M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 15, No. 2, Spring 2013, pp. 205–220
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs®
http://dx.doi.org/10.1287/msom.1120.0411
© 2013 INFORMS

A Dispatching Model for Server-to-Customer Systems That Balances Efficiency and Equity

Laura A. McLay
Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, Richmond, Virginia 23284,
lamclay@vcu.edu

Maria E. Mayorga
Department of Industrial Engineering, Clemson University, Clemson, South Carolina 29634,
mayorga@clemson.edu

The decision about which servers to dispatch to which customers is an important aspect of service systems. This decision is complicated when servers must be equitably—as well as efficiently—dispatched to customers. In this paper, we formulate a model for determining how to optimally dispatch distinguishable servers to prioritized customers given a set of equity constraints. 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 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. Four types of equity constraints are considered—two of which reflect customer equity and two of which reflect server equity—all of which draw upon the decision analytic and social science literature to compare the effects of different notions of equity on the resulting dispatching policies. The Markov decision processes are formulated as equity-constrained linear programming models. A computational example is applied to an EMS system to compare the different equity models.

Key words: emergency medical services; server-to-customer systems; public health; equity; Markov decision processes; linear programming
History: Received: May 9, 2011; accepted July 7, 2012. Published online in *Articles in Advance* December 14, 2012.

1. Introduction

Equity is a critical and controversial factor when deciding how to allocate public resources (Stone 2002). Despite the important role that modeling equity naturally occupies in the allocation of public resources, little consensus exists concerning which equity measures researchers should employ. Emergency medical service (EMS) systems are a public service in which there is an expectation of equitable service. Balancing efficiency and equity is challenging because improving efficiency often introduces inequities (McLay and Mayorga 2013). However, most of the literature surrounding EMS systems has examined how to efficiently allocate resources in terms of the location or distribution of emergency medical services.

Emergency medical dispatch is a resource allocation problem that is a critical component in EMS systems. The dispatch center handles each 911 call, and 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. For decades, EMS systems have been measured according to how they respond to and care for high-priority patients, a measure of *efficiency*. Equity is a concern in emergency medical dispatch, where *equity* evaluates the fairness of how resources (ambulances) are allocated to patients.

Equity is an important consideration for emergency medical dispatch, because dispatching an ambulance to a patient makes the ambulance temporarily unavailable to other patients. The importance of equity is evident in the following example. Consider the decision about which ambulance to dispatch to a low-priority patient who is not experiencing life-threatening symptoms (e.g., a sprained ankle) in an area where high-priority patients experiencing life-threatening symptoms (e.g., cardiac arrest) are likely to require service in the near future. Dispatching a nearby ambulance to the low-priority patient may be undesirable because the ambulance would then be unavailable for future nearby high-priority patients experiencing life-threatening symptoms, possibly leading to the future patient's death. However, if no such high-priority patients will require service, then there would be no reason to make a low-priority patient wait for a more distant ambulance to arrive. This example suggests two types of "fairness" from the patients' perspective: fairness in patient outcomes and fairness in patient waiting times. In addition, this paper considers other notions of equity from the service providers' perspectives, such as utilizing EMS personnel in "fair" ways, which affect the long-term availability of EMS resources.

All equity measures have at least one commonality; namely, for equity to be achieved, the distribution of some resource should be equalized across a set of people or between groups of people. The literature indicates that there is no single best way to evaluate equity. Keeney and Winkler (1985) differentiate between ex ante and ex post equity, which suggests that equity can be evaluated according to how resources are initially allocated and how certain resource allocation decisions result in specific health outcomes. Savas (1978) describes four qualitative ways of measuring equity in public services and indicates how each should be selected and used. Marsh and Schilling (1994) review 20 measures for equity in facility location applications. Both Savas (1978) and Marsh and Schilling (1994) note the inconsistency in how the qualitative measures are applied to operations research models. The wide range of equity measures used in the literature highlights the challenges in identifying the trade-offs between equity and efficiency. Moreover, Marsh and Schilling (1994) found little or no consensus on the best way to measure equity and note that computational tractability is often a criterion for selecting an equity measure.

Stone (2002) demonstrates that social scientists consider a broader spectrum of how to evaluate equity than is typically considered in the operations research domain. She formulates eight definitions of equity, each of which results in an unequal allocation of resources but equality in another dimension (e.g., patient survival). The eight definitions of equity can be aligned with three categories: (1) who receives service, (2) what is being allocated, and (3) how resources are allocated. Various stakeholders simultaneously use these eight different definitions of equity to support their positions; hence, no single exemplar of enforcing equity in public processes satisfies all stakeholders. This supports the purpose of this paper, namely, to understand how differing measures of equity lead to different dispatching policies.

This paper examines how to optimally balance equity and efficiency when dispatching distinguishable servers to prioritized customers in server-to-customer systems. We formulate the model as an equity-constrained infinite-horizon, undiscounted, average reward Markov decision process (MDP) that identifies how to optimally dispatch servers to customers to maximize the long-run average customer utility (efficiency). 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. The rewards and service times are location dependent, with respect to both the server and customer locations. Because there is no standard way to evaluate equity in EMS systems (Leclerc et al. 2011), four types of equity constraints are compared for ensuring that resources are fairly allocated. Two models reflect equity from the customer (patient) point of view and two models reflect equity from the server (service provider) point of view. All equity measures are drawn from the EMS literature and interviews in the field. The comparison of all four types of equity measures is meant to illustrate how enforcing different notions of equity changes the underlying dispatching policies. In addition, the degree to which one equity measure enforces other equity measures is evaluated. A computational study is conducted with a real-world example using data from an EMS system.

This paper builds on the work of McLay and Mayorga (2013), who provide a modeling framework for dispatching servers to prioritized customers given that there are classification errors in assessing customer priority. McLay and Mayorga (2013) analyze the structural properties of the resulting Markov decision process model to provide insight into optimal dispatching policies. Their computational example suggests that improved dispatching policies can introduce inequities as compared to myopic policies, which motivates the exploration of equity in this paper. This paper extends the basic modeling framework in McLay and Mayorga (2013) to consider the impact of equity constraints by using linear programming models and algorithms. There are several key differences between the two papers. First, this paper focuses on models for solving constrained Markov decision process models rather than on the structural properties of the unconstrained model. Second, and more importantly, this paper draws on multidisciplinary research to identify four potential equity measures. These measures are novel in that they simultaneously consider multiple measures of equity, consider allocating resources both initially and retrospectively, and consider server equity, which has been overlooked in the literature thus far (Leclerc et al. 2011).

This paper is organized as follows. Section 2 provides a literature review on dispatching models applied to EMS systems and on equity models applied to server-to-customer systems. The proposed constrained Markov decision process model is formulated in §3 as a linear programming model. The results are applied to a scenario using real-world data collected from Hanover County, Virginia, in §4. Concluding remarks, policy implications, and directions for future research are presented in §5.

2. Background

Optimal dispatching policies shed light on when to dispatch the closest server to a customer and when to ration the closest server (by dispatching a farther server instead) in anticipation of a more emergent call. Several papers develop optimization models for dispatching problems (Carter et al. 1972, Weintraub et al. 1999, Restrepo et al. 2009), whereas others examine the related issue of how many servers to dispatch to a customer (Chelst and Barlach 1981, Swersey 1982, Ignall et al. 1982). A complementary line of research examines how to optimally reposition and relocate servers based on real-time information and forecasted demand (Henderson 2011; Berman 1981; Gendreau et al. 2001; Maxwell et al. 2010a, b).

Equity has been explicitly considered in few operations research models for EMS systems (Marsh and Schilling 1994). Erdogan et al. (2010) examine how to enforce fairness in scheduling ambulances across different time periods. They find that when they only consider efficiency, the coverage (i.e., fraction of calls expected to be responded to within a fixed timeframe) during the busiest time periods improves at the expense of the coverage during the less busy time periods. Their model thus balances the maximin coverage of each time period with the aggregate coverage over all time periods.

Leclerc et al. (2011) provide a guide to the central issues in the modeling of equity that reflects a broad, multidisciplinary perspective. They also discuss how to define and model equity with respect to a number of principal issues (e.g., selecting the time horizon over which to assess the equity of a particular allocation of public services and choosing the groups of people over which to define equity). Importantly, they note that EMS problems almost always frame equity from the patient point of view, often equalizing the expected outcomes. Other equally valid ways to frame equity focus on server equity, e.g., health care provider equity (motivated by Armony and Ward 2010 in the service sector) and the financial resources allocated to patients (motivated by Felder and Brinkmann 2002).

Issues of efficiency and equity have been simultaneously considered when locating ambulances, a problem that complements emergency medical dispatch. Chanta et al. (2011a) create three bi-objective covering location models that directly consider fairness via a secondary objective. The first objective in all three of the models is the expected coverage (efficiency), and the second objective (fairness) considers three proposed alternatives. They found that the fairness objective drastically impacts the locations of the ambulances and can yield surprising resource allocation decisions. Chanta et al. (2011b) extend this research to consider disparities in access to service between individual customers while taking into account backup coverage. Their proposed "*p*-envy" formulation is compared to other popular location models such as the *p*-center, Gini coefficient, and maximal covering location problems. The results show that the *p*-envy model not only yields the lowest total weighted envy compared to other equity models but also yields highly efficient solutions in terms of coverage. These results are unexpected because equity and coverage are usually conflicting objectives.

Much of the previous work in the area of equity in EMS and server-to-customer systems determines how to locate or relocate ambulances when equity is not a major focus. In contrast to previous work in the area, this paper investigates how to dynamically dispatch ambulances to prioritized patients, given that minimum equity standards are enforced.

3. Markov Decision Process Models

This section presents the MDP model for determining equity-constrained dispatching policies in a server-to-customer system that assigns distinguishable servers to prioritized customers. The MDP model is formulated as a linear programming model such that equity considerations can be added as constraints without expanding the Markov state space, thus allowing for efficient linear programming algorithms for identifying optimal equity-constrained policies. The customers are arriving calls for service (patients), who have an associated location and are categorized as high priority (H) or low priority (L). The servers are ambulances that are differentiated by their response and service times. The *response time* captures the time from when a server is dispatched to when it arrives at the scene, and the *service time* captures the total time to serve a customer, including the time to respond to a customer, treat a customer at the scene, transport a customer to a hospital, and return to service. The objective is to determine which server to dispatch to arriving customers to maximize the average total reward per stage given minimum standards for equity. The reward for responding to a call is interpreted as the probability that a high-priority customer is responded to within a fixed timeframe, known as a *response time threshold* (RTT), in the system-wide *coverage level*. An RTT objective function has emerged as one of the best performance measures for efficiency in EMS systems (McLay and Mayorga 2010).

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

1. If a customer arrives and a server is available, a server must be dispatched.
2. Only available servers—servers located at their home stations—can be dispatched to customers.
3. One server is assigned to each customer.
4. Servers return to their home station after servicing customers.
5. Service times do not depend on customer priority and are exponentially distributed.
6. There is a zero-length queue for customers.

We discuss how the model could be modified to lift the assumptions and which assumptions have a strong impact on the resulting policies. Requiring a server to be dispatched to a customer if a server is available—a *nonidling policy*—is reasonable, because the model here is motivated by EMS systems, and providing service to all customers is a major goal of public service systems. Furthermore, because of concerns of liability, not reserving capacity is an embedded practice in EMS dispatching and a policy constraint of some service providers.

For simplicity, all servers return to their home stations after servicing customers, when they are said to become available for servicing other customers, and servers cannot be dynamically rerouted. In practice, servers can be dispatched from locations other than their home stations, which happens when the servers complete service. To allow for ambulances being dispatched from locations other than their home stations, the state space can be expanded to include auxiliary server locations that correspond to where a server completes service (see §3.1 for a discussion of the state space). Likewise, the state space can be expanded to include auxiliary customer locations that correspond to a server who is traveling to a customer to allow a server to be dynamically rerouted up until it arrives to serve a customer and locations that correspond to servers that are nearing the end of service with another customer. Regarding the fifth assumption, although it would be trivial to incorporate service times dependent on customer priority, the exponential assumption is harder to lift because we would have to expand the state-space to consider a non-Markov model. We acknowledge that this is a strong assumption.

The zero-length queue assumption requires further discussion. Note that the customers are only lost when all of the servers are busy, and therefore each type of customer is lost at the same rate at which it enters the system. From a tractability point of view, the customer queue causes the state-space of the model to become unmanageable, and the dispatching policy could depend on the set of customers in each queue. We believe that the zero-length assumption is reasonable for two reasons. First, the objective is to maximize the expected coverage, and therefore queued customers contribute little or no value to the objective function after waiting for a server to become free. As a result, the model has an incentive to avoid this situation. Second, although incorporating queues would increase the busy probabilities of the servers by a small degree, we have conducted extensive computational experiments using simulation and find that the zero-length queue assumption does not significantly impact the coverage level or the equity measures, because in EMS systems it is unlikely that an arriving customer will find all servers busy (McLay and Mayorga 2013). In the computational example discussed in detail in §4, the proportion of customers that is lost is 0.049.

We begin by discussing the base dispatching model that does not have equity constraints.

3.1. Base Model—Unconstrained MDP Model Review

To describe the base case MDP model without equity constraints, the input parameters are summarized. Then an infinite-horizon, undiscounted, average reward MDP model is defined.

3.1.1. Input Parameters.

$n =$ total number of customer locations;

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

$h =$ set of customer priorities, $h \in \{H, L\}$, where H denotes high priority and L denotes low priority;

$\lambda =$ 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, $i = 1, 2, \ldots, n$;

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

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

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

The set of priorities 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 priorities to reflect the simplifications used in many dispatch centers in practice. 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.

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. The policy of always sending the closest server is discussed in more detail in §3.3 as well as throughout §4.

3.1.2. States. We define the state of the system $\mathbf{s}$ where $\mathbf{s}=(s_1, s_2, \dots, s_m)$ as

$$s_j = \begin{cases} 0 & \text{if server } j \text{ is free,} \\ i & \text{if server } j \text{ is busy serving a customer} \\ & \quad \text{originating at location } i. \end{cases}$$

Let the state space of the corresponding Markov chain be defined by $\mathcal{S}=\{\mathbf{s}\colon \mathbf{s}\in\{0,1,\dots,n\}^m\}$. Thus, we see that the state of the system $\mathbf{s}$ 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, $i=1,2,\dots,n$. To formulate the MDP model as a linear programming model, we expand the state space to include both the priority (h) and location (i) of incoming customers. However, throughout this paper, the configuration of servers given by $\mathbf{s}$ is separated from the priority and location of customers (h,i) for clarity, and the state only refers to the configuration of servers.

3.1.3. Actions. Without loss of generality, we only consider stationary Markov policies (Puterman 1994). Let $X(\mathbf{s},(h,i))$ denote the set of available actions in state $\mathbf{s}$ at time t given that a customer with priority h arrives at location i. If a customer arrives, one of the available servers is dispatched to the customer, resulting in at most m actions available in each state.

3.1.4. Rewards. The rewards reflect the marginal value to the objective function of the action selected. The formulation is generic in that it can reflect any performance measure that depends on the location and customer priority as well as the server that responds. In this case, the rewards reflect the coverage level. One way to compute the rewards is to evaluate the cumulative density function of the response time distribution at the value of the response time threshold (e.g., nine minutes), assuming that the response time distribution is known or can be estimated by examining historic response data. If no servers are available, then assume that a null server (0) is dispatched with a reward of zero.

3.1.5. Transition Probabilities. One of two events determines the transition probabilities: (1) one of the busy servers may complete service and become free (obtained from the expected service times μ_{ij}), or (2) a customer arrives (obtained from λ), which requires that a server be dispatched to the customer if one is available.

The model is formulated as an equivalent discrete time MDP using uniformization. To do so, the maximum rate of transitions is determined as $\gamma=\lambda+\sum_{j=1}^{m}\beta_j$, where $\beta_j=\max_{i=1,2,\dots,n}\{(\mu_{ij})^{-1}\}$. Without loss of generality, we scale $\gamma=1$.

The optimal policy yields the optimal average reward per stage g, where the length of the stage is γ^{-1}. Define $w(\mathbf{s})$ as a relative value function in state $\mathbf{s}$. Then, the optimality equations for the average reward case follows:

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

where $I_{\{s_j=i\,|\,i>0\}}$ 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. The first line in (1) captures busy servers becoming available. The second line captures new customers arriving. The third line captures neither new customers arriving nor servers becoming available. Although (1) captures the MDP model's system dynamics, value iteration cannot be applied to it when adding side constraints (including equity constraints) without expanding the state space. A more computationally efficient formulation follows.

3.2. Linear Programming Model Formulation

Linear programming models and algorithms are advantageous for solving MDPs with certain types of side constraints (Puterman 1994). First, the equivalent linear programming formulation to (1) is introduced. Then, we discuss how to augment it with equity constraints.

Recall that the state space considers the configuration of servers given by $\mathbf{s}$. The set of feasible actions depends not only on the state but also on the event that triggers the action, either a service completion or a new customer arrival. The set of events and set of actions must be augmented with a null event and null action, respectively, because of uniformization. Only customer arrivals require a server to be sent. Let CallType denote the set of customers that can arrive during each stage, where $(h,i)\in\text{CallType}$ represents the priority $h \in \{H, L\}$ and location $i = 1, 2, \ldots, n$. A null customer is considered when no customer arrives. Thus, the set of actions available when a customer of type h arrives to location i and observes state $\mathbf{s}$ is denoted by $X(\mathbf{s}, (h, i))$, which reflects the set of available servers if a customer arrives and a null server (doing nothing) otherwise. The variables in the linear programming model formulation, $y(\mathbf{s}, j, (h, i))$, represents the proportion of time the system is in state $\mathbf{s}$ and selects action j if a customer type (h, i) arrives.

First, the base model (an MDP model without equity constraints) is formulated as a linear programming model. In §3.3, four types of equity constraints are added to this linear programming model to consider different equity measures. The linear programming model for the base model is given by

$$\begin{aligned}
\max_{y} \quad & \sum_{\mathbf{s}\in S,\,(h,i)\in \text{CallType},\, j\in X(\mathbf{s},(h,i))} u_{ij}^{h}\, y(\mathbf{s}, j, (h,i)) \\
\text{s.t.} \quad & \sum_{j'\in X(\mathbf{s}',(h',i'))} y(\mathbf{s}', j', (h', i')) \\
& - \sum_{\substack{\mathbf{s}\in S,\,(h,i)\in \text{CallType},\\ j\in X(\mathbf{s},(h,i))}} p(\mathbf{s}', (h', i') \mid \mathbf{s}, j, (h,i))\, y(\mathbf{s}, j, (h,i)) = 0, \\
& \qquad \text{for } \mathbf{s}'\in S,\ (h', i')\in \text{CallType}, \\
& \sum_{\mathbf{s}\in S,\,(h,i)\in \text{CallType},\, j\in X(\mathbf{s},(h,i))} y(\mathbf{s}, j, (h,i)) = 1 \\
& y(\mathbf{s}, j, (h,i)) \geq 0, \quad \mathbf{s}\in S,\ (h,i)\in \text{CallType}, \\
& \qquad j\in X(\mathbf{s}, (h,i)). \qquad (2)
\end{aligned}$$

In (2), the objective function reflects the average reward accumulated per stage, and it is identical to g in the optimality Equation (1). The first set of constraints balances the flow in and out of each state and call type. The transition probabilities $p(\mathbf{s}', (h', i') \mid \mathbf{s}, j, (h, i))$ are captured by the value functions (1). For example, the probability that server j becomes available if it is serving a customer at location i is $(\gamma\mu_{ij})^{-1}$. This causes a transition from state $\mathbf{s}$ with $s_j = i$ to the corresponding state $\mathbf{s}'$ with $s'_j = 0$. Assigning server j to an arriving customer with priority h at location i in state $\mathbf{s}$ with $s_j = 0$ causes a transition to the corresponding state $\mathbf{s}'$ with $s'_j = i$ with probability $\lambda P_i P_{h\mid i}/\gamma$. Finally, the probability of staying in state $\mathbf{s}$ is $(1 - \lambda/\gamma - \sum_{j=1}^{m} I_{\{s_j = i \mid i>0\}}(1/(\gamma\mu_{ij})))$ or $(1 - \sum_{j=1}^{m} I_{\{s_j = i \mid i>0\}}(1/(\gamma\mu_{ij})))$ if no servers are available in $\mathbf{s}$. The second set of constraints in (2) requires that all of the action probabilities sum to one. Note that because there is a finite number of states and bounded rewards, the optimal policy is Markovian (Puterman 1994).

The optimal policy can be determined using the linear programming variables as follows. The proportion of time that action j is selected given that the system is in state $\mathbf{s}$ with call type (h, i) is given by

$$q_{(\mathbf{s},(h,i))}(j) = \frac{y(\mathbf{s}, j, (h,i))}{\sum_{j'\in X(\mathbf{s},(h,i))} y(\mathbf{s}, j', (h,i))},$$
$$\mathbf{s}\in S,\ (h,i)\in \text{CallType},\ j\in X(\mathbf{s},(h,i)).$$

3.3. Equity Measures

The base model (2) does not explicitly enforce equity of any kind. In this section, four equity measures are introduced. The first two equity measures consider equity from the customer point of view. The equity measures are akin to ex ante equity (before the case, equity of process) and ex post equity (after the fact, equity of outcomes) explored by Keeney and Winkler (1985) and Sarin (1985). The second two equity measures focus on different aspects of server equity, which have been overlooked in the EMS literature thus far. This section focuses on the mathematical formulation of each of the equity measures. We note that because the equity measures add constraints to a linear programming model, they reduce the feasible region and may lead to infeasibility. Thus, some domain knowledge of what is achievable is necessary when selecting the level at which each equity measure is satisfied.

The heuristic policy used in many settings is to always send the closest server relative to customers (see McLay and Mayorga 2013 for a detailed discussion). Therefore, the equity-constrained models are compared to the heuristic closest server model throughout the remainder of the paper. The practical policy implications of the equity-constrained models are discussed in §5.

3.3.1. Equity Measure 1: Ex Ante Customer Equity. The first equity measure captures the fraction of customers of each priority at each location that is serviced by the closest server, irrespective of whether that server is available. This equity measure considers the allocation of servers to customers prior to service. It draws on the concept of districts and zoning in public services, where customers may expect the ambulances in their district or zone to be available for service (Clawson et al. 1999, Dean 2008).

The first equity measure is subtly different from the heuristic policy of always sending the closest server. This equity measure captures only the fraction of times when the closest server in the absolute sense is sent to a customer. The heuristic, in contrast, always sends the closest server to a customer among those that are available. Moreover, the closest server heuristic does not optimize over the set of servers available as backup in each customer's district.

The closest server is defined a priori based on districts, zones, average service times (μ_{ij}), or other criteria. Let $\text{Close}(X(\mathbf{s}, (h, i)))$ denote the closest server(s) in a given state relative to customer type (h, i), which may be equal to the null set if the closest server is not one of the available actions. For simplicity, $\text{Close}(X(\mathbf{s}, (h, i)))$ captures all of the servers at the nearest station relative to patient location i, although it could certainly be defined differently, for example, as all of the servers within a given customer's district. Let $\theta_L^1(h, i)$ denote a lower bound on the fraction of customers of each priority at each location that are serviced by the closest server. Then, enforcing ex ante customer equity can be achieved by adding the following $2n$ constraints to (2):

$$Q_1(h, i) = \sum_{\mathbf{s},\, j \in \text{Close}(X(\mathbf{s},\, (h,\, i)))} y(\mathbf{s}, j, (h, i)) \geq \lambda P_i P_{h|i} \theta_L^1(h, i), \quad h \in \{H, L\},\ i = 1, 2, \ldots, n. \tag{3}$$

No server is always available, and therefore, requiring that the closest server always be sent would lead to an infeasible linear programming model. The lower bound is normalized by the probability that a customer is of type (h, i), to compare locations with different levels of customer demand.

3.3.2. Equity Measure 2: Ex Post Customer Equity. The second equity measure evaluates how the allocation of resources affects patient outcomes (or other outcomes of interest such as patient satisfaction), which takes all server responses into account. This issue is critical because patient survival rates are in part a function of ambulance response times (Larsen et al. 1993, Stiell et al. 1999). This equity measure can be contrasted with equity measure 1, which only captures the relative response times of the closest server as compared to further servers. The computational example suggests that equity measure 2 may be sensitive to differences in population density across a region and, therefore, it may not be suitable to use in all settings.

Consider the ex post customer equity interpreted as patient survival. Let $\theta_L^2(i)$ denote a lower bound on the fraction of customers having life-threatening conditions at each location that survive. For this equity measure, let S_{ij} denote the probability of survival for a customer in location i experiencing life-threatening conditions (such as cardiac arrest), who is responded to by server j. For simplicity, assume that a fixed proportion of high-priority (H) customers have life-threatening (LT) conditions regardless of the customer location, with $P_{LT} \leq P_H$, and that no low-priority customers have life-threatening conditions. Let $P_{LT|H \cap i}$ capture the proportion of life-threatening calls of those that are classified as high priority and originating in location i. Under these assumptions, $P_{LT|H \cap i} = P_{LT}/P_H$ and $P_{LT|L \cap i} = 0$, $i = 1, 2, \ldots, n$. The conditional probability that a customer has a life-threatening condition given that the customer is in location i is given by $P_{LT|i} = P_{LT|H \cap i} P_{H|i}$ (because $P_{LT|L \cap i} = 0$), $i = 1, 2, \ldots, n$.

Enforcing ex post customer equity can be achieved by adding the following n constraints to (2):

$$Q_2(i) = \sum_{\mathbf{s} \in S,\, j \in X(\mathbf{s},\, (H,\, i))} S_{ij} y(\mathbf{s}, j, (H, i)) \geq (\lambda P_i P_{LT|i}) \theta_L^2(i), \quad i = 1, 2, \ldots, n. \tag{4}$$

The lower bound is normalized by the probability that a customer having a life-threatening condition arrives at location i. Adding these equity constraints can lead to an infeasible linear programming model if $\theta_L^2(i)$ is higher than a maximum achievable level (e.g., if $\theta_L^2(i) > \max_j \{S_{ij}\}$ for some i). These constraints capture the utility of the service provided, not just the utility of the resources allocated.

3.3.3. Equity Measure 3: Server Busy Probability Equity. A third equity measure keeps server busy probabilities between lower and upper bounds to balance the workload across servers. This requires that the emergency medical technicians (EMTs) and paramedics who staff each server be busy approximately the same amount of time, because it is desirable for EMTs and paramedics to be frequently utilized to practice their skills but not utilized so frequently that they experience distress (Boudreaux et al. 1997, Key 2002). Normalizing server busy probabilities has been considered in the literature as an operating criteria in server-to-customer systems for public services (Carter et al. 1972), and therefore it is a reasonable model of fairness.

Let θ_L^3 and θ_U^3 denote the lower and upper bounds on the server busy probabilities, which are defined to be identical for each of the servers. Dispatching server j to customer type (h, i) when the system is in state $\mathbf{s}$ leads to the deterministic transition to state $\mathbf{s}'$, which is identical to state $\mathbf{s}$ except that server j is located at i. Enforcing server busy probability equity can be achieved by adding the following $2m$ constraints to (2):

$$\theta_L^3 \leq Q_3(j) = \sum_{\mathbf{s}':\, s'_j > 0} \sum_{k=1}^{m} \sum_{(h,\, i) \in \text{CallType}} y(\mathbf{s}', k, (h, i)) \leq \theta_U^3, \quad j = 1, 2, \ldots, m. \tag{5}$$

Adding these constraints could lead to an infeasible linear programming model if the bounds θ_L^3 and θ_U^3 are not consistent with the call arrival rate and average service times.

3.3.4. Equity Measure 4: Server High-Priority Dispatch Equity. The fourth equity measure requires that servers are dispatched to high-priority customers at a given minimum rate. Responding to high-priority patients is also an important way for EMTs and paramedics to practice medical techniques to maintain medical proficiency (Key 2002) and higher rates of job satisfaction (Studnek and Fernandez 2007). This equity measure differs from the third equity measure in that it constrains the number of customers serviced by each server rather than constraining the amount of time each server is busy.

Let θ_L^4 denote the lower bound on the probability in a stage that each server is dispatched to a high-priority customer at any location, which is defined to be identical for each of the servers. Enforcing a minimum bound on the rate (probability per stage) that each server is dispatched to a high-priority customer can be achieved by adding the following m constraints to (2):

$$Q_4(j) = \sum_{\mathbf{s}\in S,\, i=1,2,\ldots,n} y(\mathbf{s}, j, (H,i)) \geq \theta_L^4, \qquad j = 1, 2, \ldots, m. \quad (6)$$

Adding these constraints could lead to an infeasible linear programming model if the bound θ_L^4 is not consistent with the rate of incoming high-priority calls, $\lambda \sum_{i=1}^{n} P_i P_{H|i}$.

Although these four equity measures are not exhaustive, they reflect important equity criteria across multiple stakeholders that capture key equity considerations noted in the decision analytic and social sciences. All four equity measures are compared through a computational example in the following section.

4. Computational Examples

This section formulates the linear programming models and analyzes their solutions to two examples with four locations and four servers ($n = m = 4$). Exactly one server is stationed at each location, which corresponds to a response district surrounding a rescue station (server). The examples use data extracted from Hanover County Fire and EMS in Virginia, a semi-rural, semi-suburban county in the metropolitan Richmond area. A data set using one year of CAD data was provided to assist in the computational portion of this paper. The CAD data set contains calls during a one-year period, including 9,708 calls for service (customers). Each record includes information regarding the location, response time, and service times for all calls. Representative input parameter values are used when there are too few data points to accurately estimate the actual values.

The first example (Example 1) is illustrated for the time period 12 P.M.–6 P.M. on Saturdays and Sundays, with $\lambda = 1.2$ customers/hour and $P_1 = 0.169$, $P_2 = 0.423$, $P_3 = 0.106$, and $P_4 = 0.302$. The second example (Example 2) performs a sensitivity analysis over P_i, $i = 1, 2, 3, 4$ (see §4.3). The time period used for these examples was selected for two reasons. First, the data analysis suggests that during these times the system operates 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. Resource allocation decisions such as locating ambulances are predetermined by volunteer preferences, and hence optimal dispatching policies have the potential to have the greatest impact during these times.

Table 1 **Average Service Times Between Customers, $i = 1, 2, 3, 4$, and Servers, $j = 1, 2, 3, 4$ (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 2 **Priorities Associated with the Customer Locations**

	$P_{h\|i}$			
h	$i=1$	$i=2$	$i=3$	$i=4$
H	0.583	0.514	0.613	0.509
L	0.417	0.486	0.387	0.491

Tables 1–4 summarize the remaining input parameters. Table 1 reports the average service times based on location of the customer and responding ambulance, used for the values of μ_{ij}, $i = 1, 2, \ldots n$, $j = 1, 2, \ldots, m$. Likewise, Table 2 reports the proportion of customers that are high- and low-priority for each of the locations. Table 3 reports the marginal reward

Table 3 **Rewards (Probability That a Call Is Responded to Within the RTT) Associated with Servers and High-Priority Customers**

	u_{ij}			
j	$i=1$	$i=2$	$i=3$	$i=4$
1	0.413	0.238	0.100	0.250
2	0.238	0.559	0.050	0.154
3	0.100	0.050	0.263	0.149
4	0.313	0.175	0.149	0.665

Table 4 **Probability of Survival Between Servers and Customers Having Life-Threatening Conditions**

	s_{ij}			
j	$i=1$	$i=2$	$i=3$	$i=4$
1	0.1093	0.0536	0.0265	0.0559
2	0.0607	0.1378	0.0265	0.0338
3	0.0252	0.0222	0.0709	0.0328
4	0.0811	0.0391	0.0394	0.1699

Table 5 **Example 1 Equity Measure Bounds**

Equity measure	Parameter	Bound	Value when active
1	θ_L^1	0.45	0.130
2	θ_L^2	0.06	0.0498
3	θ_L^3	0.28	0.279
3	θ_U^3	0.36	0.485
4	θ_L^4	0.03	0.0123

associated with dispatching server j to a high-priority customer at location i, where the reward represents the proportion of high-priority customers that are responded to in fewer than nine minutes. Recall that the rewards associated with dispatching servers to low-priority customers are zero. The closest server relative to customers at location i is defined as $j = i$, $j = 1, 2, \ldots, m$, such that the closest server relative to each customer yields the largest reward. The survival probabilities are computed by using the distribution of response times and the patient survival model provided by Larsen et al. (1993), who perform multiple linear regression for data from King County, Washington, which has similar demographics to Hanover County. McLay and Mayorga (2010) summarize the computational details. The resulting survival probabilities S_{ij} are summarized in Table 4.

The Example 1 bounds for each of the four equity measures are reported in Table 5 ("Bound"), where $\theta_L^1 = \theta_L^1(h, i)$, $h \in \{H, L\}$, $i = 1, 2, \ldots, n$ and $\theta_L^2 = \theta_L^2(i)$, $i = 1, 2, \ldots, n$. When each of these bounds is set to 0 (or 1 in the case of θ_U^3), the resulting formulation is not equity-constrained. Nonzero bounds do not necessarily change the optimal policy when only considering efficiency. Therefore, Table 5 reports the values of the equity bounds that make at least one of the equity constraints binding in the base model ("Value when active").

The linear programming models are formulated and solved using Matlab 10.0 and run on an Intel Xeon 3.00 GHz processor with 3.25 GB of RAM. The base case (unconstrained) linear programming model (2) has 5,626 variables and 6,673 constraints. Enforcing each equity measure adds at most eight constraints and no variables. Each linear programming model is solved in less than one minute of CPU time. All combinations of equity measures are considered, and their objective function values are reported in Table 6. Several combinations of equity measures result in an infeasible linear programming model, and hence enforcing multiple equity measures may not be possible.

In Table 6 and all subsequent results, the objective function value (g in (1)) is rescaled to be conditional on a high-priority customer arriving to report the coverage level for a nine-minute response time threshold.

Table 6 **Coverage When Applying Multiple Equity Measures for Example 1**

Equity measures	Coverage	Coverage ratio
None	0.418	1.000
1	0.409	0.979
2	0.405	0.969
3	0.407	0.974
4	0.393	0.939
1, 2	0.402	0.962
1, 3	0.407	0.974
1, 4	0.392	0.939
2, 3	Infeasible	—
2, 4	0.3896	0.932
3, 4	0.391	0.936
1, 2, 3	Infeasible	—
1, 2, 4	Infeasible	—
1, 3, 4	0.391	0.936
2, 3, 4	Infeasible	—
1, 2, 3, 4	Infeasible	—

Table 6 reports the maximum coverage level for the constrained problem as well as the ratio of the coverage level between the constrained and unconstrained problems to illustrate the degree to which the objective function value (efficiency) decreases to provide an equitable policy. As many as three equity measures could be simultaneously enforced while achieving a coverage ratio value of at least 0.936. This suggests that while enforcing an equity measure requires a decrease in efficiency, the resulting equitable policy may be quite efficient. This observation mirrors one made by Chanta et al. (2011b), who note that not all equity measures conflict with efficiency. The reduction does not seem drastic because it corresponds to no more than a 6.4% reduction from optimal. When taken in the context of number of calls covered, this corresponds to approximately 150 high-priority customers per year in the example setting that motivated this research (based on 10,000 calls per year, 53% of which are high priority). Thus, the decision maker is faced with the decision of trading efficiency for fairness.

A solution to the model prescribes which server to send to an incoming call based on the call location and priority. The entire set of policies cannot be succinctly summarized in a table because the optimal policies can depend on the full system state information (i.e., the location of the busy ambulances). Therefore, Table 7 reports the most preferred server to dispatch, which occurred in state (0; 0; 0; 0). The policies for the four equity-constrained models always first dispatch the same server as do the closest server and base model policies. There are some differences between the policies when considering the second to fourth most preferred servers to send, which are not reported by Table 7. There are considerable differences, however, across models between

Table 7 Most Preferred Server to Dispatch Across the Policies

Priority	Location	Closest server	Base model	Equity 1	Equity 2	Equity 3	Equity 4
High	1	1	1	1	1	1	1
	2	2	2	2	2	2	2
	3	3	3	3	3	3	3
	4	4	4	4	4	4	4
Low	1	1	3	1	1	1	1
	2	2	3	2	2	1	1
	3	3	3	3	3	3	1
	4	4	3	4	1	1	1

the policies for low-priority customers. Enforcing an equity measure often results in and modest changes for high-priority patients (compared to the optimal base model policy). This result is somewhat expected because the objective function only considers high-priority patients. If the objective included a coverage level (maybe at a higher RTT) for low-priority customers, then the equity constraints may be trivially met as a result of the objective function. The second and fourth equity measures are the exceptions because they are ultimately determined by the response to only high-priority patients. In sum, the four equity measure models result in policies that are the same as the base model policy between 0.94–0.99 and 0.78–0.97 of the decisions regarding high-priority and low-priority patients, respectively, when at least two ambulances are available.

4.1. Sensitivity Analysis on Equity Bounds

To fully understand how each equity measure affects the objective function values, a sensitivity analysis is performed on the equity measure bounds in (3)–(6) and in Table 5 in Example 1. Figure 1 illustrates the coverage levels. In Figure 1 and the subsequent figures, the x-axes are aligned such that the problem becomes more equity constrained when moving from left to right, with the leftmost point in each figure representing the coverage level for the unconstrained problem. For reference, the value of the closest server model is shown in Figure 1. All equity measures decrease the coverage level once an equity constraint becomes active. However, the resulting policies may still offer improvements over other myopic policies that could be used, and hence, optimal equity-constrained dispatching policies may offer benefits above similar fair heuristic policies.

Figure 2 illustrates the values associated with each of the equity measures. The equity measure 1

Figure 1 Response Time Threshold Coverage Level (Objective Function Values) When Enforcing Each Equity Measure

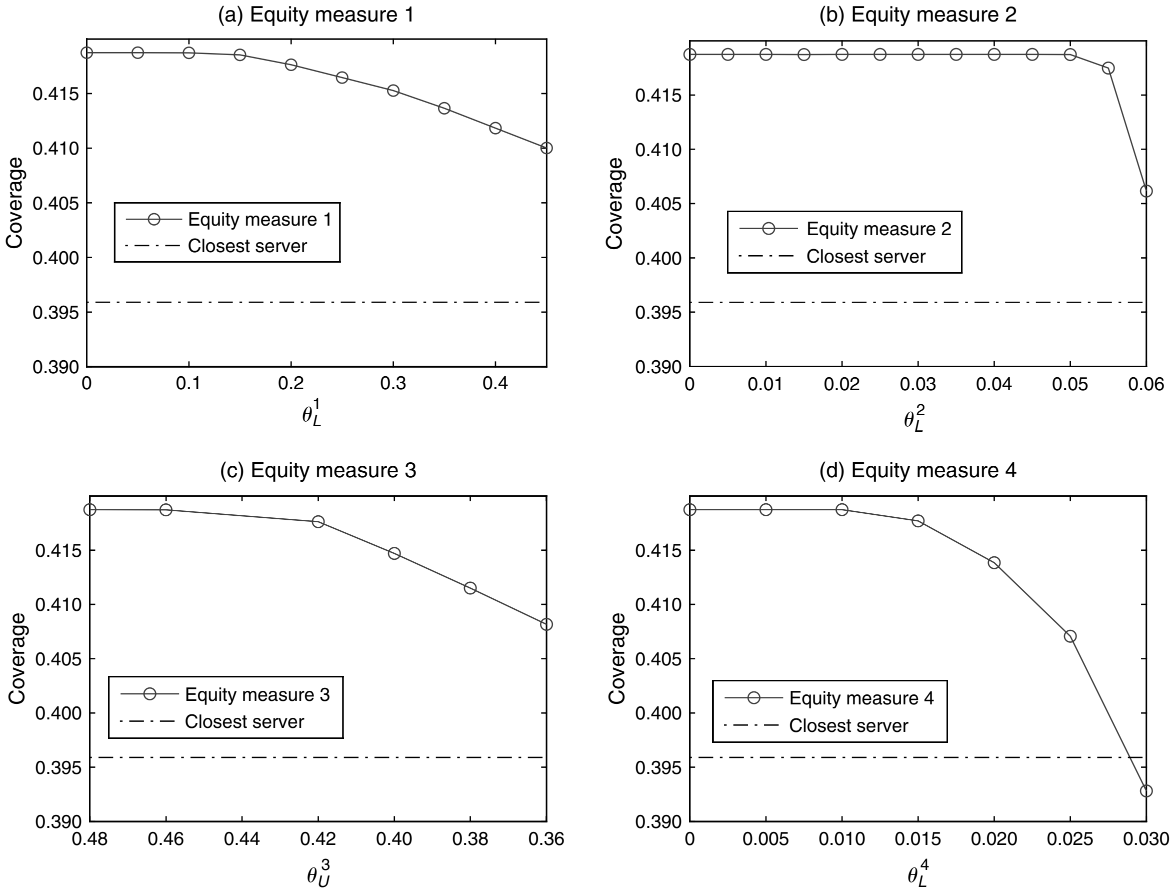
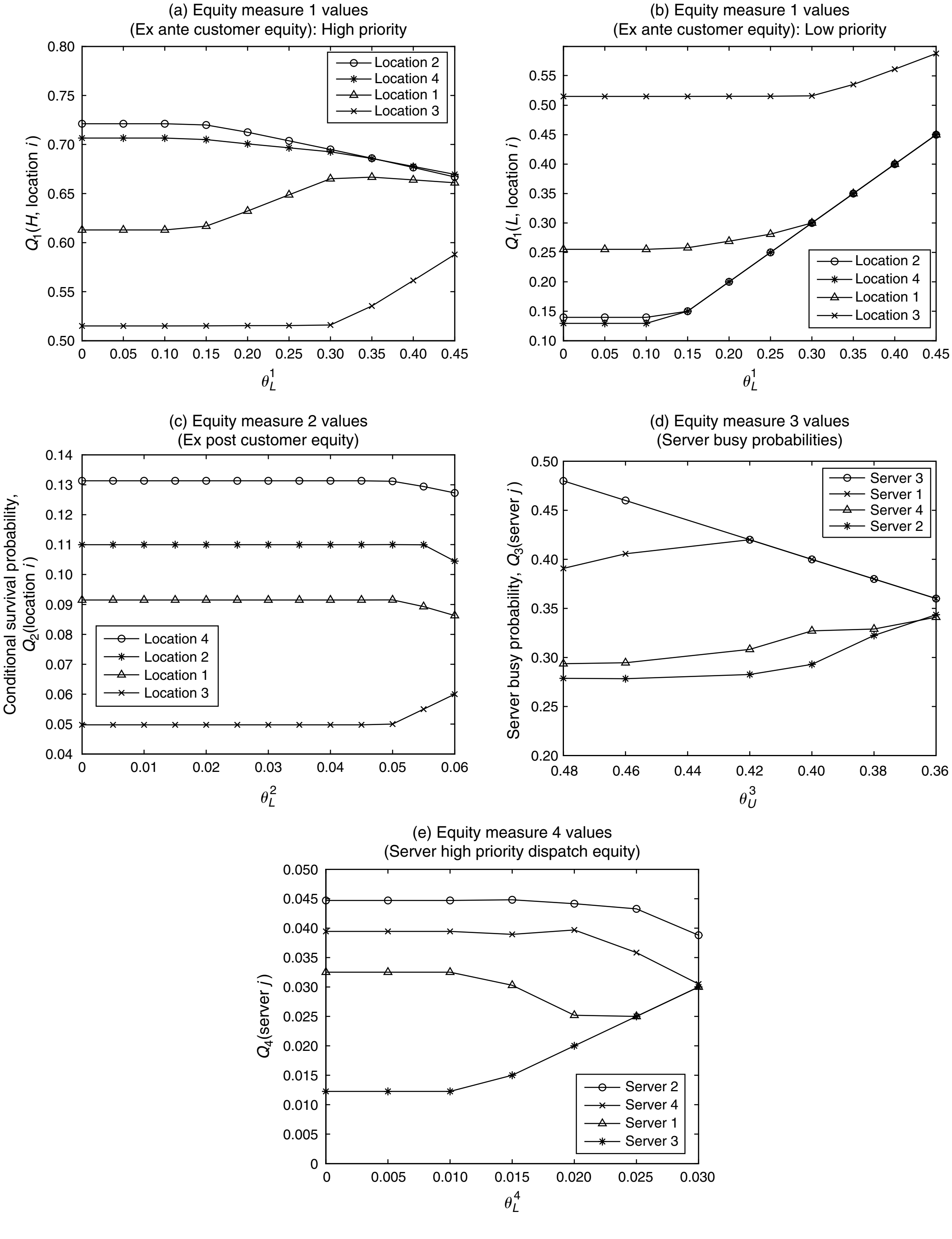


Figure 2 Equity Measure Values



(a) Equity measure 1 values (Ex ante customer equity): High priority

(b) Equity measure 1 values (Ex ante customer equity): Low priority

(c) Equity measure 2 values (Ex post customer equity)

(d) Equity measure 3 values (Server busy probabilities)

(e) Equity measure 4 values (Server high priority dispatch equity)

model solution does not vary from the optimal unconstrained model solution when $\theta_L^1 \leq 0.1$ (see Figure 2(b)). The equity constraints increase the proportion of time that the closest servers respond to low-priority customers and, to a lesser extent, the high-priority customers. For the most part, low-priority customers at locations 2 and 4 necessitate changes in the dispatching policy. Equitably serving customers at these locations is achieved by decreasing the proportion of time that the closest server is dispatched to high-priority customers at these two locations, illustrated in Figure 2(a). Figure 2(a) shows a nonlinear, nonmonotonic relationship between θ_L^1 and the proportion of time that the closest server is dispatched to high-priority customers at location 1. This observation suggests that it is not always clear whether selected values for the equity measure bounds result in relative improvements or declinations for all stakeholders.

Figure 2(c) illustrates values of equity measure 2 (ex post customer equity), and it indicates that the customer survival probabilities are largely a function of location. Certain locations are difficult for servers to reach (e.g., location 3), and therefore it is difficult to enforce minimum levels of patient survival at each location without leading to an infeasible solution, which happens when θ_L^2 increases beyond 0.06 (see the survival probabilities in Table 4). Even when equity is enforced, it is impossible to equalize patient survival probabilities across the four locations unless patients at locations 1, 2, and 4 are not treated, which would be problematic from a public service point of view. Figure 2(d) illustrates values of equity measure 3 (server busy probability equity) as a function of θ_U^3. In contrast to the survival probabilities in Figure 2(c), Figure 2(d) shows that the server busy probabilities are able to be equalized.

Figure 2(e) illustrates values of equity measure 4 (server high-priority dispatch equity). Although it is possible to equalize the rates at which the servers are dispatched to high-priority customers, doing so results in a sharp decrease in the objective function value (see Figure 1(d)). This occurs because the resulting policies tend to dispatch servers to distant high-priority patients, which have almost no chance of meeting a nine minute response time threshold.

4.2. Evaluating Equity Proxies

The degree to which one equity measure is a proxy for a second equity measure is useful when determining which single equity measure may be the most effective in satisfying multiple notions of equity. For comparison, the base case model and closest server model are also considered, thus yielding six models for comparison. Only one equity measure is applied to the model at a time. Average values of $Q_1(h, i)$, $Q_2(i)$, $Q_3(j)$, and $Q_4(j)$ are obtained to identify representative values across 30 replications. For each replication of equity measure k (with $k = 1, 2, 3, 4$), a random equity bound is generated that is uniformly between the equity bound and the value when one of the equity constraints becomes active (see Table 5). For example, for equity measure 1, θ_L^1 is uniformly distributed between 0.13 and 0.45.

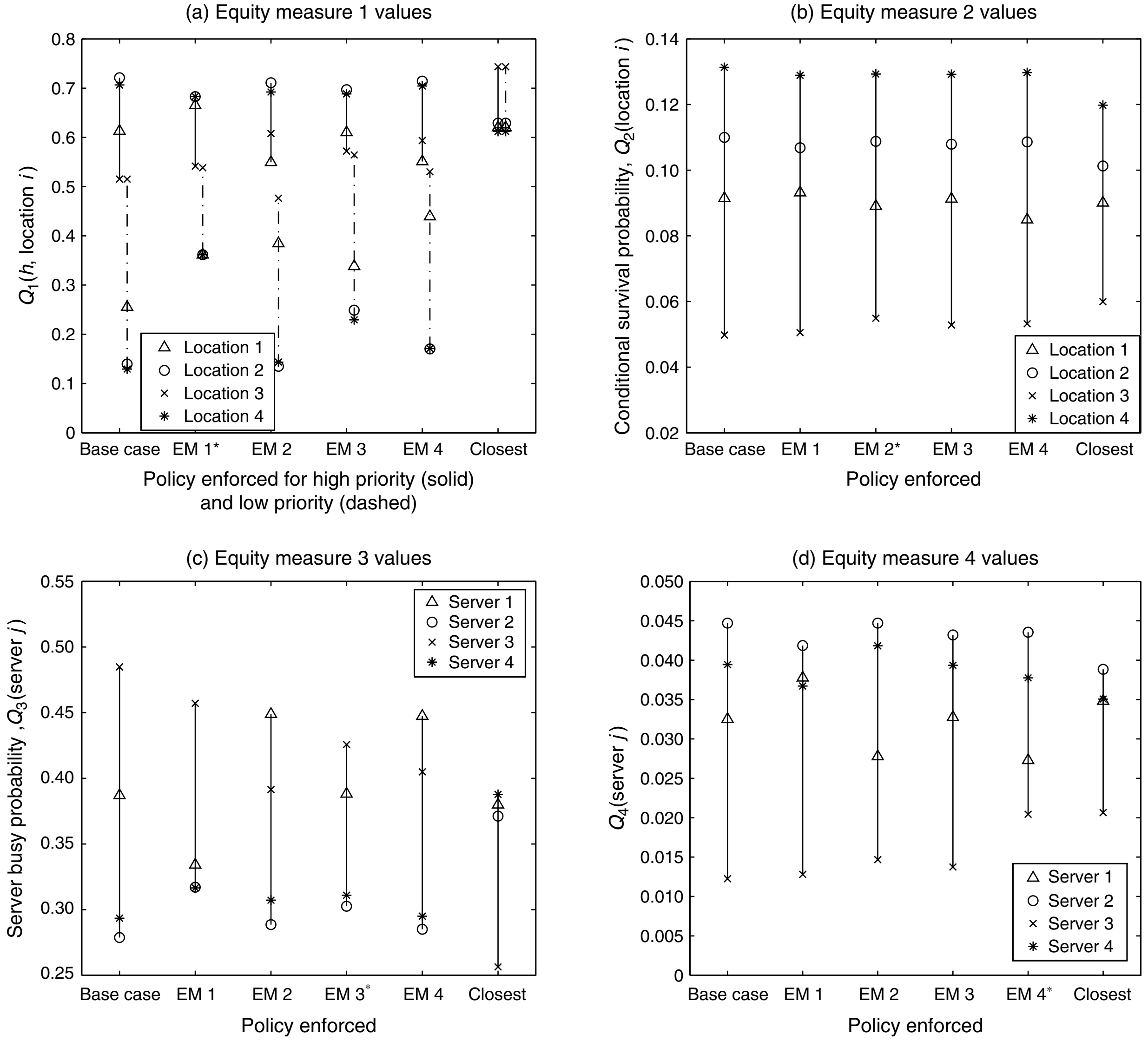
Figure 3 illustrates the average values of $Q_1(h, i)$, $Q_2(i)$, $Q_3(j)$, and $Q_4(j)$ for each of the six models across 30 replications of each equity bound (when appropriate). The equity measure models are abbreviated as "EM," and an asterisk next to a model indicates whether it matches the equity measure of interest. The bars show the ranges and the points show the average value for each location/server. Note that the average equity measures appear to be similar to those of the base model because some of the random equity bounds are not much more constraining than the base model. Refer to Figure 2 to see the possible ranges for each equity measure.

First consider the equity measures that evaluate customer fairness. Figure 3(a) shows the average values of $Q_1(h, i)$, $h \in \{H, L\}$, $i = 1, 2, 3, 4$, with high-priority customers captured by the solid line and low-priority customers captured by the dashed line. Each of the models except for the closest server model dispatches nearby servers to high-priority customers more frequently than to low-priority customers. Figure 3(b) shows the average values $Q_2(i)$, $i = 1, 2, 3, 4$, which are similar across all of the models (for each value of i) because patient survival is largely a function of the distance between servers and customers. The closest server model, however, significantly decreases the likelihood of survival for customers at locations 2 and 4.

Figures 3(c) and 3(d) show the two equity measures that evaluate server fairness. The average busy probabilities in Figure 3(c) show that the type of model has a large effect on the likelihood that each server is available. For example, server 3 is busiest the least often for the closest server model, whereas it is the first or second busiest server in each of the other five models. Server 2 has a high busy probability for the closest server model, whereas it is the least busy server in each of the other five models. Figure 3(d) shows that server 2 is most frequently dispatched to high-priority customers in all six models. Both equity measure 4 and the closest server model tend to equalize the rates at which the four servers are dispatched to high-priority patients, which is shown by the smaller ranges as compared to the other three models.

It is notable that the unconstrained base model yields the largest ranges in all four equity measure models. Enforcing an equity measure may reduce these ranges, thus yielding beneficial effects on other equity measures that are not explicitly taken into account via optimization. This supports the notion that enforcing different notions of equity is not analogous to a zero-sum game, in which enforcing one equity measure results in worse values/ranges of other equity measures. The closest server model reduces the ranges associated with equity measures 1, 2, and 4 as compared to the other models. In general settings, the closest server policy may introduce—rather than reduce—inequities. This issue is explored in more detail in the next section.

Figure 3 **Average Values of the Equity Measure Values When Implementing the Unconstrained Base Case Policy, the Closest Server Policy, and 30 Replications of the Four Equity Measure (EM) Policies**



4.3. Sensitivity Analysis on Customer Locations

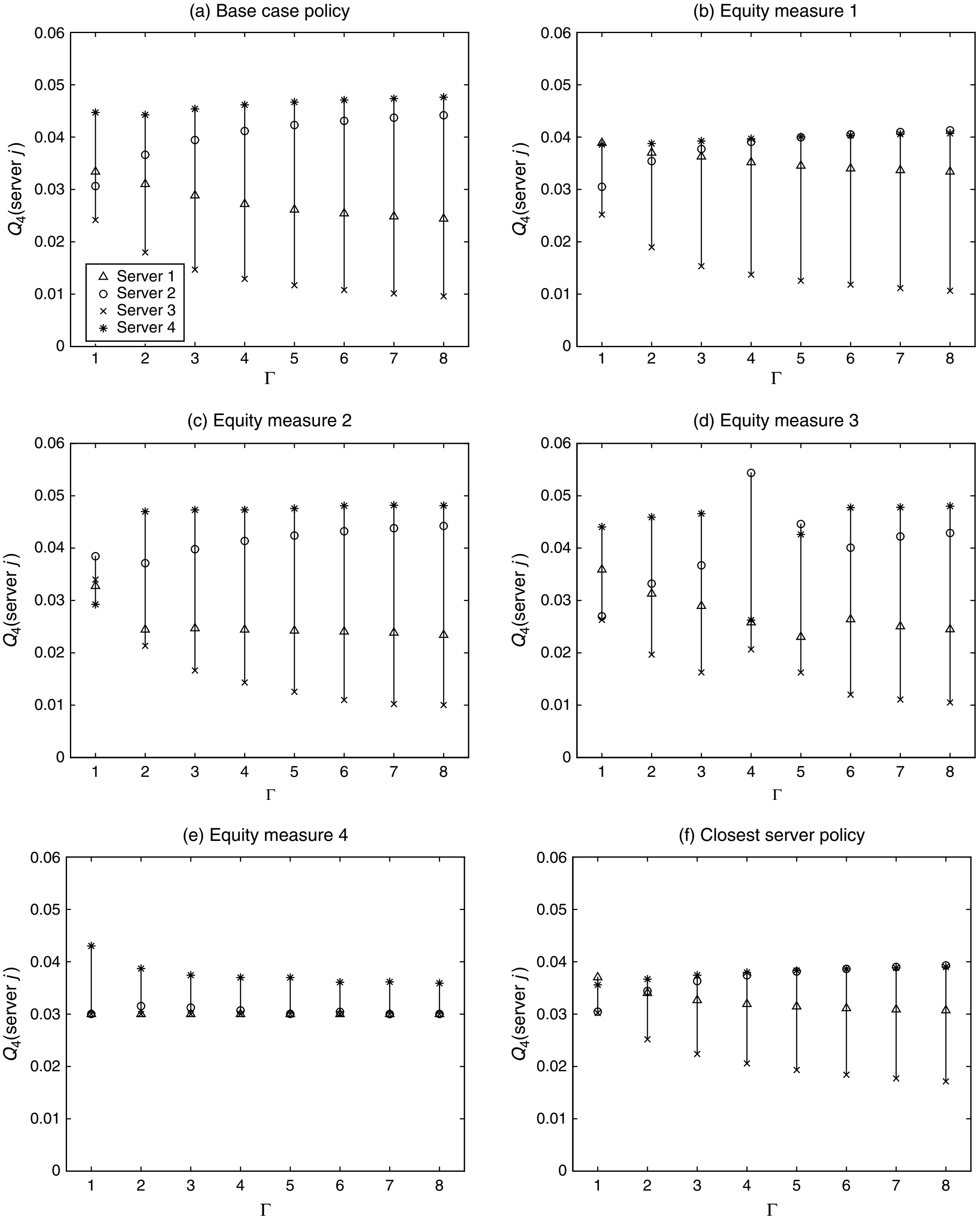
A second example ("Example 2") is analyzed to examine whether the observations in the previous section are valid in other settings. In particular, we study the effect that customer locations (geographic dispersion) have on equity measure 4 while utilizing the realistic parameters in Example 1. Although not reported here, similar observations can be made for equity measures 1–3. We use the same input parameters as reported in Tables 1–4 while varying the conditional probability that a customer arrives at location i. In Example 1, customers are more likely to arrive at locations 2 and 4 than they are at locations 1 and 3. We vary the ratio of customers at locations 2 and 4 to the customers at 1 and 3 as follows. Let $P_2 = P_4$ and $P_1 = P_3$, and let

$$\Gamma = P_2/P_1 = P_4/P_3,$$

with $1 \leq \Gamma \leq 8$. When $\Gamma = 1$, the customers are uniformly located at each of the four locations. When $\Gamma = 8$, customers are eight times more likely to be located at locations 2 and 4 as compared to locations 1 and 3. Deterministic equity bounds of $\theta_L^1 = 0.45$, $\theta_L^2 = 0.055$, $\theta_L^3 = 0.3$, $\theta_U^3 = 0.4$, and $\theta_L^4 = 0.03$ are used across all values of Γ. The equity bounds are more constraining for some values of Γ than for others, and therefore, the feasible regions vary in size across Γ. Figure 4 illustrates $Q_4(j)$—the rate at which each of the servers is dispatched to high-priority patients—across Γ for each of the six models. As in Figure 3, the bars show the ranges and the points show the values of $Q_4(j)$ for each server.

Figure 4(a) shows that under the base model, the range of $Q_4(j)$ values across the servers grows as

Figure 4 Values of the Equity Measure 4 Values When Implementing the Unconstrained Base Case, the Closest Server, and the Four Equity Measure (EM) Models, $\Gamma = 1, 2, \ldots, 8$



Γ increases. In particular, servers 2 and 4 respond to more high-priority patients as Γ increases where servers 1 and 3 respond to fewer high-priority patients. This trend is observed for all of other models except the equity measure 4 model (see Figures 4(b–d, f)). On the other hand, the equity measure 4 model provides a floor for $Q_4(j)$, thus reducing the range of $Q_4(j)$ values as Γ increases (see Figure 4(e)). In sum, Figure 4 illustrates that equity measure 4 is not enforced by other equity measures.

5. Discussion and Conclusions

This paper models and analyzes equity-constrained dispatching policies in server-to-customer systems with customer priorities. A constrained, undiscounted, infinite horizon, average-cost Markov decision process model is formulated as a linear programming model to identify optimal dispatching policies. This methodology is novel in that it investigates how various equity formulations degrade the objective function value (efficiency) and change the underlying dispatching policies. Four equity measures are considered. Two of the equity measures consider equity from the customer (patient) perspective, whereas the other two equity measures consider equity from the server (service provider) point of view.

The analysis of two examples suggests that one notion of equity could simultaneously improve equity for both customers and servers in some scenarios. This observation is important from a modeling point of view, because simultaneously enforcing multiple equity measures often leads to infeasible solutions. The equity modeling paradigm selected should be congruent with the characteristics of the setting under consideration. We acknowledge that it may not be realistic to attempt to equalize some aspects of system performance across heterogeneous regions, such as across urban and rural regions. However, the first equity measure that focuses on ex post resource allocation may be appropriate to use in regions with large population density differences because it evaluates a relative measure of equity, whereas the other three equity measures may be difficult to "equalize" in these regions. The analysis also suggests that it may be difficult to impose minimum survival probabilities for patients in different locations via dispatching alone. This suggests that it could also be difficult to enforce minimum levels of service across various demographics and groups, when service reflects ex post customer equity. When considering patient outcomes such as survival, solutions with disparate outcomes across different different groups are undesirable from a political point of view.

Implementing the policies requires some discussion. The optimal policies almost always dispatch servers in a priority list, meaning that the order of ambulances to send to a patient with a specified priority at a specified location does not depend on the full Markov state. A priority list is often called a "contingency table" in EMS systems, where they are widely used because they are intuitive and easy to implement. The closest server heuristic, a policy used in many EMS systems, is itself a priority list (closest, next closest if closest is not available, and so forth). Implementing another policy that conforms to a priority list would not be any more difficult to implement in practice. All of the equity-constrained policies in Example 1 correspond to a priority list more than 99.9% of the time. Therefore, the equity-constrained policies could be retrospectively adjusted to conform to a priority list all of the time.

In addition to improving or equalizing service, positive outcomes from implementing the server equity measures could include more satisfied staff members and better-trained personnel (Key 2002, Studnek and Fernandez 2007). EMS departments often report service at the district level (e.g., district-level coverage levels may be published on county websites). Introducing equitable policies that improve performance at the district level could increase support for emergency medical services. However, there are several drawbacks of implementing an equity-constrained policy. The first is that service can be worse for all, as was noted in the discussion surrounding equity measure 4. Second, negative outcomes will occur no matter which policies are used. It may be more difficult to justify negative patient outcomes when using anything other than the most efficient policy.

The models in this paper are based on low-congestion settings under "normal" conditions and restrict their attention to nonidling policies with a zero-length queues. The policies may be different if the nonidling policy assumption is lifted. In the example considered, the optimal policy for the base model would be to ignore all low-priority calls because they do not contribute to the objective function value. The optimal policy may be a nonidling policy when the rewards are strictly positive. In the equity-constrained models, the optimal policies may allow for idling because it may be better to allow servers to idle more frequently to ensure that the equity constraints are satisfied. One extension is to understand the impact of idling policies on equity, and another extension is to examine the impact of allowing customer queues. Queues should be included in the model in congested systems when the Erlang loss rate exceeds an acceptable threshold (e.g., 10%). An $M/M/m$ queue can be used to estimate the loss rate without solving an optimization model to provide guidance as to when to include queues in the model.

An EMS system's Standard Operating Procedures, which prescribe operations under "normal" and emergency scenarios, may also be used to provide guidance for when queues and idling should be included in the model. Typically, a nonidling policy with a priority queue is used until the congestion level becomes too high, when an EMS system switches to an operating paradigm that allows for priority queuing and idling. This congestion level can be used to inform the modeling choices. This also motivates models for measuring and enforcing fairness during emergency conditions because notions of fairness during emergency conditions may be different than notions of fairness in normal conditions.

In this paper, equity is enforced via side constraints within a linear programming model for an MDP rather than via a secondary objective. Another extension is to analyze a multiobjective model that balances equity and efficiency, as has been done in the context of locating ambulances (Chanta et al. 2011a). We reiterate that the equity measures considered are not exhaustive. For example, this paper does not consider equity across all potential stakeholders, process equity, or how equity is achieved over time, which would certainly lead to new types of resource allocation decisions. Finally, the simplifying assumptions can be lifted to consider patient queues, rerouting ambulances that have been dispatched prior to when they arrive at the scene, and limited look-ahead policies that would dispatch ambulances that are about to leave from the hospital or another patient location. Work is in progress to address these extensions.

Acknowledgments

The authors acknowledge Chief Fred C. Crosby II, Battalion Chief Henri Moore Jr., Lawrence Roakes, and 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. In addition, the authors thank the anonymous reviewers and the journal's area editor for their valuable comments and suggestions that have resulted in a substantially improved manuscript. The first author was supported by the U.S. Department of the Army [Grant Award W911NF-10-1-0176] and by the National Science Foundation [CMMI-1054148]. 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 or the National Science Foundation.

References

Armony M, Ward A (2010) Fair dynamic routing in large-scale heterogeneous-server systems. *Oper. Res.* 58(3):624–637.

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

Boudreaux E, Mandry C, Brantley PJ (1997) Stress, job satisfaction, coping, and psychological distress among emergency medical technicians. *Prehospital Disaster Management* 12(4):242–249.

Carter GM, Chaiken JM, Ignall E (1972) Response areas for two emergency units. *Oper. Res.* 20(3):571–594.

Chanta S, Mayorga ME, McLay LA (2011a) Improving emergency service in rural areas: A bi-objective covering location model for EMS systems. *Ann. Oper. Res.* DOI:10.1007/s10479-011-0972-6.

Chanta S, Mayorga ME, Kurz ME, McLay LA (2011b) The minimum *p*-envy location problem: A new model for equitable distribution of emergency resources. *IIE Trans. Healthcare Systems Engrg.* 1(2):101–115.

Chelst KR, Barlach Z (1981) Multiple unit dispatches in emergency services: Models to estimate system performance. *Management Sci.* 27(12):1390–1409.

Clawson JJ, Martin RL, Cady GA, Sinclair R (1999) EMD: Making the most of EMS. *Fire Chief* (June 1), http://firechief.com/ems/firefighting_emd_making_ems/.

Dean SF (2008) Why the closest ambulance cannot be dispatched in an urban emergency medical services system. *Prehospital Disaster Medicine* 23(2):161–165.

Erdogan G, Erkut E, Ingofsson A, Laporte G (2010) Scheduling ambulance crews for maximin coverage. *J. Oper. Res. Soc.* 61(4):543–550.

Felder S, Brinkmann H (2002) Spatial allocation of emergency medical services: Minimising the death rate or providing equal access? *Regional Sci. Urban Econom.* 32(1):27–45.

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

Henderson SG (2011) Operations research tools for addressing current challenges in emergency medical services. Cochran JJ, ed. *Wiley Encyclopedia of Operations Research and Management Science* (John Wiley & Sons, Hoboken, NJ).

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

Keeney RL, Winkler RL (1985) Evaluating decision strategies for equity of public risks. *Oper. Res.* 33(5):955–970.

Key CB (2002) Operational issues in EMS. *Emergency Medical Clinics North America* 20:913–927.

Larsen MP, Eisenberg MS, Cummins RO, Hallstrom AP (1993) Predicting survival from out-of-hospital cardiac arrest—A graphic model. *Ann. Emergency Medicine* 22(11):1652–1658.

Leclerc PD, McLay LA, Mayorga ME (2011) Modeling equity for allocating public resources. Johnson MP, ed. *Community-Based Operations Research: Decision Modeling for Local Impact and Diverse Populations* (Springer, New York), 97–118.

Marsh MT, Schilling DA (1994) Equity measurement in facility location analysis: A review and framework. *Eur. J. Oper. Res.* 74(1):1–17.

Maxwell MS, Henderson SG, Topaloglu H (2010a) Tuning approximate dynamic programming policies for ambulance redeployment via direct search. Technical report, Cornell University, Ithaca, NY.

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

McLay LA, Mayorga ME (2010) Evaluating emergency medical service performance measures. *Health Care Management Sci.* 13(2):124–136.

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

Puterman ML (1994) *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (John Wiley & Sons, New York).

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

Sarin RK (1985) Measuring equity in public risk. *Oper. Res.* 33(1):210–217.

Savas ES (1978) On equity in providing public services. *Management Sci.* 24(8):800–808.

Stiell IG, Wells GA, Field BJ (1999) 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. *New England J. Medicine* 351(7):647–656.

Stone D (2002) *Policy Paradox: The Art of Political Decision Making* (W. W. Norton & Company, New York).

Studnek J, Fernandez AR (2007) Non-urgent is no fun. *J. Emergency Medical Services* 32(10):38.

Swersey AJ (1982) A Markovian decision model for deciding how many fire companies to dispatch. *Management Sci.* 28(4):352–365.

Weintraub A, Aboud J, Fernandez C, Laporte G, Ramirez E (1999) An emergency vehicle dispatching system for an electric utility in Chile. *J. Oper. Res. Soc.* 50(7):690–696.