European Journal of Operational Research 279 (2019) 557–571

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

Robust and stochastic formulations for ambulance deployment and dispatch

Dimitris Bertsimas[a,*], Yeesian Ng[b]

[a] Sloan School of Management, Massachusetts Institute of Technology, 77 Massachusetts Avenue, E40-130, Cambridge, MA02139, USA
[b] Operations Research Center, Massachusetts Institute of Technology, 77 Massachusetts Avenue, E40-130, Cambridge, MA02139, USA

ARTICLE INFO

Article history:
Received 21 December 2018
Accepted 10 May 2019
Available online 16 May 2019

Keywords:
OR in Health Systems
Emergency medical systems
Ambulance location
Robust optimization
Data-driven optimization

ABSTRACT

In Emergency Medical Systems, operators deploy a fleet of ambulances to a set of locations before dispatching them in response to emergency calls, with the goal of minimizing the fraction of calls with late response times. We propose stochastic and robust formulations for the ambulance deployment problem that use data on emergency calls to model uncertainty. By incorporating advances in column and constraint generation, our formulations are solved to exact optimality within minutes. In extensive computational experiments on Washington DC, our approach outperforms previous approaches (i.e. the MEXCLP and MALP) that rely on probabilistic assumptions about the availability of ambulances. Our formulations achieve a reduction of 19 to 28% in number of shortfalls, requiring only 70% of the total number of ambulances required in probabilistic models to attain comparable out-of-sample performance.

1. Introduction

Emergency Medical Services (EMS) systems have drawn a great deal of attention from researchers. While the public expects the availability of EMS facilities to provide timely services, this expectation is hard to realize due to limited available resources and stringent governmental budgets. Rising costs of medical equipment, increasing call volumes, and worsening traffic conditions have placed EMS providers under increasing pressure to meet performance goals set by regulators. A key indicator of these performance goals is medical response time, due to its relationship to specific time sensitive conditions such as out-of-hospital cardiac arrest, stroke and severe trauma cases.

Prior to receiving any emergency calls, ambulances are usually positioned within a set of pre-determined locations such as parking lots, hospitals, fire stations, or on the move when returning from servicing a call. When emergency calls arrive, ambulance operators might have to elicit the location of the call from the caller through landmarks or street descriptions. Therefore, the emergency calls are often modeled as arising from a fixed set of demand regions, after which the ambulances might be required to re-stock on emergency supplies.

Operational planning by EMS providers considers short-term decisions such as for ambulance dispatch and dynamic ambulance relocations. Tactical planning involves medium term decision horizons which typically establish baseline deployment plans and manpower shift schedules. Strategic planning involves longer term decision horizons such as the location of ambulance stations and ambulance fleet dimensioning. In this work, we focus on ambulance deployment at the tactical level and guide operational decisions to be robust against short term uncertainties.

1.1. Previous work

Early ambulance deployment models focus on static policies for tactical planning. Toregas, Swain, ReVelle, and Bergman (1971) formulated the set covering location problem (SCLP), which aims to minimize the number of ambulances needed to cover a given region. Church and ReVelle (1974) formulated the maximal covering location problem (MCLP), which aims to maximize the demand that can be covered given a fixed number of ambulances. Both the SCLP and MCLP consider single coverage in which a given point is covered if it can be reached within a response time threshold by an ambulance. However, both the SCLP and MCLP do not account for the possibility that a particular ambulance will be "busy" in the event of concurrent emergency calls. Therefore, subsequent formulations such as backup coverage models by Hogan and ReVelle (1986) and double coverage models by Hogan and ReVelle (1986) and Gendreau, Laporte, and Semet (1997), were introduced. There are also formulations that model ambulance availability from a probabilistic perspective. Daskin (1983) approximate the expected value of coverage by introducing a "busy fraction"

[*] Corresponding author.
E-mail address: dbertsim@mit.edu (D. Bertsimas).

https://doi.org/10.1016/j.ejor.2019.05.011
0377-2217/© 2019 Elsevier B.V. All rights reserved.

as a proxy for the probability that a given ambulance will be unavailable, and formulated the maximum expected covering location problem (MEXCLP). However, the “busy fraction” was assumed to be constant across all sites in the MEXCLP, which is not a realistic assumption. Therefore, it was extended by Batta, Dolan, and Krishnamurthy (1989), ReVelle and Hogan (1989), and Ball and Lin (1993) to account for site-specific probabilities. This resulted in nonconvex formulations based on steady-state probabilities of queuing systems which are heuristically solved through approximations. For an extensive review of models in the ambulance deployment literature, we refer the interested reader to Brotcorne, Laporte, and Semet (2003), Li, Zhao, Zhu, and Wyatt (2011) and Bélanger, Ruiz, and Soriano (2019).

A related question is on the choice of dispatch rules: when an incident occurs, should the ambulance that is closest be dispatched? Although the “closest-idle policy” is suboptimal (Carter, Chaiken, & Ignall, 1972), there are few alternatives that have been suggested. One exception is the notion of a regionalized response by Swoveland, Uyeno, Vertinsky, and Vickson (1973), where ambulances serve their allocated region first and the closest-idle ambulance is sent only if none in the region are unavailable. Bandara, Mayorga, and McLay (2012); McLay and Mayorga (2013) developed dispatch policies with prioritized patients to reduce response times for urgent patients at the expense of longer times for non-urgent requests. Nonetheless, it does not fully address how to dispatch vehicles to minimize late arrivals when *all* patients have high priority. Recent models by Alanis, Ingolfsson, and Kolfal (2013); Gendreau, Laporte, and Semet (2001, 2005); Maxwell, Restrepo, Henderson, and Topaloglu (2010); Naoum-Sawaya and Elhedhli (2013) focus on operational planning that repositions “idle” ambulances in real-time to better respond to future calls. This led Maxwell et al. (2014) to establish bounds on the performance of an optimal ambulance redeployment policy. However, none of them address the question of whether repositioning was required because of a suboptimal initial allocation of ambulances: it could be that an “optimal” static allocation of ambulances might attain similar benefits, obviating the need to reposition ambulances in real-time.

More recently, Beraldi and Bruni (2009); Beraldi, Bruni, and Conforti (2004); Zhang and Jiang (2014) revisited the static ambulance location problem and modelled the assignment of vehicles to emergency demands using chance constraints or robust optimization. However, since the models rely on strong restrictions on the dispatch policy for tractability, and often result in deployment plans that are overly conservative. On the other hand, fully adaptive models of dispatch often suffer from the curse of dimensionality and are typically computationally intractable (Dyer & Stougie, 2006; Shapiro & Nemirovski, 2005), and rely on methods that exploit problem structure for tractability (Bertsimas & Dunning, 2016; Bertsimas & Georghiou, 2018; Billionnet, Costa, & Poirion, 2014; Hanasusanto, Kuhn, & Wiesemann, 2015; Postek & Hertog, 2016).

With recent advances in exact solution techniques using column and constraint generation (Zeng & Zhao, 2013), large instances of fully adaptive robust formulations can now be solved within minutes. Given improvements in the quality of EMS data available, we are interested whether stochastic and robust formulations of deploy-and-dispatch models might outperform previous formulations (such as the MEXCLP and MALP) that are based on probabilistic assumptions about ambulance availability. In contrast to the literature on ambulance redeployment, we are interested in minimizing the fraction of late-arrivals, based on the closest-idle dispatch policy *without requiring ambulances to be repositioned*. Given the observation by Jagtenberg, Bhulai, and van der Mei (2017) of a strong tradeoff between minimizing the fraction of late-arrivals versus response times, we demonstrate improvements in both measures via an exact mathematical optimization approach that we have not seen in the literature so far.

Although there is a connection in the two-stage integer formulation between our models, and the stochastic models by Beraldi and Bruni (2009); Beraldi et al. (2004), their models are based on probabilistic constraints, whereas our models are based on adaptive recourse functions. This makes our approach amenable to a robust formulation different from the approach by Zhang and Jiang (2014), that allows for integer uncertainty and recourse. Gabrel, Lacroix, Murat, and Remli (2014) studied a similar recourse function in the setting of location-transportation problems, and took the same approach of linearizing the inner bilevel maximation problem. Our work differs from theirs in the choice of application, uncertainty sets, and solution algorithm.

1.2. Our contribution

This paper makes the following contributions:

1. We propose tractable formulations for static ambulance deployment, namely: stochastic and robust two-stage planning models with fully adaptive recourse.
2. We adopt a data-driven approach to construct structured uncertainty sets that take into account demand interactions across multiple local and regional levels jointly.
3. We provide extensions to our deployment models, that depart from traditional models of threshold coverage, to include notions of partial coverage based on response times.
4. Through realistic computational experiments, we demonstrate improvements in model performance over competitive models in the literature across multiple regimes of ambulance demand and availability, and provide reasons for the observed improvements.

The rest of the paper is organized as follows. We review the literature in Section 1.1 on both ambulance deployment models, and multi-stage optimization. Next, we introduce the notation and preliminary concepts that are used throughout the paper in Section 2. We describe the models of ambulance deployment with recourse, and provide an algorithm for solving it in Section 3. We perform a realistic case study on the EMS system in Washington DC, and present an assessment of the proposed models through computational results in Section 4. We provide an explanation of the observed improvements in Section 4.2, and conclude in Section 5.

2. From online dispatch to ambulance deployment

We use boldface letters for vectors and matrices; e.g. $\mathbf{1}$ and $\mathbf{0}$ to denote the vector of 1’s and 0’s accordingly. We use $\mathbf{x}^\top$ to refer to the transpose of $\mathbf{x}$. In addition, we define $(\cdot)^+ := \max\{\cdot, \mathbf{0}\}$, $[n] := \{1, \dots, n\}$, $|S|$ as the cardinality of set S, and the α-quantile $\mathrm{V@R}_\alpha(x) := \inf\{\ell \in \mathbb{R} : \mathbb{P}(x > \ell) \le 1 - \alpha\}$.

In this section, we introduce some notation (see Table 1) and review the basic setup of an EMS system. We define ambulance deployment models on directed graphs $G = (V, E)$, where the set of vertices V can be partitioned $V = (I, J)$ into the set of deployment locations I, and city regions J. We represent the network structure of regional accessibility through the node-arc adjacency matrix $\mathbf{B} \in \{-1, 0, 1\}^{|V| \times |E|}$, where

$$b_{ik} = \begin{cases} 1, & \text{if } i \text{ is the start of the } k\text{th edge}, \\ -1, & \text{if } i \text{ is the end of the } k\text{th edge}, \\ 0, & \text{otherwise.} \end{cases}$$

Given the partition $V = (I, J)$, we can decompose $\mathbf{B}$ into submatrices $\mathbf{B}_I$ and $\mathbf{B}_J$ (See Appendix A for an example). Our models aim to find the optimal way to locate n ambulances at a set of potential locations I, to minimize the expected shortfall, by making

Table 1
Summary of the notation.

Symbol	Description	Indices
I	Set of ambulance deployment locations	
J	Set of ambulance demand regions	
m	Number of scenarios/time-periods	
t_{ij}	Time (minutes) to travel from i to j	$i \in I, j \in J$
$\bar{t}$	Response time threshold (minutes)	
E	Set of directed edges (i, j) satisfying $t_{ij} \le \bar{t}$	
I_j	Set of locations $i \in I$ connected to region j	$j \in J$
J_i	Set of regions connected to location i	$i \in I$
δ_j	Set of regions geographically adjacent to j	$j \in J$
n	Maximum number of ambulances	
τ	Maximum time (minutes) to complete a run	
d_j	The number of calls from region j	$j \in J$
γ_i	Bound on total calls reachable from i	$i \in I$

here-and-now decisions $\mathbf{x} = (x_i)_{i \in I}$, where x_i is the number of ambulances to be deployed at node $i \in I$, on the basis of *wait-and-see* variables $\mathbf{y} = (y_{ij})_{(i,j) \in E}$, where y_{ij} is the number of ambulances to dispatch from location $i \in I$ to region $j \in J$.

2.1. Ambulance scheduling and dispatch

We consider an arrival process $A : [0, t^{\max}] \mapsto \mathbb{Z}_+^{|J|}$, where $t^{\max} = m\tau$ is the duration of the entire time period, and $A(t)$ denotes the cumulative number of emergency calls received by time t. All ambulances take $\bar{t}$ minutes to reach the site of each call, and τ minutes to become available after being dispatched. We are interested in the performance of the static deployment policy $\mathbf{x}$ over the entire time period $[0, t^{\max}]$, and proceed by partitioning it into smaller time periods

$$(0, \tau], (\tau, 2\tau], \ldots, (t^{\max} - \tau, t^{\max}],$$

before optimizing for the demand $A(i \cdot \tau) - A((i-1) \cdot \tau)$ in each time period $i = 1, 2, \ldots, m$.

Focusing on a single time period $[t^{(0)}, t^{(0)} + \tau] \subset [0, t^{\max}]$, let $\mathbf{x}^{(0)} \in \mathbb{Z}_+^{|I|}$ be the initial assignment of ambulance availability, and $\mathbf{d}^{(1)}, \ldots, \mathbf{d}^{(k)} \in \mathbb{Z}_+^{|J|}$ be a sequence of requests at times $t^{(1)}, \ldots, t^{(k)}$ such that $t^{(0)} \le t^{(1)} < t^{(2)} < \cdots < t^{(k)} \le t^{(0)} + \tau$. If an ambulance is dispatched at any time t in $[t^{(0)}, t^{(0)} + \tau]$, it will remain unavailable for the rest of the time period $[t, t^{(0)} + \tau]$. Therefore, if we run out of ambulances and "queue" a call for the next available ambulance, it will take more than τ minutes to respond to the call. Correspondingly, for each emergency request $\mathbf{d}^{(i)}$, with a response of $\mathbf{y}^{(i)}$ (satisfying $\mathbf{B}_I \mathbf{y}^{(i)} \le \mathbf{x}^{(i-1)}$) ambulances dispatched, there will be a non-negative *shortfall* of $\mathbf{d}^{(i)} + \mathbf{B}_J \mathbf{y}^{(i)}$ incurred, and

$$\begin{aligned} \mathbf{x}^{(i)} &= \mathbf{x}^{(i-1)} - \mathbf{B}_I \mathbf{y}^{(i)} \\ &= (\mathbf{x}^{(i-2)} - \mathbf{B}_I \mathbf{y}^{(i-1)}) - \mathbf{B}_I \mathbf{y}^{(i)} \\ &= \ldots \\ &= \mathbf{x}^{(0)} - \mathbf{B}_I (\mathbf{y}^{(1)} + \cdots + \mathbf{y}^{(i)}), \end{aligned}$$

remaining ambulances for the time period $(t^{(i)}, t^{(0)} + \tau]$. Defining $\mathbf{d} = \sum_{i=1}^k \mathbf{d}^{(i)}$, and $\mathbf{y} = \sum_{i=1}^k \mathbf{y}^{(i)}$, we have $\mathbf{B}_I \mathbf{y} = \mathbf{B}_I (\mathbf{y}^{(1)} + \cdots + \mathbf{y}^{(k)}) = \mathbf{x}^{(0)} - \mathbf{x}^{(k)} \le \mathbf{x}^{(0)}$. Therefore, to find the sequence of ambulance dispatches $\mathbf{y}^{(1)} \ldots \mathbf{y}^{(k)}$ that minimizes the total shortfall, we solve the following problem:

$$Q(\mathbf{x}^{(0)}, \mathbf{d}) = \min_{\mathbf{y} \in \mathbb{Z}_+^{|E|} : \mathbf{B}_I \mathbf{y} \le \mathbf{x}^{(0)}} \mathbf{1}^\top (\mathbf{d} + \mathbf{B}_J \mathbf{y})^+. \tag{1}$$

Namely, for any sequence of dispatch decisions $\mathbf{y}^{(1)} \ldots \mathbf{y}^{(k)}$, we have

$$\sum_{i=1}^k \left(\mathbf{d}^{(i)} + \mathbf{B}_J \mathbf{y}^{(i)}\right)^+ \ge Q\left(\mathbf{x}^{(0)}, \sum_{i=1}^k \mathbf{d}^{(i)}\right).$$

since $\mathbf{y} = \sum_{i=1}^k \mathbf{y}^{(i)}$ is a feasible solution to (1).

2.2. Gradual coverage

Most of the ambulance deployment models make a distinction between demands that are "covered", and those that are not. Karasakal and Karasakal (2004) developed a partial coverage version of MCLP (MCLP-P) by using a sigmoid function to model the gradual decline of coverage along with the distance increase. Drezner, Drezner, and Goldstein (2010) proposed a gradual coverage model with a stochastic distance threshold, using probabilistic analysis to calculate the expected coverage. See Eiselt and Marianov (2009) for a review on other coverage decay functions.

To model gradual coverage, we introduce a dummy location i_0 to the set of locations I, add the constraint $x_0 \le n$ to $\mathbb{X}$, and introduce dummy edges $(0, j)$ with response times $t_{0,j} > \bar{t}$ for all $j \in J$. Then (1) can be re-written as

$$Q_\phi(\mathbf{x}, \mathbf{d}) = \min_{\mathbf{y} \in \mathbb{Z}_+^{|E|}} \phi^\top \mathbf{y} \tag{2}$$

$$\text{s.t. } \mathbf{d} + \mathbf{B}_J \mathbf{y} \ge \mathbf{0} \tag{3}$$

$$\mathbf{B}_I \mathbf{y} \le \mathbf{x}, \tag{4}$$

with $\phi \in \mathbb{Z}_+^{|E|}$ defined as $\phi = (\phi_{ij})_{(i,j) \in E}$, where

$$\phi_{ij} = \begin{cases} 0 & \text{if } t_{ij} \le \bar{t}, \\ 1 & \text{otherwise.} \end{cases} \tag{5}$$

By modifying the cost vector ϕ, we can account for different notions of partial coverage based on the estimated travel time t_{ij} from each location i to each region j. For example, instead of the $0-1$ coverage, we can consider a cost function that reports the travel time in seconds by setting $\phi_{ij} = \lfloor 60 t_{ij} \rfloor$. For consistency with established models of coverage (e.g. MEXCLP and MALP) in the literature, we perform a comparison for the formulations based on $0-1$ coverage, but report the performance of the models based on both response times and coverage (see Section 4).

3. Ambulance deployment with recourse

In this section, we are interested in the tactical decision of deploying ambulances to a fixed set of locations. Following the setup in Section 2.1, we have samples $(\mathrm{d}^i)_{i=1}^m$ of emergency calls corresponding to aggregated demands for time periods $i = 1, \ldots, m$. A natural way of modelling the problem is to consider the following formulation:

$$\min_{\mathbf{x} \in \mathbb{X}} \mathbf{c}^\top \mathbf{x} + Q_\phi(\mathbf{x}), \tag{6}$$

where $\mathbf{c}^\top \mathbf{x}$ is the cost of deployment, $\mathbb{X}$ is the set of feasible ambulance deployments, and $Q_\phi(\cdot)$ is the *recourse function* that we use to measure the performance of any given deployment $\mathbf{x}$ using the data $(\mathrm{d}^i)_{i=1}^m$.

The setup is fairly general, and we provide some examples of $\mathbb{X}$ below:

(a) $\mathbb{X}^{(1)} := \{\mathbf{x} \in \mathbb{Z}_+^{|I|} \mid \mathbf{1}^\top \mathbf{x} \le n\}$ could incorporate a constraint on the total number of ambulances and drivers available.
(b) $\mathbb{X}^{(2)} := \{\mathbf{x} \in \mathbb{Z}_+^{|I|} \mid \mathbf{x} \le \mathbf{u}\}$ could incorporate upper bounds $\mathbf{u}$ on the number of ambulances that can be deployed at each location.
(c) $\mathbb{X}^{(3)} := \{\mathbf{x} \in \mathbb{Z}_+^{|I|} \mid \mathbf{c}^\top \mathbf{x} \le b\}$ could incorporate an operating budget b on the cost $\mathbf{c}^\top \mathbf{x}$ of deploying the ambulances.
(d) $\mathbb{X}^{(4)} := \mathbb{X}^{(1)} \cap \mathbb{X}^{(2)} \cap \mathbb{X}^{(3)}$ could incorporate multiple considerations in the set of feasible ambulance deployments

In the paradigm of two-stage optimization, we compare stochastic and robust models, by modelling $Q_\phi(\cdot)$ as either

$$Q_\phi^{\text{stochastic}}(\mathbf{x}) = \mathbb{E}_{\hat{\mathbb{P}}(\mathbf{d})}\left[Q_\phi(\mathbf{x}, \mathbf{d})\right], \tag{7}$$

where $\hat{\mathbb{P}}(\mathbf{d})$ is a sample distribution over the possible scenarios, or

$$Q_{\phi}^{\text{robust}(\alpha)}(\mathbf{x}) = \max_{\mathbf{d} \in \mathbb{D}(\alpha)} \left[Q_{\phi}(\mathbf{x}, \mathbf{d}) \right], \tag{8}$$

where $\mathbb{D}(\alpha)$ is an appropriately chosen "uncertainty set" parameterized by $\alpha \in (0, 1)$.

For the stochastic approach in (6) and (7), we construct the discrete distribution $\hat{\mathbb{P}}(\mathbf{d} = \mathtt{d}^i) = 1/m$ for all $i = 1, \ldots, m$. Then a deterministic equivalent can be formulated and solved over both first stage decisions $\mathbf{x} \in \mathbb{Z}_+^{|I|}$ and second stage decision variables $(\mathbf{y}^i)_{i=1}^m$ (Ahmed, 2011; Birge & Louveaux, 2011).

For the robust approach in (6) and (8), there are two issues. First, the demand for ambulances is often sparse (i.e. predominantly zero for any region in any given hour), and merits discussion to motivate its construction in (9). Second, conventional methods of solving robust optimization problems through a robust counterpart (Ben-Tal & Nemirovski, 1999; Bertsimas, Brown, & Caramanis, 2011), or cutting-plane approach (Kelley, 1960; Thiele, Terry, & Epelman, 2009) does not immediately apply, since the recourse function is an integer optimization problem. Therefore, we develop an appropriate uncertainty set in Section 3.1, and provide a method of linearization in Section 3.2, before describing an algorithm for solving it to optimality in Section 3.3.

3.1. Structured uncertainty set

In this section, we develop uncertainty sets that can jointly model the interactions in emergency call demand across both local and regional levels. Existing models of uncertainty sets do not perform the desired function. The polyhedral uncertainty set in Bertsimas and Sim (2004) is overly-conservative for column-wise uncertainty, but recent models of moment-driven uncertainty sets (Delage & Ye, 2010) and data-driven uncertainty sets (Bertsimas, Gupta, & Kallus, 2018) do not deal with sparse integer uncertainty. If we only provide bounds on groups of regions, adversarial clusters of demand tend to concentrate in local regions. On the other hand, if we only provide bounds on local regions, the uncertainty set will be overly conservative in its estimation of the overall demand. By incorporating both types of bounds (see Fig. 1), we can construct an uncertainty set that is representative of the scenarios of interest.

To construct our uncertainty sets, we observe that emergency calls tend to follow a Poisson process that is inhomogeneous in both time and space. Therefore, sums of regional aggregated demands can be well-approximated by Poisson distributions with parameters $\hat{\gamma}$ estimated from data $\mathtt{d}^1, \ldots, \mathtt{d}^m \in \mathbb{Z}_+^{|J|}$, where

$$\begin{aligned}
\hat{\gamma}_j^{\text{single}} &= \frac{1}{m} \sum_{l=1}^{m} \left[\mathtt{d}_j^l \right] \quad \forall j \in J, \\
\hat{\gamma}_j^{\text{local}} &= \frac{1}{m} \sum_{l=1}^{m} \left[\sum_{k \in \delta_j} \mathtt{d}_k^l \right] \quad \forall j \in J, \\
\hat{\gamma}_i^{\text{regional}} &= \frac{1}{m} \sum_{l=1}^{m} \left[\sum_{j \in J_i} \mathtt{d}_k^l \right] \quad \forall i \in I, \\
\hat{\gamma}^{\text{global}} &= \frac{1}{m} \sum_{l=1}^{m} \left[\sum_{j \in J} \mathtt{d}_j^l \right].
\end{aligned}$$

For ambulance operators to control the degree of risk aversion, we introduce a parameter $\alpha \in (0, 1)$ for scaling the uncertainty set, and define the bounds

$$\begin{aligned}
\gamma_j^{\text{single}}(\alpha) &= \text{V@R}_{\alpha}(-\text{Poisson}(\hat{\gamma}_j^{\text{single}} + \epsilon)) \quad \forall j \in J, \\
\gamma_j^{\text{local}}(\alpha) &= \text{V@R}_{\alpha}(-\text{Poisson}(\hat{\gamma}_j^{\text{local}} + \epsilon)) \quad \forall j \in J, \\
\gamma_i^{\text{regional}}(\alpha) &= \text{V@R}_{\alpha}(-\text{Poisson}(\hat{\gamma}_i^{\text{regional}} + \epsilon)) \quad \forall i \in I, \\
\gamma^{\text{global}}(\alpha) &= \text{V@R}_{\alpha}(-\text{Poisson}(\hat{\gamma}^{\text{global}} + \epsilon)).
\end{aligned}$$

where the inclusion of the $\epsilon > 0$ is introduced to deal with cases where the estimated demand is 0. We let $\epsilon = 10^{-6}$ in our computational experiments. The resulting uncertainty set is as follows:

$$\begin{aligned}
\mathbb{D}(\alpha) = \Bigg\{ \mathbf{d} \in \mathbb{Z}_+^{|J|} \mid \quad & d_j \leq \gamma_j^{\text{single}}(\alpha) \quad \forall j \in J, \\
& \sum_{k \in \delta_j} d_k \leq \gamma_j^{\text{local}}(\alpha) \quad \forall j \in J, \\
& \sum_{j \in J_i} d_j \leq \gamma_i^{\text{regional}}(\alpha) \quad \forall i \in I, \\
& \sum_{j \in J} d_j \leq \gamma^{\text{global}}(\alpha) \Bigg\}.
\end{aligned} \tag{9}$$

As the polyhedral nature of the uncertainty set does not depend on the functional form of $\gamma(\cdot)$, the choice of the Poisson model here is not fundamental. In practice, it can be replaced by the negative binomial distribution to deal with overdispersion in the count data, or with any other distributions that describes the dataset. The approach is data-driven, and does not rely on model-driven assumptions to provide any probabilistic guarantees on the resulting coverage provided by the model. In practice, ambulance operators should evaluate model performance through realistic simulations, and use cross validation to choose an appropriate α (see Section 4.1.2). The following result shows that it is possible to include any scenario inside the uncertainty set (9) (at the expense of conservativeness).

Proposition 1. *For any given scenario $\mathbf{d} \in \mathbb{Z}_+^{|J|}$, there exists a sufficiently small $\alpha > 0$ such that $\mathbf{d} \in \mathbb{D}(\alpha)$.*

3.2. The recourse function

In this section, we provide a tractable formulation for the recourse function under the robust formulation (8). Observe that $\mathbf{y} = \mathbf{0}$ is always a feasible solution to (1) for all $\mathbf{x} \in \mathbb{Z}_+^{|I|}$ and $\mathbf{d} \in \mathbb{Z}_+^{|J|}$, the recourse function $Q_{\phi}(\mathbf{x}, \mathbf{d})$ is finite for all $\mathbf{x} \in \mathbb{X}$ and $\mathbf{d} \in \mathbb{D}(\alpha)$, and can be reformulated into a maximization problem for any given deployment $\mathbf{x} \in \mathbb{Z}_+^{|I|}$ and demand $\mathbf{d} \in \mathbb{Z}_+^{|J|}$.

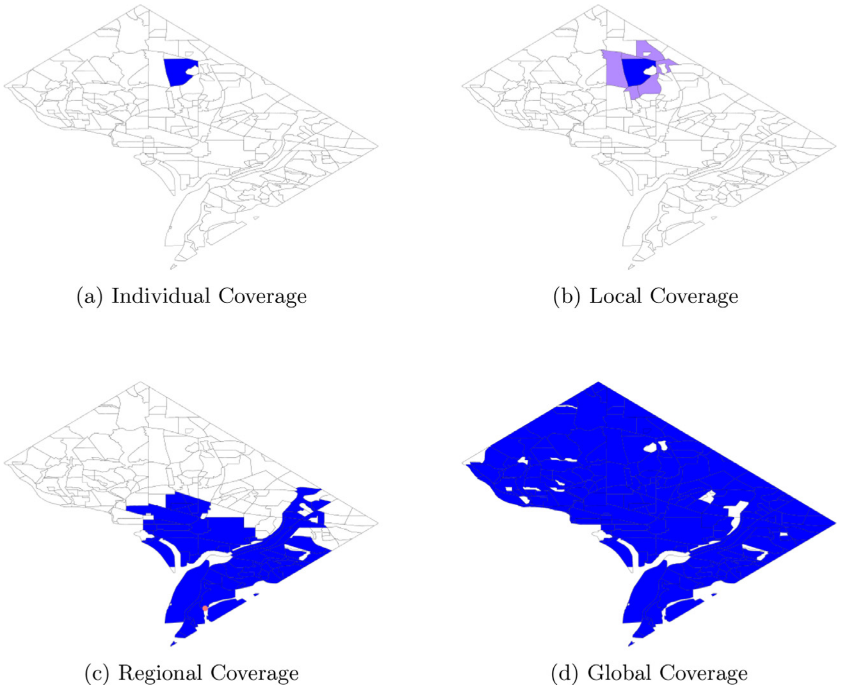
Proposition 2 (Recourse Duality)**.** *For a given $\mathbf{x} \in \mathbb{Z}_+^{|I|}$ and $\mathbf{d} \in \mathbb{Z}_+^{|J|}$, we have*

$$\begin{aligned}
Q_{\phi}(\mathbf{x}, \mathbf{d}) = \min_{\mathbf{y}} \ & \phi^{\top} \mathbf{y} \\
\text{s.t.} \ & \begin{bmatrix} \mathbf{B}_I \\ \mathbf{B}_J \end{bmatrix} \mathbf{y} \leq \begin{bmatrix} \mathbf{x} \\ -\mathbf{d} \end{bmatrix} \\
& \mathbf{y} \in \mathbb{Z}_+^{|E|} \\
= \max_{\mathbf{p}, \mathbf{q}} \ & \mathbf{x}^{\top} \mathbf{q} - \mathbf{d}^{\top} \mathbf{p} \\
\text{s.t.} \ & \mathbf{q}^{\top} \mathbf{B}_I + \mathbf{p}^{\top} \mathbf{B}_J \leq \phi^{\top} \\
& \mathbf{p} \leq \mathbf{0}, \mathbf{q} \leq \mathbf{0}.
\end{aligned}$$

This results in a mixed integer quadratic maximization problem

$$\begin{aligned}
Q_{\phi}^{\text{robust}(\alpha)}(\mathbf{x}) &= \max_{\mathbf{d} \in \mathbb{D}(\alpha)} Q_{\phi}(\mathbf{x}, \mathbf{d}) \\
&= \max_{\mathbf{d}, \mathbf{p}, \mathbf{q}} \ \mathbf{x}^{\top} \mathbf{q} - \mathbf{d}^{\top} \mathbf{p} \\
&\quad \text{s.t.} \ \mathbf{q}^{\top} \mathbf{B}_I + \mathbf{p}^{\top} \mathbf{B}_J \leq \phi^{\top} \\
&\quad \phantom{\text{s.t.}} \ \mathbf{p} \leq \mathbf{0}, \mathbf{q} \leq \mathbf{0}, \mathbf{d} \in \mathbb{D}(\alpha),
\end{aligned}$$

which is hard to solve in general. When $D(\alpha)$ is bounded, we can represent the integer vectors $\mathbf{d}$ as the sum of a small number of



(a) Individual Coverage

(b) Local Coverage

(c) Regional Coverage

(d) Global Coverage

Fig. 1. Different hierarchies of coverage for Washington D.C. Illustrates the geographical regions corresponding to (a) individual regions, (b) each region with those adjacent to it, (c) drivetime coverages from each station, and (d) the whole of Washington D.C.

binary vectors $\mathbf{d}^{(1)}, \ldots, \mathbf{d}^{(k)} \in \{0, 1\}^{|J|}$, and linearize the quadratic terms

$$-\mathbf{d}^\top \mathbf{p} = -\mathbf{p}^\top (\mathbf{d}^{(i)} + \cdots + \mathbf{d}^{(k)}) = -\mathbf{1}^\top (\mathbf{r}^{(1)} + \cdots + \mathbf{r}^{(k)}),$$

where the logical constraints

$$r_j^{(i)} = \begin{cases} 0 & \text{if } d_j^{(i)} = 0 \\ p_j & \text{otherwise.} \end{cases}$$

are enforced at optimality for $i = 1, \ldots, k$. We arrive at the following integer linear optimization model:

$$\begin{aligned} Q_\phi^{\text{robust}(\alpha)}(\mathbf{x}) = \max_{\mathbf{d}^{(1)},\ldots,\mathbf{d}^{(k)},\mathbf{p},\mathbf{q}} \quad & \mathbf{x}^\top \mathbf{q} - \mathbf{1}^\top \left(\sum_{i=1}^{k} \mathbf{r}^{(i)} \right) \\ \text{s.t.} \quad & \mathbf{q}^\top \mathbf{B}_I + \mathbf{p}^\top \mathbf{B}_J \leq \phi^\top \\ & \mathbf{r}^{(i)} \geq \mathbf{p} \quad \forall i = 1, \ldots, k \\ & \mathbf{r}^{(i)} \geq -M\mathbf{d}^i \quad \forall i = 1, \ldots, k \\ & \mathbf{r}^{(i)} \leq \mathbf{0} \quad \forall i = 1, \ldots, k \\ & \mathbf{q}^\top \mathbf{B}_I + \mathbf{p}^\top \mathbf{B}_J \leq \phi^\top \\ & \mathbf{p} \leq \mathbf{0}, \ \mathbf{q} \leq \mathbf{0} \\ & \sum_{i=1}^{k} \mathbf{d}^{(i)} \in \mathbb{D}(\alpha) \\ & \mathbf{d}^{(i)} \in \{0, 1\}^{|J|}. \end{aligned}$$

In practice, $\mathbb{D}(\alpha)$ is bounded and we determine a small value for k from data.

3.3. Column and constraint algorithm

We now state a procedure for solving the robust formulation (6) with (8). It begins with an approximation $\hat{Q}^{\text{robust}(\alpha)}(\cdot)$ of the recourse function, and generates a deployment plan $\mathbf{x}^*$ by solving

$$\mathbf{x}^* \in \operatorname*{argmin}_{\mathbf{x} \in \mathbb{X}} \ \mathbf{c}^\top \mathbf{x} + \hat{Q}^{\text{robust}(\alpha)}(\mathbf{x}),$$

before evaluating $\mathbf{x}^*$ on the worst case scenario $\mathbf{d}^*$ over $\mathbb{D}(\alpha)$ by solving

$$\mathbf{d}^* \in \operatorname*{argmax}_{\mathbf{d} \in \mathbb{D}(\alpha)} \ Q(\mathbf{x}^*, \mathbf{d}).$$

The algorithm terminates with the deployment plan $\mathbf{x}^*$ if $\hat{Q}^{\text{robust}(\alpha)}(\mathbf{x}^*) = Q^{\text{robust}(\alpha)}(\mathbf{x}^*)$. Otherwise, it uses the generated scenario $\mathbf{d}^*$ to improve the approximation function $\hat{Q}^{\text{robust}(\alpha)}(\cdot)$, and repeats the process. The algorithm is as follows:

1. Set LB $= -\infty$, UB $= \infty$ and $i = 0$.
2. Solve the following (restricted) master problem:
3. Update LB $= \eta^*$. If UB $-$ LB $\leq \epsilon$, return $\mathbf{x}^*$.
4. Solve

   $$Q_\phi^{\text{robust}(\alpha)}(\mathbf{x}^*) = \max_{\mathbf{d} \in \mathbb{D}(\alpha)} \ Q_\phi(\mathbf{x}^*, \mathbf{d}),$$

   to obtain

   $$\mathbf{d}^* \in \operatorname*{argmax}_{\mathbf{d} \in \mathbb{D}(\alpha)} \ Q_\phi(\mathbf{x}^*, \mathbf{d}).$$

5. Update UB $= \min\{\text{UB}, Q_\phi^{\text{robust}}(\mathbf{x}^*)\}$.
6. Create variables $\mathbf{y}^{i+1} \in \mathbb{Z}_+^{|E|}$ and add the following constraints:

   $$\begin{aligned} & \eta \geq \phi^\top \mathbf{y}^{i+1} \\ & \begin{bmatrix} \mathbf{B}_I \\ \mathbf{B}_J \end{bmatrix} \mathbf{y}^{i+1} \leq \begin{bmatrix} \mathbf{x} \\ -\mathbf{d}^* \end{bmatrix} \\ & \mathbf{y}^{i+1} \geq \mathbf{0}, \end{aligned}$$

   to the (restricted) master problem.
7. Update $i = i + 1$ and go to Step 2.

At each iteration of the algorithm, either the model and the oracle "converge" in their evaluation of the recourse function (when UB $-$ LB $\leq \epsilon$), or the oracle generates a scenario $\mathbf{d}$ from the uncertainty set $\mathbb{D}(\alpha)$ to be added the model. When $\mathbb{D}(\alpha)$ is finite in size (e.g. bounded and integer), the algorithm is guaranteed to converge:

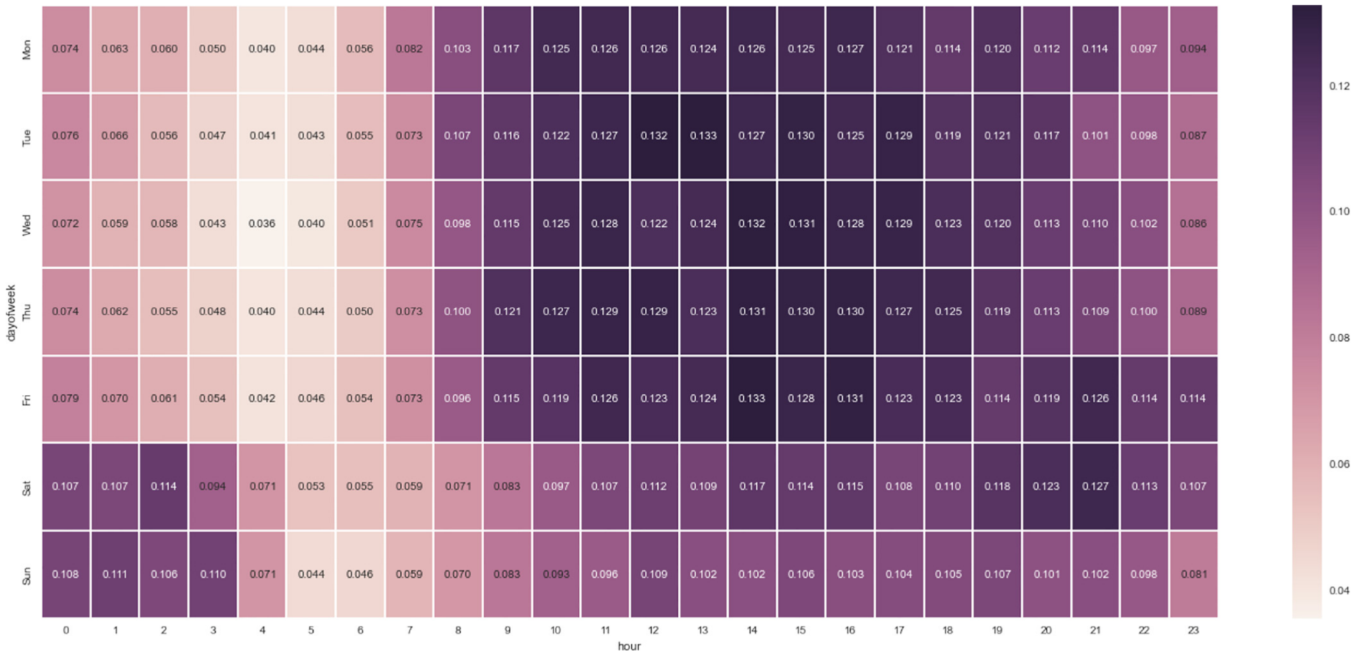


Fig. 2. Average hourly number of emergency calls per region. This is based on data of emergency calls requiring ambulatory care for DCFEMS, from Jan 2012 to Mar 2013.

Proposition 3 (Finite Convergence (Zeng & Zhao, 2013))**.** *The C&CG algorithm for the robust problem converges in a finite number of iterations.*

4. Computational results

In this section, we perform computational experiments on a realistic case study of the national EMS system for Washington D.C.. The experiments are based on data released through a Freedom of Information Act (FOIA), geo-coded by CodeForDC, and made available at https://github.com/codefordc/ERDA. The dataset includes emergency calls for both fire trucks and ambulances, and contains 179,160 relevant EMS records over the period 1 Jan 2012 to 31 March 2013. We use data from the first three months of 2012 for generating the models, and data from the remaining twelve months for evaluating the model performance (see Section 4.1.4). As commuters from the surrounding Maryland and Virginia suburbs increase the city's population during the workweek (see Fig. 2), we only consider emergency calls for the weekdays, as the same procedure can be done separately for the weekend deployment schedules.

We partition the city into 217 regions based on neighborhood zones from the Washington Post[1], and obtain geographical locations of the fire stations from http://opendata.dc.gov/. We estimate the coverage of each neighborhood from the respective fire stations, through a shortest path drivetime analysis that filters out regions that took more than 10 minutes of drivetime (see Fig. 3 for an example), using data from OpenStreetMap (Haklay & Weber, 2008). The traveling time takes into account the heterogeneity of different road types.

4.1. Experimental setup

To reflect model performance under different ambulance fleet sizes, we vary the number of ambulances from 10 to 50 in increments of 5. To evaluate the models' performances, we ran a discrete-event simulation with interarrival timings between emergency activations based on historical data. Upon activation for medical and trauma emergencies, the closest available ambulance from one of the covering locations will be dispatched to service the call[2]. In the event none of the ambulances are available, the call will be queued for the next returning ambulance.

All ambulances are assumed to have no chute time, a response time based on the estimated drivetimes, and a residual turnover time[3] that is lognormally distributed. To compare model performance, we track both the simulated response time and whether it was within 10 minutes, for each simulated emergency call. Based on a simulation spanning 12 months, we compute the monthly average and variance in both the hourly shortfall and individual response time.

4.1.1. Homogeneity of Emergency Calls

Ee generate deployment schedules for both the "peak period" from 8am to 8pm, as well as the "off-peak period" from 8pm to 8am, to pragmatically reflect the performance of the models under different demand profiles. From observation (see Fig. 2), the demand during the peak period tends to be more homogeneous, whereas the demand during the off-peak period is time-inhomogeneous, so the experiments reflect a variety of realistic situations in actual operations.

We generate the demands for both peak and off-peak periods in the following way: for the emergency calls from April 2012 to March 2013, we calculate the interarrival timings for the remaining records by taking its difference from the emergency call preceding it. For the first call of each peak (off-peak) period, we take the difference of its timing from the last call of the previous period, after subtracting the 12 hours of time difference between the two periods. This will generate a sequence of emergency calls that

[1] The Washington Post derived the neighborhood boundaries by reviewing original subdivision data, and consulting community sources.

[2] In regimes of high ambulance availability, it has been shown to be close to optimal in many studies (Lim, Mamat, & Braunl, 2011; McLay & Mayorga, 2013; Toro-Díaz, Mayorga, McLay, Rajagopalan, & Saydam, 2014).

[3] The turnover time includes the time spent at scene, conveyance to hospital, and return to base, after which it will be made available for serving the next emergency call.

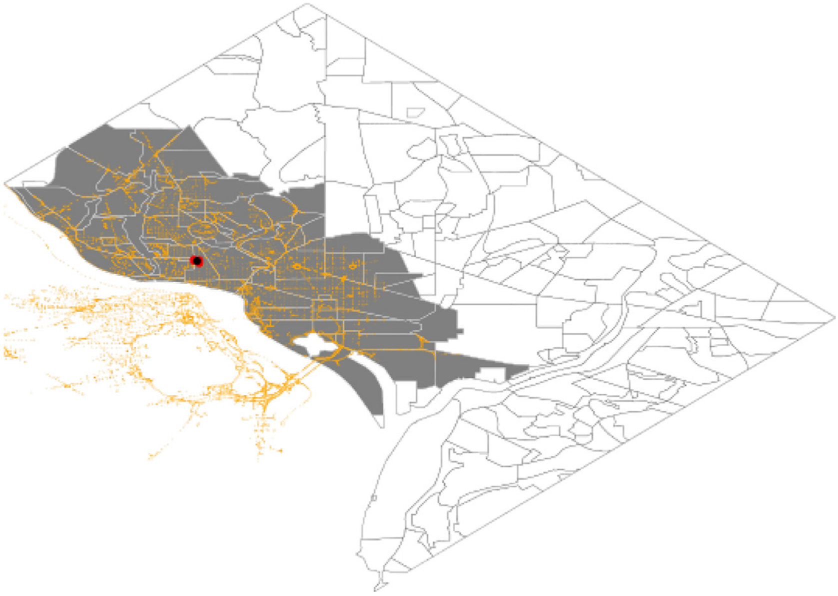


Fig. 3. Sample drivetime region. The orange dots corresponds to road nodes reached within 10 minutes, based on drivetimes using OpenStreetMap data, starting from the fire station (red dot). The grey areas indicate the coverage regions, corresponding to those that contains at least 1 orange dot. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

maintains the characteristics of both peak (off-peak) periods in our simulations.

4.1.2. Choice of uncertainty set

In our experiments, we report the performance of the robust model under different levels of α in the test set: smaller values of α corresponds to larger uncertainty sets, and correspond to a higher degree of risk aversion. The model tends to perform poorly when it does not consider enough scenarios (if α is too big), or gets biased by "extreme" scenarios (if α is too small). See Table D.1 for the deployment plans generated at different levels of α: as α becomes smaller, the size of the uncertainty set $D(\alpha)$ increases, and the more the number of ambulances the model can accomodate before it gets saturated (indicated by the shaded cells).

If the recourse function $Q^{\text{robust}(\alpha)}(\mathbf{x})$ is 0 for some $\mathbf{x} \in \mathbb{X}$, then we have some degree of freedom in deciding between different optimal deployment plans in the set $\{\mathbf{x} : Q^{\text{robust}(\alpha)}(\mathbf{x}) = 0\}$. In such cases, we shade the corresponding values to indicate that it is an invalid result. In practice, it might be appropriate to account for other considerations (such as cost or response time) as a secondary measure. Nonetheless, the issue is handled when determining the value of α through cross-validation for the robust formulation. For that reason, the optimal choice of α is data-driven and specific to each city, and we do not provide a general prescription here. Instead, we recommend that practitioners perform a cross-validation procedure on a holdout set for the best choice of α to get competitive results in practice.

4.1.3. Volatility of turnaround times

To examine the robustness of model performance to the uncertainty in the turnaround time of the ambulances, we add a lognormally distributed turnaround time $\tilde{t}^{\text{turnaround}}$ to the response times t_{ij} for each edge $(i, j) \in E$, such that $t_{ij} + \tilde{t}^{\text{turnaround}}$ is the overall time it takes for an ambulance to become available after it was dispatched for service from node $i \in I$ to node $j \in J$. By varying the parameters of the lognormal distribution for $\tilde{t}^{\text{turnaround}}$, we simulate the following three environments:

(i) volatile: $\tilde{t}^{\text{turnaround}} \sim \texttt{lognormal}(\mu = 3.57, \sigma = 0.5)$,
(ii) normal: $\tilde{t}^{\text{turnaround}} \sim \texttt{lognormal}(\mu = 3.65, \sigma = 0.3)$, and
(iii) stable: $\tilde{t}^{\text{turnaround}} \sim \texttt{lognormal}(\mu = 3.69, \sigma = 0.1)$.

The choice of a lognormal distribution was based on support from empirical data in studies (Lam, Ng, Lakshmanan, Ng, & Ong, 2016) and its use in optimization models (Alanis et al., 2013; McLay & Mayorga, 2013). The three environments correspond to turnaround distributions with the same mean of 40.25 minutes but different amounts of standard deviation (4.0 for stable, 12.3 for normal, and 21.3 for volatile; all units in minutes).

4.1.4. Models under comparison

We make comparisons of both the stochastic and robust formulations (Eq. (6) with (7) and (8), respectively) with their probabilistic counterparts MEXCLP and MALP. For a fair comparison of the models, we set the first stage deployment costs $\mathbf{c}^\top\mathbf{x}$ to 0, and used the $0 - 1$ coverage loss (described in Section 2.2). We hereby describe the formulations.

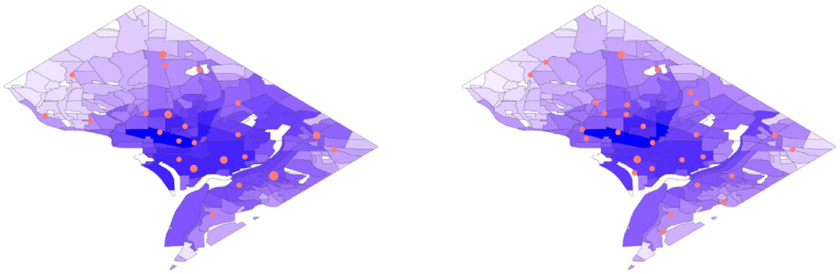
The MEXCLP by Daskin (1983) introduces the notion of a "busy fraction" $q \in (0, 1)$, as the probability that any given ambulance will be unavailable to respond to an incoming emergency call. Assuming the independence of ambulance availabilities, the resulting objective function is given by

$$\max \sum_{k=1}^{n} (1-q)q^{k-1}\mathbf{d}^\top\mathbf{z}^{(k)}, \tag{10}$$

where n is the maximum number of ambulances, and $\mathbf{z}^{(k)} \in \mathbb{B}^{|J|}$ is such that $\mathbf{z}^{(k)} = [z_1^{(k)}, \ldots, z_{|J|}^{(k)}]$ with $z_j^{(k)}$ equal to 1 if and only if region $j \in J$ is covered by at least k ambulances. As the objective function (10) is convex in k, the resulting model can be written as

$$\begin{aligned} \max\ & \sum_{k=1}^{n} (1-q)q^{k-1}\mathbf{d}^\top\mathbf{z}^{(k)} \\ \text{s.t. } & \mathbf{A}^\top\mathbf{x} \geq \mathbf{z}^{(1)} + \cdots + \mathbf{z}^{(n)} \\ & \mathbf{1}^\top\mathbf{x} \leq n \\ & \mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(n)} \in \{0,1\}^{|J|}, \quad \mathbf{x} \in \mathbb{Z}^{|I|} \end{aligned} \tag{MEXCLP}$$

ReVelle and Hogan (1989) subsequently formulated the Maximum Availability Location Problem (MALP), which introduced the


(a) Robust Deployment Coverage

(b) Stochastic Deployment Coverage

Fig. 4. Deployment Coverages for the Robust and Stochastic deployment models with 35 ambulances. The regions are shaded by counting the number of ambulances available within 10 minutes. The size of the orange circles corresponds to the number of ambulances allocated to that location. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

notion of a reliability level α, and sought to maximise the demand covered with probability α through the use of chance constraints. By linearizing the expression $1 - q^{\sum_{i\in I_j} x_i} \geq \alpha$ into $\sum_{i\in I_j} x_i \geq \lceil \log(1-\alpha)/\log q \rceil =: b$ for all $j \in J$, and defining binary vectors $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(b)}$ as in MEXCLP, they arrive at

$$\begin{aligned} \max \quad & \mathbf{d}^\top \mathbf{z}^{(b)} \\ \text{s.t.} \quad & \mathbf{A}^\top \mathbf{x} \geq \mathbf{z}^{(1)} + \cdots + \mathbf{z}^{(b)} \\ & \mathbf{z}^{(k)} \leq \mathbf{z}^{(k-1)} \quad (k = 2, \ldots, b) \\ & \mathbf{1}^\top \mathbf{x} \leq n \\ & \mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(b)} \in \{0, 1\}^{|J|}, \quad \mathbf{x} \in \mathbb{Z}^{|I|} \end{aligned} \tag{MALP}$$

In our experiments, we ran the MEXCLP and MALP with different values of q, and used cross-validation to pick the one with the best performance in out-of-sample scenarios. This came up to the value 0.654.

4.1.5. Computational setup

We implement the models in the Julia programming language (Bezanson, Edelman, Karpinski, & Shah, 2017) using the package “Julia For Mathematical Programming” (JuMP) developed by Lubin and Dunning (2015). Gurobi 6.0 was used as a solver for all the models, and experiments were run on a MacBook Pro with a 2.6 gigahertz Intel Core i5 processor, with 16 gigabtes DDR3 RAM.

4.2. Discussion of results

For both the MEXCLP and MALP formulations, the models were each solved in less than a minute for all instances. For the stochastic formulation, we formulated its deterministic equivalent with 500 scenarios, which took 1 to 2 gigabytes of RAM, and roughly a minute to solve. For the robust formulation, we begin with a naïve initial deployment plan, and trace the sequence of scenarios generated by the C&CG procedure (see Fig. C.1). In the running of the algorithm (see Section 3.3), the upper and lower bounds converges in less than 20 iterations. For example, with 35 ambulances and $\alpha = 0.01$, it took 17 iterations for the algorithm to converge to the optimal solution. The total solve time for all values of α and ambulances has been observed to take less than 10 minutes.

Although they provide a similar level of coverage for each location, the robust formulation tend to concentrate the ambulances in a few hubs, whereas the stochastic formulation distributes the ambulances more evenly distributed over the locations (see Fig. 4). In contrast, both the MEXCLP and MALP formulations tend over-concentrate their ambulances within a few “hubs” when there is an abundance of ambulances (see Table D.2), with a few locations seeing more than 10 ambulances and most locations being assigned none.

From the results (see Tables E.2 and E.1 to E.8), the largest gains in performance comes from increasing the number of ambulances available, and the differences in performance between the stochastic and robust formulations are small. On the whole, both the stochastic and robust formulations provide competitive performance with different numbers of ambulances, and demonstrate clear performance gains over the MEXCLP and MALP as the number of ambulances increases. This is expected as the stochastic and robust formulations consider the dispatch decisions, whereas the MEXCLP and MALP do not.

In particular, at $n = 50$ ambulances, both the stochastic and robust deployments have a average shortfall of less than 3.7 and 3.53 emergency calls for 360 hours of peak-hour ambulance operations, while the MEXCLP and MALP deployments have a significantly higher shortfall of 4.36 and 4.89, respectively (see Table E.2). This corresponds to an improvement (decrease) of approximately 22% of shortfalls after switching from probabilistic (MEXCLP and MALP) models to data-driven (stochastic and robust) models. Translated into operational terms, this means an average of 1 fewer incident experiencing a “late arrival” a month, which is a significant improvement given that these are rare events that should only occur for 2% of incidents during the same period.

This improvement does not come at the expense of response times. In fact, there is a corresponding improvement of approximately 34.6% in response times with 50 ambulances under peak-hour conditions. This is a surprising result, as the deployment models do not explicitly optimize for response times, even though the dispatch policy (for all deployment models) is based on the closest-idle ambulance. Similar results are observed for off-peak hour response times, with an improvement of approximately 23% in response times at 50 ambulances.

To explain the differences in performance, we look at the the proportion of time an ambulance is busy $f = ((\tilde{t}^{\text{response}} + \mathbb{E}[\tilde{t}^{\text{turnaround}}]) \cdot \bar{n}^{\text{call}})/(60 \cdot n)$, where $\bar{n}^{\text{call}}$ is the number of calls in an hour. We call f the *busy fraction*. In our experiments, f is less than one when $n \geq 25$ during peak periods and $n \geq 15$ during off-peak periods. As the data-driven models perform better than the probabilistic models when $n \geq 25$ during peak periods and $n \geq 15$ during off-peak periods, the busy fraction indicates that the data-driven models perform better when there is some slack in the system. The results are similar across both steady and volatile environments, suggesting that the volume of emergency calls matter more than service time volatility for performance comparison.

To put the improvements in perspective, it takes approximately 35 ambulances for the data-driven models to have comparable performance with the probabilistic models at 50 ambulances. This allows EMS operators to save on operating costs for up to 15 ambulances while maintaining guarantees of the same service quality. According to Heightman (2009), it costs close to half a million dollars per year to staff an advanced life support (ALS) ambulance on 24/7 basis. Therefore, this corresponds to approximately 7 million dollars in savings per year. Moreover, all of these is made possible by better deployment plans that *does not require complex technology*, and does not preclude the possibility of further improvements through the redeployment and redispatch of ambulances in real-time.

5. Conclusion

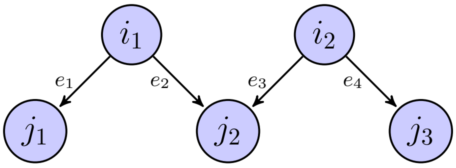
To conclude, recent economic developments have placed EMS providers under increasing pressure to meet performance goals set by regulators, often with a limited governmental budget. We provide a data-driven approach to ambulance deployment that takes into account demand interactions across regions, and improves upon probabilistic models through adaptive recourse functions that capture the essential dynamics of modelling both sparse demands and concurrent emergency calls. We demonstrate the practical tractability of the formulations, and their competitive performance across a broad range of environments in a realistic setting of Washington DC.

Acknowledgments

The authors will like to acknowledge the assistance of Sebastien Martin from the MIT Operations Research Center for suggestions on the computational study. This research was partially supported by the Office of Naval Research, Grant 021152-00001.

Appendix A. Node-arc adjacency matrix example

For the following graph:



The corresponding sub-matrices are:

$$\mathbf{B}_I = \left[\begin{array}{c|cccc} & e_1 & e_2 & e_3 & e_4 \\ i_1 & 1 & 1 & 0 & 0 \\ i_2 & 0 & 0 & 1 & 1 \end{array}\right],$$

$$\mathbf{B}_J = \left[\begin{array}{c|cccc} & e_1 & e_2 & e_3 & e_4 \\ j_1 & -1 & 0 & 0 & 0 \\ j_2 & 0 & -1 & -1 & 0 \\ j_3 & 0 & 0 & 0 & -1 \end{array}\right]$$

Appendix B. Proofs of propositions

Proof. Proposition 1 Fix $\mathbf{d} = (d_j)_{j\in J}$. By definition, we have for some $\epsilon > 0$:

$$\gamma_j^{\text{single}}(\alpha) = \text{V@R}_\alpha(-\text{Poisson}(\hat{\gamma}_j^{\text{single}} + \epsilon)) \to \infty \quad \text{as} \quad \alpha \to 0 \quad \forall j \in J,$$
$$\gamma_j^{\text{local}}(\alpha) = \text{V@R}_\alpha(-\text{Poisson}(\hat{\gamma}_j^{\text{local}} + \epsilon)) \to \infty \quad \text{as} \quad \alpha \to 0 \quad \forall j \in J,$$
$$\gamma_i^{\text{regional}}(\alpha) = \text{V@R}_\alpha(-\text{Poisson}(\hat{\gamma}_i^{\text{regional}} + \epsilon)) \to \infty \quad \text{as} \quad \alpha \to 0 \quad \forall i \in I,$$
$$\gamma^{\text{global}}(\alpha) = \text{V@R}_\alpha(-\text{Poisson}(\hat{\gamma}^{\text{global}} + \epsilon)) \to \infty \quad \text{as} \quad \alpha \to 0.$$

Therefore, let

$$\alpha_j^{\text{single}} \text{ small enough such that } \gamma_j^{\text{single}}(\alpha_j^{\text{single}}) \geq d_j \quad \forall j \in J,$$
$$\alpha_j^{\text{local}} \text{ small enough such that } \gamma_j^{\text{local}}(\alpha_j^{\text{local}}) \geq \sum_{k\in\delta_j} d_k \quad \forall j \in J,$$
$$\alpha_i^{\text{regional}} \text{ small enough such that } \gamma_i^{\text{regional}}(\alpha_i^{\text{regional}}) \geq \sum_{j\in J_i} d_j \quad \forall i \in I,$$
$$\alpha^{\text{global}} \text{ small enough such that } \gamma^{\text{global}}(\alpha^{\text{global}}) \geq \sum_{j\in J} d_j.$$

and let $\alpha = \min\{\min_{j\in J}\{\min\{\alpha_j^{\text{single}}, \alpha_j^{\text{local}}\}\}, \min_{i\in I}\{\alpha_i^{\text{regional}}\}, \alpha^{\text{global}}\}$. By construction, we can see that $\mathbf{d}$ satisfies all of the constraints in the uncertainty set

$$\mathbb{D}(\alpha) = \Big\{\mathbf{d} \in \mathbb{Z}_+^{|J|} \ \Big| \ d_j \leq \gamma_j^{\text{single}}(\alpha) \quad \forall j \in J,$$
$$\sum_{k\in\delta_j} d_k \leq \gamma_j^{\text{local}}(\alpha) \quad \forall j \in J,$$
$$\sum_{j\in J_i} d_j \leq \gamma_i^{\text{regional}}(\alpha) \quad \forall i \in I,$$
$$\sum_{j\in J} d_j \leq \gamma^{\text{global}}(\alpha)\Big\}.$$

and therefore $\mathbf{d} \in \mathbb{D}(\alpha)$. □

Proof. Proposition 2 Observe that $\mathbf{B} = \begin{bmatrix}\mathbf{B}_I \\ \mathbf{B}_J\end{bmatrix}$ is a totally unimodular matrix, and therefore

$$Q_\phi(\mathbf{x}, \mathbf{d}) = \min_{\mathbf{y}} \ \phi^\top \mathbf{y}$$
$$\text{s.t.} \begin{bmatrix}\mathbf{B}_I \\ \mathbf{B}_J\end{bmatrix}\mathbf{y} \leq \begin{bmatrix}\mathbf{x} \\ -\mathbf{d}\end{bmatrix}$$
$$\mathbf{y} \in \mathbb{Z}_+^{|E|}$$

always has an integral optimal solution whenever the right-hand side $\begin{bmatrix}\mathbf{x} \\ -\mathbf{d}\end{bmatrix}$ is integer (which is the case for all $\mathbf{x} \in \mathbb{Z}_+^{|I|}$ and $\mathbf{d} \in \mathbb{Z}_+^{|J|}$). Therefore the integer constraints can be relaxed:

$$Q_\phi(\mathbf{x}, \mathbf{d}) = Q_\phi^{\text{LP}}(\mathbf{x}, \mathbf{d}) = \min_{\mathbf{y}} \ \phi^\top \mathbf{y}$$
$$\text{s.t.} \begin{bmatrix}\mathbf{B}_I \\ \mathbf{B}_J\end{bmatrix}\mathbf{y} \leq \begin{bmatrix}\mathbf{x} \\ -\mathbf{d}\end{bmatrix}$$
$$\mathbf{y} \in \mathbb{R}_+^{|E|}$$

Taking the dual of $Q_\phi^{\text{LP}}(\mathbf{x}, \mathbf{d})$, we get

$$\max_{\mathbf{p},\mathbf{q}} \ \mathbf{x}^\top \mathbf{q} - \mathbf{d}^\top \mathbf{p}$$
$$\text{s.t. } \mathbf{q}^\top \mathbf{B}_I + \mathbf{p}^\top \mathbf{B}_J \leq \phi^\top$$
$$\mathbf{p} \leq \mathbf{0}, \mathbf{q} \leq \mathbf{0}$$

Finally, $\mathbf{y} = \mathbf{0}$ is always a feasible solution to (1) for all $\mathbf{x} \in \mathbb{Z}_+^{|I|}$ and $\mathbf{d} \in \mathbb{Z}_+^{|J|}$, with an objective function value of 0. Therefore, $Q_\phi(\mathbf{x}, \mathbf{d})$ is finite, and strong duality holds for all $\mathbf{x} \in \mathbb{X}$ and $\mathbf{d} \in \mathbb{D}(\alpha)$. □

Proof. Proposition 3 Since $\mathbb{D}(\alpha)$ is finite for any fixed $\alpha > 0$, we have convergence of the C&CG algorithm in at most $|\mathbb{D}(\alpha)|$ iterations by the following result:

Proposition 4 (Zeng & Zhao, 2013). *Let p be the number of extreme points of $\mathbb{D}(\alpha)$ if it is a polyhedron or the cardinality of $\mathbb{D}(\alpha)$ if it is a finite discrete set. Then, the C&CG algorithm will converge to the optimal value in $O(p)$ iterations, where p is the number of extreme points of $\mathbb{D}(\alpha)$ if it is a polyhedron or the cardinality of $\mathbb{D}(\alpha)$ if it is a discrete set.*

The proof for Proposition (4) can be found in Zeng and Zhao (2013). □

Appendix C. Scenario generation in the C&CG algorithm

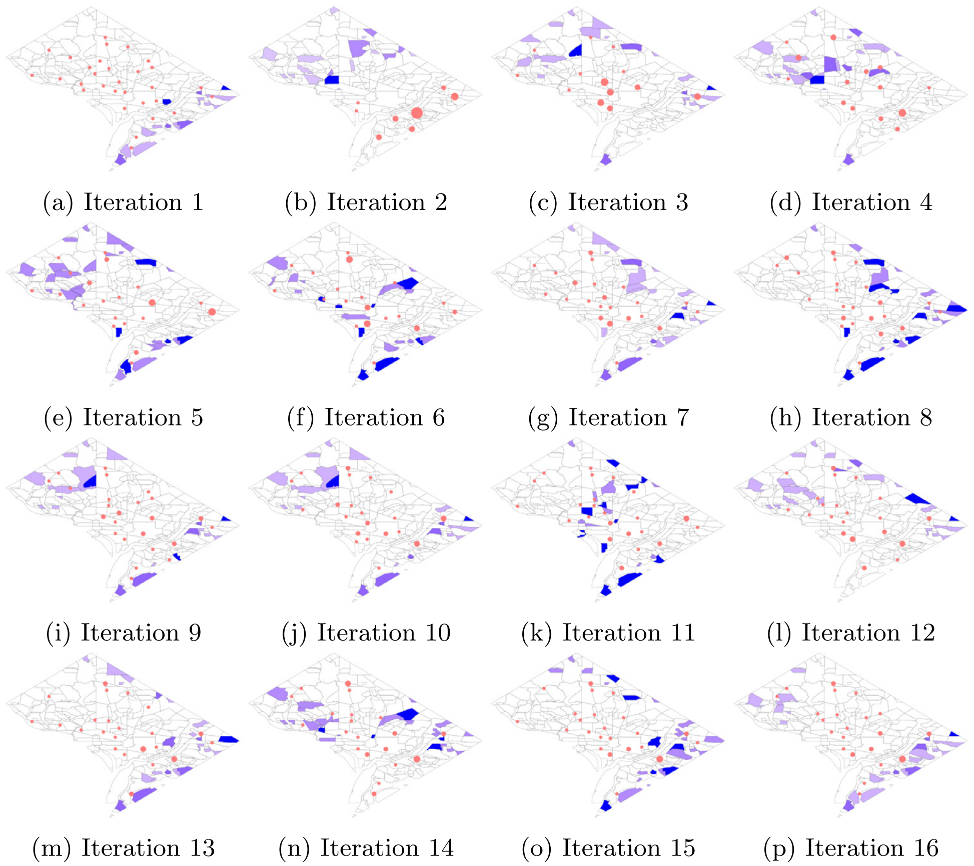


Fig. C.1. Scenarios generated by the C&CG algorithm. This is based on the robust deployment model for 35 ambulances, with parameter $\alpha = 0.01$. We illustrate the first 16 scenarios generated as chloropleths, and corresponding deployment plans as orange circles, from left to right, top to bottom. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Appendix D. Deployment plans generated

Table D.1
Robust deployment plans. Generated for different parameter values α, with varying numbers of ambulances. For values of α above 0.01, the model saturated quickly which resulted in deployment plans that were not as competitive.

Robust ($\alpha = 0.1$)

Location	Number of Ambulances								
	10	15	20	25	30	35	40	45	50
1	0	1	1	2	1	1	2	2	2
2	1	0	1	1	1	1	2	2	2
3	0	1	0	0	1	1	2	2	2
4	0	1	2	0	1	1	2	2	2
5	0	0	0	1	1	1	2	2	2
6	0	1	0	1	1	1	1	2	2
7	1	1	1	1	1	1	1	2	2
8	0	0	1	3	1	1	1	2	2
9	0	1	2	1	1	1	1	2	2
10	0	0	0	3	1	1	1	2	2
11	1	0	2	1	1	1	1	1	2
12	0	0	0	0	1	1	1	1	2
13	1	1	2	0	1	1	1	1	2
14	0	0	0	1	1	1	1	1	2
15	1	0	0	1	1	1	1	1	2
16	0	0	0	1	1	1	1	1	1
17	0	1	0	0	1	1	1	1	1
18	0	1	0	0	1	1	1	1	1
19	0	1	1	1	1	1	1	1	1
20	0	0	0	0	1	1	1	1	1
21	0	0	0	0	1	1	1	1	1
22	1	0	1	1	1	1	1	1	1
23	0	0	0	0	1	1	1	1	1
24	0	1	0	0	1	1	1	1	1
25	1	0	0	0	1	1	1	1	1
26	0	1	1	1	1	1	1	1	1
27	1	0	0	1	1	1	1	1	1
28	0	0	0	0	1	1	1	1	1
29	0	0	0	0	1	1	1	1	1
30	0	0	1	0	1	1	1	1	1
31	1	1	1	1	0	1	1	1	1
32	1	0	1	0	0	1	1	1	1
33	0	0	1	1	0	1	1	1	1
34	0	0	0	1	0	1	1	1	1
35	0	2	1	1	0	1	1	1	1

Robust ($\alpha = 0.05$)

Location	Number of Ambulances								
	10	15	20	25	30	35	40	45	50
1	0	0	0	0	1	1	2	2	2
2	0	0	0	0	0	1	2	2	2
3	1	0	1	2	0	1	2	2	2
4	1	1	1	1	0	1	2	2	2
5	0	0	1	1	1	1	2	2	2
6	0	0	0	1	0	1	1	2	2
7	0	1	2	1	2	1	1	2	2
8	1	1	1	2	0	1	1	2	2
9	1	1	1	1	3	1	1	2	2
10	1	0	0	0	0	1	1	2	2
11	0	0	0	1	1	1	1	1	2
12	0	0	0	0	0	1	1	1	2
13	0	1	1	1	0	1	1	1	2
14	0	1	1	1	2	1	1	1	2
15	0	1	0	1	0	1	1	1	2
16	1	1	1	0	0	1	1	1	1
17	1	0	0	0	1	1	1	1	1
18	0	1	1	1	1	1	1	1	1
19	0	1	2	2	1	1	1	1	1
20	0	0	0	0	1	1	1	1	1
21	0	0	0	0	0	1	1	1	1
22	1	1	1	1	3	1	1	1	1
23	0	0	0	0	0	1	1	1	1
24	0	0	0	0	2	1	1	1	1
25	0	0	0	0	0	1	1	1	1
26	0	0	0	1	1	1	1	1	1
27	1	1	1	1	3	1	1	1	1
28	0	0	0	0	0	1	1	1	1
29	0	0	0	0	0	1	1	1	1
30	0	0	0	1	2	1	1	1	1
31	1	1	1	1	1	1	1	1	1
32	0	0	1	0	2	1	1	1	1
33	0	0	1	1	2	1	1	1	1
34	0	1	1	1	0	1	1	1	1
35	0	1	1	2	0	1	1	1	1

Robust ($\alpha = 0.01$)

Location	Number of Ambulances								
	10	15	20	25	30	35	40	45	50
1	0	1	0	0	2	0	8	2	2
2	0	0	0	1	0	2	0	2	2
3	0	0	0	0	0	2	3	2	2
4	0	1	1	2	0	1	0	2	2
5	0	0	0	0	0	1	1	2	2
6	0	0	0	0	0	0	0	2	2
7	0	2	4	5	6	2	2	2	2
8	1	1	2	2	2	2	0	2	2
9	0	0	0	0	1	1	2	2	2
10	0	1	0	0	1	1	0	2	2
11	0	1	1	1	1	2	2	1	2
12	0	0	0	0	0	1	0	1	2
13	0	0	1	0	0	1	0	1	2
14	0	1	0	0	1	1	3	1	2
15	1	0	1	1	0	0	0	1	2
16	1	0	0	1	1	1	2	1	1
17	0	0	0	0	0	0	1	1	1
18	1	1	1	1	1	0	2	1	1
19	2	1	4	3	4	1	0	1	1
20	0	0	0	0	1	2	2	1	1
21	0	0	0	0	0	1	1	1	1
22	1	0	0	1	1	0	3	1	1
23	0	0	0	0	0	0	0	1	1
24	0	1	1	0	1	2	1	1	1
25	0	0	0	0	0	0	0	1	1
26	1	0	0	1	1	0	0	1	1
27	0	1	1	1	3	4	1	1	1
28	0	0	0	0	0	0	3	1	1
29	0	0	1	1	1	0	0	1	1
30	0	1	0	1	0	2	2	1	1
31	1	1	1	1	1	1	1	1	1
32	0	0	0	0	0	1	0	1	1
33	1	0	0	0	0	1	0	1	1
34	0	0	0	1	0	0	0	1	1
35	0	1	1	1	1	2	0	1	1

Robust ($\alpha = 0.001$)

Location	Number of Ambulances								
	10	15	20	25	30	35	40	45	50
1	0	0	1	2	1	4	4	6	8
2	0	0	0	1	1	0	0	1	1
3	0	0	0	0	0	1	1	1	1
4	0	0	0	0	0	0	1	4	0
5	0	0	0	0	0	1	0	1	2
6	0	0	0	0	0	0	0	0	1
7	1	2	2	3	4	5	5	3	2
8	1	1	2	1	2	1	3	0	4
9	0	1	0	1	2	1	3	7	0
10	0	0	1	0	0	1	1	0	1
11	1	1	1	1	1	1	1	0	4
12	0	0	0	0	0	0	0	1	0
13	0	0	0	0	0	0	1	0	0
14	1	0	1	0	1	1	1	3	1
15	1	0	1	1	2	1	0	2	0
16	1	0	0	2	2	2	2	1	2
17	0	0	0	0	0	0	0	1	0
18	0	1	1	1	2	2	2	0	1
19	1	3	3	4	2	3	1	3	0
20	0	0	0	0	1	1	2	2	2
21	0	0	0	0	0	0	0	0	0
22	1	0	1	1	2	2	1	1	5
23	0	0	0	0	0	0	0	0	1
24	0	1	1	2	1	1	2	3	0
25	0	1	0	0	0	1	0	0	0
26	0	1	1	1	1	1	1	0	0
27	1	1	1	1	2	1	1	0	6
28	0	0	0	0	1	0	0	0	0
29	0	1	1	0	0	0	1	1	1
30	0	0	1	1	0	1	3	0	1
31	1	1	1	1	1	2	3	3	1
32	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	1
34	0	0	0	0	0	0	0	1	3
35	0	0	0	1	1	1	0	0	1

Table D.2
Deployment plans. Generated for each model with varying numbers of ambulances. Both the Stochastic and Robust formulations evenly distribute the allocation of ambulances, but the MEXCLP and MALP clusters them in a few central locations.

Stochastic

Location	Number of Ambulances								
	10	15	20	25	30	35	40	45	50
1	0	1	1	1	1	1	1	2	2
2	0	1	1	1	1	1	1	1	1
3	0	1	1	2	2	2	2	2	3
4	0	0	1	1	1	2	2	1	1
5	0	0	1	1	1	1	1	1	2
6	0	0	0	0	0	0	0	1	1
7	0	0	1	1	1	1	1	1	3
8	1	1	1	1	1	1	1	1	3
9	1	1	1	1	1	2	1	2	2
10	0	0	0	0	0	1	1	1	1
11	0	1	1	1	1	1	1	2	1
12	0	0	0	0	0	1	1	1	1
13	0	0	0	0	1	1	1	2	1
14	0	0	0	0	1	1	1	1	1
15	0	0	1	1	1	1	1	1	2
16	1	0	0	0	0	1	1	1	1
17	0	0	0	0	0	0	1	1	1
18	1	1	1	1	0	1	1	1	1
19	1	1	1	1	3	2	2	2	2
20	0	0	0	0	0	0	1	1	1
21	0	0	1	1	1	1	1	1	2
22	0	1	1	1	1	1	1	1	1
23	0	0	0	1	1	1	1	1	1
24	1	0	0	1	1	1	1	1	1
25	0	0	0	0	1	1	1	1	0
26	1	1	1	1	1	1	1	1	1
27	1	1	1	1	1	0	1	2	1
28	0	0	0	0	0	0	1	1	1
29	0	0	0	0	0	0	0	1	1
30	0	0	0	1	1	1	1	1	1
31	1	1	1	1	1	2	3	2	3
32	0	0	0	0	0	1	1	1	1
33	0	1	1	1	1	1	2	2	3
34	0	1	0	1	2	1	2	2	1
35	1	1	2	2	2	2	1	1	1

Robust ($\alpha = 0.0001$)

Location	Number of Ambulances								
	10	15	20	25	30	35	40	45	50
1	0	0	1	2	1	1	3	4	6
2	0	1	0	1	1	0	1	1	1
3	0	0	0	1	0	0	0	0	1
4	0	0	0	0	1	0	1	1	1
5	0	0	0	0	0	0	0	0	3
6	0	0	0	0	0	0	1	0	2
7	0	2	2	3	5	6	7	7	4
8	1	1	1	1	2	3	2	2	0
9	1	2	2	1	1	2	3	3	3
10	0	0	0	0	0	1	0	0	1
11	0	0	1	1	1	0	2	3	3
12	0	0	0	0	0	0	0	0	0
13	1	0	0	0	0	0	0	0	1
14	0	0	0	0	0	1	1	1	2
15	0	1	0	0	1	2	1	2	2
16	0	0	0	0	0	1	1	1	1
17	0	0	0	0	0	0	0	0	0
18	1	0	0	1	2	2	1	0	0
19	2	3	4	4	3	3	3	1	2
20	0	0	1	1	1	2	1	2	3
21	0	0	0	0	0	0	0	0	0
22	1	1	1	2	1	1	0	3	0
23	0	0	0	0	0	0	0	1	0
24	0	0	1	1	1	0	2	1	3
25	0	0	0	0	0	0	0	0	0
26	0	1	1	1	1	1	1	1	0
27	1	1	1	1	2	1	2	1	5
28	0	0	0	0	0	0	1	1	0
29	1	0	1	1	1	1	1	0	1
30	0	0	0	1	1	2	1	3	1
31	1	1	1	1	3	4	3	4	2
32	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	1	0
34	0	0	1	0	0	0	0	0	0
35	0	1	1	1	1	1	1	1	2

MEXCLP

Location	Number of Ambulances								
	10	15	20	25	30	35	40	45	50
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	1	0	1	0
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	2	3	3	3	2	2	1	1	0
8	1	1	1	1	2	3	5	5	7
9	0	0	0	0	0	0	0	0	0
10	0	1	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0
13	1	0	0	0	0	0	0	0	0
14	0	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0
16	1	0	1	1	1	1	1	1	1
17	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0
19	3	6	8	10	12	13	14	16	17
20	0	0	0	0	0	0	2	2	3
21	0	0	0	0	0	0	0	0	0
22	0	1	2	3	5	6	7	9	10
23	0	0	0	0	0	0	0	0	0
24	1	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0
31	1	1	3	4	5	6	6	6	7
32	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0
34	0	1	1	2	2	2	3	3	4
35	0	0	0	0	0	0	0	0	0

MALP

Location	Number of Ambulances								
	10	15	20	25	30	35	40	45	50
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	3	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	9	0	0	0	0	0	0	0
8	1	1	1	1	1	1	1	11	11
9	0	0	0	0	0	0	0	0	0
10	1	0	0	0	0	0	5	0	0
11	0	0	4	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0
13	1	1	0	0	0	0	0	0	0
14	1	0	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0
16	1	0	1	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0
19	0	1	7	11	10	10	10	0	0
20	0	1	0	1	3	0	0	0	0
21	0	0	0	0	0	0	0	0	0
22	0	0	0	2	6	8	0	11	11
23	0	0	0	0	0	0	0	0	0
24	1	1	1	7	0	3	11	0	0
25	0	0	0	0	0	0	0	0	0
26	1	0	0	0	0	0	0	0	0
27	2	0	0	0	1	1	1	11	11
28	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	10	5	1	5
30	0	0	0	0	0	0	0	0	0
31	1	1	1	1	4	1	6	10	11
32	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0
34	0	0	1	1	4	0	0	0	0
35	0	0	0	0	0	0	0	0	0

Appendix E. Model performance comparisons

Table E.1
Coverage peak steady. Simulated coverages based on 12 replications, each for 360 hours of continuous peak-hour ambulance operations, with steady turnaround times. Shaded cells are results that should be ignored.

n^{amb}	Response (min)	Stochastic	Robust01	Robust005	Robust001	Robust0001	Robust00001	MEXCLP	MALP
10	shortfall ($\pm 2\sigma$)	70.19 (0.122)	74.50 (0.104)	72.23 (0.071)	73.84 (0.123)	72.96 (0.090)	73.76 (0.102)	68.81 (0.098)	73.45 (0.131)
	fraction ($\pm 2\sigma$)	0.36 (0.019)	0.39 (0.016)	0.37 (0.011)	0.38 (0.019)	0.38 (0.014)	0.38 (0.016)	0.36 (0.015)	0.38 (0.020)
15	shortfall ($\pm 2\sigma$)	45.54 (0.112)	54.69 (0.152)	44.20 (0.144)	44.28 (0.154)	48.08 (0.172)	47.37 (0.152)	44.18 (0.102)	45.96 (0.108)
	fraction ($\pm 2\sigma$)	0.24 (0.017)	0.28 (0.024)	0.23 (0.022)	0.23 (0.024)	0.25 (0.027)	0.25 (0.024)	0.23 (0.016)	0.24 (0.017)
20	shortfall ($\pm 2\sigma$)	22.36 (0.126)	20.54 (0.097)	20.61 (0.089)	23.17 (0.089)	24.67 (0.128)	26.76 (0.112)	22.66 (0.111)	25.98 (0.101)
	fraction ($\pm 2\sigma$)	0.12 (0.020)	0.11 (0.015)	0.11 (0.014)	0.12 (0.014)	0.13 (0.020)	0.14 (0.017)	0.12 (0.017)	0.13 (0.016)
25	shortfall ($\pm 2\sigma$)	10.26 (0.059)	9.62 (0.062)	9.47 (0.057)	11.85 (0.068)	9.75 (0.064)	9.62 (0.068)	10.37 (0.047)	11.58 (0.074)
	fraction ($\pm 2\sigma$)	0.05 (0.009)	0.05 (0.010)	0.05 (0.009)	0.06 (0.011)	0.05 (0.010)	0.05 (0.011)	0.05 (0.007)	0.06 (0.012)
30	shortfall ($\pm 2\sigma$)	5.31 (0.039)	6.45 (0.046)	6.30 (0.040)	5.82 (0.032)	5.76 (0.049)	5.83 (0.046)	6.69 (0.042)	8.09 (0.038)
	fraction ($\pm 2\sigma$)	0.03 (0.006)	0.03 (0.007)	0.03 (0.006)	0.03 (0.005)	0.03 (0.008)	0.03 (0.007)	0.03 (0.007)	0.04 (0.006)
35	shortfall ($\pm 2\sigma$)	4.48 (0.039)	4.86 (0.040)	4.07 (0.032)	4.52 (0.034)	4.04 (0.037)	4.14 (0.034)	5.01 (0.034)	8.65 (0.049)
	fraction ($\pm 2\sigma$)	0.02 (0.006)	0.03 (0.006)	0.02 (0.005)	0.02 (0.005)	0.02 (0.006)	0.02 (0.005)	0.03 (0.005)	0.04 (0.008)
40	shortfall ($\pm 2\sigma$)	3.98 (0.036)	4.45 (0.035)	4.45 (0.035)	3.91 (0.036)	3.77 (0.033)	3.68 (0.030)	4.64 (0.033)	6.63 (0.046)
	fraction ($\pm 2\sigma$)	0.02 (0.006)	0.02 (0.005)	0.02 (0.005)	0.02 (0.006)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.03 (0.007)
45	shortfall ($\pm 2\sigma$)	3.81 (0.035)	4.17 (0.035)	4.17 (0.035)	3.92 (0.033)	3.75 (0.032)	3.86 (0.034)	4.47 (0.034)	4.88 (0.040)
	fraction ($\pm 2\sigma$)	0.02 (0.006)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.03 (0.006)
50	shortfall ($\pm 2\sigma$)	3.74 (0.035)	4.04 (0.035)	4.04 (0.035)	4.04 (0.035)	3.54 (0.032)	3.53 (0.032)	4.36 (0.033)	4.87 (0.041)
	fraction ($\pm 2\sigma$)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.03 (0.006)

Table E.2
Coverage peak volatile. Simulated coverages based on 12 replications, each for 360 hours of continuous peak-hour ambulance operations, with volatile turnaround times. Shaded cells are results that should be ignored.

n	Response (min)	Stochastic	Robust01	Robust005	Robust001	Robust0001	Robust00001	MEXCLP	MALP
10	shortfall ($\pm 2\sigma$)	70.53 (0.083)	74.69 (0.139)	71.71 (0.106)	73.04 (0.105)	72.76 (0.134)	73.44 (0.078)	68.64 (0.078)	72.95 (0.092)
	fraction ($\pm 2\sigma$)	0.36 (0.013)	0.39 (0.022)	0.37 (0.016)	0.38 (0.016)	0.38 (0.021)	0.38 (0.012)	0.36 (0.012)	0.38 (0.014)
15	shortfall ($\pm 2\sigma$)	46.45 (0.143)	54.66 (0.140)	44.77 (0.153)	45.06 (0.178)	48.74 (0.184)	47.38 (0.166)	44.58 (0.197)	45.68 (0.135)
	fraction ($\pm 2\sigma$)	0.24 (0.022)	0.28 (0.022)	0.23 (0.024)	0.23 (0.028)	0.25 (0.028)	0.25 (0.026)	0.23 (0.031)	0.24 (0.021)
20	shortfall ($\pm 2\sigma$)	22.84 (0.118)	20.54 (0.107)	20.97 (0.122)	23.02 (0.124)	25.27 (0.135)	26.57 (0.141)	22.15 (0.147)	25.97 (0.113)
	fraction ($\pm 2\sigma$)	0.12 (0.018)	0.11 (0.017)	0.11 (0.019)	0.12 (0.019)	0.13 (0.021)	0.14 (0.022)	0.11 (0.023)	0.13 (0.017)
25	shortfall ($\pm 2\sigma$)	10.43 (0.064)	9.63 (0.051)	9.61 (0.046)	11.86 (0.051)	9.95 (0.070)	9.70 (0.044)	10.31 (0.067)	11.37 (0.074)
	fraction ($\pm 2\sigma$)	0.05 (0.010)	0.05 (0.008)	0.05 (0.007)	0.06 (0.008)	0.05 (0.011)	0.05 (0.007)	0.05 (0.010)	0.06 (0.012)
30	shortfall ($\pm 2\sigma$)	5.31 (0.038)	6.27 (0.044)	6.44 (0.036)	5.77 (0.035)	5.79 (0.038)	5.77 (0.041)	6.66 (0.050)	8.00 (0.048)
	fraction ($\pm 2\sigma$)	0.03 (0.006)	0.03 (0.007)	0.03 (0.006)	0.03 (0.005)	0.03 (0.006)	0.03 (0.006)	0.03 (0.008)	0.04 (0.008)
35	shortfall ($\pm 2\sigma$)	4.48 (0.034)	4.93 (0.033)	4.01 (0.034)	4.53 (0.036)	4.04 (0.034)	4.15 (0.035)	5.04 (0.039)	8.67 (0.060)
	fraction ($\pm 2\sigma$)	0.02 (0.005)	0.03 (0.005)	0.02 (0.005)	0.02 (0.006)	0.02 (0.005)	0.02 (0.005)	0.03 (0.006)	0.04 (0.009)
40	shortfall ($\pm 2\sigma$)	3.93 (0.035)	4.56 (0.034)	4.56 (0.034)	3.82 (0.034)	3.73 (0.032)	3.67 (0.031)	4.67 (0.033)	6.53 (0.048)
	fraction ($\pm 2\sigma$)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.03 (0.008)
45	shortfall ($\pm 2\sigma$)	3.79 (0.034)	4.20 (0.031)	4.20 (0.031)	3.93 (0.035)	3.76 (0.032)	3.85 (0.033)	4.49 (0.034)	4.90 (0.034)
	fraction ($\pm 2\sigma$)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.03 (0.005)
50	shortfall ($\pm 2\sigma$)	3.70 (0.033)	4.04 (0.033)	4.04 (0.033)	4.04 (0.033)	3.52 (0.031)	3.53 (0.032)	4.36 (0.034)	4.89 (0.034)
	fraction ($\pm 2\sigma$)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.02 (0.005)	0.03 (0.005)

Table E.3
Coverage off-peak steady. Simulated coverages based on 12 replications, each for 360 hours of continuous off-peak ambulance operations, with steady turnaround times. Shaded cells are results that should be ignored.

n^{amb}	Response (min)	Stochastic	Robust01	Robust005	Robust001	Robust0001	Robust00001	MEXCLP	MALP
10	shortfall ($\pm 2\sigma$)	24.99 (0.067)	30.08 (0.080)	24.64 (0.073)	27.39 (0.088)	26.71 (0.106)	26.00 (0.084)	24.02 (0.071)	28.09 (0.089)
	fraction ($\pm 2\sigma$)	0.27 (0.021)	0.32 (0.025)	0.26 (0.023)	0.29 (0.028)	0.28 (0.034)	0.28 (0.027)	0.26 (0.023)	0.30 (0.028)
15	shortfall ($\pm 2\sigma$)	11.36 (0.099)	12.05 (0.084)	11.13 (0.107)	11.53 (0.075)	11.96 (0.086)	10.91 (0.107)	10.59 (0.103)	14.08 (0.083)
	fraction ($\pm 2\sigma$)	0.12 (0.032)	0.13 (0.027)	0.12 (0.034)	0.12 (0.024)	0.13 (0.027)	0.12 (0.034)	0.11 (0.033)	0.15 (0.026)
20	shortfall ($\pm 2\sigma$)	4.98 (0.073)	6.13 (0.065)	4.64 (0.056)	4.78 (0.059)	5.10 (0.069)	4.87 (0.064)	5.00 (0.061)	5.71 (0.054)
	fraction ($\pm 2\sigma$)	0.05 (0.023)	0.07 (0.021)	0.05 (0.018)	0.05 (0.019)	0.05 (0.022)	0.05 (0.020)	0.05 (0.020)	0.06 (0.017)
25	shortfall ($\pm 2\sigma$)	2.78 (0.040)	4.14 (0.046)	3.10 (0.044)	2.71 (0.041)	2.44 (0.032)	3.07 (0.049)	3.14 (0.041)	3.41 (0.037)
	fraction ($\pm 2\sigma$)	0.03 (0.013)	0.04 (0.015)	0.03 (0.014)	0.03 (0.013)	0.03 (0.010)	0.03 (0.016)	0.03 (0.013)	0.04 (0.012)
30	shortfall ($\pm 2\sigma$)	2.08 (0.025)	2.48 (0.033)	2.48 (0.033)	2.32 (0.025)	2.08 (0.024)	2.05 (0.026)	2.41 (0.030)	2.63 (0.030)
	fraction ($\pm 2\sigma$)	0.02 (0.008)	0.03 (0.011)	0.03 (0.011)	0.02 (0.008)	0.02 (0.008)	0.02 (0.008)	0.03 (0.010)	0.03 (0.010)
35	shortfall ($\pm 2\sigma$)	1.89 (0.021)	2.13 (0.031)	2.13 (0.031)	2.02 (0.027)	1.80 (0.019)	1.84 (0.021)	2.25 (0.026)	3.34 (0.039)
	fraction ($\pm 2\sigma$)	0.02 (0.007)	0.02 (0.010)	0.02 (0.010)	0.02 (0.009)	0.02 (0.006)	0.02 (0.007)	0.02 (0.008)	0.04 (0.012)
40	shortfall ($\pm 2\sigma$)	1.80 (0.023)	2.03 (0.027)	2.03 (0.027)	2.03 (0.027)	1.91 (0.027)	1.86 (0.024)	2.16 (0.027)	2.03 (0.023)
	fraction ($\pm 2\sigma$)	0.02 (0.007)	0.02 (0.009)	0.02 (0.009)	0.02 (0.009)	0.02 (0.008)	0.02 (0.008)	0.02 (0.008)	0.02 (0.007)
45	shortfall ($\pm 2\sigma$)	1.73 (0.020)	1.93 (0.027)	1.93 (0.027)	1.93 (0.027)	1.93 (0.027)	1.75 (0.023)	2.15 (0.027)	2.01 (0.024)
	fraction ($\pm 2\sigma$)	0.02 (0.006)	0.02 (0.009)	0.02 (0.009)	0.02 (0.009)	0.02 (0.009)	0.02 (0.007)	0.02 (0.008)	0.02 (0.008)
50	shortfall ($\pm 2\sigma$)	1.81 (0.021)	1.88 (0.025)	1.88 (0.025)	1.88 (0.025)	1.88 (0.025)	1.70 (0.021)	2.13 (0.027)	2.01 (0.024)
	fraction ($\pm 2\sigma$)	0.02 (0.007)	0.02 (0.008)	0.02 (0.008)	0.02 (0.008)	0.02 (0.008)	0.02 (0.007)	0.02 (0.009)	0.02 (0.008)

Table E.4
Coverage off-peak volatile. Simulated coverages based on 12 replications, each for 360 hours of continuous off-peak ambulance operations, with volatile turnaround times. Shaded cells are results that should be ignored.

n^{amb}	Response (min)	Stochastic	Robust01	Robust005	Robust001	Robust0001	Robust00001	MEXCLP	MALP
10	shortfall ($\pm 2\sigma$)	24.85 (0.102)	30.75 (0.112)	25.19 (0.087)	27.53 (0.066)	26.59 (0.098)	26.28 (0.084)	24.23 (0.096)	28.29 (0.101)
	fraction ($\pm 2\sigma$)	0.26 (0.033)	0.33 (0.036)	0.27 (0.028)	0.29 (0.021)	0.28 (0.031)	0.28 (0.027)	0.26 (0.031)	0.30 (0.032)
15	shortfall ($\pm 2\sigma$)	11.69 (0.099)	12.22 (0.099)	11.24 (0.102)	11.61 (0.110)	12.05 (0.113)	11.24 (0.097)	10.81 (0.103)	14.21 (0.090)
	fraction ($\pm 2\sigma$)	0.12 (0.032)	0.13 (0.032)	0.12 (0.033)	0.12 (0.035)	0.13 (0.036)	0.12 (0.031)	0.11 (0.033)	0.15 (0.029)
20	shortfall ($\pm 2\sigma$)	4.90 (0.055)	6.06 (0.061)	4.49 (0.058)	4.69 (0.063)	5.15 (0.075)	4.64 (0.065)	4.78 (0.067)	5.68 (0.054)
	fraction ($\pm 2\sigma$)	0.05 (0.018)	0.06 (0.019)	0.05 (0.019)	0.05 (0.020)	0.05 (0.024)	0.05 (0.021)	0.05 (0.021)	0.06 (0.017)
25	shortfall ($\pm 2\sigma$)	2.63 (0.035)	4.17 (0.057)	3.10 (0.042)	2.57 (0.033)	2.35 (0.029)	3.00 (0.042)	3.00 (0.034)	3.36 (0.026)
	fraction ($\pm 2\sigma$)	0.03 (0.011)	0.04 (0.018)	0.03 (0.014)	0.03 (0.011)	0.03 (0.009)	0.03 (0.013)	0.03 (0.011)	0.04 (0.008)
30	shortfall ($\pm 2\sigma$)	2.06 (0.026)	2.42 (0.032)	2.42 (0.032)	2.26 (0.028)	2.03 (0.023)	2.05 (0.027)	2.36 (0.028)	2.60 (0.025)
	fraction ($\pm 2\sigma$)	0.02 (0.008)	0.03 (0.010)	0.03 (0.010)	0.02 (0.009)	0.02 (0.007)	0.02 (0.009)	0.03 (0.009)	0.03 (0.008)
35	shortfall ($\pm 2\sigma$)	1.87 (0.024)	2.08 (0.029)	2.08 (0.029)	1.96 (0.027)	1.80 (0.021)	1.84 (0.023)	2.19 (0.026)	3.28 (0.031)
	fraction ($\pm 2\sigma$)	0.02 (0.008)	0.02 (0.009)	0.02 (0.009)	0.02 (0.009)	0.02 (0.007)	0.02 (0.007)	0.02 (0.008)	0.03 (0.010)
40	shortfall ($\pm 2\sigma$)	1.79 (0.024)	2.02 (0.027)	2.02 (0.027)	2.02 (0.027)	1.93 (0.025)	1.84 (0.024)	2.15 (0.027)	2.02 (0.026)
	fraction ($\pm 2\sigma$)	0.02 (0.008)	0.02 (0.009)	0.02 (0.009)	0.02 (0.009)	0.02 (0.008)	0.02 (0.008)	0.02 (0.009)	0.02 (0.008)
45	shortfall ($\pm 2\sigma$)	1.73 (0.020)	1.94 (0.026)	1.94 (0.026)	1.94 (0.026)	1.94 (0.026)	1.75 (0.022)	2.14 (0.025)	2.03 (0.024)
	fraction ($\pm 2\sigma$)	0.02 (0.006)	0.02 (0.008)	0.02 (0.008)	0.02 (0.008)	0.02 (0.008)	0.02 (0.007)	0.02 (0.008)	0.02 (0.008)
50	shortfall ($\pm 2\sigma$)	1.77 (0.021)	1.89 (0.025)	1.89 (0.025)	1.89 (0.025)	1.89 (0.025)	1.70 (0.021)	2.13 (0.025)	2.03 (0.024)
	fraction ($\pm 2\sigma$)	0.02 (0.007)	0.02 (0.008)	0.02 (0.008)	0.02 (0.008)	0.02 (0.008)	0.02 (0.007)	0.02 (0.008)	0.02 (0.008)

Table E.5
Response peak steady. Simulated response times based on 12 replications, each for 360 hours of continuous peak-hour ambulance operations, with steady turnaround times. Shaded cells are results that should be ignored.

n^{amb}	Response (min)	Stochastic	Robust01	Robust005	Robust001	Robust0001	Robust00001	MEXCLP	MALP
10	mean ($\pm 2\sigma$)	9.98 (0.244)	10.42 (0.199)	10.19 (0.207)	10.18 (0.227)	10.16 (0.221)	10.28 (0.269)	9.81 (0.203)	10.47 (0.222)
	quantile ($\pm 2\sigma$)	22.36 (1.744)	23.69 (1.136)	22.96 (1.750)	22.48 (1.351)	22.34 (1.237)	22.26 (1.534)	21.68 (1.256)	25.21 (1.087)
15	mean ($\pm 2\sigma$)	8.13 (0.262)	8.86 (0.290)	8.15 (0.323)	8.15 (0.338)	8.36 (0.377)	8.24 (0.296)	8.04 (0.296)	8.78 (0.231)
	quantile ($\pm 2\sigma$)	19.07 (0.852)	20.82 (0.891)	19.08 (1.203)	18.33 (0.787)	18.67 (1.036)	17.86 (0.658)	17.48 (0.911)	22.77 (2.228)
20	mean ($\pm 2\sigma$)	6.26 (0.317)	6.07 (0.352)	6.14 (0.317)	6.52 (0.269)	6.88 (0.308)	6.94 (0.260)	6.58 (0.330)	6.85 (0.257)
	quantile ($\pm 2\sigma$)	13.95 (0.989)	13.73 (1.016)	13.82 (1.143)	13.98 (0.931)	14.25 (0.948)	14.72 (0.958)	13.42 (0.947)	15.41 (1.435)
25	mean ($\pm 2\sigma$)	4.94 (0.204)	4.88 (0.213)	4.84 (0.180)	5.35 (0.180)	5.24 (0.225)	5.12 (0.190)	5.52 (0.217)	6.00 (0.208)
	quantile ($\pm 2\sigma$)	10.23 (0.725)	10.01 (0.701)	9.95 (0.698)	10.75 (0.619)	10.01 (0.573)	10.03 (0.626)	10.21 (0.410)	10.51 (0.616)
30	mean ($\pm 2\sigma$)	3.90 (0.140)	4.53 (0.179)	4.26 (0.129)	4.28 (0.127)	4.43 (0.160)	4.21 (0.173)	5.10 (0.169)	5.42 (0.178)
	quantile ($\pm 2\sigma$)	8.21 (0.352)	8.56 (0.522)	8.75 (0.306)	8.66 (0.293)	8.80 (0.374)	8.69 (0.431)	9.23 (0.188)	9.55 (0.372)
35	mean ($\pm 2\sigma$)	3.48 (0.125)	3.60 (0.138)	3.83 (0.125)	3.94 (0.095)	3.71 (0.123)	3.88 (0.116)	4.77 (0.130)	5.78 (0.126)
	quantile ($\pm 2\sigma$)	7.51 (0.324)	7.88 (0.341)	7.39 (0.219)	8.09 (0.253)	7.54 (0.223)	7.77 (0.194)	8.67 (0.219)	9.82 (0.291)
40	mean ($\pm 2\sigma$)	3.16 (0.116)	3.36 (0.113)	3.36 (0.113)	3.33 (0.104)	3.51 (0.106)	3.66 (0.101)	4.66 (0.115)	4.83 (0.108)
	quantile ($\pm 2\sigma$)	6.72 (0.268)	7.49 (0.295)	7.49 (0.295)	6.90 (0.244)	7.03 (0.168)	7.19 (0.237)	8.53 (0.159)	9.17 (0.285)
45	mean ($\pm 2\sigma$)	3.07 (0.104)	3.21 (0.098)	3.21 (0.098)	3.36 (0.093)	3.41 (0.110)	3.47 (0.097)	4.58 (0.108)	5.46 (0.128)
	quantile ($\pm 2\sigma$)	6.47 (0.224)	7.16 (0.313)	7.16 (0.313)	7.20 (0.234)	6.86 (0.301)	7.41 (0.252)	8.36 (0.228)	8.45 (0.287)
50	mean ($\pm 2\sigma$)	2.95 (0.098)	3.03 (0.097)	3.03 (0.097)	3.03 (0.097)	3.50 (0.108)	3.39 (0.100)	4.60 (0.112)	5.43 (0.125)
	quantile ($\pm 2\sigma$)	6.24 (0.236)	6.81 (0.275)	6.81 (0.275)	6.81 (0.275)	6.76 (0.203)	6.87 (0.219)	8.31 (0.228)	8.43 (0.303)

Table E.6
Response peak volatile. Simulated response times based on 12 replications, each for 360 hours of continuous peak-hour ambulance operations, with volatile turnaround times. Shaded cells are results that should be ignored.

n^{amb}	Response (min)	Stochastic	Robust01	Robust005	Robust001	Robust0001	Robust00001	MEXCLP	MALP
10	mean ($\pm 2\sigma$)	9.98 (0.267)	10.41 (0.237)	10.19 (0.273)	10.13 (0.256)	10.12 (0.271)	10.19 (0.206)	9.73 (0.214)	10.41 (0.240)
	quantile ($\pm 2\sigma$)	22.08 (1.213)	23.32 (1.007)	22.81 (1.268)	22.01 (1.510)	21.85 (1.554)	21.69 (1.393)	20.87 (1.418)	24.22 (1.505)
15	mean ($\pm 2\sigma$)	8.20 (0.284)	8.87 (0.316)	8.16 (0.233)	8.20 (0.297)	8.40 (0.304)	8.26 (0.299)	8.03 (0.375)	8.87 (0.342)
	quantile ($\pm 2\sigma$)	18.93 (1.183)	20.68 (0.826)	18.67 (1.441)	18.23 (1.125)	18.81 (0.943)	17.49 (0.777)	17.10 (1.132)	21.99 (1.890)
20	mean ($\pm 2\sigma$)	6.30 (0.314)	6.09 (0.340)	6.15 (0.341)	6.55 (0.252)	6.89 (0.283)	6.95 (0.287)	6.54 (0.357)	6.84 (0.274)
	quantile ($\pm 2\sigma$)	14.01 (1.168)	13.68 (1.188)	13.77 (1.343)	14.10 (0.786)	14.23 (0.788)	14.54 (0.851)	13.29 (1.154)	15.14 (1.031)
25	mean ($\pm 2\sigma$)	4.94 (0.216)	4.89 (0.198)	4.84 (0.172)	5.35 (0.169)	5.24 (0.216)	5.11 (0.168)	5.49 (0.225)	5.98 (0.174)
	quantile ($\pm 2\sigma$)	10.27 (0.723)	9.96 (0.575)	9.99 (0.517)	10.76 (0.454)	10.07 (0.641)	10.01 (0.490)	10.19 (0.610)	10.44 (0.570)
30	mean ($\pm 2\sigma$)	3.89 (0.131)	4.51 (0.165)	4.26 (0.118)	4.27 (0.126)	4.42 (0.142)	4.20 (0.148)	5.08 (0.155)	5.40 (0.158)
	quantile ($\pm 2\sigma$)	8.21 (0.339)	8.53 (0.462)	8.75 (0.287)	8.65 (0.400)	8.82 (0.338)	8.66 (0.348)	9.22 (0.211)	9.53 (0.436)
35	mean ($\pm 2\sigma$)	3.48 (0.114)	3.60 (0.123)	3.82 (0.135)	3.94 (0.100)	3.70 (0.123)	3.88 (0.108)	4.76 (0.116)	5.79 (0.128)
	quantile ($\pm 2\sigma$)	7.48 (0.356)	7.87 (0.411)	7.37 (0.352)	8.07 (0.248)	7.54 (0.233)	7.77 (0.178)	8.67 (0.197)	9.85 (0.377)
40	mean ($\pm 2\sigma$)	3.15 (0.114)	3.36 (0.111)	3.36 (0.111)	3.32 (0.110)	3.51 (0.096)	3.65 (0.101)	4.64 (0.109)	4.82 (0.096)
	quantile ($\pm 2\sigma$)	6.70 (0.283)	7.46 (0.315)	7.46 (0.315)	6.88 (0.248)	7.04 (0.200)	7.18 (0.213)	8.51 (0.142)	9.14 (0.253)
45	mean ($\pm 2\sigma$)	3.06 (0.102)	3.21 (0.098)	3.21 (0.098)	3.36 (0.094)	3.40 (0.101)	3.46 (0.098)	4.57 (0.111)	5.45 (0.120)
	quantile ($\pm 2\sigma$)	6.48 (0.301)	7.13 (0.353)	7.13 (0.353)	7.22 (0.277)	6.82 (0.283)	7.41 (0.260)	8.33 (0.269)	8.42 (0.229)
50	mean ($\pm 2\sigma$)	2.95 (0.101)	3.02 (0.094)	3.02 (0.094)	3.02 (0.094)	3.50 (0.108)	3.38 (0.100)	4.59 (0.105)	5.42 (0.115)
	quantile ($\pm 2\sigma$)	6.24 (0.242)	6.80 (0.296)	6.80 (0.296)	6.80 (0.296)	6.73 (0.229)	6.85 (0.185)	8.29 (0.249)	8.40 (0.218)

Table E.7
Response off-peak steady. Simulated response times based on 12 replications, each for 360 hours of continuous off-peak ambulance operations, with steady turnaround times. Shaded cells are results that should be ignored.

n^{amb}	Response (min)	Stochastic	Robust01	Robust005	Robust001	Robust0001	Robust00001	MEXCLP	MALP
10	mean ($\pm 2\sigma$)	8.59 (0.341)	9.68 (0.301)	8.88 (0.388)	9.27 (0.366)	8.92 (0.385)	8.93 (0.399)	8.45 (0.346)	9.63 (0.380)
	quantile ($\pm 2\sigma$)	21.00 (1.346)	24.25 (1.313)	22.51 (1.935)	22.26 (0.780)	21.67 (1.215)	21.81 (1.483)	20.66 (1.381)	26.12 (1.526)
15	mean ($\pm 2\sigma$)	6.23 (0.468)	6.55 (0.437)	6.29 (0.505)	6.48 (0.385)	6.59 (0.391)	6.18 (0.481)	6.23 (0.464)	7.65 (0.487)
	quantile ($\pm 2\sigma$)	15.11 (1.374)	15.50 (1.389)	14.59 (1.786)	14.79 (1.604)	14.97 (1.322)	14.21 (1.982)	13.87 (1.602)	22.70 (3.715)
20	mean ($\pm 2\sigma$)	4.98 (0.334)	5.52 (1.374)	4.71 (0.355)	4.80 (0.313)	5.04 (0.375)	4.70 (0.377)	5.27 (0.307)	5.44 (0.294)
	quantile ($\pm 2\sigma$)	10.27 (1.776)	11.12 (1.430)	9.92 (1.424)	10.03 (1.423)	10.37 (1.477)	10.04 (1.423)	10.14 (1.074)	10.57 (0.950)
25	mean ($\pm 2\sigma$)	4.16 (0.258)	4.50 (1.073)	4.93 (0.255)	4.37 (0.199)	3.92 (0.205)	4.15 (0.259)	4.85 (0.241)	5.42 (0.193)
	quantile ($\pm 2\sigma$)	8.32 (0.957)	9.58 (1.121)	8.65 (0.801)	8.57 (0.676)	8.07 (0.819)	8.89 (0.832)	9.02 (0.519)	9.23 (0.608)
30	mean ($\pm 2\sigma$)	4.14 (0.186)	3.72 (1.001)	3.72 (1.001)	3.98 (0.185)	3.86 (0.153)	3.51 (0.189)	4.63 (0.147)	4.81 (0.176)
	quantile ($\pm 2\sigma$)	7.58 (0.476)	7.92 (0.905)	7.92 (0.905)	7.98 (0.574)	7.98 (0.374)	7.34 (0.653)	8.51 (0.446)	8.77 (0.467)
35	mean ($\pm 2\sigma$)	3.76 (0.164)	3.20 (0.180)	3.20 (0.180)	3.73 (0.180)	3.90 (0.156)	3.53 (0.127)	4.55 (0.106)	5.61 (0.168)
	quantile ($\pm 2\sigma$)	7.05 (0.251)	7.28 (0.711)	7.28 (0.711)	7.66 (0.529)	7.29 (0.442)	7.13 (0.337)	8.37 (0.315)	9.29 (0.574)
40	mean ($\pm 2\sigma$)	3.60 (0.151)	3.08 (0.164)	3.08 (0.164)	3.08 (0.164)	3.57 (0.166)	3.42 (0.148)	4.57 (0.094)	4.86 (0.125)
	quantile ($\pm 2\sigma$)	6.90 (0.251)	7.02 (0.664)	7.02 (0.664)	7.02 (0.664)	7.23 (0.472)	7.00 (0.502)	8.21 (0.247)	8.09 (0.265)
45	mean ($\pm 2\sigma$)	3.40 (0.131)	3.00 (0.151)	3.00 (0.151)	3.00 (0.151)	3.00 (0.151)	3.38 (0.135)	4.48 (0.094)	5.37 (0.121)
	quantile ($\pm 2\sigma$)	6.48 (0.263)	6.82 (0.544)	6.82 (0.544)	6.82 (0.544)	6.82 (0.544)	6.98 (0.344)	8.21 (0.231)	8.20 (0.203)
50	mean ($\pm 2\sigma$)	3.99 (0.145)	2.87 (0.134)	2.87 (0.134)	2.87 (0.134)	2.87 (0.134)	3.56 (0.148)	4.46 (0.095)	5.37 (0.119)
	quantile ($\pm 2\sigma$)	6.98 (0.247)	6.59 (0.517)	6.59 (0.517)	6.59 (0.517)	6.59 (0.517)	7.01 (0.310)	8.18 (0.247)	8.20 (0.199)

Table E.8
Response off-peak volatile. Simulated response times based on 12 replications, each for 360 hours of continuous off-peak ambulance operations, with volatile turnaround times. Shaded cells are results that should be ignored.

n^{amb}	Response (min)	Stochastic	Robust01	Robust005	Robust001	Robust0001	Robust00001	MEXCLP	MALP
10	mean ($\pm 2\sigma$)	8.59 (0.386)	9.82 (0.355)	8.98 (0.432)	9.28 (0.242)	8.95 (0.380)	8.93 (0.286)	8.45 (0.414)	9.77 (0.493)
	quantile ($\pm 2\sigma$)	20.93 (1.757)	24.35 (1.005)	22.39 (2.229)	21.83 (1.047)	21.81 (1.460)	21.47 (0.756)	20.33 (1.455)	26.31 (2.161)
15	mean ($\pm 2\sigma$)	6.29 (0.436)	6.57 (0.466)	6.29 (0.472)	6.49 (0.463)	6.61 (0.483)	6.20 (0.481)	6.23 (0.468)	7.87 (0.511)
	quantile ($\pm 2\sigma$)	15.21 (1.576)	15.55 (1.833)	14.48 (1.992)	14.85 (1.952)	14.85 (1.877)	14.25 (1.926)	13.89 (1.552)	22.90 (4.269)
20	mean ($\pm 2\sigma$)	4.99 (0.340)	5.54 (1.391)	4.69 (0.371)	4.80 (0.345)	5.04 (0.378)	4.70 (0.398)	5.24 (0.358)	5.43 (0.310)
	quantile ($\pm 2\sigma$)	10.23 (1.315)	11.20 (1.794)	9.87 (1.488)	10.01 (1.374)	10.42 (1.757)	9.95 (1.559)	10.00 (1.321)	10.65 (1.064)
25	mean ($\pm 2\sigma$)	4.15 (0.255)	4.50 (1.056)	4.93 (0.264)	4.35 (0.222)	3.91 (0.205)	4.14 (0.268)	4.82 (0.235)	5.40 (0.202)
	quantile ($\pm 2\sigma$)	8.20 (0.853)	9.67 (1.139)	8.65 (0.842)	8.49 (0.651)	8.00 (0.716)	8.81 (0.834)	8.96 (0.476)	9.19 (0.497)
30	mean ($\pm 2\sigma$)	4.13 (0.186)	3.71 (0.979)	3.71 (0.979)	3.97 (0.207)	3.85 (0.165)	3.50 (0.185)	4.62 (0.146)	4.79 (0.183)
	quantile ($\pm 2\sigma$)	7.59 (0.634)	7.87 (0.849)	7.87 (0.849)	7.94 (0.624)	7.97 (0.463)	7.31 (0.649)	8.47 (0.377)	8.75 (0.467)
35	mean ($\pm 2\sigma$)	3.75 (0.176)	3.19 (0.174)	3.19 (0.174)	3.71 (0.186)	3.88 (0.153)	3.52 (0.144)	4.54 (0.129)	5.60 (0.162)
	quantile ($\pm 2\sigma$)	7.01 (0.283)	7.20 (0.761)	7.20 (0.761)	7.59 (0.522)	7.23 (0.457)	7.09 (0.360)	8.32 (0.317)	9.24 (0.419)
40	mean ($\pm 2\sigma$)	3.59 (0.161)	3.08 (0.157)	3.08 (0.157)	3.08 (0.157)	3.56 (0.173)	3.41 (0.154)	4.56 (0.105)	4.85 (0.126)
	quantile ($\pm 2\sigma$)	6.84 (0.219)	6.99 (0.723)	6.99 (0.723)	6.99 (0.723)	7.20 (0.506)	6.95 (0.394)	8.20 (0.211)	8.07 (0.338)
45	mean ($\pm 2\sigma$)	3.40 (0.142)	2.99 (0.137)	2.99 (0.137)	2.99 (0.137)	2.99 (0.137)	3.38 (0.139)	4.47 (0.093)	5.37 (0.136)
	quantile ($\pm 2\sigma$)	6.43 (0.262)	6.77 (0.537)	6.77 (0.537)	6.77 (0.537)	6.77 (0.537)	6.97 (0.437)	8.20 (0.241)	8.19 (0.198)
50	mean ($\pm 2\sigma$)	3.98 (0.162)	2.86 (0.131)	2.86 (0.131)	2.86 (0.131)	2.86 (0.131)	3.55 (0.152)	4.45 (0.094)	5.37 (0.135)
	quantile ($\pm 2\sigma$)	6.99 (0.283)	6.55 (0.532)	6.55 (0.532)	6.55 (0.532)	6.55 (0.532)	7.00 (0.321)	8.17 (0.235)	8.18 (0.199)

References

Ahmed, S. (2011). Two-stage stochastic integer programming: a brief introduction. *Wiley Encyclopedia of Operations Research and Management Science*. American Cancer Society. doi:10.1002/9780470400531.eorms0092.

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

Ball, M. O., & Lin, F. L. (1993). A reliability model applied to emergency service vehicle location. *Operations Research, 41*(1), 18–36.

Bandara, D., Mayorga, M. E., & McLay, L. A. (2012). Optimal dispatching strategies for emergency vehicles to increase patient survivability. *International Journal of Operational Research, 15*(2), 195–214.

Batta, R., Dolan, J. M., & Krishnamurthy, N. N. (1989). The maximal expected covering location problem: Revisited. *Transportation Science, 23*(4), 277–287.

Bélanger, V., Ruiz, A., & Soriano, P. (2019). Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles. *European Journal of Operational Research, 272*(1), 1–23.

Ben-Tal, A., & Nemirovski, A. (1999). Robust solutions of uncertain linear programs. *Operations Research Letters, 25*(1), 1–13.

Beraldi, P., & Bruni, M. E. (2009). A probabilistic model applied to emergency service vehicle location. *European Journal of Operational Research, 196*(1), 323–331.

Beraldi, P., Bruni, M. E., & Conforti, D. (2004). Designing robust emergency medical service via stochastic programming. *European Journal of Operational Research, 158*(1), 183–193.

Bertsimas, D., Brown, D. B., & Caramanis, C. (2011). Theory and applications of robust optimization. *SIAM Review, 53*(3), 464–501.

Bertsimas, D., & Dunning, I. (2016). Multistage robust mixed-integer optimization with adaptive partitions. *Operations Research, 64*(4), 980–998.

Bertsimas, D., & Georghiou, A. (2018). Binary decision rules for multistage adaptive mixed-integer optimization. *Mathematical Programming, 167*(2), 395–433.

Bertsimas, D., Gupta, V., & Kallus, N. (2018). Data-driven robust optimization. *Mathematical Programming, 167*(2), 235–292.

Bertsimas, D., & Sim, M. (2004). The price of robustness. *Operations Research, 52*(1), 35–53.

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review, 59*(1), 65–98.

Billionnet, A., Costa, M.-C., & Poirion, P.-L. (2014). 2-stage robust MILP with continuous recourse variables. *Discrete Applied Mathematics, 170*, 21–32.

Birge, J. R., & Louveaux, F. (2011). *Introduction to stochastic programming*. Springer Science & Business Media.

Brotcorne, L., Laporte, G., & Semet, F. (2003). Ambulance location and relocation models. *European Journal of Operational Research, 147*(3), 451–463.

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

Church, R., & ReVelle, C. (1974). The maximal covering location problem. *Papers in Regional Science, 32*(1), 101–118.

Daskin, M. S. (1983). A maximum expected covering location model: Formulation, properties and heuristic solution. *Transportation Science, 17*(1), 48–70.

Delage, E., & Ye, Y. (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research, 58*(3), 595–612.

Drezner, T., Drezner, Z., & Goldstein, Z. (2010). A stochastic gradual cover location problem. *Naval Research Logistics (NRL), 57*(4), 367–372.

Dyer, M., & Stougie, L. (2006). Computational complexity of stochastic programming problems. *Mathematical Programming, 106*(3), 423–432.

Eiselt, H., & Marianov, V. (2009). Gradual location set covering with service quality. *Socio-Economic Planning Sciences, 43*(2), 121–130.

Gabrel, V., Lacroix, M., Murat, C., & Remli, N. (2014). Robust location transportation problems under uncertain demands. *Discrete Applied Mathematics, 164, Part 1*, 100–111.

Gendreau, M., Laporte, G., & Semet, F. (1997). Solving an ambulance location model by tabu search. *Location Science, 5*(2), 75–88.

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

Gendreau, M., Laporte, G., & Semet, F. (2005). The maximal expected coverage relocation problem for emergency vehicles. *Journal of the Operational Research Society, 57*(1), 22–28.

Haklay, M. M., & Weber, P. (2008). Openstreetmap: User-generated street maps. *IEEE Pervasive Computing, 7*(4), 12–18. doi:10.1109/MPRV.2008.80.

Hanasusanto, G. A., Kuhn, D., & Wiesemann, W. (2015). k-adaptability in two-stage robust binary programming. *Operations Research, 63*(4), 877–891.

Heightman, A. J. (2009). Resource overkill: We can do more for less. *JEMS, 34*(3), 16.

Hogan, K., & ReVelle, C. (1986). Concepts and applications of backup coverage. *Management Science, 32*(11), 1434–1444.

Jagtenberg, C., Bhulai, S., & van der Mei, R. (2017). Dynamic ambulance dispatching: is the closest-idle policy always optimal? *Health Care Management Science, 20*(4), 517–531.

Karasakal, O., & Karasakal, E. K. (2004). A maximal covering location model in the presence of partial coverage. *Computers & Operations Research, 31*(9), 1515–1526.

Kelley, J. E. (1960). The Cutting-Plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics, 8*(4), 703–712.

Lam, S. S. W., Ng, Y. S., Lakshmanan, M. R., Ng, Y. Y., & Ong, M. E. H. (2016). Ambulance deployment under demand uncertainty. *Journal of Advanced Management Science, 4*(3), 187–194.

Li, X., Zhao, Z., Zhu, X., & Wyatt, T. (2011). Covering models and optimization techniques for emergency response facility location and planning: A review. *Mathematical Methods of Operations Research, 74*(3), 281–310.

Lim, C. S., Mamat, R., & Braunl, T. (2011). Impact of ambulance dispatch policies on performance of emergency medical services. *IEEE Transactions on Intelligent Transportation Systems, 12*(2), 624–632.

Lubin, M., & Dunning, I. (2015). Computing in operations research using Julia. *INFORMS Journal on Computing, 27*(2), 238–248. doi:10.1287/ijoc.2014.0623.

Maxwell, M. S., Ni, E. C., Tong, C., Henderson, S. G., Topaloglu, H., & Hunter, S. R. (2014). A bound on the performance of an optimal ambulance redeployment policy. *Operations Research, 62*(5), 1014–1027.

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

McLay, L. A., & Mayorga, M. E. (2013). A model for optimally dispatching ambulances to emergency calls with classification errors in patient priorities. *IIE Transactions, 45*(1), 1–24.

Naoum-Sawaya, J., & Elhedhli, S. (2013). A stochastic optimization model for real-time ambulance redeployment. *Computers & Operations Research, 40*(8), 1972–1978.

Postek, K., & Hertog, D. d. (2016). Multistage adjustable robust mixed-integer optimization via iterative splitting of the uncertainty set. *INFORMS Journal on Computing, 28*(3), 553–574.

ReVelle, C., & Hogan, K. (1989). The maximum availability location problem. *Transportation Science, 23*(3), 192–200.

Shapiro, A., & Nemirovski, A. (2005). On complexity of stochastic programming problems. In *Continuous optimization* (pp. 111–146). Springer.

Swoveland, C., Uyeno, D., Vertinsky, I., & Vickson, R. (1973). Ambulance location: A probabilistic enumeration approach. *Management Science, 20*(4-part-ii), 686–698.

Thiele, A., Terry, T., & Epelman, M. (2009). Robust linear optimization with recourse. *Technical report*. Bethlehem, PA, USA: Lehigh University.

Toregas, C., Swain, R., ReVelle, C., & Bergman, L. (1971). The location of emergency service facilities. *Operations Research, 19*(6), 1363–1373.

Toro-Díaz, H., Mayorga, M. E., McLay, L. A., Rajagopalan, H. K., & Saydam, C. (2014). Reducing disparities in large-scale emergency medical service systems. *Journal of the Operational Research Society, 66*(7), 1169–1181.

Zeng, B., & Zhao, L. (2013). Solving two-stage robust optimization problems using a column-and-constraint generation method. *Operations Research Letters, 41*(5), 457–461.

Zhang, Z.-H., & Jiang, H. (2014). A robust counterpart approach to the bi-objective emergency medical service design problem. *Applied Mathematical Modelling, 38*(3), 1033–1040.