and Marathe (1978) use ''busy period T_k,'' and Sharma (1990, Chap. 4.4) uses ''k-server busy period'' to refer to k-partial busy periods. Artalejo and Lopez-Herrero (2001) use ''partial busy periods'' for what we refer to as 1-partial busy periods—that is, at least one server is busy—and they use ''full busy period'' for what we refer to as c-partial busy periods—that is, all servers are busy. Other authors (Chan et al. 2017, for example) have followed Artalejo and Lopez-Herrero in using the term ''partial busy period,'' and we extend that term in defining k-partial busy periods.

Omahen and Marathe (1978) and Sharma (1990) studied k-partial busy periods for the $M/M/c$ and $M/M/c/N$ (with queue capacity $= N - c$) systems, respectively. Bountourelis et al. (2013) observe that k-partial busy periods have not been studied for loss systems, except as a special case of the $M/M/c/N$ system. Our focus on loss systems allows us to obtain stronger results than those in Sharma (1990). Bountourelis et al. (2013) discuss applications of loss models in modeling hospital intensive care units (ICU) and highlight the importance of studying the length of periods during which ICUs are full, that is, c-partial busy period durations. We thoroughly investigate k-partial busy period durations, for $k = 1, \ldots, c$, for Erlang loss systems with state-dependent service rates and provide formulas to calculate their moments.

Shortage strategies: Protocols for managing ED capacity shortage have been formalized in an ED Surge Capacity Protocol in Alberta (Alberta Health Services 2010) and elsewhere (Viccellio and Santora 2012, The College of Emergency Medicine 2014) and medical researchers have investigated the impact of such protocols on ED crowding (Cha et al. 2009, Watase et al. 2012). Modelers have studied how to shift focus between triage and treatment when congestion in an ED exceeds a threshold (Zayas-Caban et al. 2019). Alert periods are conceptually similar to low-inventory periods for a retailer or a manufacturer, or periods where almost all beds in a hospital ward are occupied. Lawson and Porteus (2000), Duran et al. (2004), and Veeraraghavan and Scheller-Wolf (2008) discuss the use of expediting during low-inventory periods; that is, placing orders with a shortened lead time. Chan et al. (2014) investigate the use of speedup, modeled as a service rate increase, in an ICU in order to accommodate new patients that need to enter the ICU. Such short-term actions are not without risk—for example, KC and Terwiesch (2012) show that speedup can increase the chance of ICU readmission and decrease an ICU's peak capacity. We provide methods to compare the impacts of the adding and expediting actions on the expected residual Yellow Alert duration and the number of lost calls during the Yellow Alert.

3. Partial Busy Period Modeling

In this section, we model an EMS system where no actions are taken as a multi-server loss system with Poisson arrivals and either Markovian state-dependent service rates ($M/M(k)/c/c$), or a general service time ($M/G/c/c$).

We start with the $M/M(k)/c/c$ system, which has a Poisson arrival process with rate λ, and has service rate μ_k, when there are k busy ambulances in the system. We recursively obtain the first and second moments of the distribution of k-partial busy period durations. Then, we study the $M/G/c/c$ system that has a generally distributed service time, T, with mean $1/\mu$, CDF $F_T(t)$ and complementary CDF $\overline{F}_T(t)$. For $M/G/c/c$, we show that the first moment (but not the higher moments) is independent of the shape of service time distribution beyond its mean.

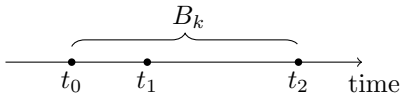


Figure 3 A schematic view of B_k and its components. Note that $t_2 - t_1 > 0$ only if the event at t_1 is an arrival.

A k-partial busy period, with duration B_k, begins when an arrival increases the number of busy servers to k and ends when a departure leaves $k-1$ busy servers. With this notation, the duration of a Yellow Alert is $B_{c-\theta+1}$ and the duration of a Red Alert is B_c.

3.1. The $M/M(k)/c/c$ Model

As illustrated in Figure 3, we decompose B_k into (1) the time from t_0, when the k-partial busy period begins, until the next event occurs at t_1, and (2) the time from t_1 until t_2, when the k-partial busy period ends. The duration of the second component, $t_2 - t_1$, can be positive or zero, depending on the event at t_1. That is:

$$\begin{aligned}\mathbf{E}[B_k] = &\Pr(t_1 \text{ event is a departure} | t_0 \text{ event is an arrival}) \mathbf{E}[t_1 - t_0] \\ &+ \Pr(t_1 \text{ event is an arrival} | t_0 \text{ event is an arrival}) \mathbf{E}[t_2 - t_0]. \end{aligned} \tag{1}$$

We explicitly include conditioning in (1) to stress that a k-partial busy period always begins with an arrival. If the t_1 event is a service completion, then the k-partial busy period ends at t_1, and $t_2 - t_1 = 0$. If the t_1 event is an arrival, however, then a $(k+1)$-partial busy period begins at t_1 with duration B_{k+1}. At epoch $t_1 + B_{k+1}$, a service completion occurs and the number of busy ambulances decreases to k. Depending on the next event, either the k-partial busy period ends, or another $(k+1)$-partial busy period begins. It follows that, if the event at t_1 is an arrival, the k-partial busy period ends after a geometrically-distributed number of $(k+1)$-partial busy periods. Following this logic, we provide recursive formulas for $\mathbf{E}[B_k]$ and $\mathbf{E}[B_k^2]$ in the following theorem:

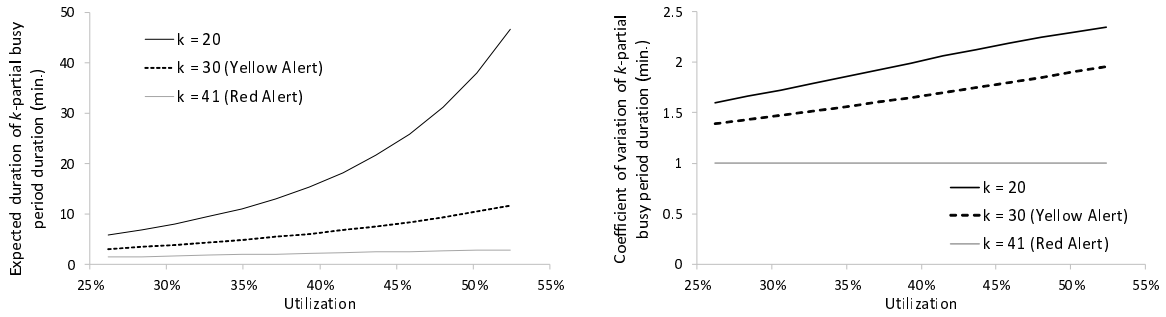
THEOREM 1. *The first two moments of B_k for an $M/M(k)/c/c$ system satisfy:*

$$\mathbf{E}[B_c] = \frac{1}{c\mu_c}, \qquad \mathbf{E}[B_k] = \frac{\lambda}{k\mu_k}\mathbf{E}[B_{k+1}] + \frac{1}{k\mu_k}, \qquad k = c-1, \ldots, 1. \tag{2}$$

$$\mathbf{E}[B_c^2] = 2\mathbf{E}\left[B_c\right]^2, \ \mathbf{E}\left[B_k^2\right] = \frac{\lambda}{k\mu_k}\mathbf{E}\left[B_{k+1}^2\right] + 2\mathbf{E}\left[B_k\right]^2, \ k = c-1, \ldots, 1. \tag{3}$$

Proof: See Appendix D.1.

The $M/M/c/c$ system is a special case of $M/M(k)/c/c$, with μ_k replaced with μ. We show in Figure 4 how the mean and coefficient of variation vary with k and with the system utilization, for a base case with $\lambda = 10.74$ calls per hour, $c = 41$ ambulances, and $1/\mu = 90$ minutes (these parameters are realistic for the Calgary EMS). We varied $1/\mu$ from 60 to 120 minutes to obtain utilization values ranging from 26% to 52%. We see (Fig. 4(a)) that the mean alert durations



(a) (b)

Figure 4 Impact of utilization and k on mean and coefficient of variation of alert durations.

increase with utilization and the increase is faster for lower k (hence, faster for Yellow Alerts than for Red Alerts). We also see (Fig. 4(b)) that alert durations become more variable as utilization increases, and are more variable for lower k.

3.1.1. The Relationship Between the Loss Probability and Partial Busy Periods

The stationary loss probability for an $M/G/c/c$ system is obtained from the Erlang B formula, the probability that all servers are busy (Gross and Harris 1998, p. 81). Here, we investigate the probability of experiencing Yellow and Red Alert periods during a k-partial busy period. Let U_k be the event that the system experiences at least one Yellow Alert and V_k be the event that the system experiences at least one Red Alert, within the current k-partial busy period. We calculate $\Pr(U_k)$ and $\Pr(V_k)$ recursively. If the system is already within a Yellow, or Red Alert, then the probability of experiencing that alert is 1; that is, $\Pr(U_k)=1$, for $k=c,\ldots,c-\theta+1$, and $\Pr(V_k)=1$, for $k=c$. For other k values, we obtain the following recursions, by conditioning on the number of times the system enters a $(k+1)$-partial busy period, without experiencing Yellow or Red Alert periods, before the end of the k-partial busy period.

THEOREM 2. *For an $M/M(k)/c/c$ system, the probabilities $\Pr(U_k)$ and $\Pr(V_k)$ satisfy:*

$$\Pr(U_k)=1,\ k=c,\ldots,c-\theta+1,\ \Pr(U_k)=1-\frac{k\mu_k}{\lambda\Pr(U_{k+1})+k\mu_k},\ k=c-\theta,\ldots,1. \tag{4}$$

$$\Pr(V_k)=1,\ k=c, \qquad \Pr(V_k)=1-\frac{k\mu_k}{\lambda\Pr(V_{k+1})+k\mu_k},\ k=c-1,\ldots,1. \tag{5}$$

Proof: See Appendix D.2.

Figure 5 shows how $\Pr(U_k)$ and $\Pr(V_k)$ vary with k when $\mu_k=(74.83+0.84k)^{-1}$, $\lambda=10.74$, and $c=41$ (the equation and parameters are realistic for the Calgary EMS). We observe that as k increases, both probabilities first decrease and then increase, reflecting the competing influences of two factors: (1) the distance from k to the state that triggers a Yellow or Red Alert (the longer it

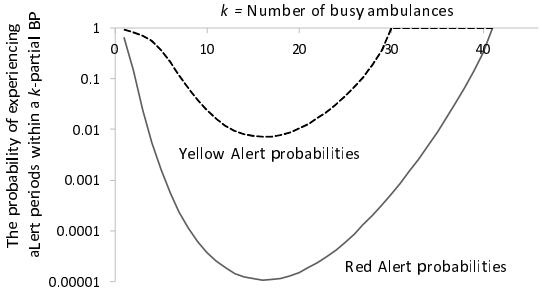


Figure 5 The probability of experiencing Yellow or Red Alert within a k-partial busy period as a function of k.

is, the less likely an Alert period) and (2) the duration B_k (which increases rapidly with k, and the longer B_k is, the more likely an Alert period).

3.2. The $M/G/c/c$ Model

Our calculations in Section 3.1 relied heavily on the Markovian property. We develop our $M/G/c/c$ formulas for $\mathbf{E}[B_k]$ and $\mathbf{E}[B_c^2]$ by observing the system only at arrival and departure epochs, and condition our calculations on the last event. A detailed discussion of our calculations is provided in Appendix D.3; here, we only present a summary of our results.

If the system has k busy ambulances, then we use R_k to denote the sojourn time; L_k to indicate that the last event was an arrival (L_k^c indicates a departure), and N_k to indicate that the next event is an arrival (N_k^c indicates a departure). Let $\tilde{T}$ be a residual service time, which follows the stationary excess distribution $F_{\tilde{T}}(t) = \mu \int_0^t \overline{F}_T(s)\, ds$. The complementary CDF of R_k, given the last event, can be calculated as:

$$\overline{F}_{R_k|L_k}(t) = e^{-\lambda t 1\{k<c\}} \overline{F}_T(t) \overline{F}_{\tilde{T}}(t)^{k-1},\ k = 1, ..., c. \tag{6}$$

$$\overline{F}_{R_k|L_k^c}(t) = e^{-\lambda t} \overline{F}_{\tilde{T}}(t)^k,\ k = 1, ..., c-1, \tag{7}$$

where $1\{k<c\}$ is an indicator function. Using (6) and (7), we obtain all moments of $R_k|L_k$ and $R_k|L_k^c$ by applying the following general result for the moments of a non-negative random variable X (Wolff 1989, p. 37):

$$\mathbf{E}[X^n] = \int_0^\infty \overline{F}_X(x^{1/n})\, dx, \tag{8}$$

When there are k busy ambulances in the system, we calculate the probability of the next event by conditioning on the last event:

$$\Pr(N_k|L_k) = \lambda \mathbf{E}[R_k|L_k],\ \Pr(N_k|L_k^c) = \lambda \mathbf{E}[R_k|L_k^c],\ k = 1, ..., c-1. \tag{9}$$

We use equations (1) and (6)-(9) to calculate the first moment of B_k. Theorem 3 summarizes our findings for the first two moments.

Table 1 Fleet size and utilization in Edmonton in 2008 and in Calgary in 2009.

	Edmonton	Calgary
Yellow Alert threshold (θ available ambulances)	8	12
Minimum number of scheduled ambulances	19	28
Maximum number of scheduled ambulances	36	54
Average number of scheduled ambulances	25	41
Average utilization	57%	43%
Average hospital time (minutes)	68.87	69.44

THEOREM 3. *The following hold for a stationary $M/G/c/c$ system:*

(a) The first moment of B_k is insensitive to the shape of the service time distribution and satisfies:

$$\mathbf{E}[B_c] = \frac{1}{c\mu},\ \mathbf{E}[B_k] = \frac{\lambda}{k\mu}\mathbf{E}[B_{k+1}] + \frac{1}{k\mu},\ k = c-1,\ldots,1. \tag{10}$$

The first moment of B_k is strictly decreasing in k and has the closed-form expression:

$$\mathbf{E}[B_k] = \sum_{i=0}^{c-k} \frac{(k-1)!}{\mu(k+i)!}\left(\frac{\lambda}{\mu}\right)^i,\ k = 1,...,c-1. \tag{11}$$

(b) The second moment of B_c is:

$$\mathbf{E}\left[B_c^2\right] = \int_0^\infty \overline{F}_T\left(t^{1/2}\right)\overline{F}_{\tilde{T}}\left(t^{1/2}\right)^{c-1} dt. \tag{12}$$

(c) The higher moments $\mathbf{E}[B_c^n]$, $n \geq 2$, are sensitive to the shape of the service time distribution.

See Appendix D.3 for a precise stationary regime definition and a Theorem 3 proof.

4. Validation

We validate Equations (2)-(3) and (10)-(12) for the first and second moments of partial busy period durations for $M/M(k)/c/c$ and $M/G/c/c$ systems against data sets from Calgary and Edmonton, two cities with population of about 1 million in Alberta, Canada. The Calgary data set has 93,734 calls from 2009 and the Edmonton data set has 64,267 calls from 2008. Tables 1-2 provide fleet size, utilization, and descriptive statistics for EMS alert periods in these two cities. Utilization is computed as the average of the ratio of busy ambulances to scheduled ambulances. Yellow and Red Alerts were more frequent in Edmonton than Calgary, consistent with the higher average ambulance utilization in Edmonton.

We show the validation results for Calgary in this section and discuss the Edmonton results in the online supplement, Section F. We use the first 6 months as a *training* sample (used to compute all parameter estimates) and the second 6 months as a *testing* sample, for both data sets.

As inherent in standard Erlang loss models, in using (2)-(3) and (10), we implicitly assume that (1) the arrival rate and the number of scheduled ambulances are constants that do not vary with time or system state, (2) the service rates do not vary with time, and (3) all ambulances have the

Table 2 Descriptive statistics for alert periods in Edmonton in 2008 and in Calgary in 2009.

	Yellow Alert		Red Alert	
	Edmonton	Calgary	Edmonton	Calgary
Sample size	1,349	703	587	9
Mean (min.)	106.41	7.09	7.20	1.37
Standard deviation (min.)	120.26	11.53	11.32	1.32
Maximum (min.)	1,012.02	127.28	138.93	4.53
Squared coefficient of variation	1.28	2.64	2.47	0.94
Avg. number of calls per alert period	25.55	2.35	2.23	1.11

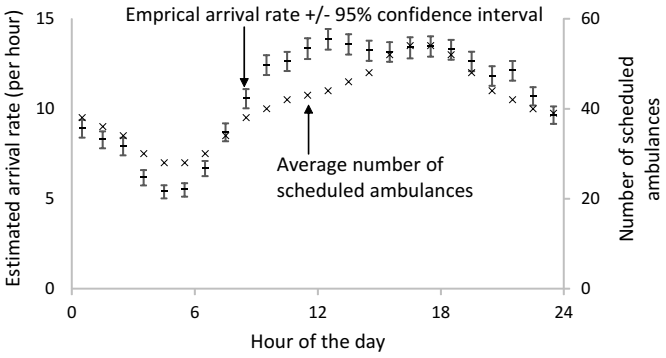


Figure 6 The arrival rate and the number of scheduled ambulances vs. time of the day (Calgary 2009 data).

same service time distribution. In a real EMS system, however, arrival rates vary systematically by time of the day and day of the week (Channouf et al. 2007, Setzler et al. 2009, Kim and Whitt 2014); the number of scheduled ambulances changes by time of the day and day of the week; and the service time of an ambulance depends on its location and on the number and locations of other available ambulances. Figure 6 illustrates how the arrival rate and the number of scheduled ambulances vary by time of the day in Calgary.

We address these real-world complications as follows:

- Instead of explicitly incorporating time-varying parameters in our model (as Ignall and Walker (1977) did), we divide each week into 168 1-hour segments and apply our model separately for each segment. We index the 1-hour time segments using τ, with τ ranging from 1 for Sundays between midnight and 1 am to $\tau = 168$ for Saturdays between 11 pm and midnight. We aggregate model outputs for the 1-hour segments to obtain global model outputs. At the end of Subsection 4.1, we outline reasons that suggest this method will be effective.
- We do not explicitly model ambulance locations, but the number of busy ambulances provides information about ambulance locations, because EMS dispatchers in Calgary and Edmonton use compliance tables to reposition ambulances (Alanis et al. 2013), that is, they try to achieve a pre-specified configuration of locations for each number of busy ambulances. Service rates

are likely to vary depending on the number of busy ambulances and the configuration of ambulance locations, and our $M/M(k)/c/c$ model captures part of this dependence via the state-dependent service rates.

Modeling of service times: We compare two modeling approaches for service times within each segment: general service times (G), and Markovian state-dependent service times ($M(k)$).

For the general distribution approach, we fit a mixture of lognormal distributions with separate components for calls that are transported and not transported to hospital (Appendix A), separately for each segment.

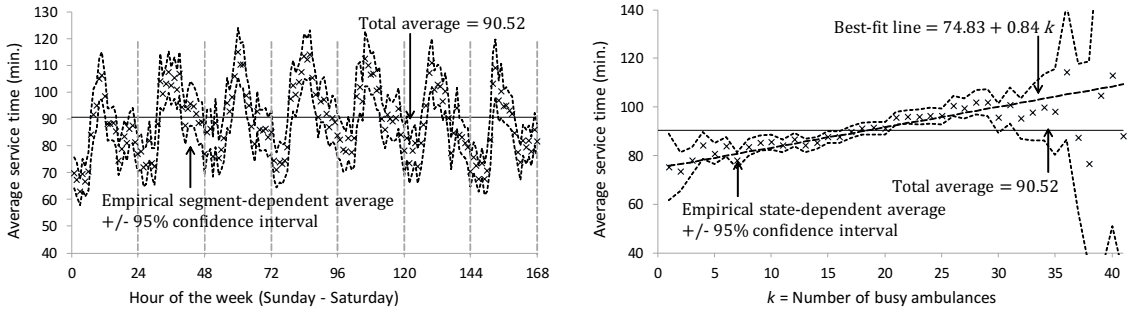
For the state-dependent approach, we use a regression model to capture how mean service times vary with segment (Figure 7(a)) and with the number of busy ambulances (Figure 7(b)). The mean service time change by segment can be partially attributed to the change in traffic and transportation speed. By increasing the number of busy ambulances, the mean service time tends to increase; Alanis et al. (2013) hypothesize that this ''slowdown'' effect occurs because a large number of ambulance patient arrivals causes ED crowding, which increases the time that ambulances are tied up in EDs, which translates to longer average service times. Delasay et al. (2016) also argue that as the number of busy ambulances increases, the average travel distance from available ambulances to call locations increases, which increases the mean service time. Delasay et al. (2016) discuss additional mechanisms through which the components of EMS service time vary with the number of busy ambulances.

To build our regression model, we compute the sample path for $\nu(t)$, the number of busy ambulances at t, by adding one at each call arrival epoch and subtracting one at each service completion epoch. We remove the data for 1 January and initialize $\nu(t)$ with the number of active calls at 0 am on 2 January, based on the assumption that none of these active calls arrived more than 24 hours before. KC (2013) used a similar approach to initialize a sample path for the number of busy physicians in an emergency department.

We estimate state-dependent mean service time for Segment τ using a simple continuous piecewise linear regression model with a cutoff point at the Yellow Alert threshold, θ:

$$\mathbf{E}\left[T_k^{(\tau)}\right] = \alpha_0 + \alpha_1 k + \alpha_2 \left(k - c^{(\tau)} + \theta\right) H + \sum_{i=1}^{24} \beta_i I_i + \sum_{i=1}^{23} \gamma_i J_i,\ \tau = 1, ..., 168,\ k = 1, ..., c, \quad (13)$$

where $T_k^{(\tau)}$ is the service time for a call that arrives during Segment τ when there are $\nu(t) = k$ busy ambulances; the dummy variable H is for Yellow Alert periods; the I_i dummy variables are for the hours of the day on weekends (Saturday-Sunday); the J_i dummy variables are for the hours of the day on weekdays (Monday-Friday); and we use 11 pm – 12 am on weekdays as the base case. In this model, we aggregate the 168 segments into 48 segments (24 for weekends and 24 for weekdays) and



(a) (b)

Figure 7 Average service time vs. number of busy ambulances and hour of the week (Calgary 2009 data).

we assume that the slope of the mean service time with respect to k (i.e. $\alpha_1 + \alpha_2 H$) is independent of the segment. We fit other models (Appendix E) where we did not aggregate the segments and where we allowed the slope with respect to k to vary by segment, but the parsimonious model in (13) was the one that minimized the Akaike Information Criterion (AIC).

The estimated intercept, slope, and Yellow Alert effect are $\hat{\alpha}_0 = 70.27$, $\hat{\alpha}_1 = 0.51$, and $\hat{\alpha}_2 = -2.21$, respectively, indicating that the estimated mean service time varies in the base case as follows: from 70.27 to 85.06 minutes outside Yellow Alert periods, when the number of busy ambulances varies from 0 to 29, and from 83.36 to 64.66 minutes inside Yellow Alert periods, when the number of busy ambulances varies from 30 to 41. The estimated increase in mean service time with k outside yellow alert periods is consistent with the slowdown finding by Alanis et al. (2013) and Delasay et al. (2016) that we discussed earlier. We attribute the negative $\hat{\alpha}_2$ to actions that EMS staff take during Yellow Alerts, and these actions cause the service times to decrease with k.

4.1. Validation Using the Entire Sample

In this subsection, we perform out-of-sample validation using the entire data set. In Subsection 4.2, we perform out-of-sample validation for each time segment separately. The validation process (both for the entire data set and separately for each time segment) consists of three steps: (1) estimate *model primitives* from the training sample, (2) use the primitives to compute *model outputs*, and (3) compare the model outputs to *empirical outputs* from the testing sample.

For validation, we take a *weighted-average approach*[2], in which we first estimate model primitives separately for each time segment in the training sample, and use these segment-specific primitives along with (2)-(3) and (10)-(12) to compute model outputs for each time segment, and then compute a weighted-average of the segment-specific model outputs to obtain global model outputs. We perform these steps for both ways of modeling services times ($M(k)$ and G).

[2] We also tried a *naive approach*, in which we estimated model primitives from the training sample, and used (2)-(3) and (10)-(12) to compute model outputs. This approach resulted in a poor fit and we do not discuss this approach further.

Model primitives: We use the superscript (τ) to indicate a notation is associated with Segment τ. For the weighted-average approach, we estimate segment-specific arrival rates $\hat{\lambda}^{(\tau)}$ and number of ambulances $\hat{c}^{(\tau)}$, which is the rounded average number of scheduled ambulances. We estimate two sets of segment-specific service time primitives: $\hat{\mu}_k^{(\tau)} = 1/\mathbf{E}\left[T_k^{(\tau)}\right]$, using (13), for $M(k)$ and lognormal mixture parameters estimated using only calls that arrived in Segment τ for G.

Model outputs: We use (2)-(3) and (10)-(12) to compute segment-specific model outputs $\mathbf{E}\left[B_k^{(\tau)}\right]$ and $\mathbf{E}\left[B_k^{(\tau)^2}\right]$, using the model primitives for Segment τ, for each of the two ways to model service times. (For G service times, we compute the second moment only for $k = c$.) We approximate the overall expected value as a weighted average of segment-specific expected values:

$$\mathbf{E}[B_k] \approx \sum_{\tau=1}^{168} \Pr(B_k \text{ begins in Segment } \tau)\mathbf{E}\left[B_k^{(\tau)}\right] \approx \frac{\sum_{\tau=1}^{168} \mathbf{E}\left[N_k^{(\tau)}\right]\mathbf{E}\left[B_k^{(\tau)}\right]}{\sum_{\tau=1}^{168} \mathbf{E}\left[N_k^{(\tau)}\right]}, \tag{14}$$

where $\mathbf{E}\left[N_k^{(\tau)}\right]$ is the expected number of k-partial busy periods that begin in Segment τ, which we approximate as:

$$\mathbf{E}\left[N_k^{(\tau)}\right] \approx l^{(\tau)}\lambda^{(\tau)}\pi_{k-1}^{(\tau)}, \tag{15}$$

where $l^{(\tau)}$ is the total duration of Segment τ, and $\pi_{k-1}^{(\tau)}$ is the steady-state probability that the Segment-τ Erlang loss system has $k-1$ busy ambulances. At the end of this subsection, we relax the assumption that the system reaches steady state within each segment, and calculate transient expected values and standard deviations for B_k, and show that our steady-state method provides close approximations, especially when k is within the Yellow Alert region, $30 \le k \le 41$. We use standard formulas (Gross and Harris 1998, p. 80, for example) with $\left(\hat{\lambda}^{(\tau)}, \hat{\mu}^{(\tau)}, \hat{c}\right)$ to obtain $\pi_{k-1}^{(\tau)}$. The rationale for the approximation is that $\lambda^{(\tau)}\pi_{k-1}^{(\tau)}$ is the steady-state rate at which k-partial busy periods begin in the Segment-τ system. Combining approximations (14)-(15), we obtain the following weighted average:

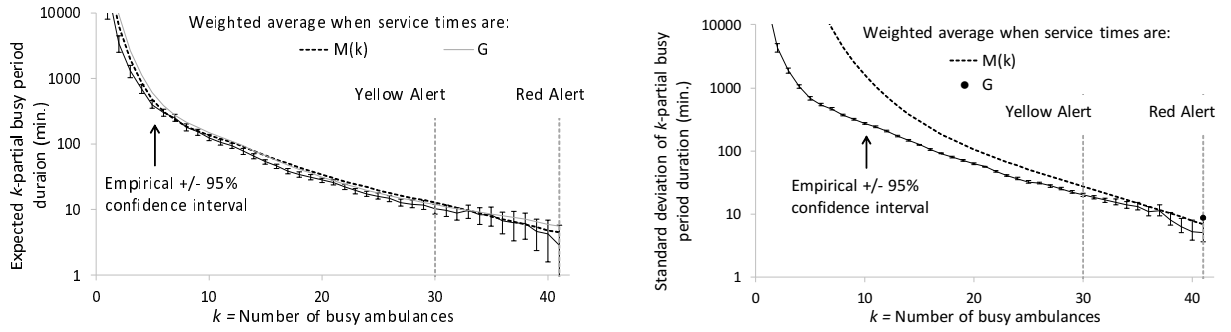
$$\mathbf{E}[B_k] \approx \sum_{\tau=1}^{168} w_k^{(\tau)}\mathbf{E}\left[B_k^{(\tau)}\right], \; w_k^{(\tau)} = \frac{l^{(\tau)}\lambda^{(\tau)}\pi_{k-1}^{(\tau)}}{\sum_{\tau=1}^{168} l^{(\tau)}\lambda^{(\tau)}\pi_{k-1}^{(\tau)}}. \tag{16}$$

We use the same approach to approximate the second moment:

$$\mathbf{E}\left[B_k^2\right] \approx \sum_{\tau=1}^{168} w_k^{(\tau)}\mathbf{E}\left[B_k{}^{(\tau)^2}\right], w_k^{(\tau)} = \frac{l^{(\tau)}\lambda^{(\tau)}\pi_{k-1}^{(\tau)}}{\sum_{\tau=1}^{168} l^{(\tau)}\lambda^{(\tau)}\pi_{k-1}^{(\tau)}}. \tag{17}$$

We use the first and second moments of B_k to compute $\mathbf{S}[B_k] = \sqrt{\mathbf{E}\left[B_k^2\right] - \mathbf{E}[B_k]^2}$. (For G service times, we can only compute $\mathbf{S}[B_c]$.)

Empirical outputs: We use the $\nu(t)$ sample path of the testing data to compute samples of empirical k-partial busy period durations $\{b_{ki}, k = 1, \ldots, c, i = 1, \ldots, n_k\}$, where n_k is the total



(a) Expected values

(b) Standard deviations

Figure 8 Empirical and model outputs for the entire sample (Calgary 2009 data), $\hat{c} = 41$.

number of k-partial busy periods. We calculate sample means, $\overline{b}_k$, and sample standard deviations, s_{b_k}, as $\overline{b}_k = (1/n_k)\sum_{i=1}^{n_k} b_{ki}$ and $s_{b_k} = \sqrt{(1/(n_k-1))\sum_{i=1}^{n_k}(b_{ki}-\overline{b}_k)^2}$, respectively.

Figure 8 illustrates our out-of-sample validation outcomes (Table 17 in the online supplement shows the numerical values). Figure 8(a) compares the model and empirical means. It shows that model outputs for $M(k)$ fit better than the ones for G, for small k, and they have very similar performance for large k, including those in the Yellow Alert region, which is the region of primary interest.

Figure 8(b) compares the model and empirical standard deviations. Here, we have G model outputs only for $k = c$. The $M(k)$ model outputs overestimate the empirical outputs but the deviation is smaller in the Yellow Alert region, which is the region of primary interest.

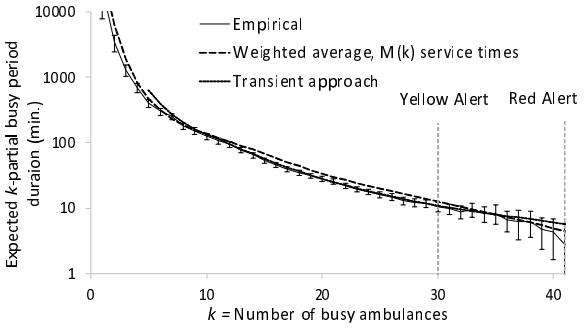
To better understand the validation results, we note that the average arrival rate decreased by 0.84% from 10.74 in the training sample to 10.65 calls per hour in the testing sample, and the average service time decreased by 6.40% from 90.52 to 84.73 minutes, while the average number of scheduled ambulances remained at 41. Based on the decrease in workload, we would expect the periods when the system has a large number of busy ambulances to be shorter in the testing sample than in the training sample. Indeed, in Figure 8(a), we see that the weighted average model outputs tend to be at the upper empirical confidence limits. In Figure 8(b), we further see that the model outputs overestimate the empirical standard deviations, indicating that the k-partial busy periods for high k in the testing sample were not only shorter, on average, but also less variable.

Perhaps the most important finding from this validation exercise is that even though the service time distribution is far from exponential, and service rates depend strongly on the number of busy ambulances, the two service time models ($M(k)$ and G) result in model outputs that are very close (the $M(k)$ model is slightly better). Our findings for the Edmonton data are similar. We already know from Theorem 3(a) that the first moments of partial busy periods are insensitive to the shape

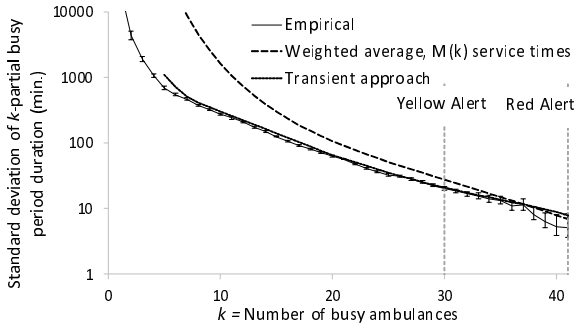
of the service time distribution beyond the mean, when service times are modeled as i.i.d. random variables. Our validation results supplement Theorem 3(a) with the numerical results that, for our data sets, the first moment is relatively insensitive to whether service times are modeled as state-dependent ($M(k)$) or not (G). It appears that for the purpose of developing valid models of partial busy periods, changes over time, such as the reduction in workload we see between our training and testing samples, could be more important than how service times are modeled.

Transient vs. steady state analysis and the effectiveness of time segmentation: We use (16)-(17) to approximate $\mathbf{E}[B_k]$ and $\mathbf{E}[B_k^2]$ by assuming that the system reaches the steady state within each time segment τ. Our method of approximating the queueing system in each time segment with a steady state system is similar to the ''stationary-independent-period-by-period'' (SIPP) approach for predicting staffing requirements for $M/M/c$ systems. Green et al. (2001) use numerical methods and show that the SIPP approach may provide inaccurate staffing levels in many realistic cases, which have time-varying parameters. We use a similar numerical method, as described in Appendix G, and investigate how the assumption of reaching the steady state within each segment impacts our $\mathbf{E}[B_k]$ and $\mathbf{E}[B_k^2]$ approximations. (The empirical and weighted average curves in Figure 9 are the same as those in Figure 8.) Figure 9(a) illustrates that model outputs for $\mathbf{E}[B_k]$, calculated based on steady-state and transient approaches are almost the same for all ks. Figure 9(b) illustrates that while the steady state model outputs for $\mathbf{S}[B_k]$ are close to those of the transient outputs for large k values (including k values within the Yellow Alert range, $30 \leq k \leq 41$), the difference between the two approaches grows when k is small. These figures illustrate that while our steady-state approximation performs well in the Yellow Alert region, which is the main focus of this paper, it is responsible for the poor performance of predicting the standard deviation of k-partial busy periods when k is small. It is important to mention that while the steady-state calculations are almost instantaneous, for the transient approach the computation time grows extremely fast when k decreases; for example, it took 2 days to complete calculations when $k = 5$.

We expect (consistent with results illustrated in Figure 9) our time segmentation approach to be more reliable than what Green et al. (2001) results suggest, for the following three reasons: 1) EMS systems typically have lower utilizations than the systems studied by Green et al. (2001). For example, based on our data, the utilization is 43% in Calgary and 57% in Edmonton (Table 1). 2) Our model is a loss system, whereas Green et al. (2001) SIPP approach is for delay systems. Therefore, in our model, queues do not build up and queues do not propagate from one time segment to future time segments. 3) Green et al. (2001) note that there is a lag between a peak in the arrival rate and a peak in congestion. Consequently, when the arrival rate is increasing, one expects steady-state models for delay systems to overestimate the probability of delay and when the arrival rate is decreasing, one expects steady-state models to underestimate the probability of delay. We



(a) Outputs of the two methods almost match.



(b) Outputs of the two methods are close within the Yellow Alert region.

Figure 9 Comparing the steady state and transient approaches using the entire sample (Calgary 2009 data).

aggregate model outputs for the 1-hour segments to obtain global model outputs, which implies that the overestimation and underestimation errors cancel out to some extent.

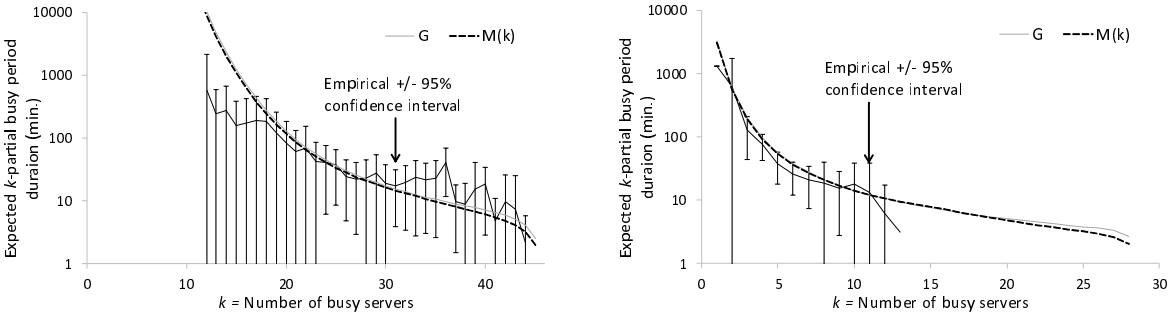
4.2. Validation by Segment

Our findings so far are that global model outputs computed using a weighted-average of segment-specific model outputs fit well with global empirical outputs. We would also like to know, however, how well the model outputs for a particular segment match empirical outputs for that segment, as that comparison provides an indication of the utility of our models for predicting alert period durations in real time. In this subsection, we perform this comparison, as a further validation step.

We perform out-of-sample validation for our model against each segment by using the same three validation steps listed in Section 4.1, for the entire sample. To demonstrate our validation steps, however, we focus on the busiest segment in the training data, Wednesday 1-2 pm, $\tau = 86$ (with 344 calls and 27.91 busy ambulances on average), and the least busy segment in the training data, Tuesday 4-5 am, $\tau = 53$ (with 113 calls and 6.52 busy ambulances on average). We focus on these two segments to show that our model provides good results both for segments with a high and a low number of busy ambulances.

Model primitives: For the busiest segment, we estimate the arrival rate to be $\hat{\lambda}^{(86)} = 13.77$ calls per hour, and the number of ambulances to be $\hat{c}^{(86)} = 45$ ambulances. We estimate the constant mean service time to be $\hat{T}^{(86)} = 114.12$ minutes, and we estimate the smoothed state-dependent mean service times as $\hat{T}_k^{(86)} = 91.79 + 0.51\,k$ minutes, for $k = 0, \ldots, 33$, and $\hat{T}_k^{(86)} = 164.79 - 1.70\,k$ minutes, for $k = 34, \ldots, 45$, from (13). The parameters of the lognormal mixture service time distribution are shown in Appendix A, Table 8.

For the least busy segment, our estimates are $\hat{\lambda}^{(53)} = 4.36$ calls per hour, $\hat{c}^{(53)} = 28$ ambulances, $\hat{T}^{(53)} = 75.49$ minutes, $\hat{T}_k^{(53)} = 70.11 + 0.51\,k$ minutes, for $k = 0, \ldots, 16$, and $\hat{T}_k^{(53)} = 105.50 - 1.70\,k$



(a) Busiest segment; $\hat{c} = 45$. (b) Least busy segment; $\hat{c} = 28$.

Figure 10 Empirical and model outputs for the first moment (Calgary 2009 data).

minutes, for $k = 17, \ldots, 28$, from (13). The parameters of the lognormal mixture service time distribution are shown in Appendix A, Table 9.

Model outputs: We compute the segment-specific model outputs as described in Subsection 4.1.

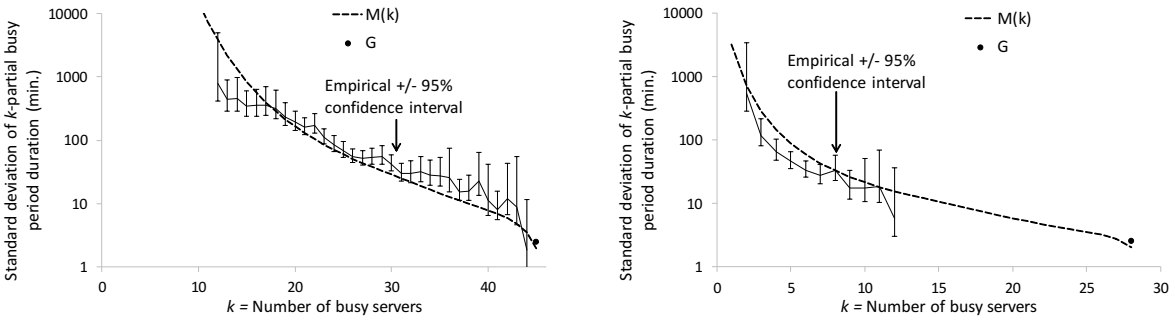
Empirical outputs: We would like the empirical outputs for Segment τ to be representative of conditions during that time segment. Therefore, we do not use empirical outputs based on the entire sample. A possible approach would be to use only k-partial busy periods that are fully contained within Segment τ, but this approach would severely bias the analysis, because partial busy periods that span more than one hour tend to be longer, on average, than those contained within an hour. Instead, we construct separate sample paths for each segment, by concatenating the 26 1-hour periods of observations that we have for each segment.

Appendix B describes the concatenation procedure. We follow the same process as we did for the entire sample to construct the sample path and calculate the empirical outputs for each segment. The concatenated sample paths are not used in the estimation of model primitives, the calculation of model outputs, or in the aggregate validation in Section 4.1—they are only used to calculate segment-specific empirical outputs.

Figures 10-11 compare model and empirical outputs for the busiest and least busy segments. The model outputs are generally within the 95% confidence intervals for the empirical outputs, especially for high values of k, which correspond to Yellow and Red Alerts. The model outputs for the two service time models are almost indistinguishable for high k values.

5. Modeling the Impact of Actions

As discussed in Section 1, once a Yellow Alert period begins, EMS staff face the uncertainty of whether the alert will be naturally short-lived, or whether the system will operate with a shortage of available ambulances for an extended period that can lead to longer response times and possibly a Red Alert. In this section, we extend the loss model with Markovian state-dependent service



(a) Busiest segment; $\hat{c} = 45$.

(b) Least busy segment; $\hat{c} = 28$.

Figure 11 Empirical and model outputs for the standard deviation (Calgary 2009 data).

rates, $M/M(k)/c/c$, to incorporate add and expedite actions that staff could take to handle alert periods. Our choice of the $M/M(k)/c/c$ model (as opposed to the $M/G/c/c$ model) follows our findings in Section 4 that while both of these models provide similar capabilities of predicting the first moment of B_k for large k, specifically when k is within the Yellow Alert range, the former provides better predictions when k is small. In addition, because of the Markovian property, the $M/M(k)/c/c$ model is more tractable.

We assume that add and expedite actions are taken within a Yellow Alert when $k' \in \{c-\theta+1, ..., c\}$ ambulances are busy. We use s_0 to denote the initial state when we take an action, $\tilde{B}_{s_0}$ to denote the remaining Yellow Alert duration and H_{s_0} to denote the number of lost calls during $\tilde{B}_{s_0}$.

Two outcomes that EMS staff would like to avoid are: (1) longer response times, which happen when the number of available ambulances decreases, and (2) lost calls (those that arrive during a Red Alert). We define two performance measures that correspond to these outcomes: (1) the expected remaining Yellow Alert duration, $\mathbf{E}\left[\tilde{B}_{s_0}\right]$, and (2) the expected number of lost calls, $\mathbf{E}\left[H_{s_0}\right]$, during the remainder of the Yellow Alert. Figure 1 illustrates how the average response time increases during Yellow Alert periods, which motivates our use of the first measure. The second measure is directly related to the second outcome of interest, the number of calls during Red Alerts.

We first model the $M/M(k)/c/c$ system as a continuous-time Markov chain (CTMC) for the number of busy servers, $\{\nu(t), t \geq 0\}$, and calculate $\mathbf{E}\left[\tilde{B}_{s_0}\right]$ and $\mathbf{E}\left[H_{s_0}\right]$ assuming no action is taken, which we use as the base case. We then augment the state variable with additional variables to model the add and expedite actions. We calculate $\mathbf{E}\left[\tilde{B}_{s_0}\right]$ and $\mathbf{E}\left[H_{s_0}\right]$ using standard results for absorbing CTMCs. We use Ω to denote the CTMC state space, A for the set of absorbing states, A^c for the set of transient states ($A \cup A^c = \Omega$), and $A^r \subseteq A^c$ for the set of Red Alert states, in which all ambulances are busy. We specify s_0, Ω, A, and A^r when no action is taken, and when add and expedite actions are taken.

5.1. No Action (Base Case)

We model the $M/M(k)/c/c$ system as a CTMC, $\{\nu(t), t \geq 0\}$. Assuming the system is currently within a Yellow Alert and the number of busy ambulances is k', for $k' \in \{c-\theta+1, ..., c\}$, the initial state becomes $s_0 = k'$, and we calculate $\mathbf{E}[\tilde{B}_{s_0}]$ and $\mathbf{E}[H_{s_0}]$.

To compute $\mathbf{E}[\tilde{B}_{s_0}] = \mathbf{E}[\tilde{B}_{k'}]$, we decompose the residual Yellow Alert duration as $\tilde{B}_{k'} = \Upsilon_{k'} + \Upsilon_{k'-1} + \cdots + \Upsilon_{c-\theta+1}$, $k' \in \{c-\theta+1, ..., c\}$, where Υ_i is the time it takes for the number of busy ambulances to decrease from i to $i-1$. For the $M/M(k)/c/c$ system, because of the memoryless property, Υ_i equals B_i in distribution. Therefore:

$$\mathbf{E}\left[\tilde{B}_{k'}\right] = \sum_{i=c-\theta+1}^{k'} \mathbf{E}[B_i], \; k' = c-\theta+1, \ldots, c, \tag{18}$$

where the $\mathbf{E}[B_i]$ values can be calculated using (2).

To compute $\mathbf{E}[H_{s_0}] = \mathbf{E}[H_{k'}]$, we collapse all states outside the Yellow Alert period into a single absorbing state and create a modified $M/M(k)/c/c$ system such that $\Omega = \{c-\theta, ..., c\}$ and $A = \{c-\theta\}$, as depicted in Figure 12(a). The initial state is $s_0 = k'$ and the set of Red Alert states is $A^r = \{c\}$. The infinitesimal generator matrix Q in canonical form (Kao 1996) is:

$$Q = \begin{array}{c|cc} & A & A^c \\ \hline A & 0 & 0 \\ A^c & Y & Z \end{array}.$$

The fundamental matrix (Kao 1996, p. 256) for this Markov chain is

$$V = -Z^{-1},$$

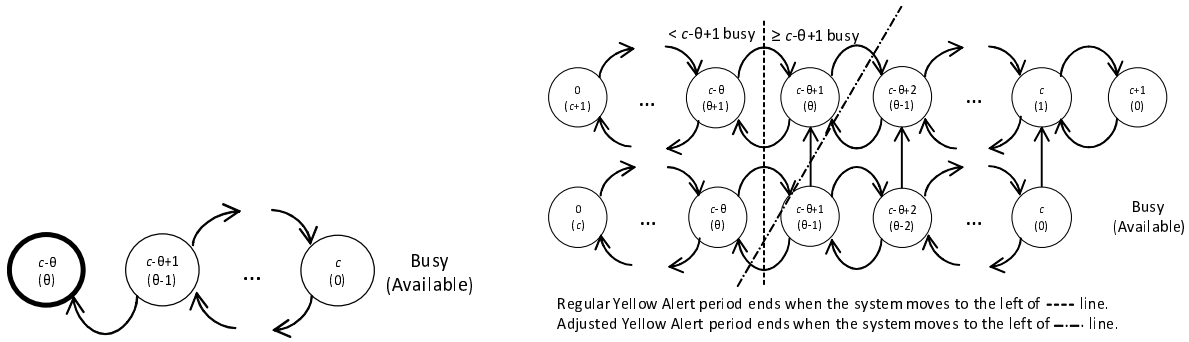
where v_{ij} is the expected time spent before absorption in transient state j, given that the chain begins in transient state i. As summarized in Proposition 1 below, the fundamental matrix provides an alternative way to obtain $\mathbf{E}\left[\tilde{B}_{s_0}\right]$, as the total expected time spent in transient states prior to absorption, and a way to obtain $\mathbf{E}[H_{s_0}]$, which equals the arrival rate times the expected time spent in state A^r:

Proposition 1.

$$\mathbf{E}\left[\tilde{B}_{s_0}\right] = \sum_{s \in A^c} v_{s_0 s}. \tag{19}$$

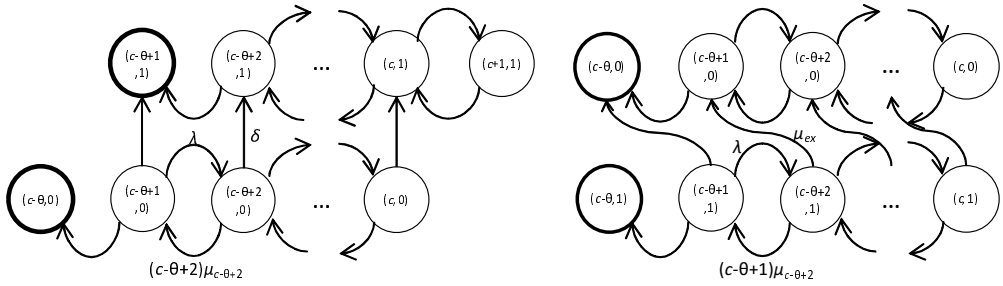
$$\mathbf{E}[H_{s_0}] = \lambda \sum_{s \in A^r} v_{s_0 s}. \tag{20}$$

In the absence of any action, (19) reduces to $\mathbf{E}\left[\tilde{B}_{k'}\right] = \sum_{s \in A^c} v_{k's}$, $A^c = \{c-\theta+1, \ldots, c\}$ and (20) reduces to $\mathbf{E}[H_{k'}] = \lambda v_{k'c}$, where $k' \in \{c-\theta+1, \ldots, c\}$.



(a) No action is taken. (b) Regular and Adjusted Yellow Alert periods when $c_{\text{add}} = 1$.

Figure 12 **Absorbing states (indicated by a thicker border) when no action is taken and adjusted Yellow Alert period when $c_{\text{add}} = 1$.**



(a) Add 1 ambulance. (b) Expedite 1 ambulance.

Figure 13 **Modified Markov chains (absorbing states have thicker borders) when $c_{\text{add}} = 1$ and $c_{\text{ex}} = 1$.**

5.2. Add Ambulances

We assume that c_{add} ambulances are added. We model the arrival times of the c_{add} ambulances as independent exponential random variables with rate δ. We augment the state of the CTMC $\{\nu(t), t \geq 0\}$ by adding $w(t)$ for the number of requested ambulances that have already arrived by time t, and form a bivariate CTMC $\{(\nu(t), w(t)), t \geq 0\}$. We also define an *adjusted* Yellow Alert period, which begins when the number of busy ambulances increases to $c - \theta + 1$ and ends when the number of *available* ambulances increases to θ for the first time. Figure 12(b) illustrates the difference between adjusted and regular Yellow Alert periods when $c_{\text{add}} = 1$. Both of these periods begin when the number of busy ambulances increases to $c - \theta + 1$. The regular partial busy period ends when the system enters a state with less than $c - \theta + 1$ busy ambulances (left of the vertical dashed line) while the adjusted one ends when the system enters a state with θ available ambulances (left of the diagonal dashed-dot line).

We modify the Markov chain $\{(\nu(t), w(t)), t \geq 0\}$ such that $\Omega = \{(i,j) | i = c - \theta + j, ..., c + j,\ j = 0, ..., c_{\text{add}}\}$ and $A = \{(i,j) | i = c - \theta + j,\ j = 0, ..., c_{\text{add}}\}$. Transitions are as follows: *Call arrival:* $(i,j) \to (i+1, j)$, if $c - \theta + j < i < c + j$; *call departure:* $(i,j) \to (i-1, j)$, if $i > c - \theta + j$; and

ambulance arrival: $(i,j) \to (i,j+1)$, if $j < c_{\text{add}}$. This system begins from the initial state $s_0 = (k',0)$, and its Red Alert states are $A^r = \{(i,j) | i = c+j,\ j = 0,...,c_{\text{add}}\}$. Figure 13(a) shows Ω, A, A^c, and transitions when $c_{\text{add}} = 1$. We use Proposition 1 to calculate $\mathbf{E}\left[\tilde{B}_{s_0}\right]$ and $\mathbf{E}[H_{s_0}]$.

5.3. Expedite Ambulances

We assume that c_{ex} ambulances that are tied up in EDs are expedited and released faster. We model remaining service times of these ambulances as independent exponential distributions with expedited rate μ_{ex}. Expedited ambulances operate with normal service rate μ_k after being released. We augment the state of the CTMC $\{\nu(t), t \geq 0\}$ by adding $y(t)$ for the number of expedited ambulances that are still tied up in the EDs (expedite process has not finished yet) at time t, and form a bivariate CTMC $\{(\nu(t), y(t)), t \geq 0\}$.

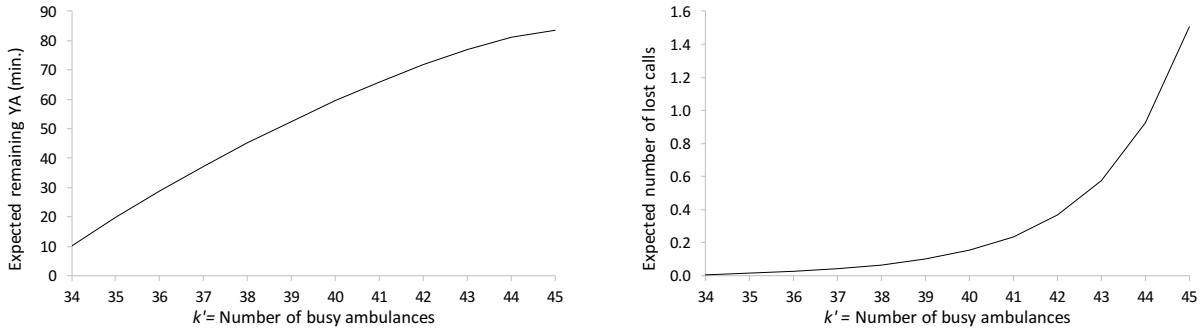
We modify the Markov chain $\{(\nu(t), y(t)), t \geq 0\}$ such that $\Omega = \{(i,j) | i = c-\theta, ..., c,\ j = c_{\text{ex}}, ..., 0\}$ and $A = \{(i,j) | i = c-\theta,\ j = c_{\text{ex}}, ..., 0\}$. Transitions are as follows: *Call arrival:* $(i,j) \to (i+1,j)$, if $c-\theta < i < c$; *normal call departure:* $(i,j) \to (i-1,j)$, if $i > c-\theta$; and *expedited call departure:* $(i,j) \to (i-1,j-1)$, if $i > c-\theta$ and $j > 0$. This system begins from the initial state $s_0 = (k', c_{\text{ex}})$, and its Red Alert states are $A^r = \{(i,j) | i = c,\ j = c_{\text{ex}}, ..., 0\}$. Figure 13(b) shows Ω, A, A^c, and transitions when $c_{\text{ex}} = 1$. We use Proposition 1 to calculate $\mathbf{E}\left[\tilde{B}_{s_0}\right]$ and $\mathbf{E}[H_{s_0}]$.

6. Numerical Results and Managerial Insights

In this paper, we do not attempt to model the cost of actions, and therefore, we do not suggest optimal actions. Instead, we use our models from Section 5 to obtain insights on when EMS staff should take actions, and how these actions compare with each other with respect to the two performance measures, the expected remaining Yellow Alert duration and the expected number of lost calls. We focus on the busiest Calgary segment ($\tau = 86$, with $\hat{\lambda}^{(86)} = 13.77$/hour, $\hat{c}^{(86)} = 45$, and $\hat{T}_k^{(86)} = 164.79 - 1.70\,k$) and use model primitives for that segment in the numerical calculations in this section. We repeated the calculations for other segments and observed patterns consistent with those that we discuss in this section. We systematically compare adding $c_{\text{add}} = 0, 1$, or 2 ambulances, and expediting $c_{\text{ex}} = 0, 1$, or 2 ambulances.

6.1. When Should Staff Take Actions?

For EMS staff, it is important to understand what would be the implications of waiting, and not taking actions in the hope that the Yellow Alert period will end soon. To gain insights on this question, we keep all parameters related to add and expedite actions fixed and vary the number of busy ambulances at the action epoch, k'. Figure 14 illustrates how the performance measures change by varying k', assuming no action is taken.



(a) Expected remaining Yellow Alert if no action is taken (b) Expected number of lost calls if no action is taken

Figure 14 How performance measures escalate when staff decide to wait and not to take actions (Busiest segment, Calgary 2009 data).

Observation 1: Staff should not wait until a Red Alert to take actions. There is an almost-linear relationship between k' and the expected remaining Yellow Alert duration as shown in Figure 14(a). However, the relationship between k' and the expected number of lost calls is convex and rapidly increasing, as shown in Figure 14(b), which is consistent with the convex and rapidly-increasing relationship between the loss probability and partial busy periods, shown in Figure 5. With respect to the expected remaining Yellow Alert duration, staff should take an action as soon as possible, because the performance measure constantly deteriorates when k' increases. With respect to the expected number of lost calls, however, they can wait at early stages of the Yellow Alert, as outcomes do not change significantly initially (for this segment, $k' = 34, ..., 39$). Nonetheless, they should be careful as the performance measure escalates quickly after a certain point (for this segment, when $k' \geq 42$).

6.2. How Do Actions Compare With Each Other?

Even when staff decide to take an action, the question of "which action should be taken?" is not easy to answer, especially when the expected action realization times (*expected action times*, for short) are different (that is, $1/\delta \neq 1/\mu_{\text{ex}}$).

6.2.1. Pairwise Comparison of Actions With Equal Expected Action Times

THEOREM 4. *Adding c_{add} ambulances is always at least as beneficial as expediting c_{ex} ambulances when $c_{\text{add}} \geq c_{\text{ex}}$, with respect to both performance measures, if the arrival time of the added ambulances and the remaining service time of the expedited ambulances have the same distribution.*

Proof: See online supplement Subsection D.4. Consider a special case when $c_{\text{add}} = c_{\text{ex}} = 1$. The proof relies on the coupling method, where we assume that the arrival of the new ambulance and the release of the expedited ambulance occur simultaneously. The proof evolves around the idea that the added ambulance stays in the system and has a long-term impact while the expedited

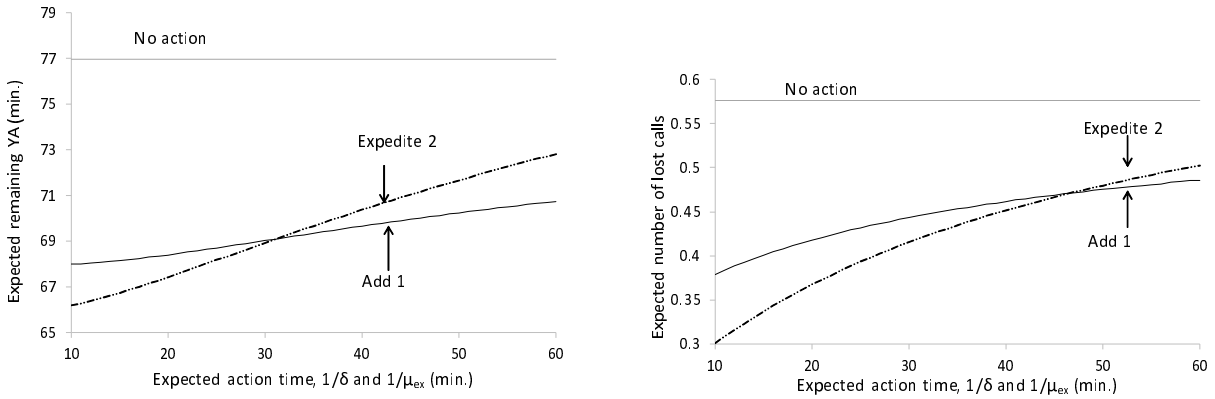
Table 3 **Threshold values for comparing add ($c_{\text{add}} = 1$) and expedite ($c_{\text{ex}} = 2$) actions, with respect to the two performance measures, for $k' = 34, ..., 45$.**

k'	34	35	36	37	38	39	40	41	42	43	44	45
E[Rem. YA]	50	48	46	45	43	41	39	36	34	31	28	25
E[Lost calls]	43	41	41	41	41	42	43	44	45	47	48	49

ambulance has a one-time impact. When $c_{\text{add}} = c_{\text{ex}} = n > 1$, the coupling idea is extended under the assumption that the remaining service time distribution of the ith expedited ambulance is the same as the arrival time distribution of the ith requested ambulance, for $i = 1, \ldots, n$. Extending the proof from the $c_{\text{add}} = c_{\text{ex}}$ case to the $c_{\text{add}} > c_{\text{ex}}$ case is straightforward as the additional number of added ambulances can only improve the system.

As discussed in Section 1, expediting is realized through actions that lead to a faster transfer of the patient's care from the EMS to the ED, or to other EMS crews (care consolidation). Adding, however, is realized by borrowing ambulances from neighboring municipalities, from another service (an interfacility-transfer ambulance fleet, for example), or asking new ambulance crews to come on duty. Decisions associated with some actions are solely made within the EMS system (care consolidation, for example), making them easier to manage, but some actions require collaborations with other health care systems (borrowing ambulances, for example), making them more challenging to manage. Depending on action types, there might be operational difficulties that staff must take into account; for example, neighboring EMS systems may not have available ambulances to lend at a given time.

Observation 2: When $c_{\text{add}} < c_{\text{ex}}$, the comparison of expedite and add actions is of a threshold type. The staff should expedite if the expected action time is small, and they should add if the expected action time is large. We calculate the expected action time threshold numerically. For $c_{\text{add}} = 1$ and $c_{\text{ex}} = 2$, we vary the expected action times between 10 and 60 minutes, by 1-minute increments, and calculate performance measures. Figure 15 illustrates linearly interpolated results when $k' = 43$. With respect to both performance measures, the actions have strictly increasing curves, and the expedite action outperforms the add action for small expected action times. With respect to the expected remaining Yellow Alert duration, the two actions perform equally (the two curves cross) when the expected action time is 31 minutes (see Fig. 15(a)), and with respect to the expected number of lost calls, they perform equally when the expected action time is 47 minutes (see Fig. 15(b)). Therefore, when the expected action time is smaller than the threshold 31 (47) minutes, the expedite action outperforms the add action, with respect to the expected remaining Yellow Alert duration (the expected number of lost calls). We recreated Figure 15 for all $k' = 34, ..., 45$. Table 3 shows the threshold values for different k'.



(a) (b)

Figure 15 **Comparing $c_{\text{add}} = 1$ and $c_{\text{ex}} = 2$ actions when we set $k' = 43$, and vary the expected action time.**

6.2.2. Pairwise Comparison of Actions With Different Expected Action Times

In the same fashion as in Figure 15, we generated performance curves for $c_{\text{add}} = 1, 2$; $c_{\text{ex}} = 1, 2$; and $k' = 34, ..., 45$, by varying the expected action times from 10 to 60 minutes. We use these curves to identify actions that can be taken to achieve a desired improvement over the base case (no action). For example, if $k' = 43$, and no action is taken, then as illustrated in Figure 15, the expected remaining Yellow Alert duration will be 76.95 minutes and the expected number of lost calls will be 0.58. If, however, staff decide to take an action and wish to improve the expected number of lost calls by 30% (reduce it to 0.4), then either $c_{\text{add}} = 1$ ambulance must be added, with $1/\delta \leq 16$ minutes, or $c_{\text{ex}} = 2$ ambulances must be expedited, with $1/\mu_{\text{ex}} \leq 27$ minutes.

Table 4 summarizes the largest expected action times that achieve 5% to 50% improvement in the expected remaining Yellow Alert duration of the base case. Table rows show desired improvements in the base case performance measure. For each desired improvement-action pair in the table, we vary the expected action time between 10 and 60 minutes and record the largest expected action time that satisfies the desired improvement. For example, when $k' = 43$, if no action is taken (base case), the expected remaining Yellow Alert duration will be 76.95 minutes. If staff want to achieve a 10% improvement (to reduce the performance measure to $0.9 \times 76.95 = 69.26$ minutes), they should add 1 ambulance with $1/\delta \leq 34$ minutes, 2 ambulances with $1/\delta \leq 60$ minutes, or expedite 2 ambulances with $1/\mu_{\text{ex}} \leq 32$ minutes. It is not possible to achieve the desired improvement by expediting only 1 ambulance. The largest improvement shown in this table is 50%, and we use increments of 5%, which indicates that we cannot achieve more than 55% improvement by adding $c_{\text{add}} = 1, 2$, or expediting $c_{\text{ex}} = 1, 2$ ambulances. Table columns show different actions, grouped by k' values. Table 5 summarizes similar results with respect to the expected number of lost calls (which cannot be improved more than 85% compared to the base case, by adding $c_{\text{add}} = 1, 2$, or expediting $c_{\text{ex}} = 1, 2$ ambulances).

Table 4 **The largest expected action time (min.) that achieves the target percentage improvement in the expected remaining Yellow Alert duration, compared to the base case.**

Imp.%	$k'=35$; Base case = 19.83				$k'=39$; Base case = 52.54				$k'=43$; Base case = 76.95			
	Add 1	Add 2	Ex. 1	Ex. 2	Add 1	Add 2	Ex. 1	Ex. 2	Add 1	Add 2	Ex. 1	Ex. 2
5%	60	60	59	60	60	60	48	60	60	60	32	60
10%	60	60	37	58	56	60	20	47	34	60	—	32
15%	41	60	24	44	21	60	—	31	—	60	—	—
20%	25	60	16	34	—	48	—	19	—	27	—	—
25%	16	46	11	27	—	29	—	—	—	—	—	—
30%	11	34	—	21	—	16	—	—	—	—	—	—
35%	—	25	—	17	—	—	—	—	—	—	—	—
40%	—	19	—	13	—	—	—	—	—	—	—	—
45%	—	14	—	10	—	—	—	—	—	—	—	—
50%	—	11	—	—	—	—	—	—	—	—	—	—

Table 5 **The largest expected action time (min.) that achieves the target percentage improvement in the expected number of lost calls, compared to the base case.**

Imp.%	$k'=35$; Base case = 0.01				$k'=39$; Base case = 0.10				$k'=43$; Base case = 0.58			
	Add 1	Add 2	Ex. 1	Ex. 2	Add 1	Add 2	Ex. 1	Ex. 2	Add 1	Add 2	Ex. 1	Ex. 2
5%	60	60	60	60	60	60	60	60	60	60	60	60
10%	60	60	60	60	60	60	60	60	60	60	45	60
15%	60	60	58	60	60	60	47	60	60	60	30	54
20%	60	60	47	60	60	60	34	60	39	60	20	43
25%	60	60	37	60	60	60	24	52	25	60	13	34
30%	60	60	29	55	46	60	15	43	16	55	—	27
35%	58	60	21	49	31	60	—	36	—	41	—	21
40%	43	60	15	42	19	60	—	29	—	31	—	16
45%	31	60	—	36	—	60	—	23	—	23	—	12
50%	21	60	—	31	—	48	—	17	—	17	—	—
55%	12	60	—	25	—	37	—	10	—	12	—	—
60%	—	52	—	20	—	28	—	—	—	—	—	—
65%	—	41	—	15	—	19	—	—	—	—	—	—
70%	—	31	—	10	—	—	—	—	—	—	—	—
75%	—	22	—	—	—	—	—	—	—	—	—	—
80%	—	14	—	—	—	—	—	—	—	—	—	—

Observation 3: If staff wait too long before taking action, then the action becomes less effective, even if the expected action time is short. Tables 4-5 show that the later an action is taken, the less its *marginal* improvement in comparison with taking no action becomes, with respect to both performance measures. For example, according to Table 4, while adding 2 ambulances can improve the base case up to 30% when $k'=39$, this action cannot bring more than, or equal to, 25% improvements if the action is taken when $k'=43$.

Expediting, through the consolidation of care of several waiting EMS patients under a single paramedic crew, is perhaps the action that requires the least coordination with other agencies, and therefore could be realized quickly. Here, we provide more insights on this action by comparing

Table 6 The largest expected action time (min.) for expediting 2 ambulance that outperforms adding 1 ambulance that is available instantaneously, with respect to the expected number of lost calls.

k'	34	35	36	37	38	39	40	41	42	43	44	45
Base Case	0.01	0.01	0.03	0.04	0.06	0.10	0.15	0.23	0.36	0.58	0.92	1.51
Inst. Add	0.00	0.00	0.01	0.02	0.03	0.05	0.08	0.13	0.21	0.33	0.53	0.86
(Imp. %)	(100%)	(67%)	(56%)	(51%)	(48%)	(46%)	(45%)	(44%)	(43%)	(43%)	(43%)	(43%)
Largest $1/\mu_{\text{ex}}$	—	13.15	19.61	21.99	22.48	21.95	20.74	18.94	16.58	13.60	—	—

Table 7 The largest expected action time (min.) for expediting 2 ambulance that outperforms adding 1 ambulance that is available instantaneously, with respect to the expected remaining Yellow Alert duration.

k'	34	35	36	37	38	39	40	41	42	43	44	45
Base case	10.26	19.83	28.79	37.21	45.12	52.54	59.50	65.95	71.82	76.95	81.07	83.62
Inst. Add	0.00	9.59	18.58	27.04	35.01	42.52	49.61	56.26	62.46	68.13	73.11	77.12
(Imp. %)	(100%)	(52%)	(35%)	(27%)	(22%)	(19%)	(17%)	(15%)	(13%)	(11%)	(10%)	(8%)
Largest $1/\mu_{\text{ex}}$	—	—	12.68	16.40	19.02	20.81	21.99	22.76	23.43	24.65	28.16	37.81

expediting $c_{\text{ex}} = 2$ ambulances with small $1/\mu_{\text{ex}}$, to adding $c_{\text{add}} = 1$ ambulance with instantaneous arrival—that is, $1/\delta$ approaches 0. Although it may not be realistic for $1/\delta$ to approach 0, it provides a bound on the amount of improvement that EMS can expect from the add action.

Observation 4: When $c_{\text{add}} < c_{\text{ex}}$, the expedite action with a sufficiently-small expected action time can outperform the add action, even if the arrival of new ambulances is instantaneous. Table 6 shows the largest $1/\mu_{\text{ex}}$ for an expedite action with $c_{\text{ex}} = 2$ that outperforms an instantaneous add action with $c_{\text{add}} = 1$, for $k' = 34, ..., 45$, with respect to the expected number of lost calls. According to this table, if $k' = 43$ and no action is taken, then the expected number of lost calls will be 0.58. At $k' = 43$, the instantaneous add action with $c_{\text{add}} = 1$ reduces the expected number of lost calls to 0.33 (43% improvement). The same improvement can be achieved by expediting 2 ambulances, when $1/\mu_{\text{ex}} \leq 13.60$ minutes. Table 7 shows similar results with respect to the expected remaining Yellow Alert duration.

7. Conclusion

This paper provides an understanding of capacity shortage periods in mission critical systems like fire, police, and EMS. We focus on EMS systems and model these systems as Erlang loss systems with service times modeled as either Markovian and state-dependent ($M(k)$) or general (G). We show that the expected duration of periods during which at least k out of c ambulances are busy is independent of the shape of the service time distribution beyond its mean, but this is not true of the higher moments. We obtain closed-form formulas and easy-to-use recursions to calculate the expected duration and (for $M(k)$ service times) the standard deviation of ambulance-shortage periods. We validate our formulas for the mean and standard deviation of partial busy periods against empirical data from the Calgary and Edmonton EMS systems. Our validation results show

that although ambulance service times are far from exponential, the loss model with Markovian state-dependent service time slightly outperforms the loss model with general service time, with respect to predicting the first moment of partial busy periods.

We expand the $M/M(k)/c/c$ model and use the theory of absorbing Markov chains to quantify the impact of adding or expediting ambulances on two performance measures: The expected remaining duration of a Yellow Alert (a proxy for periods with long response times) and the expected number of lost calls during this residual duration. We show that, regardless of the action type, the expected number of lost calls increases rapidly when the number of available ambulances at action epoch decreases, and that the escalation is almost linear with respect to the other performance measure. Based on our results, both performance measures are monotonic functions of the expected value of the time until actions are realized ($1/\delta$ and $1/\mu_{\text{ex}}$) and their comparison is of a threshold type. We show the two actions may be ranked differently with respect to the two performance measures.

Several related issues could benefit from further study, including: Models and algorithms to choose the Yellow Alert threshold θ so as to balance mitigation of capacity shortages and the added workload from operating in alert mode; investigating whether θ should vary with time; investigating the action of delaying response to low-priority calls during alert periods; analytically investigating when the stationary approach, as opposed to the transient approach, works well in modeling loss systems and providing simple modifications for improvements, if needed; and empirically investigating whether the time required to expedite (add) one additional ambulance becomes progressively larger, as one would expect.

Acknowledgement

The authors would like to thank the Department Editor, Associate Editor, and three anonymous referees, for their thoughtful comments that led to considerable improvements in this paper. We are also grateful to Alberta Health Services (AHS) for providing the data and to AHS staff for their empirical insights. This work was partially supported by the Canadian Natural Science and Engineering Research Council (Discovery Grant 203534).

Appendix

Additional supporting information may be found in the online Appendix section.

References

ABC News (2015) Union warning of reduced ambulance coverage in the central west. *ABC News* 22 July. Accessed November 27, 2019, `https://ab.co/2OU8pEL`.

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

Alberta Health Services (2010) Emergency department surge capacity protocols. Accessed June 15, 2016, https://bit.ly/2IPbuV6.

Almehdawe E, Jewkes B, He QM (2013) A Markovian queueing model for ambulance offload delays. *European Journal of Operational Research* 226(3):602–614.

Artalejo JR, Lopez-Herrero MJ (2001) Analysis of the busy period for the $M/M/c$ queue: An algorithmic approach. *Journal of Applied Probability* 38(1):209–222.

Bountourelis T, Ulukus MY, Kharoufeh JP, Nabors SG (2013) The Modeling, Analysis, and Management of Intensive Care Units. *Handbook of Healthcare Operations Management: Methods and Applications* (New York: Springer).

Brown D (2018) Paramedic union tweet has city councillor 'gravely concerned' about shortage of ambulances. *CBC News* 29 November. Accessed April 21, 2019, https://bit.ly/2GwtiBK.

Cha WC, Shin SD, Song KJ, Jung SK, Suh GJ (2009) Effect of an independent-capacity protocol on overcrowding in an urban emergency department. *Academic Emergency Medicine* 16(12):1277–1283.

Chan CW, Farias VF, Escobar GJ (2017) The impact of delays on service times in the intensive care unit. *Management Science* 63(7):2049–2072.

Chan CW, Yom-Tov G, Escobar G (2014) When to use speedup: An examination of service systems with returns. *Operations Research* 62(2):462–482.

Channouf N, L'Ecuyer P, Ingolfsson A, Avramidis AN (2007) The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta. *Health Care Management Science* 10(1):25–45.

Chong KC, Henderson SG, Lewis ME (2015) The vehicle mix decision in emergency medical service systems. *Manufacturing & Service Operations Management* 18(3):347–360.

Delasay M, Ingolfsson A, Schultz K (2016) Inventory is people: How load affects service times in emergency response. Gavirneni S, ed., *Cross-Functional Inventory Research* (World Scientific).

Duran A, Gutierrez G, Zequeira RI (2004) A continuous review inventory model with order expediting. *International Journal of Production Economics* 87(2):157–169.

Fitch JJ, Keller RA, Raynor D, Zalar C (1993) *EMS Management: Beyond the Streets* (Carlsbad, CA: JEMS Communications), 2nd edition.

Green LV, Kolesar PJ, Soares J (2001) Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research* 49(4):549–564.

Gross D, Harris CM (1998) *Fundamentals Of Queueing Theory (Wiley Series In Probability And Statistics)* (Wiley-Interscience), 3rd edition.

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

Ignall E, Walker WE (1977) An analysis of the deployment of ambulances in Washington DC. *Journal of Urban Analysis* 4(1):59–92.

Ingolfsson A (2013) EMS planning and management. Zaric GS, ed., *Operations Research and Health Care Policy*, 105–128 (New York, NY: Springer New York).

Jain R, Smith JM (1997) Modeling vehicular traffic flow using $M/G/C/C$ state dependent queueing models. *Transportation Science* 31(4):324–336.

Kao EPC (1996) *An Introduction to Stochastic Processes* (New York: Cengage Learning).

KC DS (2013) Does multitasking improve performance? Evidence from the emergency department. *Manufacturing & Service Operations Management* 16(2):168–183.

KC DS, Terwiesch C (2012) An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management* 14(1):50–65.

Kim SH, Whitt W (2014) Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing & Service Operations Management* 16(3):464–480.

Kolesar PJ, Rider KL, Crabill TB, Walker WE (1975) A queuing-linear programming approach to scheduling police patrol cars. *Operations Research* 23(6):1045–1062.

Lawson DG, Porteus EL (2000) Multistage inventory management with expediting. *Operations Research* 48(6):878–893.

Li AA, Whitt W (2014) Approximate blocking probabilities in loss models with independence and distribution assumptions relaxed. *Performance Evaluation* 80:82–101.

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

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

Omahen K, Marathe V (1978) Analysis and applications of the delay cycle for the $M/M/c$ queueing system. *Journal of the ACM* 25(2):283–303.

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

Rumbolt R (2017) EMS declare 'Code Red' in Calgary after responding to more than 200 incidents. *Calgary Herald* 17 March. Accessed November 27, 2019, `https://bit.ly/2Dmw7En`.

Schmid V (2012) Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. *European Journal of Operational Research* 219(3):611–621.

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

Sharma O (1990) *Markovian Queues* (New York: Ellis Horwood).

Taylor PG (2011) Insensitivity in stochastic models. Boucherie RJ, van Dijk NM, eds., *Queueing Networks: A Fundamental Approach*, 121–140 (Boston, MA: Springer US).

The College of Emergency Medicine (2014) Crowding in emergency departments. Accessed April 23, 2019, `https://bit.ly/2PsILXj`.

Veeraraghavan S, Scheller-Wolf A (2008) Now or later: A simple policy for effective dual sourcing in capacitated systems. *Operations Research* 56(4):850–864.

Viccellio P, Santora C (2012) Transferring admitted emergency department patients to hallway beds leads to lower length of stay and higher patient satisfaction. Accessed April 23, 2019, `https://bit.ly/2GCsUTg`.

Watase T, Fu R, Foster D, Langley D, Handel DA (2012) The impact of an ED-only full-capacity protocol. *The American Journal of Emergency Medicine* 30(8):1329–1335.

Weiss A, Williams L, Smith JM (2012) Performance & optimization of $M/G/c/c$ building evacuation networks. *Journal of Mathematical Modelling and Algorithms* 11(4):361–386.

Wolff RW (1989) *Stochastic Modeling and the Theory of Queues* (Prentice Hall).

Zayas-Caban G, Xie J, Green LV, Lewis ME (2019) Policies for physician allocation to triage and treatment in emergency departments. *IISE Transactions on Healthcare Systems Engineering* 9(4):342–356.

Zekman P (2014) 2 Investigators, BGA: Ambulance response times getting worse as memo aims to quiet dispatchers. *CBS Chicago* Accessed April 21, 2019, `https://cbsloc.al/2XCJGrd`.