An optimal dispatching model for server-to-customer systems with classification errors

Laura A. McLay
Department of Statistical Sciences & Operations Research
1015 Floyd Avenue, Box 843083
Richmond, Virginia 23284
Virginia Commonwealth University
lamclay@vcu.edu

Maria E. Mayorga
Freeman Hall Box 340920
Department of Industrial Engineering
Clemson University
Clemson, South Carolina 29631
mayorga@clemson.edu

July 15, 2010

Abstract

The decision of which servers to dispatch to which customers is an important aspect of server-to-customer systems. Such decisions are complicated when servers have different operating characteristics, customers are prioritized, and there are errors in assessing customer priorities. In this paper, we formulate a model for determining how to optimally dispatch distinguishable servers to prioritized customers given that dispatchers make classification errors in assessing the true customer priorities. These issues are examined through the lens of emergency medical service (EMS) dispatch, for which a Markov decision process model is developed that captures how to optimally dispatch ambulances (servers) to prioritized patients (customers). It is assumed that customers arrive sequentially, with the priority and location of each customer becoming known upon arrival, with classification errors in these customer priorities. The proposed model determines how to optimally dispatch heterogeneous servers to customers to maximize the long run average utility in a Markov decision process. The utilities and transition probabilities are location-dependent, with respect to both the server and customer locations. The analysis considers two cases for approaching the classification errors that correspond to over- and under-responding to perceived customer priority. A computational example is applied to an EMS system in order to determine the optimal policy for dispatching ambulances to patients in order to maximize patient survival.

1 Introduction

Determining how to optimally dispatch servers to customers is an important aspect of server-to-customer systems. Such decisions are complicated when servers have different operating characteristics, customers are prioritized, and there are errors in assessing customer priorities. Server-to-customer systems have a wide scope of applicability, as they are used to model many problems in the public domain. This paper examines optimal dispatching policies through the lens of emergency medical service (EMS) systems, in which we formulate a Markov decision process model that captures how to optimally dispatch ambulances (servers) to prioritized patients (customers).

Responding to emergency cardiac arrest 911 calls is a major focus of emergency medical service (EMS) systems. EMS systems are often evaluated by how effectively they respond to and perform care for cardiac arrest calls because: (1) effective treatment is available (i.e., CPR, early defibrillation); (2) treatment is highly time dependent (Wik et al.[40] states that the likelihood of survival decreases by approximately 10% for each minute delay in providing defibrillation after the onset of cardiac arrest); (3) each element in the community emergency response (early access, early CPR, early defibrillation, and early advanced life support) must be strong for the highest likelihood of survival to occur; and (4) if the EMS system can respond effectively to cardiac arrest and achieve a good survival rate, the same elements will usually yield similar outcomes for other life or death conditions such as major traumatic injury or stroke. This has been well-documented by medical research [35]. Furthermore, the EMS response time interval is highly correlated with cardiac arrest patient survival, whereas the link between the EMS response time interval and patient survival from other types of emergency medical conditions is unclear [11].

For decades, EMS systems have been measured according to how they respond to and care for cardiac arrest patients, with the link between EMS response and patient outcomes being well-understood [11, 8]. However, the survival to hospital discharge rate after out-of-hospital cardiac arrest is low, with approximately five to seven percent of cardiac arrest patients surviving [29]. One way to improve both EMS responses and patient outcomes is through improved resource allocation decisions. A number of operations research modeling efforts have examined how to optimally locate ambulances

in order to improve patient outcomes [10, 26, 28]. However, these papers address one resource allocation problem, namely, how to locate ambulances.

Emergency medical dispatch is another resource allocation problem that is a critical component in EMS systems. The dispatch center handles each 911 call, where a dispatcher determines the nature and priority of the call and dispatches the appropriate medical units. The nature of the call reflects the type of call (such as motor vehicle accident, trauma, or difficulty breathing). The priority assigned to the call reflects the operator's perception of whether the call is an emergency. Most frequently, calls are prioritized as Priority 1, 2, 3, where Priority 1 calls are life-threatening, Priority 2 calls are emergencies that may be life-threatening, and Priority 3 calls do not appear to be life-threatening. There are classification errors associated with the call types and priorities, which complicates the decision of which medical units to dispatch to a call. Determining the optimal way to dispatch medical units to calls is critical for improving patient outcomes.

This paper examines how to optimally dispatch distinguishable servers to prioritized customers in servers-to-customer systems given that dispatchers make classification errors in assessing the true customer priorities. We formulate the model as an infinite-horizon, undiscounted, average reward Markov decision process (MDP) which determines how to optimally dispatch servers to customers in order to maximize the long-run average customer utility. The model is applied to dispatching ambulances to patients in an EMS system. It is assumed that customers arrive sequentially, with the priority and location of each customer becoming known upon arrival, with classification errors in the customer priorities. The customer utilities and transition probabilities are location-dependent, with respect to both the server and customer locations. The analysis considers two cases for approaching the classification errors that correspond to over-responding and under-responding to perceived customer priorities. A computational study is conducted with a real-world example using data from an EMS system in order to determine the optimal policy for dispatching ambulances to patients in order to maximize patient survival. The results indicate that it is not always best to send the closest server, particularly for low-priority customers that are not likely to be life-threatening. Surprisingly, optimal policies do not always send the closest server to high-priority customers. The examples suggest that policies aimed at over-responding

to lower-priority patients are preferable when there is a higher rate of classification errors while policies aimed at under-responding to low-priority patients are preferable when there is a lower rate of classification errors.

This paper is organized as follows. Section 2 provides a literature review on dispatching models applied to EMS systems. The proposed Markov decision process model is formulated in Section 3. Uniformization, value iteration, and the impact of classification errors are also discussed. Section 4 reports the structural properties of the proposed model as well as the interpretation of the optimal policy using Markov chain limiting distributions. The results are applied to a scenario using real-world data collected from Hanover County, Virginia in Section 5. Concluding remarks, including policy implications as well as directions for future research are presented in Section 6.

2 Background

Optimizing dispatching protocols—in addition to locating medical units—potentially have large effects on patient outcomes. To maximize patient survival rates, it is necessary to understand when to dispatch the closest server to a customer and when to ration the closest server (dispatch a farther server instead) in anticipation of a more emergent call. However, little attention has been given to identifying optimal dispatching policies in EMS systems. Carter et al. [4] use a queuing optimization model to determine the locations of two EMS response areas (i.e., beats) to balance the workload between two servers. Each server is assumed to always respond to customers within its response area, if the server is available. If both servers are busy, customers are served by servers outside of the area. They show that it is not always best to dispatch the closest server. Jarvis [16] introduces a Markov decision process model for the Hypercube queuing model for determining optimal dispatching policies for a single type of server, and it is illustrated with an example that minimizes the average distance traveled when responding to a customer. Weintraub et al. [39] develop a dispatching model for an electric utility with prioritized customers, although the priorities are not analogous to EMS priorities. Other research examines the complementary issue of how many servers to dispatch to a customer. Chelst and Barlach [6], Swersey [36], and Ignall et al. [15] provide models that determine how many servers—ambulances or fire engines—to

dispatch to a customer.

A number of papers examine how to determine or approximate server busy probabilities for server-to-customer systems given a known dispatching policy. The Hypercube model is the most well-known of these models, which analyzes vehicle location and response district design for urban environments [19, 20, 22]. The Hypercube model assumes the underlying server-to-customer system dynamics are that of a multi-server queuing system with indistinguishable servers. The Hypercube model has been extended several times [7, 6, 21, 17, 3]. Estimating busy probabilities is important for determining how to locate servers using facility location models, for example.

Henderson [14, 13] proposes using System Status Management (SSM) and approximate dynamic programming for relocating ambulances. The dispatching decisions considered in this paper differ from those in SSM in that our approach assumes that servers return to their home station when not servicing customers (static locations) whereas SSM relocates servers (dynamic locations) based on real-time information and forecasted demand. These relocation models are different than the ideas considered in this paper in that ambulances do not typically have a home station that they respond to between serving customers. Rather, they focus on where to dynamically locate ambulances after every customer arrives [1, 2, 12, 25]. While deployment plans would theoretically change hourly with incoming customers and while resources could be constantly repositioned to meet demand as efficiently as possible, such an approach is costly and sometimes impractical as repositioning ambulances takes time and often ignores crew fatigue. In contrast, this paper considers a set of ambulances at a fixed set of home stations (without relocating) and finds the optimal way to dispatch ambulances to customers. Thus, our approach adds to the literature by examining how to dispatch a heterogenous set of servers according to customer priority. The decision context explored in this paper complements those considered in the literature, such as optimally locating servers using facility location models or dynamically relocating servers using System Status Management (SSM). This paper lifts these assumptions by providing a methodology for determining the optimal dispatching policy under more realistic assumptions.

Other papers from the medical literature illustrate the importance of taking dispatcher classification errors into account when developing dispatching policies. Dunford [9] provides a review of the role of EMS dispatch centers. He indicates that each EMS

system must determine dispatching standards that reflect acceptable classification errors in assessing the nature and priority of customers. Roppolo et al. [31] indicate that reducing the Type II errors associated with recognizing cardiac arrest patients increases their chance of survival. This paper seeks to identify actionable policies that optimally dispatch servers to customers despite imperfect classification.

Much of the previous work in this area has determined how to relocate ambulances when servicing unprioritized patients. In contrast to previous work in the area, this paper investigates how to dynamically dispatch ambulances to prioritized patients, where there are classification errors in assessing the true severity of the patient.

3 Markov decision process model

This section presents the Markov decision process model (MDP) for determining optimal dispatching policies in a server-to-customer system that assigns distinguishable servers to prioritized customers. The servers are ambulances that are differentiated by their response and service times. The customers are arriving calls for service, who have an associated location and are categorized into risk groups, high (H) or low (L), depending on their priority classification (this is discussed in detail in Section 3.2). The objective is to optimally determine which server to dispatch to arriving customers in order to maximize the average utility per stage. The utility is interpreted as the conditional probability of survival for cardiac arrest patients, who represent a given proportion of the customers. The optimal policy is compared to the myopic policy of always sending the closest server, a greedy approach.

In the model, customers arrive according to a Poisson process with rate λ. As soon as a customer arrives, its location and risk are evaluated, and one of the available servers is dispatched to it. The model makes several assumptions:

- One server is assigned to each customer.
- Service times do not depend on customer priority.
- If a customer arrives and a server is available, a server must be dispatched.
- There is a zero-length queue for customers.

We note that the model could be trivially modified to lift the first three assumptions. Requiring a server to be dispatched to a customer if a server is available is reasonable,

since the model here is motivated by EMS systems, where providing service to all customers is a major goal of public service systems. The zero-length queue assumption requires further discussion. We point out that without this assumption, the state-space of the model quickly becomes unmanageable as the dispatching policy could depend on the number of customers in each queue. However, we believe this assumption is practical for two reasons. First, the objective is to maximize the average utility per stage, where the utility is given by the survival probability of cardiac arrest patients. This probability quickly drops as a function of time and it is essentially zero for response times greater than 11 minutes [37, 34]. Since a customer joining a queue would have to wait for an ambulance to become available before a server is dispatched to it, these customers contribute little or no value to the objective function. Second, while incorporating a non-zero queue length would increase the busy probabilities of the servers, we have conducted extensive computational experiments using simulation and find that the zero-length queue assumption does not significantly impact the long-run average utilities or the optimal dispatching policy using real-world data, since in EMS systems it is highly unlikely that an arriving customer will find all ambulances busy.

Next, the input parameters are summarized. Then, an infinite-horizon, undiscounted, average reward MDP model is defined.

Input parameters:

n = total number of customer locations.

m = the number of servers, each at a fixed location.

h = set of customer risk groups, with $h \in \{H, L\}$, where H denotes high-risk and L denotes low-risk.

λ = expected number of customers that arrive per unit of time (Poisson parameter).

P_i = the conditional probability that a customer arrives at customer location i, given that a customer arrives, $i = 1, 2, ..., n$.

$P_{h|i}$ = the conditional probability that a customer has risk $h \in \{H, L\}$ given that a customer has arrived at location i, $i = 1, 2, ..., n$.

μ_{ij} = the expected service time when server j is dispatched to a customer at location i (distributed exponentially), $i = 1, 2, ..., n,\ j = 1, 2, ..., m$.

u_{ij}^h = the expected utility when server j is dispatched to a customer with risk $h \in \{H, L\}$ at location i, $i = 1, 2, ..., n,\ j = 1, 2, ..., m$.

Note that the set of risk groups can be generalized to consider a wide range of customer classifications without any modification to the model. Based on interviews in the field, we present the model for two risk groups to reflect the simplifications used in many dispatch centers in practice. The discussion of classification errors in Section 3.2 is framed with respect to having two risk categories. The customer parameters λ, P_i, and $P_{h|i}$ reflect call volume and geographic differences in the volume and prioritization of customers. These parameters can be estimated using computer-aided dispatch (CAD) data. The service times, μ_{ij}, capture the time that a server drives to a customer, treats the customer, and returns to service. The utilities capture the performance measure used by the system. In EMS systems, performance measures most frequently capture either the probability of survival or the proportion of customers that are responded to within a fixed timeframe [28, 27], and they are defined generally here. The link between the utilities and classification errors is made explicit in Section 3.2.

Time. An undiscounted, infinite horizon is assumed to study the optimal dispatching policy in steady-state. Although this assumption is not realistic, it is useful for providing insight into dispatching policies that may be superior to always sending the closest server during the peak hours of operation, where the customer arrival rate is essentially constant over several hours.

States. We define the state of the system $\mathbf{s}(t)$ at time t where $\mathbf{s}(t) = (s_1(t), s_2(t), \ldots, s_m(t))$ as

$$s_j(t) = \begin{cases} 0, & \text{if server } j \text{ is free at time } t; \\ i, & \text{if server } j \text{ is busy at time } t \text{ serving a customer originating as location } i. \end{cases}$$

Let the state space of the corresponding Markov chain at time t be defined by $\mathcal{S} = \{\mathbf{s}(t) : \mathbf{s}(t) \in \{0, 1, \ldots, n\}^m\}$; therefore, the total number of states is $|\mathcal{S}| = (n+1)^m$. Thus, we see that the state of the system describes the combinations of busy and free servers at any point in time where each server can either be free (state 0) or busy treating a customer in location i, $i = 1, 2, ..., n$).

Actions. Let $X(t) = X(\mathbf{s}(t))$ denote the set of available actions in state $\mathbf{s}$ at time t if a customer arrives. If a customer arrives, one of the available servers is dispatched

to the customer. A zero length queue is assumed such that if no servers are available, then the customer is lost. If a customer does not arrive, then no servers are dispatched. Therefore, there are at most m actions available in each state.

Rewards. The rewards reflect the utility of the action selected—patient survival, in this case—which depends on the location and risk group of the customer as well as the server that responds. The interpretation of the rewards are given in Section 3.2.

Transition Probabilities. Two elements are needed to determine the transition probabilities: (1) one of the busy servers may have completed service and is free, and (2) a customer arrives, which requires that a server is dispatched to the customer if one is available.

The next section discusses how the model is formulated as an equivalent discrete time MDP that can be solved using value iteration.

3.1 Uniformization formulation and value iteration algorithm

This section describes how uniformization is used to determine the optimal policy in an equivalent discrete time MDP. The goal of the value functions is to maximize the system-wide utilities associated with responding to emergency calls. Since the state transitions are Markovian, it suffices to consider policies where decisions are made at discrete time epochs. Thus, instead of a continuous time control problem, uniformization is applied (see [24]) and used to solve an equivalent discrete time problem. To do so, the maximum rate of transitions is determined as

$$\gamma = \lambda + \sum_{j=1}^{m} \beta_j$$

where

$$\beta_j = \max_{i=1,2,\ldots,n} \{(\mu_{ij})^{-1}\}$$

Without loss of generality, we scale $\gamma = 1$.

The optimal policy yields the optimal average utility per stage g, where the length of the stage is determined by γ. When uniformization is used, the state is denoted as $\mathbf{s}$ (and not $\mathbf{s}(t)$), since when a continuous time problem is converted to an equivalent discrete time problem, decisions are made at discrete time epochs. A value iteration algorithm is used to find the optimal policy, where the iterations in the algorithm are interpreted as stages in a corresponding finite-horizon MDP. Define $v_k(\mathbf{s})$ as the value

of being in state $\mathbf{s} = (s_1, s_2, ..., s_m)$ during iteration k, and let $v_0(\mathbf{s}) = 0$ for all $\mathbf{s}$. Then, the optimality equations for the N-stage and average cost cases follow, respectively.

$$\begin{aligned} v_{k+1}(\mathbf{s}) &= \frac{1}{\gamma}\left[\sum_{j=1}^{m} I_{\{s_j=i|i>0\}}\frac{1}{\mu_{ij}}v_k(s_1, s_2, ..., s_{j-1}, 0, s_{j+1}, ..., s_m)\right. \\ &+ \sum_{i=1}^{n}\sum_{h\in\{H,L\}}\lambda P_i P_{h|i}\max_{j\in X(\mathbf{s})}\{I_{\{s_j=0\}}v_k(s_1, s_2, ..., s_{j-1}, i, s_{j+1}, ..., s_m) + \gamma u_{ij}^h\} \\ &+ \left.\left(\gamma - \lambda - \sum_{j=1}^{m} I_{\{s_j=i|i>0\}}\frac{1}{\mu_{ij}}\right)v_k(\mathbf{s})\right], \qquad (1) \\ &\text{for } k = 0, \ldots, N-1, \text{ and } v_0(\mathbf{s}) = 0 \text{ for all } \mathbf{s} \in \mathcal{S} \end{aligned}$$

$$\begin{aligned} g + w(\mathbf{s}) &= \frac{1}{\gamma}\left[\sum_{j=1}^{m} I_{\{s_j=i|i>0\}}\frac{1}{\mu_{ij}}w(s_1, s_2, ..., s_{j-1}, 0, s_{j+1}, ..., s_m)\right. \\ &+ \sum_{i=1}^{n}\sum_{h\in\{H,L\}}\lambda P_i P_{h|i}\max_{j\in X(\mathbf{s})}\{I_{\{s_j=0\}}w(s_1, s_2, ..., s_{j-1}, i, s_{j+1}, ..., s_m) + \gamma u_{ij}^h\} \\ &+ \left.\left(\gamma - \lambda - \sum_{j=1}^{m} I_{\{s_j=i|i>0\}}\frac{1}{\mu_{ij}}\right)w(\mathbf{s})\right], \qquad (2) \end{aligned}$$

where g is the optimal average utility per stage and $w(\mathbf{s})$ is a relative value function in state $\mathbf{s}$. Furthermore, $I_{\{s_j=i|i>0\}}$ is an indicator variable that indicates if server j is servicing a customer at location i, and $I_{\{s_j=0\}}$ is an indicator variable that indicates if server j is available. Both (1) and (2) are analogous. The first line in (1) captures busy servers becoming available. The second line captures new customers arriving, where $X(\mathbf{s})$ denotes the set of servers available. It is assumed that when there are no available servers in a state and a customer arrives, no server is dispatched and a reward of zero is received. The third line captures neither new customers arriving nor servers becoming available. Note that (2) captures the exact, infinite horizon, average cost optimality equations, whereas (1) captures the optimality equations for the finite horizon case, the latter of which can exploited to analyze the MDP structural properties.

The optimality equations converge to a steady-state average utility per stage g. To achieve, the optimal policy, the relative value iteration algorithm (see [30]) is run until the upper and lower bounds converge to g as follows:

$$L_k \leq L_{k+1} \leq g \leq U_{k+1} \leq U_k$$

with lower bound $L_k = \min_{\mathbf{s}\in S}[v_{k+1}(\mathbf{s}) - v_k(\mathbf{s})]$ and upper bound $U_k = \max_{\mathbf{s}\in S}[v_{k+1}(\mathbf{s}) - v_k(\mathbf{s})]$. Value iteration is executed until $U_{k+1} - L_{k+1} \leq \varepsilon$ for a given tolerance ε.

3.2 Interpretation of rewards with classification errors

The value iteration algorithm for finding the optimal policy yields the optimal average utility per stage g. Without loss of generality, the utility is interpreted as the average number of lives saved per stage. When considering the customer arrival rate λ, the average utility per stage can be reformulated to capture the average conditional probability of survival for cardiac arrest patients.

To define the probability of survival, consider two classification schemes in a server-to-customer system. Customers are classified into K priority groups based on their perceived acuity. Each priority type is then mapped into a risk group, e.g. high-risk (H) or low-risk (L). We consider $K = 3$ priorities to be consistent with the majority of EMS systems throughout the country, where Priority 1 is anticipated to be life-threatening, Priority 2 may be life-threatening, and Priority 3 is anticipated not to be life-threatening. However, the approach can trivially be modified to consider an arbitrary number of customer priorities. The dispatcher assigns a priority to incoming customers, while the risk categories are predetermined by the EMS system. Since EMS systems are evaluated according to how they respond to cardiac arrest customers, we interpret the life-threatening customers as cardiac arrest customers. Although other customers are life-threatening, the EMS response is not as instrumental in patient survival (such as cancer or stroke) [11]. Consider the following events:

- LT = a customer is life-threatening
- $Pr1$ = a customer is Priority 1
- $Pr2$ = a customer is Priority 2
- $Pr3$ = a customer is Priority 3

We assume that the proportion of customers who are life-threatening is smaller than the proportion of customers who are classified as Priority 1 ($P_{LT} \leq P_{Pr1}$) and that no Priority 3 patients are life-threatening ($P_{LT|Pr3} = 0$). Let α denote the ratio of the proportion of Priority 1 customers that are life-threatening to the proportion of Priority

Table 1: Classification errors in assessing patient priorities

	Case 1 $H = \{Pr1\},\ L = \{Pr2, Pr3\}$	**Case 2** $H = \{Pr1, Pr2\},\ L = \{Pr3\}$
$P_{LT\|H\cap i}$	$\frac{P_{LT}}{P_{Pr1\cap i}+(1/\alpha)P_{Pr2\cap i}}$	$\frac{P_{LT\|Pr1}P_{Pr1\cap i}+P_{LT\|Pr2}P_{Pr2\cap i}}{P_{Pr1\cap i}+P_{Pr2\cap i}}$
$P_{LT\|L\cap i}$	$\frac{P_{LT}P_{Pr2\cap i}}{(\alpha P_{Pr1}+P_{Pr2})(P_{Pr2\cap i}+P_{Pr3\cap i})}$	0

2 customers that are life-threatening,

$$\alpha = P_{LT|Pr1}/P_{LT|Pr2},$$

with $\alpha \geq 1$. Two cases are used for mapping Priority 1, 2, 3 customers into risk groups:
Case 1: Priority 1 customers are classified as high-risk and Priority 2 and 3 customers are classified as low-risk (Under-responding).
Case 2: Priority 1 and 2 customers are classified as high-risk and Priority 3 customers are classified as low-risk (Over-responding).

Cases 1 and 2 result in differing conditional probabilities that a customer is life-threatening, given that the customer is high-risk. Table 1 reports the conditional probabilities that a customer is life-threatening given its location and risk category, $P_{LT|H\cap i}$ and $P_{LT|L\cap i}$. The values in Table 1 are obtained by considering that $P_{LT|Pr2} = P_{LT}/(\alpha P_{Pr1} + P_{Pr2})$ and $P_{LT|Pr1} = \alpha P_{LT|Pr2}$ using Bayes rule. Note that Case 2 results in constant values of $P_{LT|h\cap i}$ across all values of α, $h \in \{H, L\}$, $i = 1, 2, ..., n$, since it is assumed that no Priority 3 patients are life-threatening.

The values in Table 1 are used to compute the utilities as follows. It is assumed that the probability of survival, S_{ij} when ambulance j is dispatched to a life-threatening customer at location i, with $0 \leq S_{ij} \leq 1$. Since the utilities can be defined as the probability that a life is saved, given that a customer of risk $h \in \{H, L\}$ at location $i = 1, 2, ..., n$, the utilities are defined as

$$u_{ij}^{h} = P_{\text{Survive}|LT\cap h\cap i\cap j} P_{LT|h\cap i\cap j} = S_{ij} P_{LT|h\cap i}. \tag{3}$$

The final statement is obtained by noting that survival is independent of high- or low-risk classification and that whether a customer is life-threatening is independent of which server is dispatched.

Note that the quality of the classifications is determined by α, which is implicit in (3) when considering Table 1. These utilities can be used to compare Case 1 policies to

Case 2 policies given various levels of α. In addition, both Case 1 and Case 2 policies can be compared to other policies, such as a policy that always sends the closest server. This will be explored in detail through two examples in Section 5.

4 Structural properties and Markov Chain Interpretation

This section summarizes the structural properties of the proposed model as well as interprets the MDP policies in terms of their resulting Markov chain limiting distributions.

First, note that the MDP model has a finite number of states and that the rewards are bounded. Second, note that the chain is positive recurrent, since any server can be dispatched to any type of customer if all other servers are busy. Therefore, the value functions exist and their solution yields an optimal policy.

Recall that the state is defined as $\mathbf{s} = (s_1, s_2, ..., s_m)$. We denote $\mathbf{e}_j$ as a vector of zeros, except for a value of 1 in the j^{th} location, e.g. $\mathbf{e}_1 = (1, 0, \ldots, 0)$.

There are two main results in this section. Proposition 1 indicates that the average utility per stage when in a state with a given server being free is at least the average utility per stage in the corresponding state with a server dispatched to a customer. In other words, it is more beneficial to have a server idle than busy.

Proposition 1. *Suppose* $\mathbf{s}$ *is such that* $s_{j*} = 0$ *for some* $j* = 1, 2, ..., m$. *Then*

$$v_k(\mathbf{s}) - v_k(\mathbf{s} + i^*\mathbf{e}_{j*}) \geq 0, i^* = 1, 2, ..., n \text{ for all } k \geq 0. \tag{4}$$

Proof. The result is shown by induction using value iteration. Note that $v_0(\mathbf{s}) = 0$ for all $\mathbf{s}$, and thus, the claim is trivial for $k = 0$. Suppose that (4) holds for k. We will show that the inequality is preserved by the optimality equation. Thus consider the optimality equation at iteration $k+1$ for any $\mathbf{s}$ such that $s_{j*} = 0$, for some $j* = 1, 2, \ldots, m$. Define an alternate state $\hat{\mathbf{s}} = \mathbf{s} + i^*\mathbf{e}_{j*}$, $i^* = 1, 2, ..., n$. From (1) we obtain.

$$
\begin{aligned}
&\gamma(v_{k+1}(\mathbf{s}) - v_{k+1}(\hat{\mathbf{s}})) = \\
&\sum_{j=1}^{m} I_{\{s_j=i|i>0\}} \frac{1}{\mu_{ij}} v_k(\mathbf{s} - s_j\mathbf{e}_j) - \sum_{j=1}^{m} I_{\{\hat{s}_j=i|i>0\}} \frac{1}{\mu_{ij}} v_k(\hat{\mathbf{s}} - s_j\mathbf{e}_j) \\
+ \quad &\sum_{i=1}^{n} \sum_{h\in\{H,L\}} \lambda P_i P_{h|i} \left(\max_{j\in X(\mathbf{s})} \{I_{\{s_j=0\}} v_k(\mathbf{s} + i\mathbf{e}_j) + \gamma u_{ij}^h\} - \max_{j\in X(\hat{\mathbf{s}})} \{I_{\{s_j=0\}} v_k(\hat{\mathbf{s}} + i\mathbf{e}_j) + \gamma u_{ij}^h\} \right) \\
+ \quad &\left(\gamma - \lambda - \sum_{j=1}^{m} I_{\{s_j=i|i>0\}} \frac{1}{\mu_{ij}} \right) v_k(\mathbf{s}) - \left(\gamma - \lambda - \sum_{j=1}^{m} I_{\{\hat{s}_j=i|i>0\}} \frac{1}{\mu_{ij}} \right) v_k(\hat{\mathbf{s}})
\end{aligned}
$$

Since $\hat{\mathbf{s}} = \mathbf{s} + i^*\mathbf{e}_{j*}$, and since the set of available actions in the second maximization statement is a subset of the set of the available actions in the first maximization statement (i.e., $X(\hat{\mathbf{s}}) = X(\mathbf{s}) \cup j*$), we can rewrite this as

$$
\begin{aligned}
&\gamma(v_{k+1}(\mathbf{s}) - v_{k+1}(\hat{\mathbf{s}})) = \\
&\sum_{j=1}^{m} I_{\{s_j=i|i>0\}} \frac{1}{\mu_{ij}} (v_k(\mathbf{s} - s_j\mathbf{e}_j) - v_k(\mathbf{s} + i^*\mathbf{e}_{j*} - s_j\mathbf{e}_j)) - \frac{1}{\mu_{i^*j*}} v_k(\mathbf{s} + i^*\mathbf{e}_{j*} - i*\mathbf{e}_{j*}) \quad (5) \\
+ \quad &\sum_{i=1}^{n} \sum_{h\in\{H,L\}} \lambda P_i P_{h|i} \left[\max \left(\max_{j\in X(\hat{\mathbf{s}})} \{I_{\{\hat{s}_j=0\}} v_k(\mathbf{s} + i\mathbf{e}_j) + \gamma u_{ij}^h\}, v_k(\mathbf{s} + i\mathbf{e}_{j*}) + \gamma u_{ij*}^h \right) \right] \quad (6) \\
- \quad &\sum_{i=1}^{n} \sum_{h\in\{H,L\}} \lambda P_i P_{h|i} \left(\max_{j\in X(\hat{\mathbf{s}})} \{I_{\{\hat{s}_j=0\}} v_k(\hat{\mathbf{s}} + i\mathbf{e}_j) + \gamma u_{ij}^h\} \right) \quad (7) \\
+ \quad &\left(\gamma - \lambda - \sum_{j=1}^{m} I_{\{s_j=i|i>0\}} \frac{1}{\mu_{ij}} \right) (v_k(\mathbf{s}) - v_k(\mathbf{s} + i^*\mathbf{e}_{j*}) + \frac{1}{\mu_{i^*j*}} v_k(\mathbf{s} + i^*\mathbf{e}_{j*} - i^*\mathbf{e}_{j*}) \quad (8)
\end{aligned}
$$

The last terms in (5) and (8) cancel while each of the first terms in (5) and (8) are non-negative from the induction hypothesis. (Note that $\gamma - \lambda - I_{\{s_j=i|i>0\}} \frac{1}{\mu_{ij}} \geq 0$ according to the definition of γ). Thus, to prove the result it suffices to show that for all i, h.

$$
\max \left(\max_{j\in X(\hat{\mathbf{s}})} \{I_{\{\hat{s}_j=0\}} v_k(\mathbf{s} + i\mathbf{e}_j) + \gamma u_{ij}^h\}, v_k(\mathbf{s} + i\mathbf{e}_{j*}) + \gamma u_{ij*}^h \right) \geq \left(\max_{j\in X(\hat{\mathbf{s}})} \{I_{\{\hat{s}_j=0\}} v_k(\hat{\mathbf{s}} + i\mathbf{e}_j) + \gamma u_{ij}^h\} \right)
$$

From the induction hypothesis $v_k(\mathbf{s} + i\mathbf{e}_j) \geq v_k(\hat{\mathbf{s}} + i\mathbf{e}_j) = v_k(\mathbf{s} + i^*\mathbf{e}_{j*} + i\mathbf{e}_j)$, for any i, and the result follows and v_k satisfies (4) for all $k \geq 0$. □

The second result makes the following two assumptions about the relationship between distances and the input parameters. Distance is a key component in the second result as well as in the examples considered in Section 5. However, distance is not explicitly captured in the parameters in Section 3. In order to relate distance to these parameters implicitly, we make the following assumption.

Assumption 1: For a fixed server j, location i_1 is closer to server j than location i_2 if $u^H_{i_1,j} \geq u^H_{i_2,j}$ and $u^L_{i_1,j} \geq u^L_{i_2,j}$.

Assumption 1 suggests that the parameters are structured in a way that the utilities for a particular server's response are non-decreasing in its distances to the customer locations. This reflects the medical literature that suggests that cardiac arrest survival probabilities are a function of the response times (a proxy for distance) [18, 37, 38]. Although Assumption 1 is not explicitly used by Propositions in this section, it helps us to understand the structural properties by formalizing the concept of distance.

Assumption 2: A fixed server j is closer to location i_1 than to location i_2 if $\mu_{i_1,j} < \mu_{i_2,j}$.

We note that Assumption 2 is more restrictive than Assumption 1, as it claims that the utilities are non-increasing in customer distance from the fixed server and that the service times are non-decreasing in customer distance from the fixed server. While the response times—one component of the service time—are likely non-decreasing with distance, the entire service time captures times that are not likely to be a function of the distance between the customer location and server, such as the travel time to the closest hospital. When Assumption 2 holds, Proposition 2 holds, which indicates that the average utilities obtained per stage are non-increasing in customer distance from a fixed server. In other words, it is more beneficial for a server to be busy serving a customer that is closer to his fixed location than one that is farther away. The same notation is used as in Proposition 1. Without loss of generality, we assume that we can order the locations relative to any server $j*$ such that $\mu_{[1],j*} \leq \mu_{[2],j*} \leq ... \leq \mu_{[n],j*}$, $j* = 1, 2, ..., m$, with $u^h_{[1],j*} \geq u^h_{[2],j*} \geq ... \geq u^h_{[n],j*}$, $h \in \{H, L\}$, $j* = 1, 2, ..., m$.

Proposition 2. *Suppose* $\mathbf{s}$ *is such that* $s_{j*} = [i^*]$ *for some* $j* = 1, 2, ..., m$. *Then*

$$v_k(\mathbf{s}) - v_k(\mathbf{s} - [i]^*\mathbf{e}_{j*} + [i^* + 1]\mathbf{e}_{j*}) \geq 0, [i^*] = 1, 2, ..., n-1 \text{ for all } k \geq 0. \quad (9)$$

The proof is similar to that shown in Proposition 1, where we make use of Assumption 2 by noting that $\frac{1}{\mu_{[i^*],j*}} > \frac{1}{\mu_{[i^*+1],j*}}$. Thus, we omit it here for brevity.

We note that the corresponding claim from the perspective of each fixed location is not true. That is, the value functions are not non-increasing with the server distance given a customer at a fixed location. Therefore, it is not always optimal to dispatch closer servers to to a customer, when such a choice is available. Counter-examples are shown in Section 5.

Since this paper considers an infinite horizon, undiscounted MDP, each policy can be examined in terms of the Markov chain that it induces, whose transition probability matrix is denoted by P^π. The Markov chain transition probability matrices for arbitrary instances are not explicitly stated here, since the formulation of the value functions in (1) indicate that the Markov chain transition probability matrices cannot be succinctly summarized. Note that (1) suggests that the Markov chain is aperiodic and ergodic, and thus, the limiting distribution exists.

The average utility per stage can be computed by considering the limiting distribution of the Markov chain. Since an optimal policy π^* exists, we can compare policies π in terms of the limiting distribution. Let ψ^π denote the limiting distribution of the Markov chain induced by policy π, with the probability of being in state $\mathbf{s}$ denoted by $\psi(\mathbf{s})$. Then the average utility per stage associated with policy π is

$$U^\pi = \sum_{\mathbf{s}\in\mathcal{S}} \psi(\mathbf{s}) \sum_{i=1}^{n} \sum_{h\in\{H,L\}} \lambda P_i P_{h|i} u_{ij}^h x_{ij}^{h,\pi}(\mathbf{s}),$$

where $x_{ij}^{h,\pi}(\mathbf{s})$ is 1 if server j is dispatched to a customer of priority h at location i when the system is in state $\mathbf{s}$, and 0 otherwise. Let Π denote the set of all possible policies. Then the optimal policy π^* is

$$\pi^* = \arg\max_{\pi\in\Pi} U^\pi. \tag{10}$$

Consider the particular case when $n = m = 2$. In this case, there is only one state that allows multiple actions when customers arrive (corresponding to both servers being available). All policies are identical and deterministic except for when both servers are available (i.e., state $(0,0)$). Therefore, define $x_{ij}^{h,\pi} \equiv x_{ij}^{h,\pi}((0,0))$. The average utility per stage associated with policy π is

$$\begin{aligned} U^\pi \;=\; & \sum_{h\in\{H,L\}} \Big[\lambda P_1 P_{h|1} u_{11}^h \Big(\psi((0,0)) x_{11}^{h,\pi} + \psi((0,1)) + \psi((0,2))\Big) \\ & + \lambda P_1 P_{h|1} u_{12}^h \Big(\psi((0,0))(1 - x_{11}^{h,\pi}) + \psi((1,0)) + \psi((2,0))\Big) \\ & \lambda P_2 P_{h|2} u_{21}^h \Big(\psi((0,0))(1 - x_{22}^{h,\pi}) + \psi((0,1)) + \psi((0,2))\Big) \\ & + \lambda P_2 P_{h|2} u_{22}^h \Big(\psi((0,0)) x_{22}^{h,\pi} + \psi((1,0)) + \psi((2,0))\Big)\Big] \end{aligned} \tag{11}$$

Note that setting the level of four variables—$x_{11}^{H,\pi}, x_{11}^{L,\pi}, x_{22}^{H,\pi}, x_{22}^{L,\pi}$—determines the policy π. Therefore, there are sixteen deterministic policies that are obtained by setting

the variables to zero or one. The transition probability matrix for all such policies is given by:

$$P^{\pi} = \tag{12}$$

$$\begin{array}{c} \\ (0,0) \\ (0,1) \\ (0,2) \\ (1,0) \\ (1,1) \\ (1,2) \\ (2,0) \\ (2,1) \\ (2,2) \end{array}
\begin{array}{c}
\begin{array}{ccccccccc} (0,0) & (0,1) & (0,2) & (1,0) & (1,1) & (1,2) & (2,0) & (2,1) & (2,2) \end{array} \\
\begin{pmatrix}
\gamma-\lambda & \lambda P_1(P_{H|1}(1-x_{11}^{H,\pi}) & \lambda P_2(P_{H|2}x_{22}^{H,\pi} & \lambda P_1(P_{H|1}x_{11}^{H,\pi} & 0 & 0 & \lambda P_2(P_{H|2}(1-x_{22}^{H,\pi}) & 0 & 0 \\
 & +P_{L|1}(1-x_{11}^{L,\pi})) & +P_{L|2}x_{22}^{L,\pi}) & +P_{L|1}x_{11}^{L,\pi}) & & & +P_{L|2}(1-x_{22}^{L,\pi})) & & \\
\mu_{12}^{-1} & \gamma-\lambda-\mu_{12}^{-1} & 0 & 0 & \lambda P_1 & 0 & 0 & \lambda P_2 & 0 \\
\mu_{22}^{-1} & 0 & \gamma-\lambda-\mu_{22}^{-1} & 0 & 0 & \lambda P_1 & 0 & 0 & \lambda P_2 \\
\mu_{11}^{-1} & 0 & 0 & \gamma-\lambda-\mu_{11}^{-1} & \lambda P_1 & \lambda P_2 & 0 & 0 & 0 \\
0 & \mu_{11}^{-1} & 0 & \mu_{12}^{-1} & \gamma-\mu_{11}^{-1}-\mu_{12}^{-1} & 0 & 0 & 0 & 0 \\
0 & 0 & \mu_{11}^{-1} & \mu_{22}^{-1} & 0 & \gamma-\mu_{11}^{-1}-\mu_{22}^{-1} & 0 & 0 & 0 \\
\mu_{12}^{-1} & 0 & 0 & 0 & 0 & 0 & \gamma-\lambda-\mu_{21}^{-1} & \lambda P_1 & \lambda P_2 \\
0 & \mu_{21}^{-1} & 0 & 0 & 0 & 0 & \mu_{12}^{-1} & \gamma-\mu_{21}^{-1}-\mu_{12}^{-1} & 0 \\
0 & 0 & \mu_{21}^{-1} & 0 & 0 & 0 & \mu_{22}^{-1} & 0 & \gamma-\mu_{21}^{-1}-\mu_{22}^{-1}
\end{pmatrix}
\end{array}
\quad \left(\frac{1}{\gamma}\right)$$

We note that the Markov chain interpretation is not practical for identifying optimal policies, due to the large number of policies that must be compared in (10). However, the Markov chain interpretation provides a novel way for comparing policies, and it will be illustrated in the example in Section 5.1.

5 Computational example

This section analyzes the MDP model and its optimal policies for two examples. First, an illustrative example with two locations and two servers is discussed to provide intuition behind the optimal policy. Then, an example with four locations and four servers is discussed with real-world parameters to provide insight into how an optimal dispatching policy operates. Both examples provided in this section consider a set of locations with exactly one server stationed at each location, which correspond to response districts surrounding each rescue station (server). Both examples use data extracted from Hanover County Fire and EMS in Virginia. Hanover County is a semi-rural, semi-suburban county in the metropolitan Richmond area. A data set using one year of computer aided dispatch (CAD) data was provided to assist in the computational portion of this paper.

5.1 Two location example

A hypothetical example was constructed using data similar to two of the regions in Hanover County. This example illustrates the optimal policies over a simplified decision structure, which provide insight into more complex decision structures. This example considers two locations, each of which has one server located at its station. Customers

arrive according to a Poisson process with rate $\lambda = 0.5$ customers per hour. Half of all customers are Priority 1 irrespective of their locations, resulting in $P_{Pr1|1} = P_{Pr1|2} = P_{Pr1} = 0.5$. The remaining customers are equally likely to be Priority 2 and 3, irrespective of their locations, resulting in $P_{Pr2|1} = P_{Pr2|2} = P_{Pr2} = 0.25$ and $P_{Pr3|1} = P_{Pr3|2} = P_{Pr3} = 0.25$. Recall that Case 1 and Case 2 refer to the classification schemes that assign priority groups to risk groups, as detailed in section 3.2. Case 1 results in $P_{H|1} = P_{H|2} = P_{L|1} = P_{L|2} = 0.5$. Case 2 results in $P_{H|1} = P_{H|2} = 0.75$ and $P_{L|1} = P_{L|2} = 0.25$. The average service times are 60 and 65 minutes at locations 1 and 2, respectively, when a server responds to customers at its locations (an in-district response), resulting in $\mu_{11} = 1$ and $\mu_{22} = 65/60$. Recall that Case 1 treats only Priority 1 customers as high-risk, underresponding to some life-threatening Priority 2 customers that have been classified as low-risk. Case 2 treats both Priority 1 and 2 customers as high-risk, reducing Type II errors by overresponding to many customers that are not life-threatening. When a server responds from the other location (and out-of-district response), the average service time is 75 minutes, resulting in $\mu_{12} = \mu_{21} = 75/60$.

The survival probabilities for life-threatening customers are 0.15 and 0.1 at locations 1 and 2, respectively, when a server responds to customers at its locations (an in-district response), resulting in $S_{11} = 0.15$ and $S_{22} = 0.1$. When a server responds from the other location (and out-of-district response), the survival rates are $S_{12} = S_{21} = 0.05$. The survival rates for non-life threatening customers are zero. Note that α determines the utilities associated with responding to high- and low-risk customers. First consider the scenario with perfect classification (i.e. no errors), resulting in $\alpha = \infty$ and $P_{LT} = P_H$. This indicates that all high-risk patients are life-threatening and all low-risk patients are not life-threatening. For Case 1, this results in $u^H_{11} = 0.15, u^H_{22} = 0.1, u^H_{12} = u^H_{21} = 0.05$ and $u^L_{11} = u^L_{22} = u^L_{12} = u^L_{21} = 0$ (using (3) and Table 1). For Case 2, this results in $u^H_{11} = 0.10, u^H_{22} = 0.0667, u^H_{12} = u^H_{21} = 0.0333$ and $u^L_{11} = u^L_{22} = u^L_{12} = u^L_{21} = 0$.

In order to investigate the effect of the geographic dispersion of customers, the customer arrival rate at locations 1 and 2 is varied such that the customer arrival rate is at most time ten times higher at one of the two locations, resulting in $1/10 \leq P_1/P_2 \leq 10$.

The value functions are solved using value iteration to maximize the average utility per stage, and they are rescaled such that they reflect the conditional probability

of survival per stage per life-threatening customer, which is referred to as the *conditional survival probability* hereafter for simplicity. The results are compared to the myopic policy of always sending the closest server (if the servers are available). Figure 1 shows the conditional survival probabilities of the Case 1, Case 2, and closest server policies. Note that Case 1 always dominates Case 2, since there is no advantage to treating Priority 2 customers as high-risk when there is no chance that they are life-threatening. Figure 1 shows that the optimal Case 1 2 policy is identical to the closest server policy when the customer arrival rate is almost balanced between the two locations (i.e., $\log(P_1/P_2) = -0.2$), resulting in nearly identical conditional survival probabilities. However, Case 2 is not identical to the closest server policy for any of the scenarios considered. On the other hand, when the customer arrival rate is not balanced between the two locations (i.e., $\log(P_1/P_2) = 1,\ -1$, which is often realistic), the differences in the conditional survival probabilities between the optimal Case 1 or Case 2 policy and the myopic policy are greatest. Note that although the absolute difference in the conditional survival probability is low, many lives can be saved at no additional cost (in terms of the number of servers used) after a large number of customers have been received. For example, when $\log(P_1/P_2) = 1$, one additional life per every 136 life-threatening customers can be saved for Case 1, and one additional life per every 265 life-threatening customers can be saved for Case 2.

Servers are used in different ways between the three policies, depending on P_1/P_2. The optimal Case 1 and 2 policies tends to "ration" the server at the location with the higher customer arrival rate for its high-risk customers. The server at the location with the lower customer arrival rate serves more low-risk customers at both locations. Table 2 shows the decisions in the optimal Case 1 and 2 policies when both servers are available, where "Loc" is an abbreviation for location. When $\log(P_1/P_2) \leq -0.6$, both Case 1 and 2 policies send server 1 to low-risk customers at location 2, the busier location, as well as location 1. When $\log(P_1/P_2) \geq -0.2$, both Case 1 and 2 policies send server 2 to low-risk customers at locations 1 and 2.

Figure 2 shows the busy probabilities of the two servers for the Case 1 policies. The Case 2 busy probabilities are identical to the Case 1 busy probabilities except for when $\log(P_1/P_2) = -0.4$, and hence, it is omitted. The busy probabilities report the proportion of the time a server is servicing customers. Thus, the Case 1 policies results

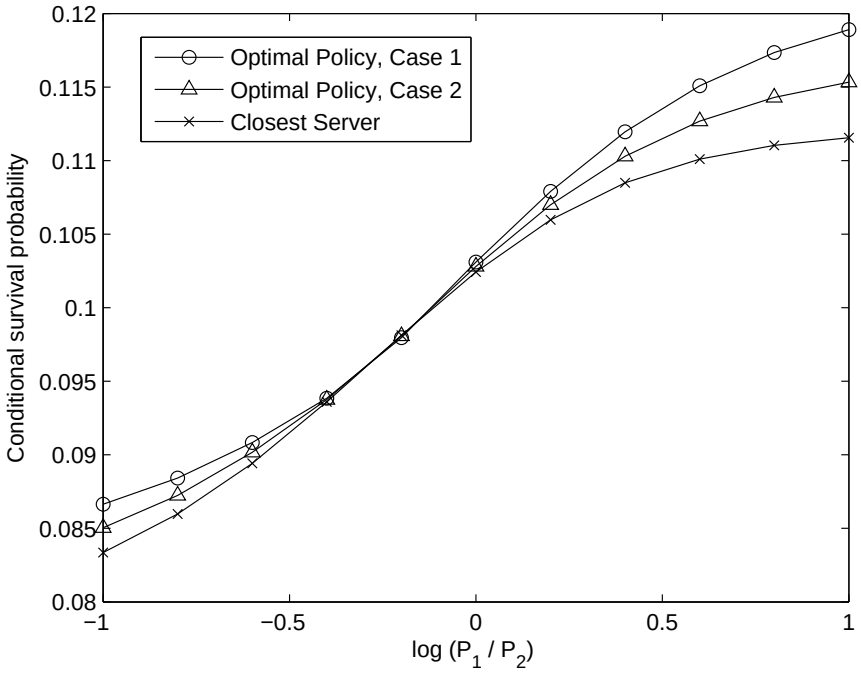


Figure 1: The conditional survival probability (averaged per customer) when using the optimal policy compared to the policy of always sending the closest server

in a more balanced workload between the two servers than the closest server policy when $\log(P_1/P_2) \leq -0.2$ and $\log(P_1/P_2) > 0.6$, when there are large disparities in the customer arrival rates between the two locations. A balanced workload is desirable, since it means that the personnel operating the servers (such as EMTs) are able to treat patients and maintain their proficiency.

Figure 3(a) shows the conditional probability that the closest server is dispatched to high- and low-risk customers for the two locations. It is desirable for these values to be high, which means that the customers wait less for a server to arrive (on average). Figure 3(a) indicates that the high-risk customers at the location with the higher customer arrival rate are responded to more frequently by the closest server. This is done at the expense of the location with the lower customer arrival rate, whose server is more likely to be treating low-risk patients, as captured by Figure 3(b). The low-risk customers at the busier location are unlikely to be serviced by the closest server.

Thus far, we have only considered the scenario with perfect information (i.e., $\alpha = \infty$). Next, we vary α, to see the effect that this has consider the scenario when the customer

Table 2: The Case 1, Case 2, and closest server policy decisions when both servers are available as a function of P_1/P_2

	Case 1				Case 2				Closest server			
	Loc 1		Loc 2		Loc 1		Loc 2		Loc 1		Loc 2	
$\log(P_1/P_2)$	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low
-1.0	1	1	2	1	1	1	2	1	1	1	2	2
-0.8	1	1	2	1	1	1	2	1	1	1	2	2
-0.6	1	1	2	1	1	1	2	1	1	1	2	2
-0.4	1	1	2	2	1	1	2	1	1	1	2	2
-0.2	1	2	2	2	1	2	2	2	1	1	2	2
0.0	1	2	2	2	1	2	2	2	1	1	2	2
0.2	1	2	2	2	1	2	2	2	1	1	2	2
0.4	1	2	2	2	1	2	2	2	1	1	2	2
0.6	1	2	2	2	1	2	2	2	1	1	2	2
0.8	1	2	2	2	1	2	2	2	1	1	2	2
1.0	1	2	2	2	1	2	2	2	1	1	2	2

arrival rate is ten times higher at location 1 than location 2, resulting in $P_1 = 10/11$ $P_2 = 1/11$. Figure 4 shows the conditional survival probability for the Case 1, Case 2, and closest server policies as a function of α. When $\alpha < 8$, Case 2 dominates Case 1. This indicates that when there is greater uncertainty in customer risk, it is best to treat Priority 2 customers as high-risk. When $\alpha \geq 8$, Case 1 dominates Case 2. This indicates that when there is greater certainty in customer risk, it is best to treat Priority 2 customers as low-risk.

Figure 4 illustrates the conditional survival probability for the policies found using value iteration. The conditional survival probability can alternatively be found by analyzing the limiting distribution of the Markov chain for each possible policy using (11). A policy is determined by setting the values of four binary values, resulting in sixteen possible policies. There are four policies when requiring that the closest server is sent to high-risk customers when it is available (i.e., $x_{11}^{H,\pi} = 1$ and $x_{22}^{H,\pi} = 1$ for all π). Figure 5 illustrates the conditional survival probabilities for the resulting four Case 1 policies when considering that the closest server is always sent to high-risk customers (when it is available) as a function of α. It shows that only two policies are ever optimal across all values of α, in terms of the underlying decisions regarding how servers are dispatched. For $\alpha < 4$, the optimal policy is to send the closest server to high-risk and low-risk customers at both locations. This suggests that when there are many classification errors, the optimal policy hedges and tends to send the closest server to

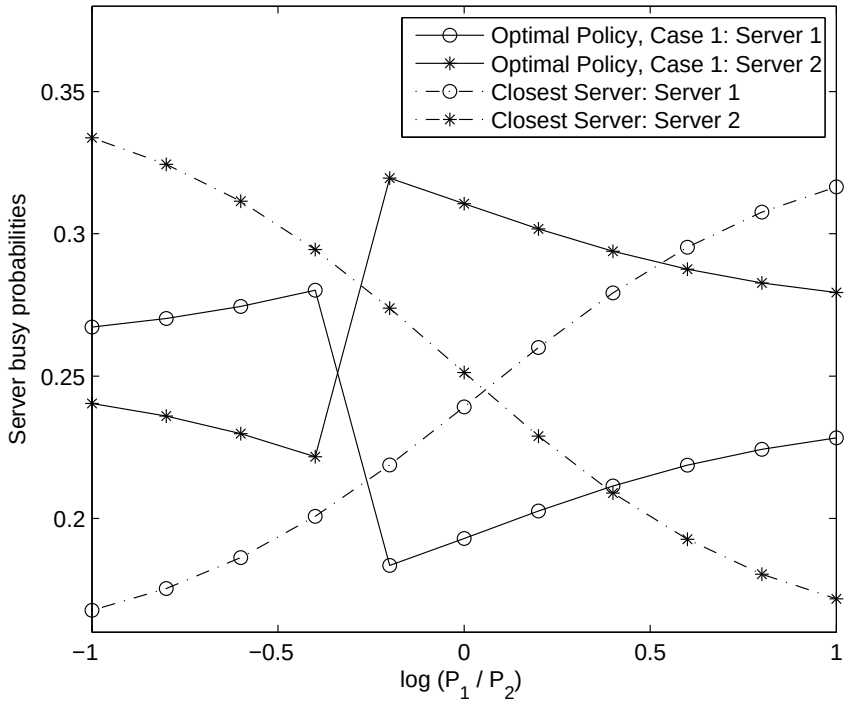
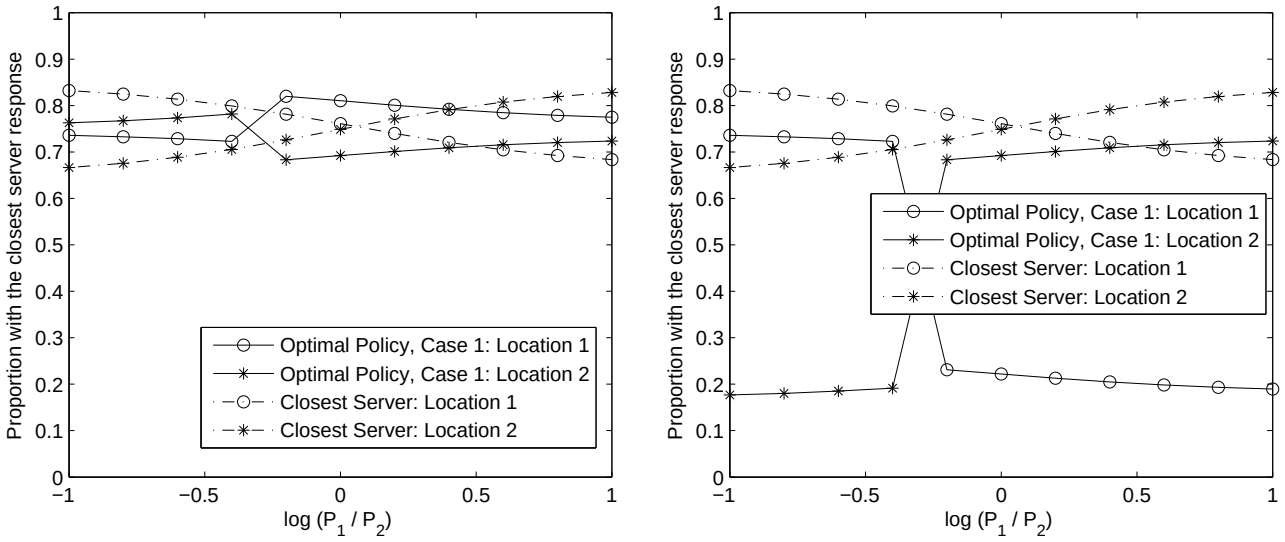


Figure 2: The server busy probabilities for the Case 1 and closest server policies as a function of $\log(P_1/P_2)$

low-risk customers, since they have a greater chance of being life-threatening. For $\alpha \geq 4$, the optimal policy is to send the closest server to high-risk customers at both locations and to send server 2 to low-risk customers at both locations. For $\alpha \geq 4$, both policies that send server 1 to low-risk customers at location 1 (the closest server for location 1) are inferior to the two policies that send server 2 to low-risk customers at location 1 (the farthest server for location 1). Note that the optimal Case 1 policy shown in Figure 4 is identical to the maximum policy in Figure 5 at each value of α.

Table 3 summarizes the preferred server to dispatch to each type of customer at each location given that both servers are available. In Table 3, a 1 indicates that server 1 would be dispatched to a customer (given the location and risk classification) when both servers are available, whereas a 2 indicates that server 2 would be dispatched to a customer. It indicates that the optimal policy sends the closest server to high-risk customers for all values of α and sends the closest server to low-risk customers for $\alpha < 4$. The optimal Case 2 policies never send the closest servers to all types of customers.



(a) High-risk customers

(b) Low-risk customers

Figure 3: The conditional probability that the closest server is dispatched to high- and low-risk customers for the optimal Case 1 and closest server policies, based on the customer location (location 1 or 2)

Table 3: The Case 1, Case 2, and closest server policy decisions when both servers are available as a function of α, when $\log(P_1/P_2) = 1$

	Case 1				Case 2				Closest server			
	Loc 1		Loc 2		Loc 1		Loc 2		Loc 1		Loc 2	
α	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low
< 4	1	1	2	2	1	2	2	2	1	1	2	2
≥ 4	1	2	2	2	1	2	2	2	1	1	2	2

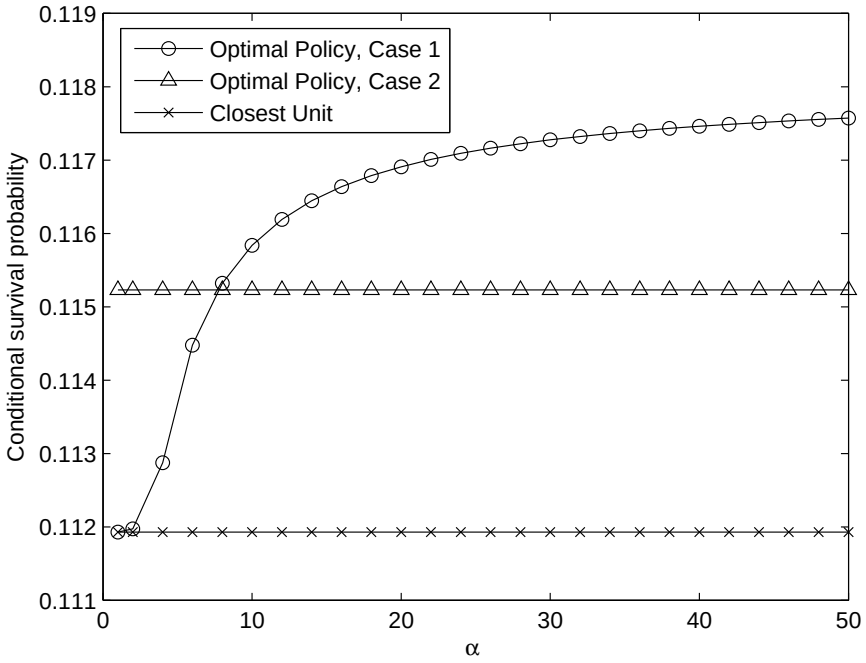


Figure 4: The conditional survival probability (averaged per customer) when using the optimal policy compared to the policy of always sending the closest server.

While the conditional survival probability is of interest, it reflects an aggregate utility over all customers in the system. In public services, fairness is often a consideration [33, 32, 23, 5]. One notion of fairness reflects the conditional probability of survival for customers at each of the locations, which are reflected in Figure 6. Four of the policies are illustrated: sending the closest server (for all values of α), the optimal Case 2 policy (for all values of α), the optimal Case 1 policy for $\alpha = 1$, and the optimal Case 1 policy for $\alpha = \infty$. Note that although the conditional survival probabilities change across the four policies, their values are relatively close to one another at a given location. This illustrates that although different policies affect patient survival, they are not likely to radically change the conditional probability of survival at specific locations, since survival is a function of the distribution of customers as well as the distances that servers travel on the transportation network, all of which are unchanged by implementing different dispatching policies.

Thus far in the analysis, it is assumed that α is known. However, in practice, it is not always likely exactly known. When this assumption is lifted, we can consider the impact

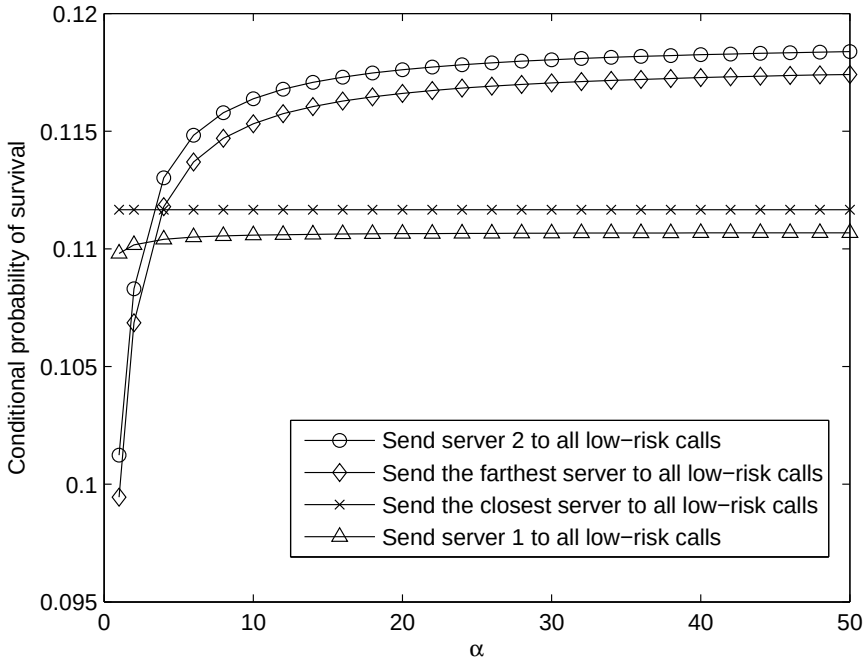


Figure 5: The conditional survival probabilities of the four Case 1 policies using the Markov chain limiting distribution

of making optimum decisions over inaccurate values of α. Let α denote the true value for the ratio of classification errors, and let α_0 denote the inaccurate value for the ratio of classification errors that is believed to be true. Note that for the example considered, there are only two Case 1 policies that are optimal across values of α (see Figure 5). As a result, we consider two values of α_0 corresponding to either of the optimal Case 1 policies with $\alpha_0 < 4$ or $\alpha_0 \geq 4$ across various values of α. Figure 7 illustrates the conditional survival probabilities when using α_0 as in an inaccurate estimate for α in the MDP model as a function of α, the true value. The conditional survival probability for the closest server is shown as a baseline, since it does not depend on α or α_0. The curve corresponding to $\alpha_0 < 4$ (which underestimates α when $\alpha \geq 4$) is virtually constant across values of α, whereas the curve corresponding to $\alpha_0 \geq 4$ (which overestimates α when $\alpha < 4$) increases sharply with α. The curves corresponding to $\alpha_0 < 4$ and to the closest server are identical. This suggests that underestimating α may result is lower conditional survival probabilities than overestimating α. It also suggests that if α is likely to be low, using the optimal Case 1 policy may not improve the conditional

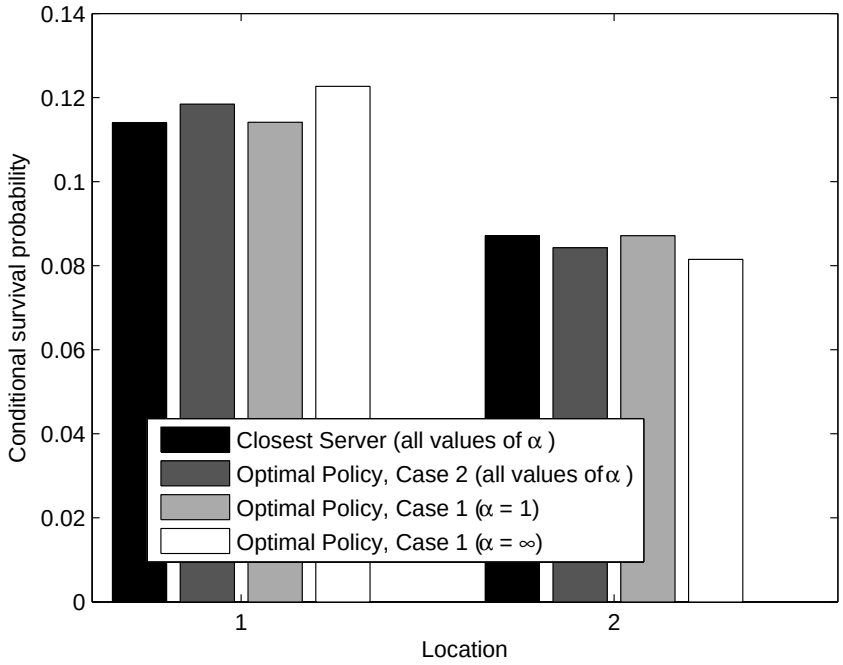


Figure 6: The conditional survival probability at each location for Case 1, Case 2, and closest server policies

survival probability beyond what would be obtained by using the closest server policy if α is not known with certainty.

5.2 Four location example

Hanover County, Virginia is divided into four response districts, which correspond to the four rescue stations in the county. Therefore, a real-world example was created using data from the four rescue stations and their respective stations. The locations and their corresponding closest station are identical, resulting in an example with $n = m = 4$. The example is illustrated for the 12pm – 6pm time period, Saturday and Sunday, with $\lambda = 1.208$ customers/hour, with $P_1 = 0.169$, $P_2 = 0.423$, $P_3 = 0.106$, and $P_4 = 0.302$. This time period was selected for two reasons. First, the data analysis suggests that these times operate in steady state, with the customer arrival rate approximately constant per unit time. Second, the EMS system is entirely operated by volunteer EMTs during these times. The ambulances are located based on volunteer preferences, and hence, the EMS administration has little control over ambulance locations. Resource

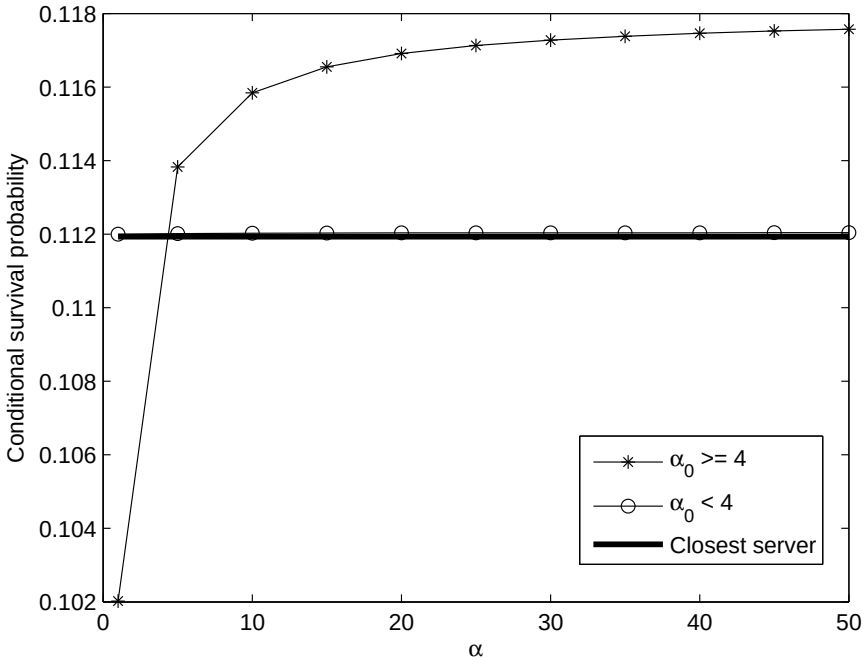


Figure 7: The conditional survival probability with inaccurate estimates for the ration of the proportion of Priority 1 customers that are life-threatening to the proportion of Priority 2 patients that are life-threatening (α_0) as a function of α.

allocation decisions such as locating ambulances are not feasible, and hence, optimal dispatching policies have the potential to have the greatest impact during these times.

The CAD data set contains calls during a one year period, including 9708 calls for service (customers). Each record includes information regarding the location, response time, and service times for all calls. Table 4 reports the average service time based on location of the customer and responding ambulance, used for the values of $\mu_{ij},\ i = 1, 2, ...n,\ j = 1, 2, ..., m$. Representative service times were used when there were too few customers to estimate the actual values. Likewise, Table 5 reports the proportion of customers that are Priority 1, 2, 3 for each of the locations. Case 1 results in $P_{H|i} = P_{Pr1|i}$ and $P_{L|i} = P_{Pr2|i} + P_{Pr3|i}$, $i = 1, 2, ..., n$. Case 2 results in $P_{H|i} = P_{Pr1|i} + P_{Pr2|i}$ and $P_{L|i} = P_{Pr3|i}$, $i = 1, 2, ..., n$.

To determine the rewards (utilities) using (3), first consider the survival probabilities $S_{ij},\ i = 1, 2, ..., n,\ j = 1, 2, ..., m$. The distribution of response times was estimated using the CAD data using lognormal distributions. The probability of survival if am-

bulance j responds to a customer at location i, S_{ij}, is determined as follows,

$$S_{ij} = \int_{t \geq 0} S(t) dF_{ij}(t), \ i = 1, 2, ..., n, \ j = 1, 2, ..., m. \tag{13}$$

where t denotes the response time (in minutes), F_{ij} is the cumulative distribution function for the response times when server j responds to customers at location i, and $S(t)$ is the survival function as a function of time, where time is measured in minutes. The survival function is assumed to represent a probability or proportion (i.e., $0 \leq S(t) \leq 1$), $t \geq 0$. The survival function is based on patient survival models developed in the medical literature. The patient survival model used in this paper is based on a study by Larsen et al. [18], which performs multiple linear regression for data from King County, WA, and has similar demographics to Hanover County. The times are measured relative to the time of collapse and are measured in minutes. The relationship between survival and response time can be defined under the following assumptions that are confirmed to be reasonable by real-world data.

- It takes exactly one minute for a call to EMS to be made and an ambulance to be dispatched after the patient collapses.
- CPR is performed and an AED is used by a paramedic or EMT immediately upon arrival, and CPR is not performed earlier by a bystander.
- Advanced life support is provided one minute after arrival.

These assumptions simplify the survival function to

$$S(t) = \max\{0.594 - 0.055t, \ 0\}. \tag{14}$$

The resulting survival probabilities S_{ij} are summarized in Table 6. Note that since there were few customers for some pairs of stations and locations (particularly for pairs that crossed the entire county), representative values for S_{ij} were selected. A list ordering of the closest to farthest stations from each location is based on the the highest to lowest values of S_{ij} across $j = 1, 2, ..., m$ for each $i = 1, 2, ..., n$. This results in utilities that are consistent with Assumption 1.

Three policies are considered: Case 1, Case 2, and always sending the closest server. Figure 8 shows the optimal solutions for the three policies considered, rescaled accordingly to reflect the conditional probability of survival given that a life-threatening customer has arrived. When $\alpha < 8$, Case 2 dominates Case 1. As with the example with

Table 4: Average service times (in hours)

	μ_{ij}			
j	$i=1$	$i=2$	$i=3$	$i=4$
1	1.15	1.30	1.73	1.33
2	1.30	1.02	1.73	1.27
3	1.22	1.26	1.57	1.56
4	1.25	1.34	1.73	1.16

Table 5: Priorities associated with the customer locations

	$P_{Pr\ j\|i}$			
j	$i=1$	$i=2$	$i=3$	$i=4$
1	0.583	0.514	0.613	0.509
2	0.229	0.212	0.236	0.237
3	0.188	0.274	0.151	0.254

two locations, this indicates that when there is greater uncertainty in patient severity, it is best to treat Priority 2 patients as high-risk. When $\alpha > 8$, Case 1 dominates Case 2. This indicates that when there is greater certainty in patient severity, it is best to treat Priority 2 patients as low-risk. Both Case 1 and 2 dominate the policy of always sending the closest server.

Figure 8 illustrates the conditional survival probability for the Case 1, Case 2, and closest server policies. It illustrates that Case 2 dominates Case 1 when $\alpha < 8$. This indicates that when there is greater uncertainty in customer risk, it is best to treat Priority 2 customers as high-risk. When $\alpha \geq 8$, Case 1 dominates Case 2. This indicates that when there is greater certainty in customer risk, it is best to treat Priority 2 customers as low-risk. Both Case 1 and Case 2 dominate the closest server policy for all values of α.

The conditional survival probability varies according to customer location. Figure 9 shows the conditional survival probability for life-threatening customers at the four locations for four policies: sending the closest server (for all values of α), the optimal

Table 6: The probability of survival between servers and customers

	S_{ij}			
j	$i=1$	$i=2$	$i=3$	$i=4$
1	0.1002	0.0557	0.0231	0.0587
2	0.0557	0.1433	0.0231	0.0355
3	0.0231	0.0231	0.0618	0.0344
4	0.0744	0.0407	0.0344	0.1785

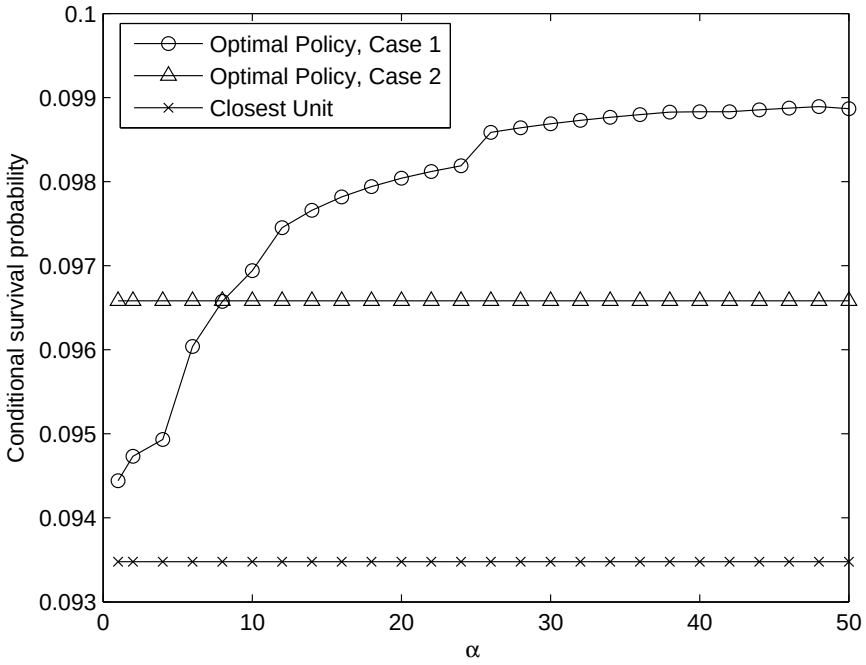


Figure 8: The conditional survival probability (averaged per customer) when using the optimal policy compared to the policy of always sending the closest server.

Case 2 policy (for all values of α), the optimal Case 1 policy for $\alpha = 1$, and the optimal Case 1 policy for $\alpha = \infty$. Figure 9 illustrates that despite the policy used, the range of the survival probabilities at each location are not overlapping, which suggests that the aggregate conditional survival probabilities across all customer locations can be changed by changing dispatching standards, the conditional survival probabilities at specific locations are largely predetermined by factors that cannot be changed (such as the distribution of customer locations and the transportation network). Given perfect classification ($\alpha = \infty$), the optimal Case 1 policy has a higher (overall) conditional survival probability than Case 2, and both the optimal Cases 1 and 2 policies have higher (overall) conditional survival probabilities than the closest server policy. Relative to the closest server policy, the optimal Cases 1 and 2 policies have higher conditional survival probabilities at the locations that already have the highest conditional survival probabilities—locations s 2 and 4—and decreases the conditional survival probability at the location with the conditional survival probabilities—location 3. This suggests that optimal dispatching policies can introduce inequities in terms of the geography

of patient survival throughout a region, where the locations receiving the best service receive even better service and the locations receiving the worst service receive even worse service.

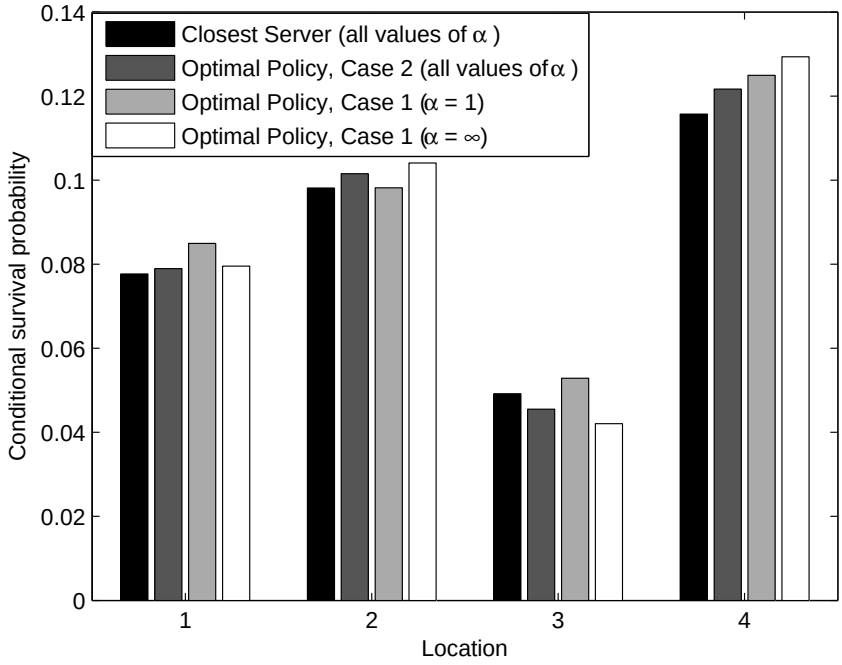


Figure 9: The conditional survival probability (averaged per customer) at the four customer locations.

Figure 10 illustrates the server busy probabilities for the same four scenarios as considered in Figure 9. It is notable that server 3—the least busy server according to the closest server policy—becomes the busiest server in the optimal Case 1 policy with perfect classification. This is a result of server 3 being dispatched more frequently to low-risk customers at all of the locations. It is also notable that the server 4—the location with the highest survival probability—becomes the least busy server in the optimal Case 1 policy with perfect classification. This is a result of server 1 being rationed to serve the high-risk customers nearby by infrequently being dispatching to low-risk customers.

Table 7 shows the optimal list of servers to dispatch to high- and low-risk customers at each of the four locations under the Case 1 and 2 policies for $\alpha = 1, 10, \infty$. To interpret this table, consider high-risk customers at Location (Loc) 1 with $\alpha = 1$. Table

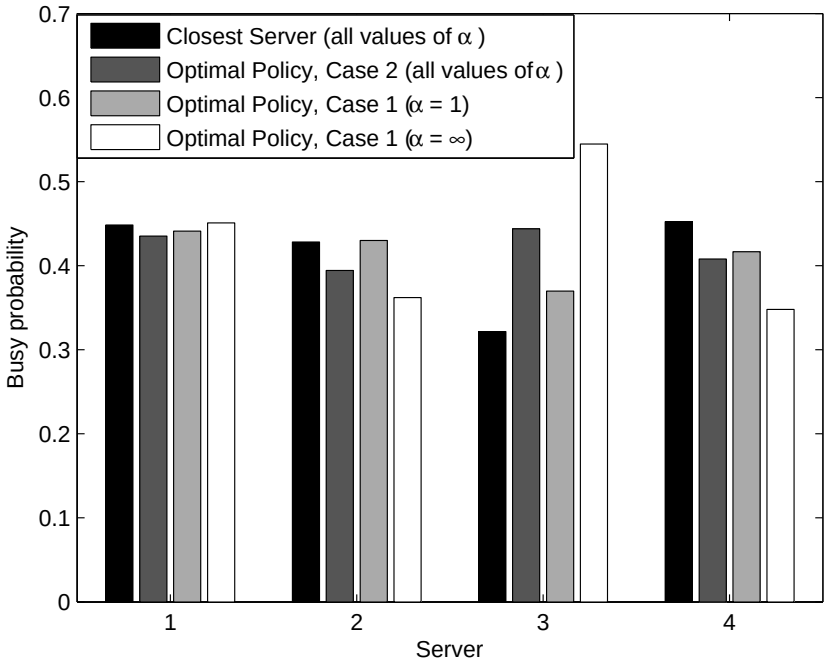


Figure 10: The busy probabilities at the four customer locations.

7 indicates that server 1 is dispatched first, followed by server 4 (if server 1 is busy), then server 2 (when servers 1 and 4 are busy), and finally server 3 (when all other servers are busy). The dispatch order reports the order of servers to be sent to customers at each location. It is assumed that all the busy servers have been dispatched to customers at their closest location. This is important to note, since the optimal policy is not always a priority list, in that the ordering of servers dispatched to a customer can depend on the location to which the busy servers have been dispatched. For the scenarios, however, the optimal policy reflects a priority list nearly all of the time, with a priority list observed in all but four or fewer Markov states. Servers with an asterisk (*) in Table 7 are not those that are the closest servers available given the set of available servers.

Table 7 indicates that the optimal Case 2 policy dispatches the same servers to customers in the $\alpha = 10,\ \infty$ scenarios. Both the optimal Case 1 ($\alpha = \infty$) and Case 2 (all values of α) scenarios always dispatch the same servers to low-priority customers, regardless of the location, with a priority list of servers 3, 1, 2, 4. This is interesting, and it suggests that a simple dispatching protocol could be established for low-risk customers that would not rely on customer location. Both the Case 1 and Case 2 policies always

Table 7: The optimal policy as a function of α for the four location example

		Dispatch	Case 1				Case 2			
Risk group	α	Order	Loc 1	Loc 2	Loc 3	Loc 4	Loc 1	Loc 2	Loc 3	Loc 4
High	1	1	1	2	3	4	1	2	3	4
		2	4	1	1*	1	4	1	4	1
		3	2	3*	4	3	2	3*	1	3
		4	3	4	2	2	3	4	2	2
High	10	1	1	2	3	4	1	2	3	4
		2	4	1	4	1	4	1	1*	1
		3	2	3*	1	3	2	3*	4	3
		4	3	4	2	2	3	4	2	2
High	∞	1	1	2	3	4	1	2	3	4
		2	4	1	4	1	4	1	1*	1
		3	2	3*	1	3	2	3*	4	3
		4	3	4	2	2	3	4	2	2
Low	1	1	1	2	3	4	3*	3*	3	3*
		2	2*	1	1*	3*	1	1*	1*	1*
		3	4	3*	4	1	2*	2	2*	2*
		4	3	4	2	2	4	4	4	4
Low	10	1	3*	3*	3	3*	3*	3*	3	3*
		2	1	2	1*	4	1	1*	1*	1*
		3	2*	1	2*	1	2*	2	2*	2*
		4	4	4	4	2	4	4	4	4
Low	∞	1	3*	3*	3	3*	3*	3*	3	3*
		2	1	1*	1*	1*	1	1*	1*	1*
		3	2*	2	2*	2*	2*	2	2*	2*
		4	4	4	4	4	4	4	4	4

dispatch the closest server to high-risk customers when all of the servers are available. However, the closest server is not always dispatched when some of the servers are busy. Several exceptions occur for customers at locations 2 and 3 when the closest server is busy. Both of these situations require the dispatching of a server across the county when only distant servers are available, with the further of the available servers sometimes being dispatched.

Table 7 indicates that the closest server is not always dispatched to a customer in the optimal solution. By computing the limiting distribution of the Markov chains corresponding to the Case 1, Case 2, and closest server policies we can compare these policies in terms of the proportion of customers to whom the closest server is dispatched from the set of servers available. This is captured by Figures 11 and 12 as a function of α, which illustrates the high-risk and low-risk scenarios, respectively. The Case 2

policies send the closest server to approximately 0.98 of high-risk customers and 0.18 of low-risk customers. The Case 1 policies send the closest server to 0.951 – 0.972 of high-risk customers and to 0.22 – 0.88 of low-risk customers, depending on α. Figure 12 shows that proportion of low-risk customers to whom the closest server is dispatched is non-increasing with α for Case 1. Figure 11 shows that the proportion of high-risk customers to whom the closest server is dispatched does not strictly increase or decrease with α. Figures 11 and 12 illustrate that the Case 1 policies tend to dispatch the closest server to both high- and low-risk customers for low values of α.

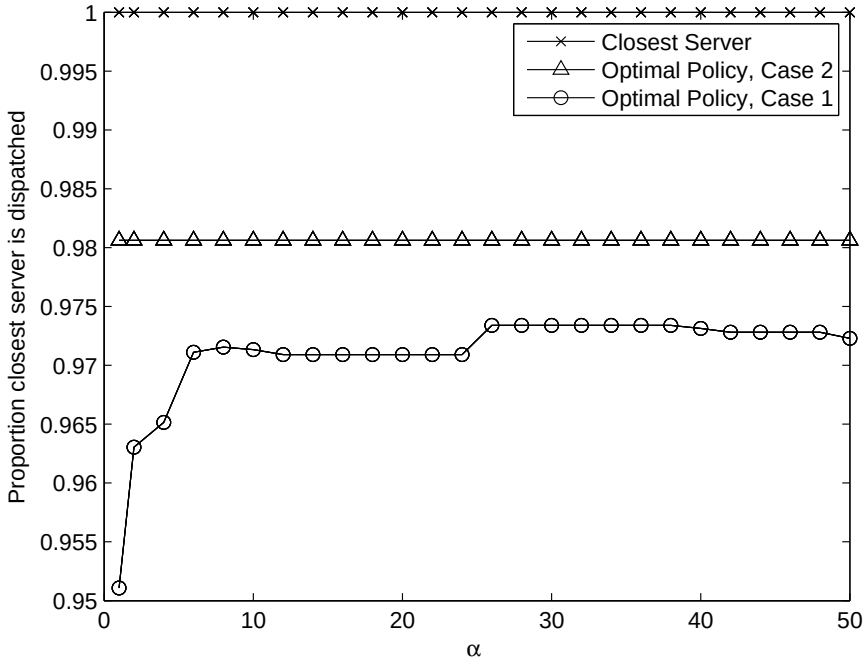


Figure 11: The conditional probability that the closest server is dispatched for high-risk customers.

Figure 13 illustrates the resulting conditional probability of survival when using an inaccurate value of α (i.e., α_0) as an input parameter and is interpreted in an analogous manner as in Figure 7. The conditional survival probability for the closest server policy is shown as a baseline, since it does not depend on α or α_0. Four values of α_0 are considered with $\alpha_0 = 1, 20, 100, \infty$. The curves corresponding to $\alpha_0 = 1, 20$ (underestimating α when α is high) are virtually constant across α, which suggests that underestimating α results in predictable survival probabilities. The curves corresponding to $\alpha_0 = 100, \infty$ (overestimating α when α is low) result in significantly lower survival probabilities when

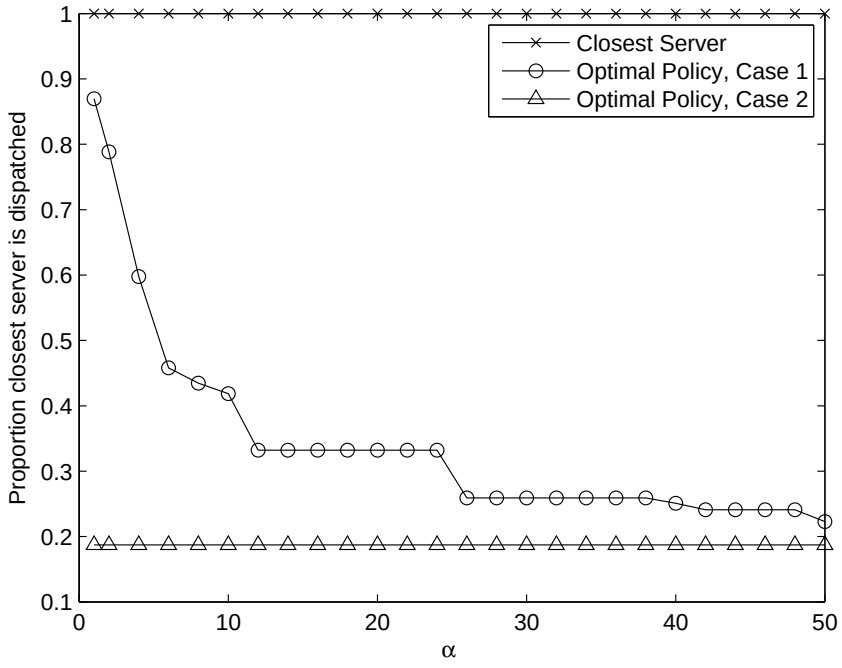


Figure 12: The conditional probability that the closest server is dispatched for low-risk customers.

α is low. Figure 13 suggests that underestimating α may result is lower conditional survival probabilities than overestimating α. It also suggests that if α is likely to be low, using the optimal Case 1 policy may not improve the conditional survival probability beyond what would be obtained by using the closest server policy. This suggests the need to accurately estimate classification errors, since they can drastically impact system performance.

6 Conclusions

This paper models and analyzes optimal dispatching policies in server-to-customer systems with customer priorities and classification errors in assessing their priorities. An undiscounted, infinite horizon average-cost Markov decision process model is developed and analyzed to identify optimal dispatching policies. The model for determining how to optimally dispatch distinguishable servers to prioritized customers given that dispatchers make classification errors in assessing the true customer priorities. Structural properties of the model are examined. The model is interpreted according to the limiting

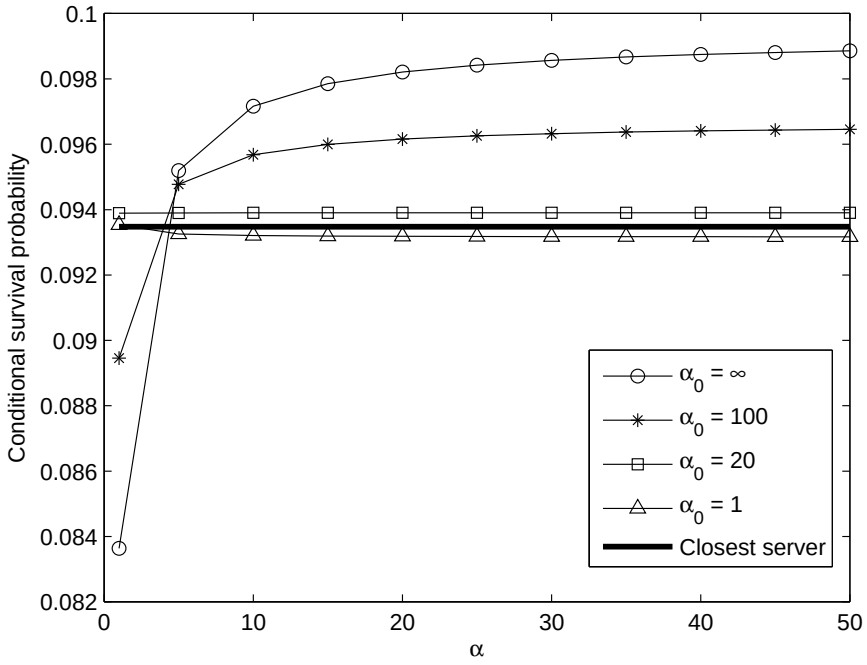


Figure 13: The conditional survival probability (averaged per customer) when using the optimal policy compared to the policy of always sending the closest server for inaccurate estimates for α.

distribution of the Markov chain that each policy induces.

The model and the impact of classification errors are motivated by emergency medical dispatch, where ambulances (servers) are dispatched to prioritized patients (customers). Two cases are considered for approaching the classification errors that correspond to over- and under-responding to perceived patient risk. Two computational examples using real-world data are examined according to maximizing patient survival. The results indicate that it is not always best to dispatch the closest server to a customer. They suggest that that over-responding to customers should be preferred when there is a high rate of classification errors and that under-responding is preferred when there is a low rate of classification errors. The existing model can be trivially modified to interpret utility in terms of performance measures other than patient survival, such as percentage of customers responded to in less than nine minutes (a traditional response time threshold).

The methodology in this paper could be extended to consider other operational issues that determine the long-term likelihood of patient survival. First, dispatching

ambulances to customers in such a way that the workload is balanced is necessary for all paramedics and EMTs to practice their skills. Second, ambulances should be dispatched to customers in such a way that no ambulances are used almost exclusively for transporting patients to the hospital, since this leads to low morale, which could lead to turnover. Both of these issues do not affect the immediate survival of a patient, but could impact the survival of future patients by degrading the workforce through the lack of skill maintenance or through turnover.

The methodology in this paper can be applied to a wide spectrum of transportation problems, including dispatching public services (police cars and fire engines), dispatching military medevacs to soldiers in the theater of operations, and responding to bioterrorism sensor alarms.

Acknowledgements

It is a pleasure to acknowledge Battalion Chief Henri Moore, Jr., Chief Fred C. Crosby, II, Mr. Lawrence Roakes, and Mr. Edward Buchanan of the Hanover County Fire and EMS Department in Hanover County, Virginia, for the knowledge, experience, and data they provided to support this research effort. Suggestions for the medical component of this research from Dr. Joseph P. Ornato, M.D., are gratefully acknowledged. The first author was supported by the U.S. Department of the Army under Grant Award Number W911NF-10-1-0176. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of the Army.

References

[1] O. Berman. Repositioning of 2 distinguishable service vehicles on networks. *IIE Transactions on Systems and Man Cybernetics*, 11:187 – 193, 1981.

[2] O. Berman. Repositioning of distinguishable urban service units on networks. *Computers & Operations Research*, 8:105 – 118, 1981.

[3] S. Budge, A. Ingolfsson, and E. Erkut. Approximating vehicle dispatch probabilities for emergency service systems with location-specific service times and multiple units per location. *Operations Research*, 57:251–255, 2009.

[4] G. M. Carter, J. M. Chaiken, and E. Ignall. Response areas for two emergency units. *Operations Research*, 20(3):571 – 594, 1972.

[5] S. Chanta, M. E. Mayorga, and L. A. McLay. The minimum p-envy location problem: a new model for equitable distribution of emergency resources. Technical report, Clemson University, South Carolina, 2010.

[6] K. R. Chelst and Z. Barlach. Multiple unit dispatches in emergency services: Models to estimate system performance. *Management Science*, 27(12):1390 – 1409, 1981.

[7] K. R. Chelst and J. P. Jarvis. Estimating the probability distribution of travel times for urban emergency service systems. *Operations Research*, 27(1):199 – 204, 1979.

[8] R. Davis. Atlanta becomes a template for improving cardiac-arrest survival rates. *USA Today*, 23 August 2007.

[9] J. V. Dunford. Emergency medical dispatch. *Emergency Medicine Climics of North America*, 20:859 – 875, 2002.

[10] E. Erkut, A. Ingolfsson, and G. Erdogan. Ambulance location for maximum survival. *Naval Research Logistics*, 55(1):42 – 55, 2008.

[11] J. Fitch. Response times: Myths, measurement and management. *Journal of Emergency Medical Services*, 9:46 – 56, 30.

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

[13] S. G. Henderson. Operations research tools for addressing current challenges in emergency medical services. In J. J. Cochran, editor, *Wiley Encyclopedia of Operations Research and Management Science*. Wiley, Hoboken, NJ, 2011.

[14] S. G. Henderson and A. J. Mason. Ambulance service planning: Simulation and data visualization. In M. L. Brandeau, F. Sainfort, and W. P. Pierskalla, editors, *Operations Research and Health Care: A Handbook of Methods and Applications*, pages 77 – 102. Kluwer Academic, Boston, MA, 2004.

[15] E. Ignall, G. Carter, and K. Rider. An algorithm for the initial dispatch of fire companies. *Management Science*, 28(4):366 – 372, 1982.

[16] J. P. Jarvis. Optimization in stochastic systems with distinguishable servers. Technical report tr-19-75, MIT, Cambridge, MA, 1975.

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

[18] M. P. Larsen, M. S. Eisenberg, R. O. Cummins, and A. Hallstrom. Predicting survival from out-of-hospital cardiac arrest-a graphic model. *Annals of Emergency Medicine*, 22:1652 – 1658, 1993.

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

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

[21] R. C. Larson and M. A. McKnew. Police patrol-initiated activities within a systems queuing model. *Management Science*, 28(7):759 – 774, 1982.

[22] R. C. Larson and A. R. Odoni. *Urban Operations Research*. Prentice-Hall, New Jersey, 1981.

[23] P. D. Leclerc, L. A. McLay, and M. E. Mayorga. Modeling equity for allocating public resources. Technical report, Virginia Commonwealth University, Richmond, Virginia, 2010.

[24] S. A. Lippman. Applying a new device in the optimization of exponential queuing systems. *Operations Research*, 23(4):687–710, 1975.

[25] M. Maxwell, M. Restrepo, S. G. Henderson, and H. Topaloglu. Approximate dynamic programming for ambulance redeployment. *INFORMS Journal on Computing*, 22:266 – 281, 2010.

[26] L. A. McLay. A maximum expected covering location model with two types of servers. *IIE Transactions*, 41(8):730 – 741, 2009.

[27] L. A. McLay. Emergency medical service systems that improve patient survivability. In *Encyclopedia of Operations Research*. John Wiley & Sons, Inc., Hoboken, NJ, 2011. forthcoming.

[28] L. A. McLay and M. E. Mayorga. Evaluating emergency medical service performance measures. *Health Care Management Science*, 13(2):124 – 136, 2010.

[29] J. P. Ornato, M. A. McBurnie, G. Nichol, M. Salive, M. Weisfeldt, B. Riegel, J. Christenson, T. Terndrup, and M. Daya. The public access defibrillation (PAD) trial: study design and rationale. *Resuscitation*, 56:135 – 147, 2003.

[30] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, 1994.

[31] L. P. Roppolo, A. Westfall, P. E. Pepe, L. Nobel, J. Cowan, J. J. Kay, and A. H. Idris. Dispatcher assessments for agonal breathing improve detection of cardiac arrest. *Resuscitation*, 80:769 – 772, 2009.

[32] R. Sarin. Measuring equity in public risk. *Operations Research*, 33:210 – 217, 1985.

[33] E. Savas. On equity in providing public services. *Management Science*, 24:800 – 808, 1978.

[34] M. Sayre, A. Travers, and et al. Measuring survival rates from sudden cardiac arrest: the elusive definition. *Resuscitation*, 62:25 – 34, 2004.

[35] I. G. Stiell, G. A. Wells, and B. J. Field. Improved out-of-hospital cardiac arrest survival through the inexpensive optimization of an existing defibrillation program: Opals study phase ii. ontario prehospital advanced life support. *The New England Journal of Medicine*, 351(7):647 – 656, 1999.

[36] A. J. Swersey. A Markovian decision model for deciding how many fire companies to dispatch. *Management Science*, 28(4):352 – 365, 1982.

[37] T. D. Valenzuela, D. J. Roe, S. Cretin, D. W. Spaite, and M. P. Larsen. Estimating effectiveness of cardiac arrest intervention-a logistic regression survival model. *Circulation*, 96:3308 – 3313, 1997.

[38] R. A. Waaelwijn, R. de Vos, J. G. P. Tijssen, and R. W. Koster. Survival models for out-of-hospital cardiopulmonary resuscitation from the perspectives of the bystander, the first responder, and the paramedic. *Resuscitation*, 51:113 – 122, 2001.

[39] A. Weintraub, J. Aboud, C. Fernandez, G. Laporte, and E. Ramirez. An emergency vehicle dispatching system for an electric utility in chile. *Journal of the Operational Research Society*, 50:690–696, 1999.

[40] L. Wik, T. Boye Hansen, F. Fylling, T. Steen, P. Vaagenes, B. H. Auestad, and P. A. Steen. Delaying defibrillation to give basic cardiopulmonary resuscitation

to patients with out-of-hospital ventricular fibrillation. *Journal of the American Medical Association*, 289(11):1389 – 1395, 2003.