Ambulance Emergency Response Optimization in Developing Countries

Justin J. Boutilier, Timothy C.Y. Chan
Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Ontario M5S 3G8, Canada
j.boutilier@mail.utoronto.ca tcychan@mie.utoronto.ca

The lack of emergency medical transportation is viewed as the main barrier to the access and availability of emergency medical care in low and middle-income countries (LMICs). In this paper, we present a robust optimization approach to optimize both the location and routing of emergency response vehicles, accounting for uncertainty in travel times and spatial demand characteristic of LMICs. We traveled to Dhaka, Bangladesh, the sixth largest and third most densely populated city in the world, to conduct field research resulting in the collection of two unique datasets that inform our approach. This data is leveraged to estimate demand for emergency medical services in a LMIC setting and to predict the travel time between any two locations in the road network for different times of day and days of the week. We combine our prediction-optimization framework with a simulation model and real data to provide an in-depth investigation into four policy-related questions. First, we demonstrate that outpost locations optimized for weekday rush hour lead to good performance for all times of day and days of the week. Second, we find that the performance of the current system could be replicated using one-third of the current outpost locations and one-half of the current number of ambulances. Lastly, we show that a fleet of small ambulances have the potential to significantly outperform traditional ambulance vans. In particular, they are able to capture approximately three times more demand while reducing the median average response time by roughly 10-18% over the entire week and 24-35% during rush hour due to increased routing flexibility offered by more nimble vehicles on a larger road network. Our results provide practical insights for emergency response optimization that can be leveraged by hospital-based and private ambulance providers in Dhaka and other urban centers in developing countries.

Key words: Robust optimization, machine learning, facility location, global health, emergency medicine.

1. Introduction

Time-sensitive medical emergencies are a major health concern in low and middle income countries (LMICs), comprising one third of all deaths (Razzak and Kellerman 2002). Examples of such emergencies include cardiac arrest, motor vehicle accidents, and maternal health issues such as childbirth. Over the last decade, researchers and international organizations have stressed the need for increased focus on emergency medical care in LMICs (Nations 2010, Organization 2013). In

particular, the 66th World Health Assembly passed a resolution (60.22) that *"recognizes the necessity of evidence-based approaches to development of emergency care and asks WHO to promote emergency medicine research"* (Sixtieth World Health Assembly 2007, Anderson et al. 2012). However, despite widespread evidence that emergency medical care in LMICs save lives (Sodemann et al. 1997, Schmid et al. 2001), poor access and availability continues to be a major problem (Kobusingye et al. 2005, Levine et al. 2007) with the lack of emergency medical transportation noted as being the main barrier (Lungu et al. 2001, Macintyre and Hotchkiss 1999).

Optimizing the transport of emergency patients in urban centers in LMICs comes with unique challenges that are not present in high-income countries. First and foremost, traffic can be extremely unpredictable, and route disruptions caused by political demonstrations or extreme congestion occur regularly (Jain et al. 2012, Pojani and Stead 2015). Second, it is not the norm, and often not possible due to congestion, for motorists to yield for emergency vehicles. As a result, route optimization (and vehicle outpost location, by extension) becomes a critical component for improving emergency vehicle response times. Third, LMICs generally do not have historical emergency call data that can be used to forecast future emergency demand. In fact, most LMICs do not have a centralized emergency response system, so the prospect of collecting a large, high-quality dataset is itself a major challenge. Together, these challenges lead to a high degree of uncertainty in both travel times and spatial demand. The nature of these uncertainties directly impacts any modeling approach, which must be compatible with "small data" environments characteristic of LMICs.

In this paper, we develop a robust optimization approach to optimize both the location and routing of emergency response vehicles, accounting for uncertainty in travel times and spatial demand characteristic of LMICs. We traveled to Dhaka, Bangladesh, the sixth largest and third most densely populated city in the world, to conduct field research resulting in the collection of two unique datasets that inform our approach. First, we obtained a field dataset that includes patient travel data associated with several thousand hospital arrivals. This data, acting as a proxy for historical call data available in all modern, high-income countries, is leveraged to develop a framework for estimating emergency medical services incidents in a LMIC setting. Second, we equipped five vehicles with custom-built GPS devices that recorded their time and location over a period of 30 days, allowing us to understand traffic and road network characteristics in Dhaka. We then developed a machine learning model that uses the GPS data, along with census data, to predict the travel time between any two locations in the road network for different times of day and days of the week. For both demand and travel times, our predictions are leveraged to create data-driven uncertainty sets that are input into our robust location-routing model. Overall, our paper highlights the opportunity to creatively combine optimization with machine learning to solve a challenging emergency response problem in a resource-limited setting.

Like many urban centers in developing countries, Dhaka does not have a fleet of ambulances that form a centralized emergency response system. Instead, patients use a variety of transportation modes to reach hospitals in emergencies, including rickshaws, auto-rickshaws (i.e., three-wheeled motorcycles), private cars, and private or hospital-based ambulance services. Our modeling framework is well-suited to handle different transportation modes, which are accounted for via differences in road network connectivity according to vehicle type. Smaller and more nimble vehicles can traverse roads that larger vehicles cannot access. Therefore, the consideration of transportation mode affects the ultimate computational tractability of our models. In this paper, we focus on traditional *van ambulances* and the locally inspired *small ambulances*, which are based on three-wheeled motorcycles that have platforms that can be used for patient transport. In Dhaka (and many developing countries), most traditional van ambulances lack advanced medical equipment and are not typically staffed by paramedics, meaning that small ambulances are essentially equivalent from a medical equipment standpoint. Small ambulances have been recently proposed in Bangladesh (Wadud 2017), but are not yet implemented and their potential impact on response times and patient outcomes has not been studied in the scientific literature.

Ambulance services in Dhaka are currently decentralized, meaning there are both private ambulance service providers, which are for-profit businesses, and ambulance fleets that belong to hospitals. Both types of organizations are incentivized to increase the number of patient transports they make, but lack appropriate decision support tools to optimize their operations. For example, hospitals do not currently strategically preposition their ambulances in the city, but rather position their entire fleet at the hospital. Savas (1969) demonstrated the potential improvements over a similar hospital-based strategy in New York City. Therefore, private or hospital-based ambulance services are natural knowledge users of our research. Until recently, contact information for these services was also decentralized and unique to each provider, providing significant access challenges for patients. However, in December 2017, Bangladesh introduced the first centralized emergency services number "999" (Tribune 2017). The insights derived from our results can inform government policy on how to build a centralized emergency response system and aid non-government organizations to determine how to best position emergency response vehicle outposts. In particular, we use our real data and a simulation framework to answer four policy-related questions and derive practical insights for emergency response optimization in Dhaka and other LMICs:

1. **Should different outpost locations be used for different times of day? (Section 5.1)** In some high-income countries, ambulance locations are adjusted spatiotemporally throughout the day, but does that value persist in LMICs?

2. **What performance improvements are possible by optimizing outpost locations? (Section 5.2)** How different would a centralized, optimized system be from the current situation

where ambulances are parked at hospitals? How does repositioning outpost locations compare to adding new locations?

3. **How much can the system be improved by using small ambulances? (Section 5.3)** Can small ambulances capture additional demand that is currently unserved (or under-served) by existing van ambulances? What is the potential value of increased routing flexibility offered by small ambulances given their ability to traverse smaller roads in the network that are inaccessible to vans?

4. **How important is it to consider uncertainty when designing an emergency response network? (Section 5.4)** What is the performance improvement of our robust optimization model compared to a deterministic model? How does our robust approach compare to the perfect information case?

The problem of optimizing emergency vehicle response has historically been cast as a facility location problem (Toregas et al. 1971). Although the facility location literature is rich, there is no unified framework for optimizing emergency vehicle response under both edge-based travel time uncertainty and demand uncertainty (Ahmadi-Javid et al. 2017). A key distinction between this paper and previous work is how we model travel time uncertainty. Our model provides a general edge-based framework for travel time uncertainty, whereas previous research has focused on modeling travel time uncertainty using a path-based approach (Snyder 2006). Edge-length uncertainty is critical for our model because many of the underlying causes of travel time uncertainty in Dhaka (e.g., intersections without signal control, floods, strikes, etc.) impact small subsets of edges as opposed to the entire path. Uncertainty on individual edges can affect multiple routes and must be accounted for during optimization. Our routing problem is effectively a robust shortest path problem and, depending on how we model edge-length uncertainty, is equivalent to a regularized shortest path problem. The equivalence between robustness and regularization has been noted in domains such as regression (Bertsimas and Copenhaver 2017), but has not been previously demonstrated for the shortest path problem.

Overall, we use the aforementioned challenges faced by LMICs and gaps in the facility location literature to motivate the development of a novel location-routing model that is tailored for emergency response optimization in developing urban centers. We make the following contributions:

- We develop a novel edge-based reformulation of the classical path-based p-median problem. The p-median problem seeks to locate P facilities relative to a set of demand nodes such that the total demand weighted distance to all demand nodes is minimized (Hakimi 1964, 1965). This reformulation forms the foundation of a two-stage robust optimization model that considers both uncertain edge lengths (travel time) and node weights (demand). Our approach generalizes previous emergency facility location models based on the p-median architecture and provides a unified

framework for emergency response optimization under travel time and demand uncertainty that is suitable for LMICs. We develop several approaches to solve our model. First, we develop an equivalent single-stage mixed-integer linear optimization problem. Second, we develop an exact scenario (i.e., row and column) generation algorithm that can improve the solution time by several orders of magnitude. For application to large-scale problems representative of the real road network in Dhaka, we develop a novel heuristic algorithm by extending a state-of-the-art p-median heuristic to work with edge-length uncertainty (Section 3). All theorem proofs can be found in the Electronic Companion.

- We develop a methodology to predict emergency demand spatially for urban centers without historical demand data by decomposing demand into components that can be estimated using census data and a regularized logistic regression model. This approach represents the first attempt to predict emergency demand in a developing urban center. Our complete dataset, including census, survey, and hospital location data is unique because, to the best of our knowledge, hospital arrival surveys and patient travel data have never been collected together previously in any LMIC (Sections 4.2 and EC.1).
- We develop and compare several machine learning models to predict travel time on the Dhaka road network by time of day and day of week, using a dataset of vehicle trips collected by our custom-made GPS devices. We find that a random forest model performs the best, with a $43.3-64.2\%$ improvement in prediction accuracy over several baseline approaches. This paper is the first to use real travel time data from a LMIC for optimization (Sections 4.3 and EC.2).
- Using a simulation framework and real data from Dhaka, we provide an in-depth investigation into the four policy-related questions posed above (Section 5):

1. In contrast to developing countries where researchers have estimated performance improvements from repositioning ambulances according to the time and day, there is little to gain in Dhaka by optimizing outpost locations spatiotemporally. Instead, using outpost locations optimized for weekday rush hour leads to good performance for all times of day and days of the week.

2. A centralized network designed from a clean slate can replicate the performance of the current system using roughly one-half of the ambulances and one-third of the outpost locations currently in use.

3. A fleet of small ambulances has the potential to significantly outperform traditional van ambulances. In particular, they can capture over three times the demand as van ambulances while reducing the median average response time by roughly 10-18% over the entire week and 24-35% during rush hour. This gain requires emergency response providers to tailor outpost locations specifically for small ambulances, instead of locating them at outposts optimized for traditional van ambulances.

4. Our robust solutions can reduce the median and worst-case response times by up to 33% and 45%, respectively, compared to a deterministic solution that does not take uncertainty into account. Furthermore, the performance of the robust solution is comparable to a solution that has access to perfect information on the uncertainty.

2. Literature review

Our work is related to three major streams of literature: 1) demand prediction in the context of emergency response optimization, 2) vehicle travel time prediction, and 3) facility location.

2.1. Demand prediction

While most papers use historical emergency call data as a direct estimate for future demand, a growing and more relevant body of literature uses that data to develop machine learning models that can predict future demand. Early approaches considered only spatial demand, using multiple linear regression to relate the magnitude of demand for ambulances with population and other socio-economic factors (e.g., Schuman et al. 1977, Kamenetzky et al. 1982). Key covariates can be summarized into three main groups: measures of population (e.g., household size), measures of economic status (e.g., employment rate, poverty level), and measures of social status (e.g., literacy rate, marriage rate). Temporal-only approaches were developed to forecast emergency calls at various time scales, including daily (Baker and Fitzpatrick 1986), multi-hour blocks (Trudeau et al. 1989), and hourly (Channouf et al. 2007, Matteson et al. 2011). Finally, there exist methods to predict future emergency demand at fine spatiotemporal resolutions (Setzler et al. 2009, Zhou et al. 2015, Zhou and Matteson 2016).

The aforementioned approaches rely on granular historical call data to train prediction models. High-income countries tend to be data-rich, so research efforts have focused on advanced demand prediction techniques using this abundant and granular data. However, in most LMICs, historical call data is not available (Bradley et al. 2017). In this paper, we develop a new approach that does not use historical call data and instead makes use of the limited spatiotemporal data available in many LMICs.

2.2. Travel time prediction

Research on predicting edge-based travel times for ambulances has focused on developing non-linear relationships between travel time and distance (Kolesar et al. 1975, Budge et al. 2010, Hofleitner et al. 2012b,a, Westgate et al. 2016). However, almost all prior research depends directly on the availability of historical emergency transport data collected by a centralized system, which typically does not exist in LMICs.

In recent years, machine learning approaches that leverage decentralized travel time data have gained popularity and demonstrated superior prediction accuracy for regular vehicle travel time estimation (Vlahogianni et al. 2014). In contrast to ambulances, regular vehicle travel times are highly dependent on the time of day and the day of the week (Kok et al. 2012, Woodard et al. 2017). Travel times for emergency vehicles and regular vehicles are similar in LMICs because road users do not yield for ambulances. As a consequence, we employ a general travel time prediction approach similar to that of Zhang and Li (2015), who use a random forest model that accounts for distance, time of day, and day of week. We extend their model by incorporating demographic and geographic characteristics for the origin and destination nodes, which encodes spatial information about the trip. ML for travel time prediction

2.3. Facility location

Facility location is a very well-studied field and we provide only a brief review of the relevant literature. For a general review of facility location, please see Owen and Daskin (1998) or Melo et al. (2009), and for a comprehensive review of facility location in the context of emergency medical services, please see Li et al. (2011), Basar et al. (2012), or Ahmadi-Javid et al. (2017).

2.3.1. Emergency response. Facility location models have been applied extensively to emergency medical services location problems with the majority of previous research focusing on ambulances. There have been many papers that investigate ambulance response optimization in *urban areas in high-income countries* (e.g., Brandeau and Larson 1986, Ingolfsson et al. 2008), in *rural areas in high-income countries* (e.g., Adenso-Diaz and Rodriguez 1997, Chanta et al. 2014), and in *rural areas in LMICs* (e.g., Bennett et al. 1982, Eaton et al. 1986). However, there have been only a few papers that consider *urban areas in LMICs* (Fujiwara et al. 1987, Basar et al. 2011, Salman and Yücel 2015, Zhang and Li 2015), and they differ from our work in several important aspects. First, these papers focus on upper-middle-income countries (China, Thailand, and Turkey), whereas we focus on a low-income country (Bangladesh). Second, previous urban ambulance response optimization research, including the papers listed above, has focused exclusively on regions that already have a centralized ambulance system. In contrast, our paper is the first to focus on a developing urban center without an existing ambulance system, which leads to new policy questions not considered in areas with an existing system.

2.3.2. Demand and travel time uncertainty. Demand uncertainty has received significant attention in general location-allocation problems (e.g., Shen et al. 2003, Atamturk and Zhang 2007, Baron et al. 2011) as well as in the specific context of ambulance response optimization (Beraldi et al. 2004, Beraldi and Bruni 2009, Noyan 2010). The ambulance-specific papers all use

chance constraints to model uncertain demand, whereas we employ a scenario-based approach that integrates a prediction model trained with our field data.

Travel time uncertainty in the context of ambulance response optimization has been focused on path-length uncertainty (Ingolfsson et al. 2008, Berman et al. 2013, Abdul Ghani and Ahmad 2017). For networks with edge-length uncertainty, previous research has focused on the 1-median problem (Carson and Batta 1990, Averbakh 2003) and networks with special structure (Mirchandani and Odoni 1979, Mirchandani and Oudjit 1980). We are the first to investigate edge-length uncertainty for the general p-median problem applied to ambulance response optimization.

Nearly all previous literature on combining both edge-length and node-weight uncertainty has focused on the special case of the 1-median problem (Chen and Lin 1998, Vairaktarakis and Kouvelis 1999), whereas we develop a methodology for the general p-median problem under uncertainty. The study by Serra and Marianov (1998), which considers the p-median problem with both uncertain *path lengths* and node weights, is the closest to our work. In contrast, we consider uncertain *edge lengths* and node weights, which can be interpreted as a generalization of their model.

2.3.3. Ambulance repositioning. Ambulance repositioning has received significant attention in the emergency response literature (Brotcorne et al. 2003, Saydam et al. 2013, Nasrollahzadeh et al. 2018). Repositioning strategies are often motivated by temporal changes in spatial demand and coverage gaps caused by busy vehicles. Real-time repositioning, which seeks to preposition ambulances in real time to better respond to future calls, leverages projected demand patterns and GPS-based ambulance location data. Repositioning strategies combined with dispatching decisions can also be used to mitigate system uncertainty (Enayati et al. 2018). However, in many LMICs, real-time repositioning strategies are unrealistic because there is no centralized emergency response system to manage the real-time repositioning decisions.

Static ambulance repositioning is a simplified version of real-time repositioning that focuses on allocating ambulances to pre-selected outposts according to shift schedules, times-of-day, or the number of available ambulances (Alanis et al. 2013, van Barneveld 2016, Sudtachat et al. 2016, van Barneveld et al. 2017). Although static approaches are typically less effective than real-time strategies (Maxwell et al. 2010), they are easy to implement and manage. For example, compliance tables can be used to inform ambulance providers which outpost locations should be used for specific times of day or when there are only a certain number of ambulances available. We investigate the value of static repositioning in Section 5.1, which is motivated by changing demand patterns and the impact of changing traffic patterns on travel times (Schmid and Doerner 2010). While traffic is less of a concern in high-income countries, emergency vehicles typically face the same traffic conditions as regular road users in LMICs since other vehicles do not (or cannot) yield to ambulances.

3. Optimization approach

We develop a two-stage robust optimization model to determine emergency response vehicle outpost locations. The outpost locations are determined based on how vehicles will be routed from the outpost to demand points (second stage), considering uncertainty in both demand and travel times.

We begin by introducing a novel edge-based location model that we prove to be equivalent to the classical p-median model. The advantage of our model is that it can handle edge-length uncertainty. Next, we introduce our models of uncertainty for emergency demand and travel times. Finally, we develop and compare several solution approaches.

1) edge-based location model
2) model of uncertainty for emergency demand
3) travel times

3.1. Network flow formulation

Let the road network be represented as the directed graph $G=(\mathcal{N},\mathcal{E})$. Let $|\mathcal{N}|=n$, $|\mathcal{E}|=m$, and $\mathbf{A}$ denote the $n\times m$ node-arc incidence matrix. Let $\mathbf{c}$ denote the vector of edge lengths (i.e., travel times) and $\mathbf{d}$ denote the vector of node weights (i.e., demand in terms of average annual emergency transports required). Let $\boldsymbol{\alpha}$ denote the supply available at each potential facility (i.e., number of trips that can be made from each outpost per year) and $\boldsymbol{\Omega}$ represent the $n\times n$ diagonal matrix whose entries are the n elements in $\boldsymbol{\alpha}$. We use P to represent the number of outposts to be located and $\mathbf{e}$ to denote the vector of all ones. The decision variable representing the vector of flows along each edge is denoted by $\mathbf{f}$ (i.e., how many trips occur on each edge annually). The outpost location variable is given by $\mathbf{y}\in\{0,1\}$ where 1 indicates an outpost is located at node $i\in\mathcal{N}$. Note that all defined vectors are column vectors. In vector form (see EC.3 for the non-vectorized version), our deterministic network flow formulation (**NFF**) is:

$$\begin{aligned} \textbf{NFF:} \quad & \underset{\mathbf{y},\mathbf{f}}{\text{minimize}} && \mathbf{c}'\mathbf{f} \\ & \text{subject to} && \mathbf{e}'\mathbf{y}=P, \\ & && \mathbf{A}\mathbf{f}\le\boldsymbol{\Omega}\mathbf{y}-\mathbf{d}, \\ & && \mathbf{f}\ge\mathbf{0}, \\ & && \mathbf{y}\in\{0,1\}^n. \end{aligned} \tag{1}$$

The second constraint accounts for supply nodes, ensures that all demand is met, and allows for transshipment flow. In scalar form, the constraint can be written as:

$$\sum_{j\in O(i)} f_{ij}-\sum_{j\in I(i)} f_{ji}\le\alpha_i y_i-d_i,\forall i\in N,$$

where $I(i)=\{j\in N|(j,i)\in\mathcal{E}\}$ and $O(i)=\{j\in N|(i,j)\in\mathcal{E}\}$. If $y_i=1$, then node i becomes a source node that produces up to α_i-d_i trips per year. If $y_i=0$, then node i becomes a demand node and the constraint reduces to $\sum_{j\in I(i)} f_{ji}-\sum_{j\in O(i)} f_{ij}\ge d_i$. This ensures that at least d_i trips flow

into node i (thereby satisfying demand), but also allows for trips to flow into and out of node i en route to another location.

To ensure that (1) is feasible for any value of P, we require the following assumption.

ASSUMPTION 1. $\alpha_i \geq \sum_{i=j}^{n} d_j$, $\forall i \in \mathcal{N}$.

This assumption states that each outpost has enough capacity to service the entire system (i.e., all demand nodes) and to ensure feasibility, we set $\alpha_i = \sum_{i=1}^{n} d_i$, $\forall i \in \mathcal{N}$. We do not consider queuing in our model because our primary focus is to determine where to strategically locate emergency response outposts, rather than determining the total number of emergency response vehicles. However, we later evaluate the tactical performance of our solutions with respect to queuing and system congestion using a simulation model. Lemma 1 follows immediately from this assumption (proof omitted).

LEMMA 1. *There exists an optimal solution to **NFF** such that each demand node is assigned to exactly one outpost.*

This result generally holds true for uncapacitated facility location models such as the p-median. Finally, using Lemma 1, we can show the equivalence between **NFF** and the p-median problem.

THEOREM 1. *A solution is optimal for **NFF** if and only if it is optimal for the p-median problem.*

The proof of Theorem 1 provides a constructive approach to obtain an optimal solution of **NFF** given an optimal solution of the p-median problem, and vice versa. Mathematically, this approach provides a polynomial-time many-one reduction between the **NFF** and the p-median problem in both directions (Post 1944, Karp 1972).

3.2. Robust optimization model

In this section, we present our two-stage robust optimization model, considering both the travel times $\mathbf{c}$ and demands $\mathbf{d}$ as uncertain with $\mathcal{C}$ and $\mathcal{D}$ representing the corresponding uncertainty sets, respectively. Our general two-stage robust network flow formulation is:

$$\begin{aligned} \textbf{R-NFF:} \quad \min_{\mathbf{y}} \max_{\mathbf{c}\in\mathcal{C},\mathbf{d}\in\mathcal{D}} \min_{\mathbf{f}} \quad & \mathbf{c}'\mathbf{f} \\ \text{subject to} \quad & \mathbf{e}'\mathbf{y} = P, \\ & \mathbf{A}\mathbf{f} \leq \boldsymbol{\alpha}\mathbf{I}\mathbf{y} - \mathbf{d}, \\ & \mathbf{f} \geq \mathbf{0}, \\ & \mathbf{y} \in \{0,1\}^n. \end{aligned} \tag{2}$$

The two-stage nature of our formulation is well-suited to the problem of emergency outpost location and vehicle routing. In the first stage, **R-NFF** determines the optimal outpost locations considering

both $\mathbf{c}$ and $\mathbf{d}$ as uncertain. Intuitively, determining these locations is a high-level strategic decision that must be made under uncertainty, before demand or traffic are realized. Then, given the realized demand and travel time conditions, the second stage determines the optimal routes from the outposts to reach each demand point (i.e., patient location). Routing is a secondary decision that is used to inform the first stage location decision because the suitability of an outpost location is influenced by the route options emanating from that outpost.

3.2.1. Demand uncertainty set ($\mathcal{D}$). To model uncertainty in emergency transport demand, we use a scenario-based uncertainty set. We use this approach to preserve tractability while still capitalizing on the richness of our demand predictions. For N scenarios, the resulting uncertainty set is defined as $\mathcal{D} = \{\mathbf{d}^1, \mathbf{d}^2, ..., \mathbf{d}^N\}$, where the dimension of $\mathbf{d}$ is equal to the number of nodes in the network. To generate the scenarios that form the uncertainty set, we employ a form of bootstrapping and simulate possible realizations of demand vectors using our framework from Sections 4.2 and EC.1.

3.2.2. Travel time uncertainty set ($\mathcal{C}$). Uncertainty in travel time is modeled using an interdiction-based uncertainty set with an overall budget constraint (Wood 1993). Intuitively, this set models an adversary who is adding traffic (i.e., increasing travel time) to the baseline traffic on each edge. The budget constraint restricts the total amount of travel time that can be added across the network. The mathematical formulation of this uncertainty set is $\mathcal{C} = \left\{ c_{ij}, (i,j) \in \mathcal{E} \;\middle|\; c_{ij} = \hat{c}_{ij} + w_{ij}, \sum_{(i,j)\in\mathcal{E}} w_{ij} \leq B, w_{ij} \geq 0, \forall (i,j) \in \mathcal{E} \right\}$. We estimate the baseline travel time $\hat{c}_{ij}$ for each edge using the final random forest model from Section EC.2.2. In our numerical experiments, we perform a detailed sensitivity analysis on the budget B.

3.3. Solution Algorithms

In this section, we present several methods to solve **R-NFF**. First, we show that there is an equivalent single-stage mixed-integer optimization model for **R-NFF**. Then, we present an exact row-and-column generation algorithm to solve this equivalent problem. Finally, for the integer master problem, we devise an efficient heuristic that is needed for large-scale instances. See EC.5 for a detailed numerical comparison of the solution times and optimality gaps between the mixed-integer model, exact solution algorithm, and heuristic solution algorithm.

3.3.1. Equivalent mixed-integer optimization model. First, we replicate $\mathbf{f}$ for each of the scenarios in the demand uncertainty set ($\mathcal{D}$). Formally, we define $\mathbf{f}^k$ as the flow decision variable for scenario $k = 1, \ldots, N$ and ζ^k to be the dual variable corresponding to scenario k for the travel time uncertainty set constraint $\sum_{(i,j)\in\mathcal{E}} w_{ij}^k \leq B$ in $\mathcal{C}$. The flow variable $\mathbf{f}^k$ corresponding to the limiting scenario for the first set of constraints in (3) is an optimal flow vector for (2).

THEOREM 2. ***R-NFF*** *is equivalent to the following mixed-integer linear optimization problem:*

$$
\begin{aligned}
\underset{\mathbf{y},\mathrm{t},\mathbf{f^k},\zeta^\mathrm{k}}{\text{minimize}} \quad & t \\
\text{subject to} \quad & t \geq \hat{\mathbf{c}}'\mathbf{f}^k + \zeta^k B, \quad k = 1, ..., N, \\
& \mathbf{A}\mathbf{f}^k \leq \boldsymbol{\alpha}\mathbf{I}\mathbf{y} - \mathbf{d}^k, \quad k = 1, ..., N, \\
& \mathbf{f}^k \leq \zeta^k \mathbf{e}, \quad k = 1, ..., N, \\
& \mathbf{f}^k \geq \mathbf{0}, \quad k = 1, ..., N, \\
& \zeta^k \geq 0, \quad k = 1, ..., N, \\
& \mathbf{e}'\mathbf{y} = P, \\
& \mathbf{y} \in \{0,1\}^n.
\end{aligned}
\tag{3}
$$

Formulation (3) quickly becomes intractable as the number of scenarios increases and the size of the graph grows. We address these two challenges in the next two subsections. First, we develop a scenario generation algorithm that scales efficiently with the number of scenarios. Similar decomposition algorithms have been developed by Atamturk and Zhang (2007), Zeng and Zhao (2013), Gabrel et al. (2014) and Chan (2017) for related two-stage problems. Second, we develop a heuristic to efficiently solve the master problem associated with the scenario generation approach.

3.3.2. Scenario Generation. Consider a subset of the demand scenarios $\mathcal{D}_{|S|} = \{\mathbf{d}^1, \mathbf{d}^2, ..., \mathbf{d}^{|S|}\} \subset \mathcal{D}$, where S is an index set for the vectors in $\mathcal{D}_{|S|}$, and the corresponding relaxation of formulation (3) with $\mathcal{D}_{|S|}$ in place of $\mathcal{D}$:

$$
\begin{aligned}
\textbf{R-NFF-MP:} \quad \underset{\mathbf{y},\mathrm{t},\mathbf{f^s},\zeta^\mathrm{s}}{\text{minimize}} \quad & t \\
\text{subject to} \quad & t \geq \hat{\mathbf{c}}'\mathbf{f}^s + \zeta^s B, \quad \forall s \in S, \\
& \mathbf{A}\mathbf{f}^s \leq \boldsymbol{\alpha}\mathbf{I}\mathbf{y} - \mathbf{d}^s, \quad \forall s \in S, \\
& \mathbf{f}^s \leq \zeta^s \mathbf{e}, \quad \forall s \in S, \\
& \mathbf{f}^s \geq \mathbf{0}, \quad \forall s \in S, \\
& \zeta^s \geq 0, \quad \forall s \in S, \\
& \mathbf{e}'\mathbf{y} = P, \\
& \mathbf{y} \in \{0,1\}^n.
\end{aligned}
\tag{4}
$$

The relaxed master problem, (4), produces a lower bound on the optimal value of (2) that can be tightened by adding additional scenarios to the set $\mathcal{D}_{|S|}$. Given an optimal solution $\bar{\mathbf{y}}$ to (4), we solve the following sub-problem, which is a linear optimization problem, for every $\mathbf{d}^k \in \mathcal{D}$:

$$\textbf{R-NFF-SP-k:} \quad Z_{SP}^k = \underset{\mathbf{f}^k,\zeta^k}{\text{minimize}} \quad \hat{\mathbf{c}}'\mathbf{f}^k + \zeta^k B$$
$$\text{subject to} \quad \mathbf{A}\mathbf{f}^k \leq \boldsymbol{\alpha}\mathbf{I}\bar{\mathbf{y}} - \mathbf{d}^k,$$
$$\mathbf{f}^k \leq \zeta^k \mathbf{e}, \tag{5}$$
$$\mathbf{f}^k \geq \mathbf{0},$$
$$\zeta^k \geq 0.$$

We choose the scenario $k^* \in \arg\max_{k=1,\ldots,N}\{Z_{SP}^k\}$ and add the decision variables $\mathbf{f}^{k^*}$ and ζ^{k^*}, plus their corresponding constraints, to (4). Hence, this approach generates both rows and columns. The scenario generation algorithm terminates when the optimal value of (4) is equal to $Z_{SP}^{k^*}$.

Finally, we comment on the structure of the subproblem (5) and connect it to a stream of research that draws an equivalence between robust optimization and regularization. Since ζ^k is being minimized in (5), the constraint $\mathbf{f}^k \leq \zeta^k \mathbf{e}$ identifies the maximum value of f_{ij}^k over all $(i,j) \in \mathcal{E}$. Thus, we can rewrite (5) as (we drop the index k for simplicity):

$$\underset{\mathbf{f}}{\text{minimize}} \quad \hat{\mathbf{c}}'\mathbf{f} + B\|\mathbf{f}\|_\infty$$
$$\text{subject to} \quad \mathbf{A}\mathbf{f} \leq \boldsymbol{\alpha}\mathbf{I}\bar{\mathbf{y}} - \mathbf{d}, \tag{6}$$
$$\mathbf{f} \geq \mathbf{0}.$$

Formulation (6) is a "regularized" shortest path problem. Without the term $B\|\mathbf{f}\|_\infty$ in the objective, (6) is exactly a shortest path problem. The extra term balances finding the shortest path with minimizing the maximum flow along any edge, which is weighted by the budget B. In our application, larger values of B correspond to higher levels of traffic uncertainty. Thus, for large B, an optimal solution to (6) would prefer to spread out the flows (smaller maximum f_{ij}), forcing nature to expend more budget to "lengthen" multiple edges. Equivalently, if flows are concentrated on a few arcs, then nature has easy targets for adding traffic to cause maximal disruption. Our reformulation elucidates a clear connection between a robust shortest path problem and a regularized shortest path problem, similar to the way equivalences have been derived in regression (Xu et al. 2010, Bertsimas and Copenhaver 2017). For example, if we replace the constraint $\sum_{(i,j)\in\mathcal{E}} w_{ij} \leq B$ in $\mathcal{C}$ with $w_{ij} \leq B, \forall (i,j) \in \mathcal{E}$, then our subproblem is equivalent to a L1-regularized (lasso) problem.

3.3.3. Master problem heuristic. To solve the large-scale, real-world instances considered in our Dhaka experiments, we require a heuristic for the master problem, which is in essence a p-median problem. Although there are many heuristics that have been developed for the p-median problem, we cannot apply these algorithms directly because they are unable to handle edge-length uncertainty. Instead, we adapt the heuristic developed by Densham and Rushton (1992) for the classical p-median problem. This heuristic, designed for large-scale problems, leverages both the

interchange heuristic proposed by Teitz and Bart (1968) and the alternate heuristic proposed by Maranzana (1964). A key benefit of this type of algorithm is that it scales well with both the size of the graph and the number of facilities (P). In fact, our heuristic represents the first tractable approach to solving large-scale instances of location problems with edge-length uncertainty. Our approach involves three main phases.

Initialization phase. We initialize our algorithm by randomly selecting P nodes to serve as initial outpost locations, encoded by $\bar{\mathbf{y}}$. We solve (5) with this $\bar{\mathbf{y}}$ for every $\mathbf{d}^k \in \mathcal{D}_{\mathcal{S}}$, and identify $k^* \in \arg\max_{k \in \mathcal{S}}\{Z_{SP}^k\}$, $\mathbf{f}^{k^*}$, and ζ^{k^*}. The corresponding cost of this solution is $\hat{\mathbf{c}}'\mathbf{f}^{k^*} + \zeta^{k^*}B$. An advantage of a random initialization phase is that our algorithm can be embedded in a meta-heuristic or a simple approach that considers multiple random starts. We investigate the impact of the number of random starts in our numerical experiments.

Interchange phase. In the interchange phase, we randomly swap a current outpost location node with a candidate node that is not currently in the solution. The new objective value is calculated as before after solving (5) for every $\mathbf{d}^k \in \mathcal{D}_{\mathcal{S}}$. Swaps that reduce the objective value are accepted. We consider ℓ random interchanges per outpost location, where ℓ is a user-chosen parameter.

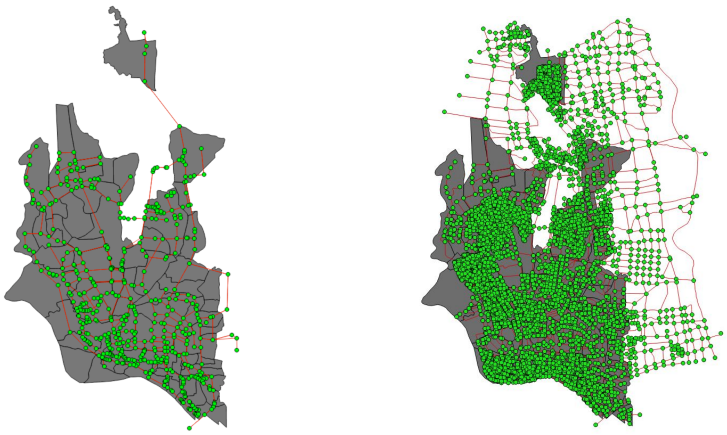
Alternate phase. In the alternate phase, we use the incumbent solution from the interchange phase to partition the network into P connected subgraphs that are disjoint from each other. Each subgraph contains exactly one outpost location and all demand nodes served by that outpost. We solve (4) for $P = 1$ (i.e., the robust 1-median problem) on each subgraph to determine the optimal outpost location. We then re-combine all subgraphs and the new optimal outpost locations to obtain an updated set of outpost locations, $\bar{y}$, in the full network. We compute the cost of this solution as before, by solving (5) for every $\mathbf{d}^k \in \mathcal{D}_{\mathcal{S}}$. The alternate phase continues to partition and re-combine outpost locations until it has reached a local optimum. The algorithm then proceeds back to the interchange phase.

Termination. The algorithm iterates between the interchange and alternate phases until a solution from the alternate phase is found that does not result in any swaps during the interchange phase. The algorithm terminates with a solution to a single instance of the master problem (4).

Integration with scenario generation algorithm. The returned solution from the heuristic either terminates the scenario generation algorithm (when the the optimal value of (4) is equal to $Z_{SP}^{k^*}$) or is used as input to the sub-problem (5).

4. Application to Dhaka

In this section, we outline the application of our methodology to Dhaka, Bangladesh. Section 4.1 describes the road networks, Section 4.2 details our approach for estimating spatiotemporal demand



(a) Ambulance (b) Complete

Figure 1 Two road networks overlaid on a ward map of Dhaka.

for emergency transportation, Section 4.3 outlines our travel time predictions, Section 4.4 presents our tactical simulation model, and Section 4.5 describes our experimental setup.

4.1. Road networks

We consider two different road networks in Dhaka. The first road network that we consider is the *ambulance network*. In consultation with a transportation engineer in Dhaka and using a detailed map of the entire city, we determined exactly which roads are feasible for ambulance travel (many roads are too narrow for a van ambulance). The ambulance network has 530 nodes and $1,280$ edges. A node is defined as the intersection of edges (i.e., roads). The second road network we consider is the *complete network*. This network – a superset of the ambulance network – includes all roads ranging from large arterial roads to small alleyways that can only be traversed by small vehicles like rickshaws, motorcycles, and auto-rickshaws. The complete road network has $5,358$ nodes and $16,538$ edges. Figure 1 displays both networks overlaid on Dhaka's 92 wards.

4.2. Demand for emergency transportation

In this section, we outline our framework for estimating spatiotemporal demand for emergency transportation. We do not have data on the total number of emergency transports as we would in North America because Dhaka does not have a centralized emergency medical system. Instead, we propose a two step process that leverages the limited data at our disposal (see EC.1.1 for a detailed description of our data). First, we provide a novel decomposition of a standard metric for emergency demand: the annual number of emergency trips from ward w at time τ via mode m (Section 4.2.1). Second, we develop a simulation framework to estimate the precise time and location for each emergency transport (Section 4.2.2).

4.2.1. Estimating the annual number of emergency trips. We decompose the estimated annual number of emergency trips for each ward w, time of day τ, and mode m, denoted by $d_{w,\tau,m}$, into three components:

$$d_{w,\tau,m} = \xi \, n_{w,\tau} \, \delta_{w,m}, \tag{7}$$

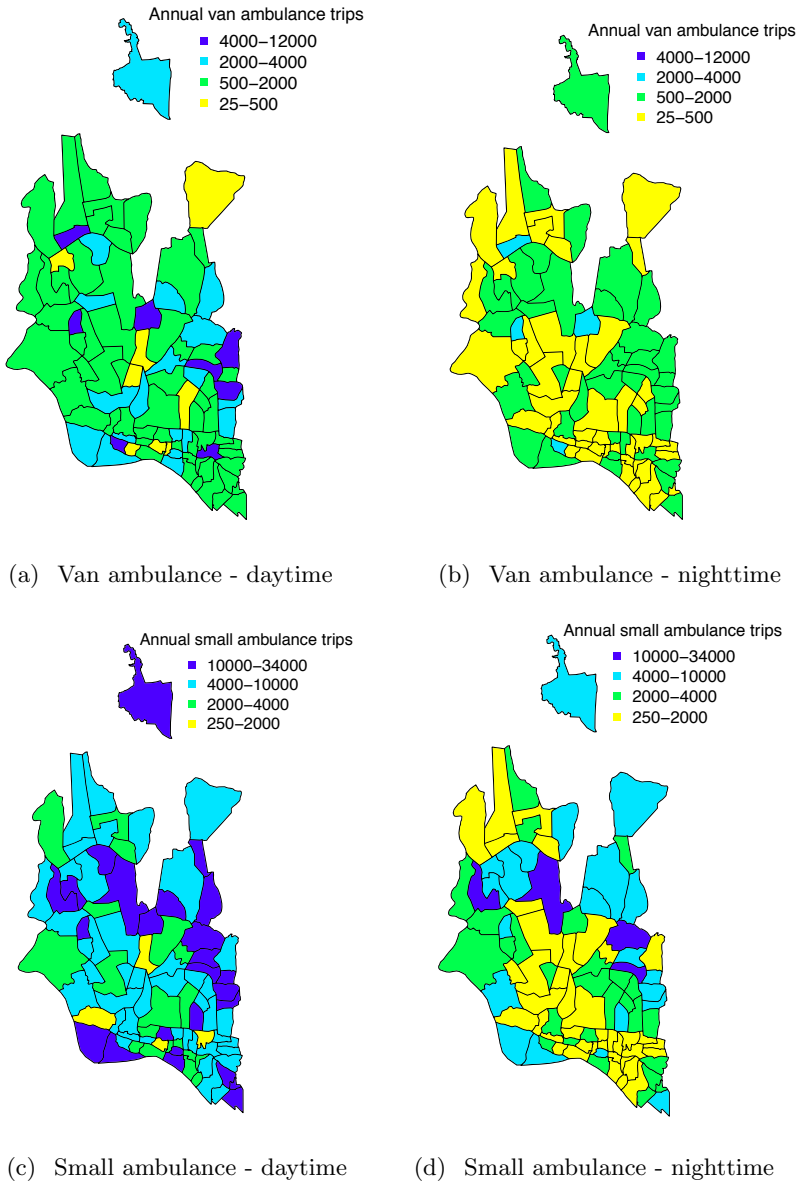
where ξ represents the average annual number of emergency trips per person, $n_{w,\tau}$ represents the population in ward w at time τ, and $\delta_{w,m}$ represents the proportion of emergency trips from ward w that arrived via mode m.

Equation (7) suggests an approach to estimating $d_{w,\tau,m}$ by estimating its constituent terms ξ, $n_{w,\tau}$, and $\delta_{w,m}$. To do this, we consider two time periods (daytime (D) and nighttime (N)) and two modes of transport (van ambulance (V) and small ambulance (S)). We consider two time periods because emergency demand is known to follow a circadian rhythm (Bagai et al. 2013, McCormack and Coates 2015), meaning that demand is much higher during the day than at night. We consider two modes of transport because of the multi-modal nature of decentralized ambulance services in LMICs.

In total, there are 369 quantities to estimate: two per ward for population ($n_{w,\tau}$), two per ward for mode ($\delta_{w,m}$), and a single value for the average annual number of ED visits per person across the entire city (ξ). The estimation of these three sets of parameters are described in EC.1.2.1, EC.1.2.2, and EC.1.2.3, respectively. Figure 2 shows the final estimation for the expected annual number of daytime and nighttime trips arising from each ward, for both van and small ambulances.

We use our estimations to simulate scenarios for the uncertainty set described in Section 3.2.1. To do this, we assume that the population in each ward follows a triangle distribution on the interval between the estimated daytime and nighttime population, with a peak at the midpoint; we assume that ξ follows a triangle distribution on the estimated interval $[0.23 - 0.46]$ (see EC.1.2.2 for details), where the peak occurs at 0.40 for conservatism; and we assume that $\delta_{w,m}$ follows a truncated normal distribution with a mean equal to the predicted ward value (Figures EC.3(c) and EC.3(d)) and a standard deviation equal to the median error (see Figures EC.4(a) and EC.4(b)).

The simulated demand vectors need to be adjusted so that the demand in each ward is spread proportionally to the road network nodes in that ward. In other words, we need to map the predicted demand based on the 92 wards to the ∼500 or ∼5,000 nodes in the ambulance and complete networks, respectively. To make this adjustment, we generate a fine grid of nodes spaced 25m apart across all 92 wards, resulting in over 200,000 grid nodes. We distribute the simulated demand in each ward uniformly among the grid nodes in that ward. Then, we assign each grid node and its corresponding demand to the closest road network node using Euclidean distance.



(a) Van ambulance - daytime (b) Van ambulance - nighttime

(c) Small ambulance - daytime (d) Small ambulance - nighttime

Figure 2 Expected annual number of van and small ambulance trips arising in each ward.

4.2.2. Simulating spatiotemporal emergency trips. As written, (7) provides a high-level approach to estimate the spatial and temporal heterogeneity in emergency demand. These annual estimations are useful for our optimization model because locating ambulance outposts is a high-level strategic decision that may be fixed for long time period. However, our simulation model, which evaluates the tactical performance of the optimization results, requires the exact time and node location for each emergency trip. To do this, we develop a novel procedure that maps annual ward-based demand to a fine spatiotemporal resolution.

We approximate the time-dependent arrival rate for emergency demand with the piecewise linear function shown in Figure 3. For each ward and mode of transportation, we partition the daily

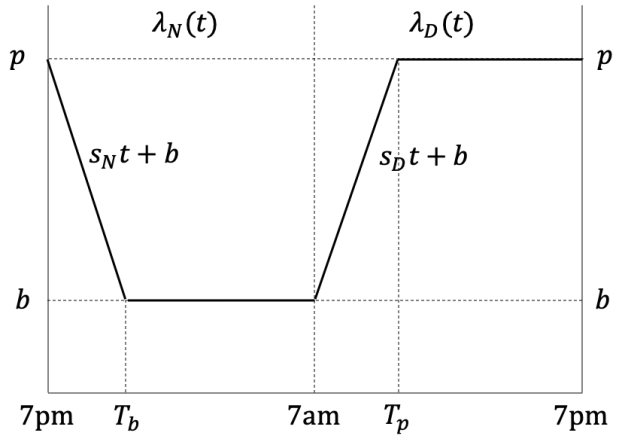


Figure 3 A visualization of $\lambda_N(t)$ and $\lambda_D(t)$ (we drop the ward and mode indices).

arrival rate function into separate daytime ($\lambda_{w,D,m}(t)$) and nighttime ($\lambda_{w,N,m}(t)$) components. We drop the w and m indices for the remainder of this section. We translate our annual estimates for emergency trips (d_τ) into daily estimates ($\hat{d}_\tau$) by assuming that each day has an equal number of expected emergency trips (i.e., $\hat{d}_\tau = \frac{d_\tau}{365}$) (Bagai et al. 2013, McCormack and Coates 2015). We also assume that $T_p = 9$am and $T_b = 11$pm, based on estimates from the literature (Bagai et al. 2013). The boundary conditions of $\lambda_D(t)$ and $\lambda_N(t)$ are set to ensure continuity of the overall function (i.e., $b = b_N = b_D$ and $p = p_N = p_D$). Hence, for each ward and mode, there are four unknowns: s_D, s_N, b, and p.

We obtain closed form solutions for the four parameters by leveraging the fact that $\hat{d}_D = \int_0^{12} \lambda_D(t)dt$ and $s_D = \frac{p-b}{T_p}$ (similar equations hold for nighttime). Once we determine the parametric form for the arrival rate function, we use the order statistic sampling method to generate the exact time for each emergency trip (Cox and Lewis 1966, Pasupathy 2011):

1. Generate the number of emergency trips: $n \sim \text{Poisson}(\hat{d}_\tau)$
2. Independently generate n random variates $t'_1, t'_2, \ldots, t'_n$ from the cumulative distribution function given by $F_\tau(t) = \frac{1}{\hat{d}_\tau} \int_0^t \lambda_\tau(t)dt$
3. Order $t'_1, t'_2, \ldots, t'_n$ and return the ordered times.

We repeat these steps for each ward (w), time of day (τ), and mode of transport (m). For each simulated 24-hour period, this procedure produces a series of times for each ward and mode that follow the distribution shown in Figure 3. Figure 4 displays one week of van ambulance emergency calls summed across all wards and binned according to the hour of the day.

We map the times to nodes on the road network using a modified version of the procedure outlined in Section 4.2.1, which maps the total ward demand to nodes in the road network using a finely spaced grid. In other words, each road network node captures the demand of g grid nodes. For a given ward, we randomly assign each trip to a road network node with a probability corresponding

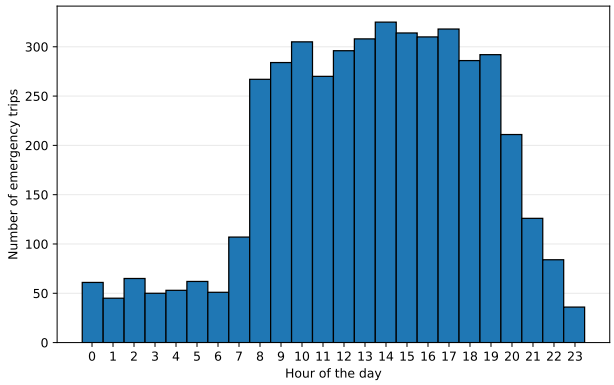


Figure 4 A histogram displaying one week of van ambulance emergency calls summed across all wards and binned according to the hour of the day.

to the proportion of grid nodes captured by that road network node. In summary, our demand simulation procedure estimates the exact time and road network node location for each van and small ambulance trip. We use the estimates as input to the tactical simulation model described in Section 4.4.

4.3. Travel time prediction

We compare four machine learning models and several baseline approaches for predicting travel time on the Dhaka road network according to time of day and day of week, using a dataset of vehicle trips collected by our custom-made GPS devices. We find that a random forest model performs the best, with a $43.3 - 64.2\%$ improvement in prediction accuracy (measured with root mean squared error) over the baseline approaches. See Section EC.2.2 for details. We use the final random forest model trained with all available data to estimate the baseline travel time $\hat{c}_{ij}$ for each edge, which is used as part of the uncertainty set described in Section 3.2.2. To the best of our knowledge, this paper is the first to use real travel time data from a LMIC for optimization.

4.4. Tactical simulation model

The main focus of the simulation model is to capture the effects of congestion (i.e., waiting time) on overall response times. Our approach is similar to the model developed by McCormack and Coates (2015) with three key differences (due to the lack of historical data):

1. We simulate the time and location of emergency trips using the procedure outlined in Section 4.2.2.

2. We simulate the travel time between the ambulance base and the patient (and between the patient and hospital) by solving the robust shortest path problem with edge lengths predicted according to the hour of the day and day of the week.

3. We simulate the scene time using an exponential distribution with an average scene time of 15 minutes because there is no data on scene time from Dhaka (Brown et al. 2016, Nagata et al. 2016).

EC.6 provides a detailed description of our simulation framework. The output of the simulation model is the waiting time, drive time, scene time, transport to hospital time, and the return to home base travel time (if applicable) for each emergency trip. We use response time to denote the summation of waiting time and drive time.

4.5. Experimental setup

For all experiments in Section 5, we solve formulation (3) using the heuristic scenario generation (HSGen) algorithm with 10 random starts and 10 random interchanges. The optimization model solutions are then input to the simulation model described in Section 4.4 to evaluate the tactical system performance over a seven day period. We use a three day warm up period to reach the steady state system.

Unless otherwise indicated, we use the uncertainty sets described in Sections 3.2.1 and 3.2.2 with 100 ambulance demand scenarios and a travel time budget (B) of 1000 seconds. Through a detailed sensitivity analysis on the travel time budget (see EC.7.2), we find that optimizing outpost locations using a budget of 1000 seconds generates solutions that perform comparably to solutions optimized for other budgets. Sections 5.1, 5.2, and 5.4 use the ambulance road network, while Section 5.3 uses both the ambulance and complete road networks. All optimization experiments were programmed using `MATLAB 2016a` and linear programming sub-problems were solved using `Gurobi 7.0`. All simulations were programming using `Python 3.5`. The HSGen algorithm was able to solve each large-scale problem instance in under one hour, and most were solved within 10 minutes. These real-world instances are comparable with the largest problems solved in the facility location literature and papers that focus on problems this large exclusively use heuristic methods (Fischetti et al. 2017).

5. Policy experiments

In this section, we demonstrate the application of our models using data from Dhaka. Each of the following subsections addresses a policy question relevant to the design of an emergency response system: 1) Should different outposts be used for different times of day? (Section 5.1) 2) What performance improvements are possible by optimizing outpost locations? (Section 5.2) 3) How much can the system be improved by using small ambulances? (Section 5.3) and 4) How important is it to consider uncertainty when designing an emergency response network? (Section 5.4). EC.7.1, EC.7.2, and EC.7.3 quantify the impact of the number of ambulances per outpost, the impact of the robust travel time budget, and the differences between the optimization and simulation results, respectively.

5.1. Should different outposts be used for different times of day?

In this section, we quantify the benefit of using different outpost locations for different times of day and days of the week, which we refer to as temporal snapshots. In many developed countries, demand is estimated at a fine spatiotemporal resolution (Zhou et al. 2015), allowing ambulances to be repositioned and response to be optimized for different snapshots (van Barneveld et al. 2017, Nasrollahzadeh et al. 2018). However, there is a second key motivation for intra-day changes in ambulance locations in LMICs, which is the impact of changing traffic patterns on travel times. We observed first-hand on several occasions during our field work the dramatic increase in travel times in different parts of the city during the evening rush hour. While traffic is less of a concern in high-income countries, emergency vehicles typically face the same traffic conditions as regular road users in LMICs since other vehicles do not (or cannot) yield to ambulances. Thus, our experiments in this section compare the performance of a system that changes outpost locations according to time of day versus a configuration that keeps the ambulance outposts static at all times.

We use baseline travel times for three different temporal snapshots: weekday rush hour (Monday between 6pm and 7pm), weekday overnight (Monday between 2am and 3am), and weekend midday (Saturday between 12pm and 1pm). We use daytime population scenarios for rush hour and we use nighttime population scenarios for weekday overnight and weekend midday. For all three snapshots, we solve (3) with $P = 20$. We simulate the performance of each set of outpost locations on all three temporal snapshots using seven ambulances per outpost, chosen based on our investigation of the impact of the number of ambulances per outpost (see EC.7.1 for details).

Figure 5 displays the distribution of response times from the simulation model corresponding to outpost locations optimized for each of the three temporal snapshots. We find that the median ambulance response time is 58.0 and 38.1 minutes longer during rush hour as compared to overnight and weekend, respectively. During rush hour, the rush hour-optimized locations have a median response time that is 14.4 min (15.0%) and 12.5 min (13.4%) and faster than the median response time of the overnight- and weekend-optimized locations, respectively. During overnight and weekend, the rush hour-optimized locations are 5.9 min (20%) and 0.2 min (0.5%) better than the best outpost locations, respectively.

5.1.1. Discussion and policy implications. Our results suggest that ambulance providers in Dhaka do not need to optimize outpost locations by time of day or day of week. Instead, providers can use static outpost locations optimized for daytime rush hour. The rush hour-optimized locations produce significant gains in response time during rush hour, while maintaining similar performance to specialized outpost locations at other times of the day. This finding is important because it supports reduced system complexity by removing the need to reposition emergency vehicles. In

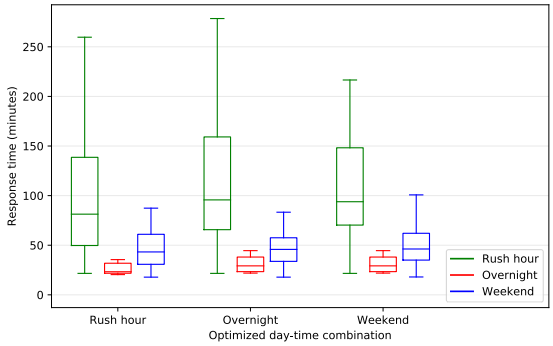


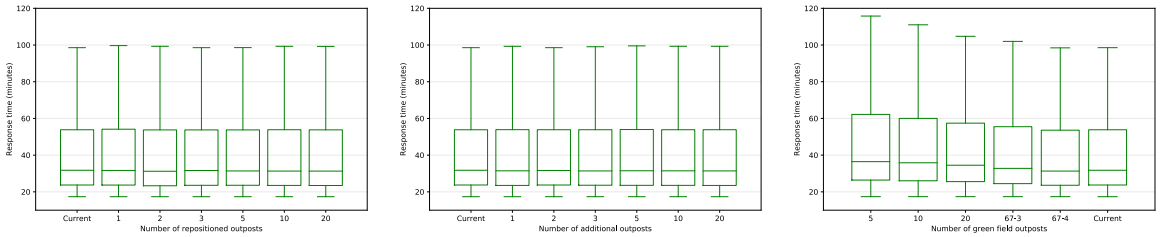
Figure 5 The response time performance of outpost locations optimized for one specific snapshot and applied to other snapshots.

LMICs, it has been shown that complex solutions are far less likely to succeed compared to simple ones (Bradley et al. 2017). Thus, our rush hour-optimized solution is recommended since it is optimal for the busiest time of day, close to optimal otherwise, and more likely to be implemented than a solution that involves regular repositioning.

5.2. What performance improvements are possible by optimizing outpost locations?

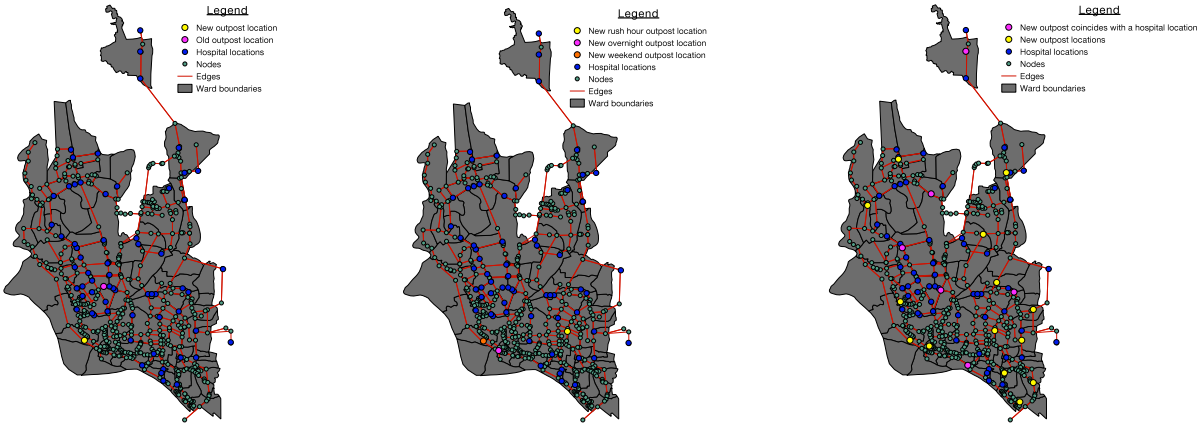
Given the results in Section 5.1, we turn our attention to designing a static ambulance emergency response network for daytime rush hour and quantifying the gain from shifting away from the current practice of having hospital-based ambulances. There are currently 87 hospitals with emergency departments. Many of these hospitals have their own ambulance services, while others rely on private services. In both cases, ambulance providers typically position their fleets at the hospitals. We estimate the total number of ambulances in Dhaka by assuming that each of the 19 government hospitals has a fleet of seven ambulances, while each of the 68 private hospitals has a fleet of two ambulances. We obtain these estimates based on the volume differences between government and private hospitals, and based on our field experience. In total, we estimate that Dhaka has approximately 269 ambulances. To calculate response times, we assign each hospital and its fleet of ambulances to the closest node on the ambulance road network, resulting in 67 unique locations. Using these locations, Section 5.2.1 determines the baseline performance of the current hospital-based outpost locations in Dhaka.

We then consider three policy experiments for improving baseline ambulance response times that may inform the decision making of existing ambulance providers interested in improving or expanding their operations, as well as possibly new entrants or the government looking to design a system from scratch. In particular, Section 5.2.2 quantifies the value of repositioning current outposts. Section 5.2.3 quantifies the value of adding additional outpost locations to the current network. Finally, Section 5.2.4 quantifies the performance of an ambulance network that is designed



(a) Repositioned outposts (b) Additional outposts (c) New outposts

Figure 6 Response time performance for different emergency response network configurations. The 67-3 and 67-4 labels in (c) correspond to 67 outposts with 3 and 4 ambulances per outpost, respectively.



(a) Single repositioned outpost during rush hour

(b) Single additional rush hour, overnight, and weekend outpost

(c) 20 new outposts during rush hour

Figure 7 A visualization of the outpost locations for different improvment policies.

from scratch, without consideration of current outpost locations. We measure performance of the ambulance networks over an entire week using our simulation model.

5.2.1. What is the baseline performance of current hospital-based outpost locations? The median response time of the current outpost locations is 47.2, 24.1, and 33.3 minutes, during rush hour, overnight, and weekend, respectively. The variability in average response time is much larger during rush hour, with a 168.6 minute difference between the best and worst response times, compared to 91.0 and 94.5 minute differences between the best and worst response times during overnight and weekend, respectively.

5.2.2. What is the value of repositioning current outpost locations? We use a modified version of HSGen for these experiments. For each random start, we randomly choose the required number of outposts to reposition from the current locations and fix all the other outposts

for the remainder of the algorithm. As a result, the problem is reduced to determining the location of a specified number of outposts given a set of incumbent outpost locations. When an outpost is repositioned, all ambulances at that outpost are also repositioned.

Figure 6(a) displays the distribution of response times for each number of repositioned outposts. Repositioning outposts provides only marginal improvements in response time. For example, relocating one outpost provides no improvement in response time, while repositioning 20 outposts provides only a 0.5 min (1.2%) response time improvement.

Figure 7(a) shows a representative solution from the one-outpost repositioning problem. The current outpost locations are blue, the old outpost location is pink, and the repositioned outpost location is yellow. Although the Euclidean distance between the old and new outpost locations is only 3.3 km, the time to travel between them is 403.8 minutes during rush hour, meaning that the new location can provide quicker service to an area that would otherwise see significant delays during rush hour.

5.2.3. What is the value of adding additional outpost locations to the current network? Figure 6(b) displays the distribution of response times as a function of outposts added. Note that additional outposts are selected from a candidate set that includes all nodes without a facility and additional outposts are staffed with two ambulances. The addition of new outposts provides nearly the same value as repositioning outposts, suggesting that some current outposts provide minimal value. Figure 7(b) displays the location of a single additional rush hour, overnight, and weekend outpost. The additional weekend outpost is the same as the repositioned outpost shown in Figure 7(a). Although the additional rush hour location is quite different, it is located in an area with many business, government offices, and universities; during rush hour, this area is particularly busy with people commuting home.

5.2.4. What is the value of designing a new emergency response network? Figure 6(c) displays the response time distribution for newly optimized networks. Ambulances are distributed uniformly over the the new network outposts so that the entire system has a total of 140 ambulances (129 fewer than the current system), unless otherwise stated. We observe steady response time improvements for all new greenfield solutions. The response time performance of 20 new outposts is only 2.9 minutes (6.3%) worse than the current 67 outposts, suggesting that similar performance can be achieved with only one-third of the current locations and roughly half as many ambulances. Figure 7(c) displays the location of 20 new outposts in relation to the current outposts. The new outposts are more strategically spread out compared to the current hospital locations that are concentrated in central Dhaka. For example, new outposts are added to the southwest and east of the city, which include low-income areas there were previously under-served by hospital-based outposts.

5.2.5. Discussion and policy implications. Our first two experiments (Sections 5.2.2 and 5.2.3) measure gains from local changes to the current network. The results suggest that policies focused on repositioning current outposts or adding additional outposts provide little value. Furthermore, the improvements from repositioning current outposts are nearly identical to adding new outposts. Practically, this result suggests that some of the current outpost locations are contributing very little to the overall response time calculation (i.e., they are rarely the fastest responding outpost to any given demand point).

If we consider a move towards centralization and a complete redesign of the current system, our third experiment shows that we can achieve roughly the performance of the current system with one-third of the outpost locations and half as many ambulances. The non-governmental organization behind the newly implemented 999-number or a formal government agency seeking to implement a centralized emergency response system may consider a complete re-design. Examining the 20 optimized locations from this experiment we find that nine of them coincide with hospital locations, while the other 11 are located off-site. Another way to view these results is that over 40 of the current hospital-based ambulance outposts can be removed without much impact on city-wide response times, or put to better use by concentrating the ambulances at fewer, more strategically located outposts.

Our experiments recommend putting outposts in the southwest, southeast, and northeast wards, suggesting that these areas are generally under-served. The southwest seems particularly underserved since both repositioned and newly added outposts are located there. Knowing the demographics of the city, this result is not particularly surprising: the southwest wards form part of Old Dhaka and encompasses very dense low-income areas (see Figure EC.1) that have poor access to emergency transportation.

Overall, the key takeaway is that the current ambulance network in Dhaka is a dominated solution: response times in Dhaka can be significantly reduced without adding new resources, or equivalently, many fewer resources can be employed to match the current level of performance. Our modeling framework can play a pivotal role in the process to help decision makers strategically position their current ambulance resources. Of course, complementary initiatives will be required to achieve these gains, such as better public education about emergency medical transport and awareness of the newly created 999 number, which became operational in December 2017.

5.3. How much can the system be improved by using small ambulances?

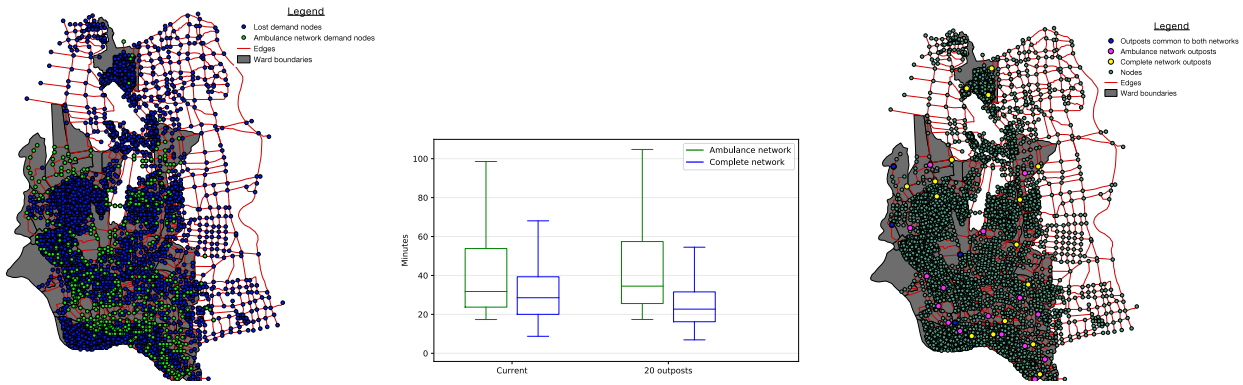
In this section, we consider the hypothetical situation where the city is served by a fleet of small ambulances that are able to traverse every road in the complete road network. Compared to the ambulance road network, the complete road network provides access to a larger portion of the

city, including many dense low-income areas that are not accessible to van ambulances. In some areas of the city, an entire sub-network of the complete road network is reduced to a single node in the ambulance road network. However, the distance between nodes in the sub-network and the nearest ambulance network node may be quite far, and it may be unrealistic to assume patients will coordinate multiple modes of transportation for different legs of their trip. As a result, we hypothesize that much of the emergency demand that arises from these low-income areas is lost or unserved. In Section 5.3.1, we quantify the potential emergency demand lost as a result of lack of access via the ambulance road network. To do this, we generate 100 demand scenarios using the prediction models for both van and small ambulance demand from Section EC.1.2 and map this demand to nodes on the complete road network. Nodes that belong to the ambulance network retain the sum of the van and small ambulance demand, while demand corresponding to complete network nodes that are not present in the ambulance network are assumed to be lost.

In addition to potentially capturing more demand, the complete road network also provides more routing options for small ambulances, which in turn may enable them to better avoid congestion and deal with travel time uncertainty. In Section 5.3.2, we quantify the value of increased routing flexibility provided by the complete network. We start by mapping demand (van plus small ambulance demand) to nodes in the ambulance network. Then, we use our simulation model to evaluate the response time performance of the current 67 hospital-based outpost locations as well as 20 new locations on both the ambulance and complete road networks. The 20 new locations are optimized for the corresponding road network, so they represent two distinct solutions.

5.3.1. How much potential demand is lost by van ambulances restricted to the ambulance road network? The complete network captures an average of 769,790 small ambulance trips per year, while the ambulance road network only captures 225,559 trips per year, representing a potential loss of 544,231 ambulance (70.7%) trips. These numbers represent an upper bound on the true number of ambulance trips because we have implicitly assumed that all available ambulance trips will be captured (in reality, some may be missed). Figure 8(a) visualizes the lost ambulance demand. The 530 green nodes are those that capture demand in both the ambulance and complete networks, while the 4,828 blue nodes only capture demand in the complete network and therefore, represent lost demand for the ambulance road network.

5.3.2. What is the value of the increased routing flexibility offered by small ambulances? Figure 8(b) displays the response time performance for the current 67 baseline outpost locations and 20 new outpost locations on both the ambulance and complete road networks. The median response time of the current 67 locations on the ambulance road network is 31.8 minutes over an entire week. Small ambulances located at the same outposts are able to reduce the median



(a) Lost demand

(b) Response time performance for baseline and new locations

(c) Twenty new outpost locations optimized for the two road networks

Figure 8 Comparing the performance of the ambulance and complete road networks

response time to 28.5 minutes (a 10.1% reduction). If we consider 20 new outpost locations, we get a larger reduction in median response time of 17.8%, from 34.5 minutes to 28.4 minutes. During the busiest time of week, rush hour, the improvements are even larger at 23.7% and 35.2% for current and new outpost locations, respectively. Figure 8(c) shows the 20 new outpost locations for both the ambulance (pink nodes) and complete (yellow nodes) road networks. The four blue nodes represent outpost locations that are the same in both networks.

5.3.3. Discussion and policy implications. The results in this section represent the first attempt to provide quantitative evidence of the potential benefit of small ambulances in an LMIC.

Note that that 23% of survey respondents indicated that traditional ambulance vans were either not available or too slow to reach their location (see Section EC.1.1.2). Small ambulances offer a potential solution to both these issues. Our results have three policy implications:

1. Smaller response vehicles can potentially capture three times more emergency demand than traditional van ambulances in Dhaka. Much of the additional demand captured is generated from nodes in hard-to-reach and low-income areas, such as urban slums (southern and western clusters of nodes in Figure 8(a)). These areas are known to already suffer from poor access and availability of emergency medical care.

2. Smaller response vehicles are able to reduce the median average response time by roughly 10-18% over the entire week and 24-35% during rush hour. These reductions are entirely due to increased routing flexibility offered by having nimbler vehicles navigating a larger road network. These results may even be somewhat conservative because we did not incorporate the fact that small ambulance are typically able to travel faster than larger ambulances.

3. Our results demonstrate that the outpost locations chosen for small ambulances are very different from those chosen for traditional van ambulances. This result emphasizes the importance of

considering small ambulances independently; we cannot assume they should be positioned alongside traditional ambulances, even if the ambulance outpost locations are themselves optimized, because they are optimized for a different road network.

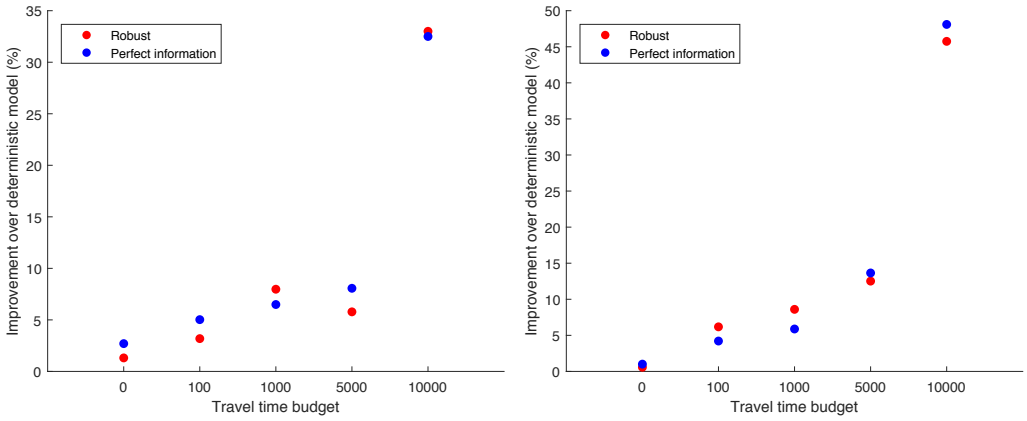
Overall, the key takeaway from these experiments is that small ambulances have the potential to not only significantly improve system efficiency through lower response times, but also simultaneously improve equity and access by capturing substantial demand in the hardest to reach areas of the city. Although both van and small ambulances have similarly limited medical capabilities and are not typically staffed by trained paramedics, further research is needed to evaluate their medical and operational impact in LMICs.

5.4. How important is it to consider uncertainty when designing an emergency response network?

In this section, we quantify the value of robustness by comparing our robust optimization model to the deterministic model (**NFF**), as well as to a perfect information formulation that solves **NFF** after the uncertainty has been realized. We focus on the situation where 20 new outposts are being located. We also examine how the performance gaps vary as the travel time budget is varied. The deterministic formulation uses the average demand and baseline travel times with no uncertainty, while the perfect information formulation finds a unique solution for each demand scenario. In this section, we directly report the optimization results, rather than evaluating them via simulation.

Figure 9 displays the response time improvement of the robust and perfect information solutions over the deterministic solution for different levels of travel time uncertainty. Because both models are solved using a heuristic, there are instances where the robust solution slightly outperforms the perfect information solution. For a travel time budget of 1000 seconds, the robust solution generates a 8.0% and 8.6% improvement over the deterministic solution in the median and worst-case average response time, respectively. Compared to the perfect information solution, the robust solution has a median average response time that is only 0.8% worse, with a worst-case average response time that is 1.9% better. As expected, the gains from the robust model increase as the size of the uncertainty set grows. For example, with a budget of 10,000 seconds, the robust solution improves upon the deterministic solution by 33.0% in median and 45.8% in worst-case average response time. At the same time, the performance of the robust solution continues to track the performance of the perfect information solution quite closely.

5.4.1. Discussion and policy implications. Our results demonstrate that a robust optimization framework tailored for the uncertainties faced by LMICs is able to produce solutions that significantly outperform solutions that do not consider uncertainty. As expected, the performance



(a) Median (b) Worst-case

Figure 9 Response time improvement of the robust and perfect information formulations over the deterministic formulation as a function of travel time uncertainty.

gains increase with the amount of uncertainty considered. Furthermore, our robust solutions are comparable to those derived from a perfect information model. Overall, these results further reinforce the importance of robustness for designing emergency response solutions in environments with substantial uncertainty characteristic of LMICs.

6. Conclusion

In this paper, we developed a comprehensive framework for emergency response optimization that combines two machine learning approaches and a simulation model with a robust optimization model tailored to address the specific challenges faced by LMICs. Our optimization model generalizes previous emergency response models in both high, middle, and low-income countries and provides a unified framework for emergency response optimization under travel time and demand uncertainty. We use two unique datasets that we collected in Dhaka, Bangladesh to train our machine learning models and build our uncertainty sets.

Using our real data and modelling framework, we address four policy questions related to the design of an emergency response system in LMICs, using Dhaka, Bangladesh as a target site. First, we demonstrated that daily population migration has a minimal impact on response times and that outpost locations optimized specifically for rush hour perform well throughout the day and week. Second, we demonstrated that a centralized network designed from a clean slate can replicate the performance of the current system using roughly half of the ambulance resources and one-third of the outpost locations currently in use. Half of the new outposts would coincide with current outpost locations, while the other half should be strategically positioned in the lower-income parts of the city. Third, we show that small ambulances may be able to capture three times more demand

than van ambulances due to their ability to access parts of the city with narrow roads such as slums. In addition, the routing flexibility offered by the larger road network available to small ambulances can reduce the median average response time by roughly 10-18% over the entire week and 24-35% during rush hour, based on our experiments. Our final experiment demonstrated that our robust optimization framework is able to produce networks with average response times that are up to 33% faster than a deterministic solution, comparable to a network designed with perfect information on the uncertainty.

Acknowledgments

The authors gratefully acknowledge Dr. Moinul Hossain for his input on early stages of the project and for leading the data collection efforts. The authors thank Mehedi Hasan for his help with demand data collection, Mahfuzur Rahman Siddiquee for his help with travel time data collection, and all those who volunteered in Dhaka. We are grateful to Prof. Yu-Ling Cheng and Dr. Laurie Morrison for support and advice throughout this project. This research was supported by Grand Challenges Canada and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Abdul Ghani N, Ahmad N (2017) Analysis of mclp, q-malp, and mq-malp with travel time uncertainty using monte carlo simulation. *Journal of Computational Engineering* .

Adenso-Diaz B, Rodriguez F (1997) A simple search heuristic for the mclp: Application to the location of ambulance bases in a rural region. *Omega* 25:181–187.

Ahmadi-Javid A, Seyedi P, Syam SS (2017) A survey of healthcare facility location. *Computers and Operations Research* 79:223–263.

Ahmed N, Rahman Siddiquee M, Karim R, Zaman M, Monzur R, Hossain M (2015) Map matching on sparse gps data: A perspective of a developing city. *10th International Conference of Eastern Asia Society For Transportation Studies.* 10.

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, URL `http://dx.doi.org/10.1111/j.1937-5956.2012.01362.x`.

Anderson PD, Suter RE, Mulligan R, Bodiwala G, Razzak JA, Mock C (2012) World health assembly resolution 60.22 and its importance as a health care policy tool for improving emergency care access and availability globally. *Annals of Emergency Medicine* 60:35–44.

Atamturk A, Zhang M (2007) Two-stage robust network flow and design under demand uncertainty. *Operations Research* 55:662–673.

Averbakh I (2003) Complexity of robust single facility location problems on networks with uncertain edge lengths. *Discrete Applied Mathematics* 127:505–522.

Bagai A, McNally BF, Al-Khatib SM, Myers JB, Kim S, Karlsson L, Torp-Pedersen C, Wissenberg M, van Diepen S, Fosbol EL, Monk L, Abella BS, Granger CB, Jollis JG (2013) Temporal differences in out-of-hospital cardiac arrest incidence and survival. *Circulation* 128(24):2595–2602.

Baker JR, Fitzpatrick KE (1986) Determination of an optimal forecast model for ambulance demand using goal programming. *The Journal of the Operational Research* 37:1047–1059.

Baron O, Milner J, Naseraldin H (2011) Facility location: A robust optimization approach. *Production and Operations Management* 20:772–785.

Basar A, Catay B, Unluyurt T (2011) A multi-period double coverage approach for locating the emergency medical service stations in istanbul. *Journal of the Operational Research Society* 62:627–637.

Basar A, Catay B, Unluyurt T (2012) A taxonomy for emergency service station location problem. *Optimization letters* 6:1147–1160.

Bennett VL, Eaton DJ, Church RL (1982) Selecting sites for rural health workers. *Social Science and Medicine* 16:63–72.

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

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

Berchet C (2015) Emergency care services: trends, drivers, and interventions to manage the demand. *Organization for Economic Co-operation and Development* .

Berman O, Hajizadeh I, Krass D (2013) The maximum covering problem with travel time uncertainty. *IIE Transactions* 45:81–96.

Bertsimas D, Copenhaver MS (2017) Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research* In Press.

Bradley BD, Jung T, Tandon-Verma A, Khoury B, Chan TCY, Cheng YL (2017) Operations research in global health: a scoping review with a focus on the themes of health equity and impact. *Health Research Policy and Systems* 15:32.

Brandeau ML, Larson RC (1986) Extending and applying the hypercube queueing model to deploy ambulances in boston. *National Emergency Training Center* .

Brotcorne L, Laporte G, Semet F (2003) Ambulance location and relocation models. *European Journal of Operational Research* 147(3):451 – 463, ISSN 0377-2217, URL `http://dx.doi.org/https://doi.org/10.1016/S0377-2217(02)00364-8`.

Brown JB, Rosengart MR, Forsythe RM, Reynolds BR, Gestring ML, Hallinan WM, Peitzman AB, Billiar TR, Sperry JL (2016) Not all prehospital time is equal: Influence of scene time on mortality. *J Trauma Acute Care Surg* 81(1):93–100, ISSN 2163-0763 (Electronic); 2163-0755 (Print); 2163-0755 (Linking), URL `http://dx.doi.org/10.1097/TA.0000000000000999`.

Budge S, Ingolfsson A, Zerom D (2010) Empirical analysis of ambulance travel times: the case of calgary emergency medical services. *Management Science* 56:716–723.

Carson YM, Batta R (1990) Locating an ambulance on the amherst campus of the state university of new york at buffalo. *Interfaces* 20:43–49.

Chan TCY (2017) Rise and shock: Optimal defibrillator placement in a high-rise building. *Prehospital Emergency Care* 21(3):309–314, URL `http://dx.doi.org/10.1080/10903127.2016.1247202`, pMID: 27858504.

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

Chanta S, Mayorga ME, McLay LA (2014) Improving emergency service in rural areas: a bi-objective covering location model for ems systems. *Annals of Operations Research* 221:133–159.

Chen B, Lin CS (1998) Minmax regret robust 1-median location on a tree. *Networks* 31:93–103.

Cox D, Lewis P (1966) *The statistical analysis of series of events* (John Wiley and Sons).

Densham PJ, Rushton G (1992) A more efficient heuristic for solving large p-median problems. *Papers in Regional Science* 71:307–329.

Eaton DJ, Héctor ML, Sánchez U, Morgan J (1986) Determining ambulance deployment in santo domingo, dominican republic. *Journal of the Operational Research Society* 113–126.

Enayati S, Mayorga ME, Rajagopalan HK, Saydam C (2018) Real-time ambulance redeployment approach to improve service coverage with fair and restricted workload for ems providers. *Omega* 79:67 – 80, ISSN 0305-0483, URL `http://dx.doi.org/https://doi.org/10.1016/j.omega.2017.08.001`.

Fischetti M, Ljubic I, Sinnl M (2017) Redesigning benders decomposition for large-scale facility location. *Management Science* 63:2146–2162.

Fujiwara O, Makjamroen T, Gupta KK (1987) Ambulance deployment analysis: a case study of bangkok. *European Journal of Operational Research* 31:9–18.

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

Goins S, Conroy MB (2015) New york state all payer emergency room visits. Technical report, New York State Department of Health Statistical Brief.

Hakimi SL (1964) Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations research* 12:450–459.

Hakimi SL (1965) Optimum distribution of switching centers in a communication network and some related graph theoretic problems. *Operations Research* 13:462–475.

Hofleitner A, Herring R, Abbeel P, Bayen A (2012a) Learning the dynamics of arterial traffic from probe data using a dynamic bayesian network. *IEEE Transactions on Intelligent Transportation Systems* 13:1679–1693.

Hofleitner A, Herring R, Bayen A (2012b) Arterial travel time forecast with streaming data: A hybrid approach of flow modeling and machine learning. *Transportation Research Part B* 46:1097–1122.

Ingolfsson A, Budge S, Erkut E (2008) Optimal ambulance location with random delays and travel times. *Health Care management science* 11:262–274.

Jain V, Sharma A, Subramanian L (2012) Road traffic congestion in the developing world. *Proceedings of the 2nd ACM Symposium on Computing for Development* .

Kamenetzky RD, Shuman LJ, Wolfe H (1982) Estimating need and demand for prehospital care. *Operations Research* 30:1148–1167.

Karim MZ, Hansen EL, Ahmad BU, Lahiry S (2009) A retrospective study of illness and admission pattern of emergency patients utilizing a corporate hospital in dhaka, bangladesh: 2006-2008. *The ORION* 32:1–7.

Karp RM (1972) *Complexity of computer computations* (Springer).

Kobusingye OC, Hyder AA, Bishai D, Hicks ER, Mock C, Joshipura M (2005) Emergency medical systems in low- and middle-income countries: recommendations for action. *Bulletin of the World Health Organization* 83:626–631.

Kok A, Hans E, Schutten J (2012) Vehicle routing under time-dependent travel times: The impact of congestion avoidance. *Computers and Operations Research* 39(5):910 – 918, ISSN 0305-0548, URL `http://dx.doi.org/https://doi.org/10.1016/j.cor.2011.05.027`.

Kolesar P, Walker W, Hausner J (1975) Determining the relation between fire engine travel times and travel distances in new york city. *Operations Research* 23:614–627.

Levine AC, Gadiraju S, Goel A, Johar S, King R, Arnold K (2007) International emergency medicine: a review of the literature. *Acad Emerg Med*, 182–1833 (14).

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, ISSN 1432-5217, URL `http://dx.doi.org/10.1007/s00186-011-0363-4`.

Lungu K, Kamfose V, Hussein J, Ashwood-Smith H (2001) Are bicycle ambulances and community transport plans effective in strengthening obstetric referral systems in southern malawi? *Malawi Medical Journal* 12:16–18.

Macintyre K, Hotchkiss DR (1999) Referral revisited: community financing schemes and emergency transport in rural africa. *Soc Sci Med*, 1473–1487 (49).

Maranzana FE (1964) On the location of supply points to minimize transport costs. *Operational Research Quarterly* 15:261–270.

Matteson DS, McLean MW, Woodard DB, Henderson SG (2011) Forecasting emergency medical service call arrival rates. *Annals of Applied Statistics* 5:1379–1406.

Maxwell MS, Restrepo M, Henderson SG, Topaloglu H (2010) Approximate dynamic programming for ambulance redeployment. *INFORMS Journal on Computing* 22(2):266–281, URL http://dx.doi.org/10.1287/ijoc.1090.0345.

McCormack R, Coates G (2015) A simulation model to enable the optimization of ambulance fleet allocation and base station location for increased patient survival. *European Journal of Operational Research* 247(1):294–309, URL http://dx.doi.org/https://doi.org/10.1016/j.ejor.2015.05.040.

Melo MT, Nickel S, Saldanha-Da-Gama F (2009) Facility location and supply chain management: A review. *European journal of operational research* 196:401–412.

Mirchandani PB, Odoni AR (1979) Locations of medians on stochastic networks. *Transportation Science* 13:85–97.

Mirchandani PB, Oudjit A (1980) Localizing 2-medians on probabilistic and deterministic tree networks. *Networks* 10:329–350.

Nagata I, Abe T, Nakata Y, Tamiya N (2016) Factors related to prolonged on-scene time during ambulance transportation for critical emergency patients in a big city in japan: a population-based observational study. *BMJ Open* 6(1), ISSN 2044-6055, URL http://dx.doi.org/10.1136/bmjopen-2015-009599.

Nasrollahzadeh AA, Khademi A, Mayorga ME (2018) Real-time ambulance dispatching and relocation. *Manufacturing & Service Operations Management* 20(3):467–480, URL http://dx.doi.org/10.1287/msom.2017.0649.

Nations U (2010) The millennium development goals report. Technical report, United Nations.

Neumann J (1928) Zur theorie der gesellschaftsspiele. *Math. Annalen* 100:295–320.

Noyan N (2010) Alternate risk measures for emergency medical service system design. *Annals of Operations Research* 181:559–589.

of Statistics BB (2010) Report of the household income and expenditure survey. Technical report, Ministry of Planning.

Organization TWH (2013) The world health report 2013: Research for universal health coverage. Technical report, The World Health Organization.

Owen SH, Daskin MS (1998) Strategic facility location: A review. *European journal of operational research* 111:423–447.

Pasupathy R (2011) *Generating Nonhomogeneous Poisson Processes* (American Cancer Society), ISBN 9780470400531, URL http://dx.doi.org/10.1002/9780470400531.eorms0356.

Pojani D, Stead D (2015) Sustainable urban transport in the developing world: beyond megacities. *Sustainability* 7:7784–7805.

PoSaw LL, Aggarwal P, Bernstein SL (1998) Emergency medicine in the new delhi area, india. *Ann Emerg Med* 32:609–615.

Post EL (1944) Recursively enumerable sets of positive integers and their decision problems. *Bulletin of the American Mathematical Society* 50:284–316.

Raftery KA (1996) Emergency medicine in southern pakistan. *Emerg Med* 27:79–93.

Razzak JA, Kellerman AL (2002) Emergency medical care in developing countries: is it worthwhile? *Bulletin of the World Health Organization* 80:900–905.

ReVelle CS, Swain RW (1970) Central facilities location. *Geographical Analysis* 2:30–42.

Salman FS, Yücel E (2015) Emergency facility location under random network damage: Insights from the istanbul case. *Computers and Operations Research* 62:266–281.

Savas ES (1969) Simulation and cost-effectiveness analysis of new york's emergency ambulance service. *Management Science* 15(12):B–608–B–627, URL `http://dx.doi.org/10.1287/mnsc.15.12.B608`.

Saydam C, Rajagopalan HK, Sharer E, Lawrimore-Belanger K (2013) The dynamic redeployment coverage location model. *Health Systems* 2(2):103–119, URL `http://dx.doi.org/10.1057/hs.2012.27`.

Schmid T, Kanenda O, Ahluwalia I, Kouletio M (2001) Transportation for maternal emergencies in tanzania: empowering communities through participatory problem solving. *American Journal of Public Health* 91:1589–1590.

Schmid V, Doerner KF (2010) Ambulance location and relocation problems with time-dependent travel times. *European Journal of Operational Research* 207(3):1293 – 1303, ISSN 0377-2217, URL `http://dx.doi.org/https://doi.org/10.1016/j.ejor.2010.06.033`.

Schuman LJ, Wolfe H, Sepulveda J (1977) Estimating demand for emergency transportation. *Med Care* 15:738–749.

Serra D, Marianov V (1998) The p-median problem in a changing network: the case of barcelona. *Location Science* 6:383–394.

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

Shen ZJM, Coullard C, Daskin MS (2003) A joint location-inventory model. *Transportation science* 37:40–55.

Sixtieth World Health Assembly (2007) Agenda item 12.14: Health systems: emergency-care systems. Technical report, World Health Organization.

Snyder LV (2006) Facility location under uncertainty: a review. *IIE Transactions* 38:537–554.

Sodemann M, Jakobsen MS, Molbak K, Alvarenga IC, Aaby P (1997) High mortality despite good care-seeking behaviour: a community study of childhood deaths in guinea-bissau. *Bulletin of the World Health Organization* 75:205–212.

Streatfield PK, Karar ZA (2008) Population challenges for bangladesh in the coming decades. *J Health Popul Nutr* 26:261–272.

Sudtachat K, Mayorga ME, Mclay LA (2016) A nested-compliance table policy for emergency medical service systems under relocation. *Omega* 58:154 – 168, ISSN 0305-0483, URL http://dx.doi.org/https://doi.org/10.1016/j.omega.2015.06.001.

Teitz MB, Bart P (1968) Heuristic methods for estimating generalized vertex median of a weighted graph. *Operations Research* 16:955–961.

Toregas CR, Swain RW, ReVelle CS, Bergman L (1971) The location of emergency service facilities. *Operations Research* 19:1363–1373.

Tribune D (2017) '999' emergency services begin. Technical report, Dhaka Tribune, URL https://www.dhakatribune.com/bangladesh/2017/12/12/999-emergency-services-begin/.

Trudeau P, Rousseau Jm, Ferland JA, Choquette J (1989) An operations research approach for the planning and operation of an ambulance service. *Information Systems and Operational Research* 27:95–113.

Vairaktarakis GL, Kouvelis P (1999) Incorporation dynamic aspects and uncertainty in 1-median location problems. *Naval Research Logistics* 46:147–168.

van Barneveld T (2016) The minimum expected penalty relocation problem for the computation of compliance tables for ambulance vehicles. *INFORMS Journal on Computing* 28(2):370–384, URL http://dx.doi.org/10.1287/ijoc.2015.0687.

van Barneveld T, van der Mei R, Bhulai S (2017) Compliance tables for an ems system with two types of medical response units. *Computers and Operations Research* 80:68 – 81, ISSN 0305-0548, URL http://dx.doi.org/https://doi.org/10.1016/j.cor.2016.11.013.

Vlahogianni EI, Karlaftis MG, Golias JC (2014) Short-term traffic forecasting: Where we are and where we are going. *Transportation Research Part C* 43:3–19.

Wadud M (2017) Cheap solar ambulances to speed into service in rural bangladesh. *Thomson Reuters Foundation* URL https://www.reuters.com/article/us-bangladesh-solar-ambulance/cheap-solar-ambulances-to-speed-into-service-in-rural-bangladesh-idUSKBN15T1AP.

Westgate BS, Woodard DB, Matteson DS, Henderson SG (2016) Large-network travel time distribution estimation for ambulances. *European Journal of Operational Research* 252:322–333.

Wood RK (1993) Deterministic network interdiction. *Mathematical and Computer Modelling* 17:1–18.

Woodard D, Nogin G, Koch P, Racz D, Goldszmidt M, Horvitz E (2017) Predicting travel time reliability using mobile phone gps data. *Transportation Research Part C: Emerging Technologies* 75:30 – 44, ISSN 0968-090X, URL http://dx.doi.org/https://doi.org/10.1016/j.trc.2016.10.011.

Xu H, Caramanis C, Mannor S (2010) Robust regression and lasso. *IEEE Transactions in Information Theory* 56:3561–574.

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

Zhang Y, Chenyang Y, Kan H, Cao J, Peng L, Xu J, Wang W (2014) Effect of ambient temperature on emergency department visits in shanghai, china: a time series study. *Environmental Health* 13:1–8.

Zhang ZH, Li K (2015) A novel probabilistic formulation for locating and sizing emergency medical service stations. *Annals of Operations Research* 229:813–835.

Zhou Z, Matteson DS (2016) Predicting melbourne ambulance demand using kernel warping. *Ann. Appl. Stat.* 10(4):1977–1996, URL `http://dx.doi.org/10.1214/16-AOAS961`.

Zhou Z, Matteson DS, Woodard DB, Henderson SG, Micheas AC (2015) A spatio-temporal point process model for ambulance demand. *Journal of the American Statistical Association* 110:6–15.

Electronic Companion

EC.1. Demand for emergency transportation

In this section, we provide a descriptive analysis of our census and survey data (EC.1.1) and present our methodology for estimating the components of emergency transport demand described in Section 4.2.1 (EC.1.2).

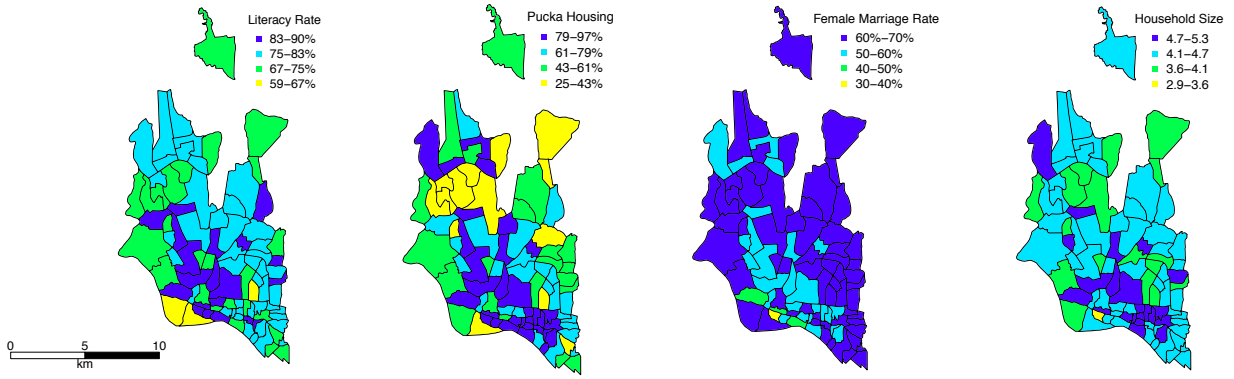
EC.1.1. Descriptive analysis

EC.1.1.1. Census data. We obtained the 2011 Dhaka census from the Bangladesh Bureau of Statistics. The census includes detailed demographic information for each of Dhaka's 92 official wards (census tracts). Dhaka occupies a very small area of roughly 300 km^2 with an estimated population of 7.35 million in 2011 (8.95 million in 2016), which is a slightly larger population than New York City in under 40% of the area. Table EC.1 provides summary statistics for key census characteristics and Figure EC.1 illustrates the variation in four demographic characteristics across Dhaka's 92 wards.

Table EC.1 Demographic summary statistics across Dhaka's 92 wards.

Characteristic	Dhaka	Individual ward		
	(all wards)	Minimum	Mean	Maximum
Population (2011 census)	7,349,324	18,170	79,884	228,870
Average household size	4.3	3.0	4.4	5.3
Male-female population ratio	1.2	0.9	1.3	2.5
Population under 19 (%)	36.2	22.9	35.5	43.5
Population over 60 (%)	4.5	2.4	4.5	7.6
Married (%)	59.0	29.4	57.4	66.7
Literacy (%)	73.7	52.5	74.9	90.4
Pukka* house (%)	58.4	24.9	66.4	96.5
Jupri* house (%)	2.1	0	1.8	11.2
Sanitary toilet (%)	58.0	7.4	60.3	98.1
Electricity (%)	98.4	92.5	98.8	99.9
Rent-free home (%)	3.4	0.4	3.4	14.7
Male-female employment ratio	2.0	0.3	2.8	12.8

*Note that a Pukka house is a solid permanent dwelling usually made from brick or stone that is reflective of higher socioeconomic status, while a Jupri house is a temporary dwelling typically made from tin and other available supplies.



(a) Literacy rate (b) Pukka housing (c) Female marriage rate (d) Household size

Figure EC.1 Ward-based rates for four demographic characteristics. A Pukka house is a solid permanent dwelling usually made from brick or stone that is reflective of higher socioeconomic status.

EC.1.1.2. Survey data. We obtained data from 2,808 surveys administered by physicians to patients arriving at emergency departments (EDs) in 16 major hospitals (9 private, 7 government) across Dhaka. The survey had 14 questions (see Table EC.2) and was administered over 30 days between July 7, 2014 and August 25, 2014. The survey data includes the chief complaint, date, time, and ward location of the emergency, mode and cost of transportation, and the time of arrival at the ED. Our survey data is unique because it includes patient travel data and as a result, we are able to provide insights on the current emergency medical system that have not been previously captured.

Figure 2(a) displays the various modes of transportation taken by patients and the costs incurred for each mode. Traditional ambulance vans were one of the least used modes of transport, comprising only 7.3% of all trips. Of the survey respondents who answered the question "Why did you not use an ambulance", 16% indicated that they tried but it was not available and 7% cited slow response times, both of which are issues that can be addressed using our approach. Another major impediment was cost. Ambulances were found to be the most expensive mode of transportation with trips typically costing more than 16 US dollars (USD). For context, the average annual income in Bangladesh is 1,260 USD (of Statistics 2010). In contrast, rickshaws and auto-rickshaws, the two cheapest modes of transportation, comprised 34% and 25% of all trips, respectively.

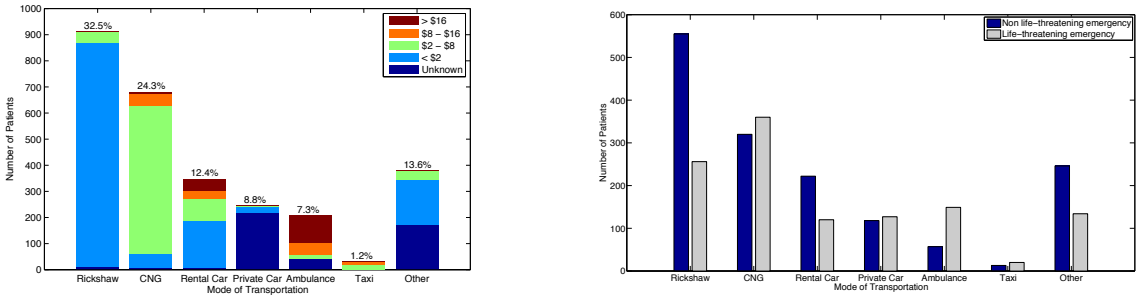
Aside from cost, another possible explanation for low ambulance utilization is that ambulances were one of the slowest modes of transportation, while rickshaws, private cars, and other, which includes walking, were the fastest modes of transportation. Table EC.3 provides a breakdown of the inter-ward distance travelled by each mode of transportation. Higher rickshaw usage, especially for

Table EC.2 Hospital survey questions and response options for all 2808 surveys.

Question	Response Type
P1. What was the approximate time of the emergency?	Free text
P2. When did you decide to leave for the hospital?	Free text
P3. Which ward did you leave from?	Ward Number
P4. What time did you leave?	Free text
P5. What was your method of transportation?	A. Own car B. Rental car C. Rickshaw D. Ambulance E. CNG F. Taxicab G. Other
P6. What was the cost of transportation (in BDT)	A. < 100 B. 100-500 C. 500-1000 D. 1000+
P7. Do you have a mobile phone?	A. Yes B. No
P8. Do you know how to contact an ambulance?	A. Yes B. No
P9. Why did you not take an ambulance?	Free text
P10. Why did you come to this hospital?	Free text
H1. Name of hospital	Free text
H2. What was the arrival time of the patient?	Free text
H3. What is the general type of injury/complication?	Free text
H4. What time did the patient first receive treatment?	Free text

short trips, is likely because rickshaws are readily available at all times and at nearly any location.

Although overall ambulance utilization in Dhaka is low, Figure 2(b) shows that it is the mode with the highest proportion of trips for life-threatening emergencies (classified by the attending physician). More than two thirds of all ambulance trips are for life-threatening emergencies, compared to less than one third of rickshaw trips. In life-threatening emergencies, ambulances become the third most common mode of transportation. This data suggests that patients recognize the importance of ambulances and are willing to use them for life-threatening emergencies. These findings also reinforce the importance of considering multiple vehicles types in LMICs.



(a) Number of trips and cost for each mode (b) Mode of transportation according to severity

Figure EC.2 Histograms of trip characteristics for each mode of transportation.

Table EC.3 A breakdown of the inter-ward distance travelled by each mode of transportation.

Mode	Trips within ward (%)	Median travel distance for all trips (m)	Median travel distance for out of ward trips (m)
Rickshaw	30.2	1358	1544
CNG	6.8	4749	5233
Rental Car	7.8	6041	7018
Private Car	41.8	1367	2462
Ambulance	9.2	3379	3670
Taxi	5.0	5737	6262
Other	24.0	2017	4041

EC.1.2. Estimating the annual number of emergency trips

In this section, we present our methodology for estimating the three components that comprise the annual number of emergency trips: the population in ward w at time τ (EC.1.2.1), the average annual number of ED visits per person (EC.1.2.2), and the proportion of ED visits from ward w arriving via mode m (EC.1.2.3).

EC.1.2.1. Estimating population ($n_{w,\tau}$). In this section, we estimate both the daytime and nighttime population for each ward (i.e., $\tau \in \{D, N\}$). Dhaka has a major difference in the spatial distribution of the daytime and nighttime population, due to daily migration. The magnitude of the daily migration out of Dhaka is estimated to be over $700,000$ as many people leave the city during the day to work in the surrounding industrial areas. We estimate the daytime population in each ward from the Earthquake Vulnerability Assessment of Dhaka, which was conducted by the Government of Bangladesh with support from the United Nations Development Programme. This assessment estimates the total population in each ward during the daytime working hours. In 2008, the total daytime population in Dhaka was estimated to be 6.63 million people. The nighttime population in each ward is obtained directly from the 2011 census and was estimated to be 7.35 million

people. The population of Dhaka has been consistently growing at a rate of approximately $320,000$ people per year (Streatfield and Karar 2008). Under the assumption that each ward is growing at a rate proportional to its population, we estimate the total 2016 daytime and nighttime population in Dhaka to be 8.24 and 8.95 million, respectively. Figures EC.3(a) and EC.3(b) illustrate the estimated geographical distribution of the daytime and nighttime populations, respectively.

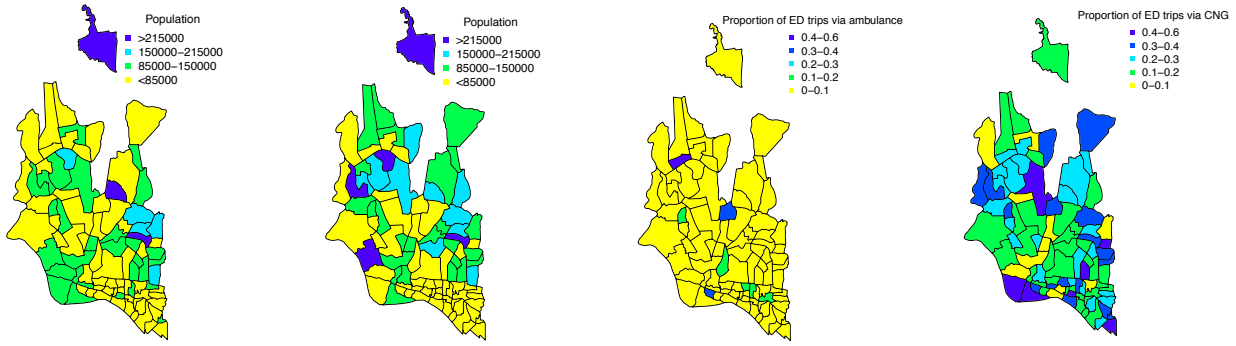
EC.1.2.2. Estimating the average annual number of ED visits per person (ξ). In this section, we leverage published research from South Asian cities to estimate the average annual number of ED visits per person. Given data limitations and the coarseness of previous studies, we cannot generate ward-specific rates and instead settle on a single estimate for the population.

A recent study of a ED arrivals at a "specialty corporate hospital" in Dhaka, found an average of $10,000$ ED visits each year (Karim et al. 2009). This result requires careful interpretation because specialty hospitals can be very expensive and serve only a limited population. To the best of our knowledge, there is no further data on ED visits in Dhaka or other cities in Bangladesh.

To supplement this lack of data, we estimate the number of ED visits using data from other similar South Asian cities. A study of three major government hospitals in Kirachi, Pakistan found an average of $70,000 - 100,000$ annual ED visits per hospital (Raftery 1996). A similar study of two major hospitals in New Dehli, India found that a private hospital with free emergency services received $30,000$ annual ED visits, while a government funded hospital with free services received over $100,000$ annual ED visits (PoSaw et al. 1998).

From these reports, we estimate the number of annual ED visits for a government funded hospital to be between $70,000 - 100,000$ and we estimate the number of annual ED visits for a private hospital to be between $10,000 - 30,000$. Dhaka has 87 hospitals with EDs, of which 19 are government funded. From this information, we estimate the number of annual ED visits to be between $2.08 - 4.15$ million. Given that Dhaka has a population of 8.95 million, the visit rate is between $230 - 460$ per 1000 persons. Therefore, the average number of annual ED visits per capita, ξ, is estimated to be between $0.23 - 0.46$.

It is difficult to put these numbers into context because most LMIC countries do not collect data on annual ED visits. However, data is available for 19 high income OECD countries and the average number of annual ED visits per capita across all countries is 0.31 with a range from 0.07 to 0.70 (Berchet 2015). These results require careful interpretation because they are from high income countries and they combine data from both rural and urban areas, which are known to have significant differences in ED visit rates. For example, in the US, urban areas have an annual rate of 0.32 ED visits per capita as compared to 0.45 ED visits per capita in rural areas. The only reliable data available for large urban areas is from New York City and Shanghai, which are



(a) $n_{w,D}$ (b) $n_{w,N}$ (c) $\delta_{w,V}$ (d) $\delta_{w,S}$

Figure EC.3 The geographical distribution of several estimated components of demand.

similar to Dhaka in terms of population, but not in terms of culture or demographics. The annual per capita ED visit rates in New York City and Shanghai are 0.37 (Goins and Conroy 2015) and 0.33 (Zhang et al. 2014), respectively.

EC.1.2.3. Estimating the proportion of ED visits from ward w arriving via mode m ($\delta_{w,m}$). In this section, we train a regularized (lasso) logistic regression model for predicting the proportion of ED visits from ward w arriving via mode m ($\delta_{w,m}$). In particular, we aim to estimate the proportion of ED visits arriving via ambulance van ($\delta_{w,V}$) and the proportion of ED visits arriving via small ambulance ($\delta_{w,S}$), for each ward in Dhaka, respectively. Although we did not explicitly assume that $\delta_{w,S} + \delta_{w,V} \leq 1$, our results naturally satisfied this inequality.

The nature of our data presents significant challenges for model training. Our survey data indicates that only 74 of Dhaka's 92 wards include at least one surveyed patient and as a result, we cannot directly estimate $\delta_{w,m}$ for each ward. Furthermore, only 32 of 92 wards include at least 20 observations. Given that the overall ambulance usage is 7%, roughly 20 observations are required to ensure sufficient granularity in our estimations. One way to overcome the lack of data is to assume that $\delta_{w,m}$ is uniform across all 92 wards and estimate a single value for each mode, δ_m. Intuitively, assuming that δ_m is uniform across all wards is analogous to using population as a proxy for the annual number of emergency trips. We use this naive approach as a benchmark for our models. A second approach to overcome the lack of data is to group patients according to the ward in which the trip originated and calculate $\delta_{w,m}$ for each ward with 20 or more patients (i.e., only 32 wards). We employ this approach to transform our data and weight each $\delta_{w,m}$ by the number of observations (i.e., patients) from ward w. As a result, our final dataset includes 32 ward observations comprising $1,843$ patients. This approach assumes that $\delta_{w,m}$ is constant across each ward and over time, which is a limitation because δ_m may be different for each individual patient and for different

times of day. However, modeling emergency mode choice decisions at a fine spatiotemporal level requires very granular data, which does not exist in many developing countries.

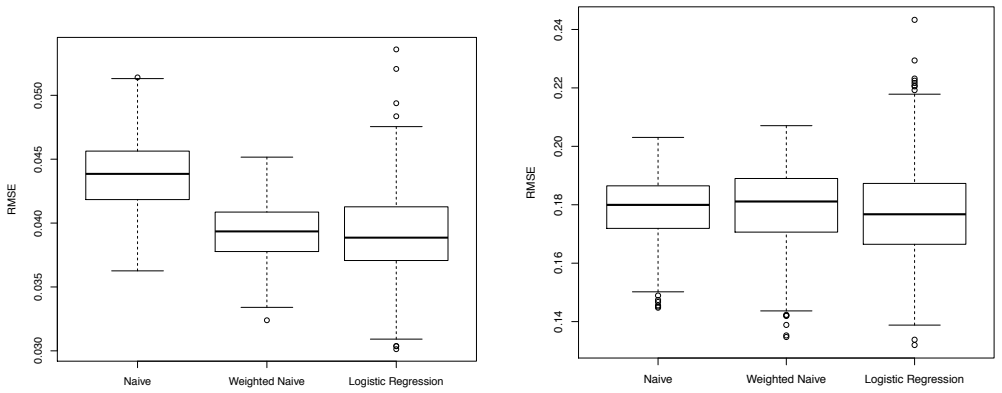
Grouping patients by ward as opposed to a patient level approach that treats each patient as a unique observation is beneficial for our application for three key reasons: 1) we are interested in estimating $\delta_{w,m}$ at the ward level, not the patient level, 2) we want our approach to be generalizable and this framework allows other regions in LMICs with only census data to apply our models, and 3) we require independent variables or features that are available for all 92 wards. The only features available to us for all 92 wards are from the census and we link this data to the survey data using the ward where the trip originated. In contrast, a patient level approach will cause all patients from the same ward to have identical independent variables, regardless of their mode choice.

The set of 27 ward-level demographic features we use in each model was selected from the census data, which contains 104 unique fields. Note that our models do not use individual-level features. To do this, we first remove all highly correlated ($R^2 > 0.85$) variables that appear to represent the same latent feature. In particular, we remove the minimum number of variables required to eliminate all pairwise correlations above 0.85. Next, we combined variables to create new features that have been previously shown to correlate with ambulance demand. For example, the original data contained male population, female population, and total population, which are all highly correlated. We kept total population and created a new variable using the ratio of male to female population. After this procedure, a final set of 27 demographic census features remained.

Our data is well suited for logistic regression because our observations can be viewed as independent Bernoulli trials and modelled using a binomial distribution. Given the large set of features and the likelihood of overfitting, we consider a logistic regression model with L1-regularization (LASSO), where γ denotes the regularization parameter. We optimize over 1000 values of γ between 0.0001 and 0.01. Recall, that the naive prediction approach mentioned above predicts a constant equal to the average δ_m across all wards. We also consider a weighed (by population) naive approach that predicts a constant equal to the weighted average δ_m across all wards.

We train our models using repeated 10-fold cross validation, which partitions the data into ten sets: eight sets of three wards and two sets of four wards. Each set is used exactly once as the testing set, while the remaining 9 are combined and used as the training set. We repeat this process 500 times to reduce the variance in our estimations of model accuracy. We measure prediction accuracy using root mean squared error. Once the value of γ that minimizes RMSE is determined through repeated cross validation, we train a final model using this γ and all available data to estimate $\delta_{w,m}$ for all 92 wards. All models were implemented using `R version 3.3.3.`

Figures EC.4(a) and EC.4(b) display box plots of the RMSE distribution across the 500 repetitions for $\delta_{w,V}$ and $\delta_{w,S}$, respectively. The solid black line indicates the median and the box



(a) $\delta_{w,v}$ (b) $\delta_{w,s}$

Figure EC.4 Comparison of RMSE between the logistic regression model and the naive approaches.

indicates the interquartile range. The whiskers extend to 1.5 times the interquartile range. For $\delta_{w,V}$, the median RMSE was 0.04385 and 0.03935 for the naive and weighted naive, respectively. The logistic regression model (with $\gamma = 0.00183$) performed the best with a median RMSE of 0.03886, corresponding to a 11.4% improvement over the naive approach and a 1.2% improvement over the weighted naive approach. For $\delta_{w,S}$, the median RMSE was 0.180 and 0.181 for the naive and weighted naive, respectively. The logistic regression model (with $\gamma = 0.00183$) preformed the best with a median RMSE of 0.176, corresponding to a 2.2% improvement over the naive approach and a 2.8% improvement over the weighted naive approach. Both improvements were found to be statistically significant at $\alpha = 0.05$ using the Wilcoxon signed-rank test. For both $\delta_{w,V}$ and $\delta_{w,S}$, the RMSE improvements from logistic regression were found to be statistically significant at $\alpha = 0.05$ using the Wilcoxon signed-rank test.

Although the logistic regression model only marginally improves upon the weighted naive approach, the model is able to provide insight into the demographic features that contribute to ambulance usage. Our final features (see Tables EC.4 and EC.5), consistent with prior literature, include measures of population (e.g., average household size), measures of social status (e.g., female marriage rate), and measures of economic status (e.g., access to electricity). Figures EC.3(c) and EC.3(d) display the model-predicted values of ED visits arriving via ambulance van ($\delta_{w,S}$) and small ambulance ($\delta_{w,S}$), respectively. We find that areas of higher socioeconomic status are more likely to use ambulance vans as compared to small ambulances. For example, the wards with the largest values of $\delta_{w,V}$ include areas with a high density of foreigners, government officials, and an area with major government offices, hospitals, and universities. In contrast, small ambulance use is highest in many of the city's outer wards, which include slums.

Table EC.4 Non-zero regression coefficients as determined by LASSO for ambulance vans ($\delta_{w,V}$).

Feature	Coefficient
Intercept	−4.065
Average household size (number of persons)	−0.085
Ratio of male to female population	1.358
Female marriage rate (%)	−0.014
Access to electricity	0.005

Table EC.5 Non-zero regression coefficients as determined by LASSO for small ambulances ($\delta_{w,S}$).

Feature	Coefficient
Intercept	−13.600
Ratio of male to female population	0.062
Male marriage rate (%)	−0.002
Population between 0-19 (%)	0.0077
Population over 60 (%)	−0.149
Disability rate (%)	−0.390
Pukka house (%)	−0.028
Access to a sanitary toilet with seal (%)	0.003
Access to electricity (%)	0.126
Ratio of male to female employment	0.019

Although we focused on predicting the probability that a patient chooses an ambulance van or small ambulance, given that they require transportation to an ED, our approach can can be readily adapted to focus on other modes of transportation, such as private cars, rickshaws, or motorcycles.

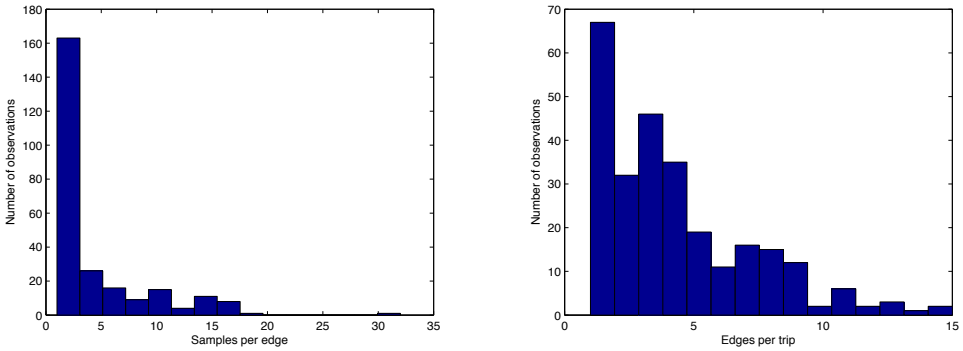
EC.2. Travel time analysis

In this section, we describe the travel time data collection methodology (Section EC.2.1) and develop machine learning models to predict the baseline travel time between any two locations in both road networks (Section EC.2.2).

EC.2.1. Travel time data

We gathered vehicle location data using custom GPS devices and an accompanying Android mobile application developed by our collaborators. These devices were used by five volunteer citizens over 16 days from March 14, 2014 to June 13, 2014 and over 14 days from February 28, 2015 to April 2, 2015. All drivers were instructed to drive normally, using typical routes and speed. A map matching algorithm was developed to map the GPS data to edges on the road network (Ahmed et al. 2015).

We obtained data for 269 unique trips. A trip is defined as a path through the network from some origin node to some destination node. A destination node is defined as either one from which



(a) Samples per edge (b) Edges per trip

Figure EC.5 Histograms for collected GPS data. Only sampled edges are included in these figures.

there is no subsequent GPS activity within 20 minutes on an edge emanating from that node or one with the last recorded GPS activity before the device was turned off by the driver. Trips ranged from 1 to 15 edges, with an average trip length of 4.1 edges. Edges in the network were present in a trip between 0 to 30 times; edges that were present in at least one trip had an average of 3.9 observations for a total of 1,103 edge observations (see Figure EC.EC.5). The median travel time of a trip was $592s$ (min: $10s$, max: $5543s$) while the median travel time on an edge was $105s$ (min: $5s$, max: $5062s$). Figures EC.5(a) and EC.5(b) display the number of data samples per sampled edge and the number of edges per trip in the ambulance road network, respectively.

To predict the travel time between two nodes, one challenge we face is limited data. In particular, if we use trip data to train our models, we are limited to only 269 observations. On the other hand, if we use edge data, which is more plentiful and includes 1,103 observations, then we are unable to capture the delays caused at nodes between edges (i.e., intersections) or the impact of traveling through a ward because most edges lie wholly within one ward. To deal with this trade-off, we develop a modified bootstrapping method that simultaneously solves the limited data issue and the issues with using edge data. This bootstrapping method expands our dataset by partitioning each trip into all continguous *sub-trips*. For example, a trip that begins at node 1, visits nodes 2, then 3, and terminates at node 4 (denoted $1-2-3-4$) would result in six sub-trips: $1-2$, $2-3$, $3-4$, $1-2-3$, $2-3-4$, and $1-2-3-4$. In other words, we include the original trip ($1-2-3-4$), all individual edges ($1-2$, $2-3$, $3-4$) and all sup-trips ($1-2-3$, $2-3-4$). This bootstrapping process results in a total of $4,086$ sub-trips, a 15 times increase in the size of the training set. The new sub-trip data is not a direct replication of trip data because each sub-trip has unique features according to the origin/destination of that sub-trip. Figure EC.6 displays a histogram of speeds for trips, edges, and sub-trips. Note that the sub-trip data includes both trip and edge data. The

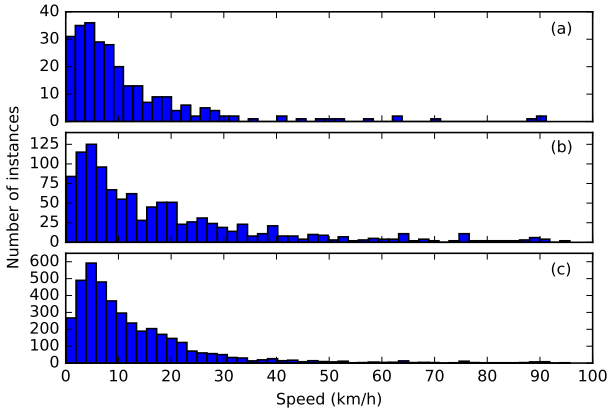
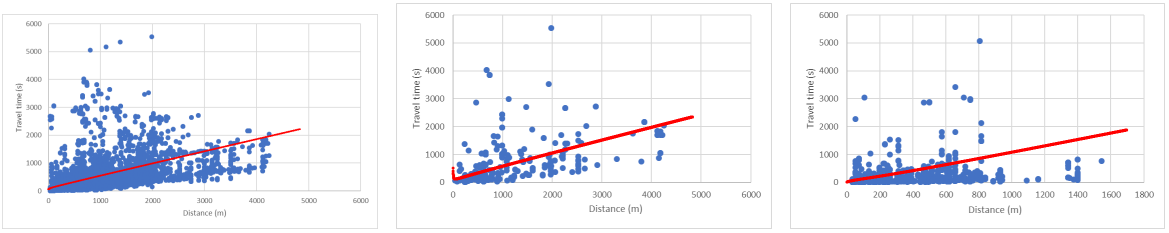


Figure EC.6 Histograms of speed for (a) trip data, (b) edge data, and (c) sub-trip data.



(a) Sub-trip (b) Trip (c) Edge

Figure EC.7 Scatter plots of distance vs. time with the fitted Kolesar model.

average speed (standard deviation) for the trip, edge, and sub-trip data is 2.05 km/h (3.90), 3.30 km/h (5.31), and 2.45 km/h (3.64), respectively.

Figure EC.7 displays a scatter plot of distance vs. travel time for the sub-trip, trip, and edge data. The red curve is the Kolesar et al. (1975) model trained using all available data. Figure EC.8 displays a boxplot of speed for each hour of the day where data was available. We find that speeds are slowest during the evening rush hour (i.e., 6pm and 7pm) and fastest in the mid-afternoon and nighttime/early morning. These results are consistent with our experience in Dhaka and with Google Maps traffic data (note that Google only started providing this service to Dhaka in late 2017/early 2018). McCormack and Coates (2015) also found very similar results using data from the London Ambulance Service and this reference had been added to the paper.

EC.2.2. Travel time prediction

Using the trip, edge, and sub-trip data, we compare four machine learning approaches for predicting the travel time in seconds between any two nodes (not necessarily adjacent) in the network. We use 73 features including the distance on the road network between the given nodes, the day of

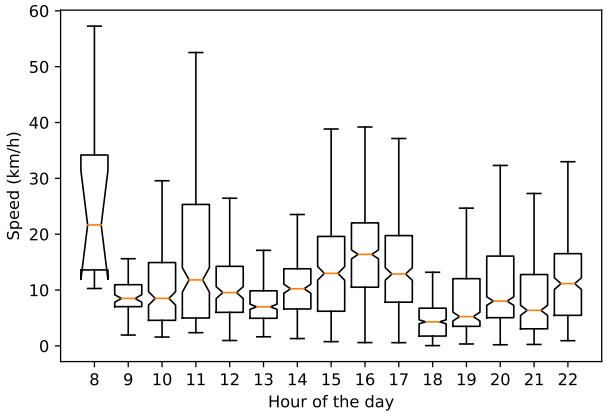


Figure EC.8 Boxplot of speed for each hour of the day. The box represents the interquartile range (i.e., $IQR = Q3 - Q1$), the red line indicates the median, the notches indicate the 95% confidence interval around the median, and the upper (lower) whiskers correspond to $Q3 + 1.5 * IQR$ ($Q1 - 1.5 * IQR$).

week, and time of day. In addition, for both the origin and destination wards, we include building-type information (e.g., the number of commercial or industrial buildings) and the 27 demographic features used to predict emergency demand (see EC.1.2.3). The target is a real number that denotes the travel time in seconds between the two nodes.

We compared the accuracy of four popular machine learning models: AdaBoost, Random Forest, linear regression with L1-regularization (LASSO), and K-nearest neighbors (KNN). For AdaBoost, we optimize the learning rate over $\{0.0001, 0.001, 0.01, 0.1, 1\}$ and number of weak learners over $\{100, 250, 500, 750, 1000\}$; for Random forest, we optimize the number of trees over $\{100, 250, 500, 750, 1000\}$; for LASSO, we optimize the regularization parameter over $\{0.0001, 0.001, 0.01, 0.1, 1\}$; for KNN we optimize the number of neighbors over $\{100, 250, 500, 750, 1000\}$. We train our models using repeated 10-fold cross validation and we repeat the cross validation process 100 times. We used a nested 3-fold cross validation loop on the training set for hyper-parameter tuning. In particular, as part of 10-fold cross validation, we partition the data into two sets: set1 and the test set. Set1 comprises 90% of the data, while the test set comprises 10%. We then split set1 into two sets: a training set and a validation set. We use 3-fold cross validation on set1 to conduct three training-validation instances and we use the average validation set error to choose our hyperparameters. Once the final hyperparameters have been selected, we apply the final model to the test set (that was not used as part of the model selection or fitting process in any way) to estimate generalization. We repeat the entire process 100 times to obtain more robust estimates.

We measure prediction accuracy using root mean squared error (RMSE). Once the final hyperparameters are determined, we train the model using all available data to obtain the final predictions

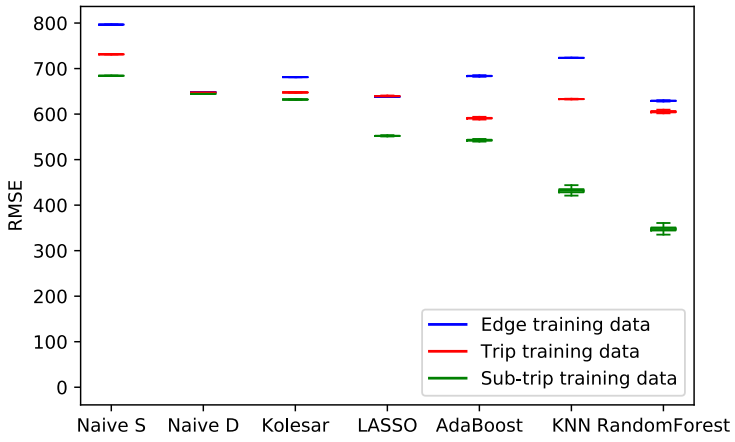


Figure EC.9 Mean-squared error results for the models tested on sub-trip data.

of travel time, which are used in our optimization model. A previously developed model and two naive approaches serve as a baseline. The first naive approach, *Naive S*, predicts a constant equal to the average travel time from the empirical data, and the second, *Naive D*, is a simple linear regression model fit to distance only. We also compared our machine learning approaches to the model developed by Kolesar et al. (1975) that we trained using the maximum likelihood methodology proposed by Budge et al. (2010). All experiments were implemented using `Python 3.5`.

Figure EC.9 displays a boxplot of the root mean squared error (RMSE) distribution across 100 repetitions for each of the prediction models tested on sub-trip data. The median RMSE for the Naive D (Naive S) approach was 648s (797s) when trained on edge data, 647s (731s) when trained on trip data, and 645s (684s) when trained on sub-trip data. The random forest model performed the best with a median RMSE of 629s, 605s, and 348s corresponding to improvements of 3% (21%), 6% (17%), and 46% (49%) over the Naive D (Naive S) approach when trained on edge, trip, and sub-trip data, respectively. All improvements were found to be statistically significant at $\alpha = 0.01$ using the Wilcoxon signed-rank test. Figures EC.10 and EC.11 display boxplots of the RMSE for each of the prediction models tested on edge and trip data, respectively. These results depict a similar finding: a random forest model trained with sub-trip data is the most accurate.

To quantify the impact of time-based and geographical census features on prediction accuracy, we trained our sub-trip models with only distance features, distance and time features, and all features. Figure EC.EC.12 displays a boxplot comparing the RMSE of our models for these experiments. The Lasso, AdaBoost, and RandomForest models improved when both time and geographic features were included, and these improvements were found to be statistically significant using the Wilcoxon signed-rank test. In particular, including time-based features for the random forest model provides a RMSE improvement of 26% (corresponding to a 134.5s reduction in RMSE), over a model with

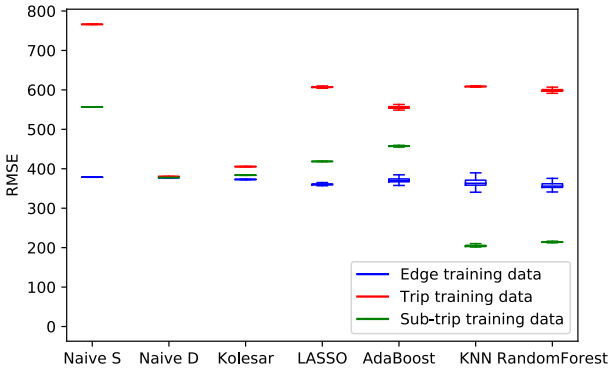


Figure EC.10 Mean-squared error results for the models tested on edge data.

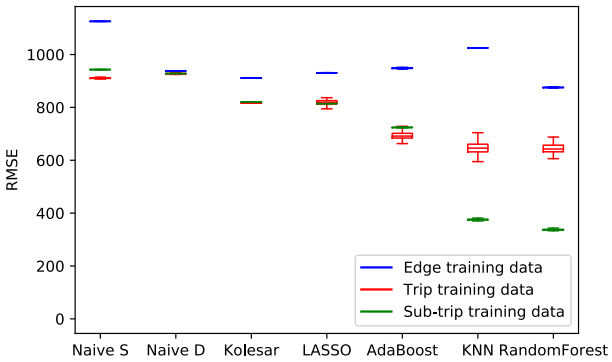


Figure EC.11 Mean-squared error results for the models tested on trip data.

access to only distance. The KNN model improved when adding time features, but it did not improve when geographical features were added because, unlike the other three models, KNN does not have an internal feature weighting process. In other words, the KNN model values all 73 census features equally. By using only the distance and time features, we are implicitly selecting the most important features for the model. These results reinforce the importance of considering time of day and day of the week for travel time estimation in urban areas in LMICs.

A random forest model comprising 1,000 decision trees was selected as the final model and trained using all 4,086 sub-trips. Each feature was available for inclusion to all 1,000 trees and relative feature importance was determined using the number of trees in the forest to which that feature contributes. Table EC.6 lists the features that had a relative importance greater than 0.01. Our results suggest that travel distance, hour of day, and day of week are the three most important features. As expected, travel distance is the most dominant feature with a relative importance of 0.4128. The hour of the day, which can be used as a proxy for peak traffic times, is the only other feature with an importance over 0.1. Our findings are consistent with the results of previous traffic studies, which also found travel distance and the time of day to be the main factors (Zhang and

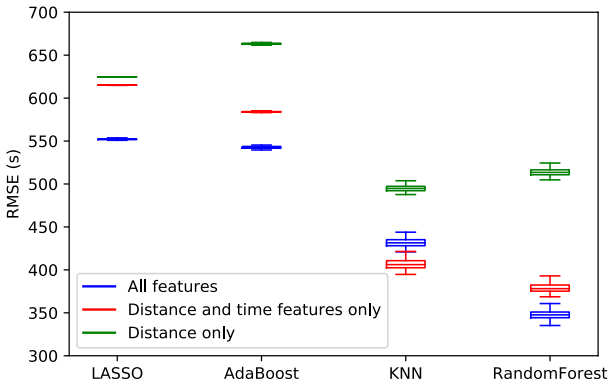


Figure EC.12 Mean-squared error results for sub-trip models with only distance features, distance and time features, and all features.

Table EC.6 Relative feature importance as determined by the random forest model.

Feature	Relative importance
Travel distance (m)	0.413
Hour of day	0.147
Day of week	0.084
Destination node medical facilities (no.)	0.028
Destination node ratio of male to female employment (%)	0.020
Destination node ratio of male to female industrial employment (%)	0.020
Destination node home owners (%)	0.020
Destination node Jupri homes (%)	0.019
Destination node population over 60 (%)	0.018
Origin node literacy rate (%)	0.015
Origin node home owners (%)	0.015
Origin node non-sanitary toilets (%)	0.011

Li 2015, Vlahogianni et al. 2014). As mentioned in Section 2.2, our approach extends previous work by incorporating demographic features for the origin and destination nodes. We found nine geographical census features with a relative importance of at least 0.01. These additional features contribute to an 8% reduction in RMSE relative to a random forest model that only has access to distance and time features.

EC.3. Proof of Theorem 1.

Proof. This proof establishes the Theorem 1 result by construction. Note that without vector notation, we can re-write NFF as:

$$\begin{aligned}
\underset{\mathbf{y},\mathbf{f}}{\text{minimize}} \quad & \sum_{(i,j)\in\mathcal{E}} c_{ij} f_{ij} \\
\text{subject to} \quad & \sum_{i\in N} y_i = P, \\
& \sum_{j\in O(i)} f_{ij} - \sum_{j\in I(i)} f_{ji} \leq \alpha_i y_i - d_i, \forall i \in N, \\
& f_{ij} \geq 0, \forall (i,j) \in \mathcal{E}, \\
& y_i \in \{0,1\}, \forall i \in N,
\end{aligned} \tag{EC.1}$$

where $I(i) = \{j \in N | (j,i) \in \mathcal{E}\}$ and $O(i) = \{j \in N | (i,j) \in \mathcal{E}\}$. Recall the classic p-median formulation. The facility location variable is defined as $x_{ii} = 1$ if a facility is located at node $i \in N$ and the assignment (routing) decision variables is denoted $x_{ij} = 1$ if demand node j has been assigned to facility node i. Using this notation, the p-median problem (ReVelle and Swain 1970) can be formulated as:

$$\begin{aligned}
\underset{\mathbf{x}}{\text{minimize}} \quad & \sum_{i\in N}\sum_{j\in N} d_j t_{ij} x_{ij} \\
\text{subject to} \quad & \sum_{i\in N} x_{ii} = P, \\
& \sum_{i\in N} x_{ij} = 1, \ \forall j \in N, \\
& x_{ii} \geq x_{ij}, \ \forall i,j, i \neq j, \in N, \\
& x_{ij} \geq 0, \forall i,j \in N, i \neq j, \\
& x_{ii} \in \{0,1\}, \forall i \in N,
\end{aligned} \tag{EC.2}$$

where t_{ij} denotes the shortest travel time between nodes $i, j \in N$ (i and j need not be adjacent) and $t_{ii} = 0$, $\forall i \in N$. The optimal solution to (EC.2) is denoted by $\hat{\mathbf{x}}$.

We first show that the p-median is polynomially reducible to NFF. That is, we show that the optimal solution from the p-median can be transformed into the optimal solution of NFF in polynomial time and both solutions have the same optimal cost. First, set $\tilde{y}_i = \hat{x}_{ii}, \forall i \in N$. By definition, the demand weighted shortest path length from $w \in N$ to $r \in N$ is given by $d_r t_{wr}$. To find the path from w to r (i.e., the sequence of nodes $i_w, ..., i_r$) along which flow must be directed,

we solve:

$$\begin{aligned}
\underset{\mathbf{f}}{\text{minimize}} \quad & \sum_{(i,j)\in E} c_{ij} f_{ij}^{rw} \\
\text{subject to} \quad & \sum_{i\in O(r)} f_{ri}^{rw} - \sum_{i\in I(r)} f_{ir}^{rw} = d_w, \\
& \sum_{j\in O(i)} f_{ij}^{rw} - \sum_{j\in I(i)} f_{ji}^{rw} = 0, \forall i \in N\setminus\{r,w\}, \\
& \sum_{i\in O(w)} f_{wi}^{rw} - \sum_{i\in I(w)} f_{iw}^{rw} = -d_w, \\
& f_{ij} \geq 0, \forall (i,j)\in E.
\end{aligned} \tag{EC.3}$$

We denote the optimal solution to (EC.3) as $\hat{\mathbf{f}}^{rw}$. For the special case when $r=w$, Formulation (EC.3) is not well-defined and we assume that the shortest path has length zero (i.e., no flow is produced and $f_{ij}^{ww} = f_{ij}^{rr} = 0, \forall i,j \in N$). Set $\tilde{f}_{ij} = \sum_{r\in N}\sum_{w\in N} \hat{f}_{ij}^{rw}\hat{x}_{rw}$ to obtain a solution $(\tilde{\mathbf{y}}, \tilde{\mathbf{f}})$ to Formulation (EC.1). We now show that the obtained solution $(\tilde{\mathbf{y}}, \tilde{\mathbf{f}})$ is feasible with respect to (EC.1).

For the first constraint, we have:

$$\sum_{i\in N} \tilde{y}_i = \sum_{i\in N} \hat{x}_{ii} = P.$$

For the second constraint, define $J = \{j \in N \,|\, y_j = 0\}$ and $I = \{i \in N \,|\, y_i = 1\}$. Note that $I \cup J = N$. Consider some $k \in J$ (i.e., $y_k = 0$),

$$\begin{aligned}
\sum_{j\in O(k)} \tilde{f}_{kj} - \sum_{j\in I(k)} \tilde{f}_{jk} &= \sum_{j\in O(k)}\sum_{r\in N}\sum_{w\in N} \hat{f}_{kj}^{rw}\hat{x}_{rw} - \sum_{j\in I(k)}\sum_{r\in N}\sum_{w\in N} \hat{f}_{jk}^{rw}\hat{x}_{rw}, \\
&= \sum_{r\in N}\sum_{w\in N} \hat{x}_{rw}\left(\sum_{j\in O(k)} \hat{f}_{kj}^{rw} - \sum_{j\in I(k)} \hat{f}_{jk}^{rw}\right), \\
&= \sum_{r\in N\setminus\{k\}} \hat{x}_{rk}\left(\sum_{j\in O(k)} \hat{f}_{kj}^{rk} - \sum_{j\in I(k)} \hat{f}_{jk}^{rk}\right) + \sum_{w\in N\setminus\{k\}} \hat{x}_{kw}\left(\sum_{j\in O(k)} \hat{f}_{kj}^{kw} - \sum_{j\in I(k)} \hat{f}_{jk}^{kw}\right) \\
&\quad + \sum_{r\in N\setminus\{k\}}\sum_{w\in N\setminus\{k\}} \hat{x}_{rw}\left(\sum_{j\in O(k)} \hat{f}_{kj}^{rw} - \sum_{j\in I(k)} \hat{f}_{jk}^{rw}\right) + \hat{x}_{kk}\left(\sum_{j\in O(k)} \hat{f}_{kj}^{kk} - \sum_{j\in I(k)} \hat{f}_{jk}^{kk}\right), \\
&= \sum_{r\in N\setminus\{k\}} \hat{x}_{rk}(-d_k) + \sum_{w\in N\setminus\{k\}} \hat{x}_{kw}(d_w) + \sum_{r\in N\setminus\{k\}}\sum_{w\in N\setminus\{k\}} \hat{x}_{rw}(0) + \hat{x}_{kk}(0), \\
&= -d_k.
\end{aligned}$$

Consider some $k \in I$ (i.e., $y_k = 1$),

$$\begin{aligned}
\sum_{j\in O(k)} \tilde{f}_{kj} - \sum_{j\in I(k)} \tilde{f}_{jk} &= \sum_{j\in O(k)} \sum_{r\in N} \sum_{w\in N} \hat{f}^{rw}_{kj} \hat{x}_{rw} - \sum_{j\in I(k)} \sum_{r\in N} \sum_{w\in N} \hat{f}^{rw}_{jk} \hat{x}_{rw}, \\
&= \sum_{r\in N} \sum_{w\in N} \hat{x}_{rw} \left(\sum_{j\in O(k)} \hat{f}^{rw}_{kj} - \sum_{j\in I(k)} \hat{f}^{rw}_{jk} \right), \\
&= \sum_{r\in N\setminus\{k\}} \hat{x}_{rk} \left(\sum_{j\in O(k)} \hat{f}^{rk}_{kj} - \sum_{j\in I(k)} \hat{f}^{rk}_{jk} \right) + \sum_{w\in N\setminus\{k\}} \hat{x}_{kw} \left(\sum_{j\in O(k)} \hat{f}^{kw}_{kj} - \sum_{j\in I(k)} \hat{f}^{kw}_{jk} \right) \\
&\quad + \sum_{r\in N\setminus\{k\}} \sum_{w\in N\setminus\{k\}} \hat{x}_{rw} \left(\sum_{j\in O(k)} \hat{f}^{rw}_{kj} - \sum_{j\in I(k)} \hat{f}^{rw}_{jk} \right) + \hat{x}_{kk} \left(\sum_{j\in O(k)} \hat{f}^{kk}_{kj} - \sum_{j\in I(k)} \hat{f}^{kk}_{jk} \right), \\
&= \sum_{r\in N\setminus\{k\}} \hat{x}_{rk}(-d_k) + \sum_{w\in N\setminus\{k\}} \hat{x}_{kw}(d_w) + \sum_{r\in N\setminus\{k\}} \sum_{w\in N\setminus\{k\}} \hat{x}_{rw}(0) + \hat{x}_{kk}(0), \\
&\leq \sum_{w\in N\setminus\{k\}} d_w = \sum_{w\in N} d_w - d_k = \alpha - d_k.
\end{aligned}$$

Lastly, we show that the objective function values of both solutions are equal,

$$\begin{aligned}
\sum_{(i,j)\in E} c_{ij} \tilde{f}_{ij} &= \sum_{(i,j)\in E} c_{ij} \sum_{r\in N} \sum_{w\in N} \hat{f}^{rw}_{ij} \hat{x}_{rw}, \\
&= \sum_{r\in N} \sum_{w\in N} \hat{x}_{rw} \left(\sum_{(i,j)\in E} c_{ij} \hat{f}^{rw}_{ij} \right), \\
&= \sum_{r\in N} \sum_{w\in N} \hat{x}_{rw} d_w t_{rw}.
\end{aligned}$$

We now prove the reverse direction. That is, a solution from NFF can be transformed into a solution for the p-median with the same optimal cost. We denote the optimal solution to NFF as $(\hat{\mathbf{y}}, \hat{\mathbf{f}})$.

First, we set $\tilde{x}_{kk} = \hat{y}_k, \forall k \in N$. Define $J = \{r \in N \mid \hat{y}_r = 0\}$ and $I = \{w \in N \mid \hat{y}_w = 1\}$. Note that $I \cup J = N$. Compute t_{rw}, $\forall r \in I$ and $\forall w \in J$ (i.e., the shortest path between nodes r and w). This can be done by using Dijkstra's algorithm or by extracting the path lengths directly from the given optimal solution to NFF. Both methods are polynomial time.

Now, consider some $k \in J$, and solve $\text{argmin}_{\text{i}\in\text{I}}\, t_{ik}$. Denote the optimal index as i^k and the optimal value as $t_{i^k k}$. Set $\tilde{x}_{i^k k} = 1$, $\tilde{x}_{kj} = 0, \forall j \in N$, and $\tilde{x}_{ik} = 0, \forall i \in N \setminus \{i^k\}$. Consider some $k \in I$, which implies that $\tilde{x}_{kk} = 1$. Set $\tilde{x}_{ik} = 0, \forall i \in N \setminus \{k\}$ to obtain the solution, $\tilde{\mathbf{x}}$. We now show that the obtained solution $\tilde{\mathbf{x}}$ is feasible for the p-median.

For the first constraint, we have:

$$\sum_{i\in N} \tilde{x}_{ii} = \sum_{i\in N} \hat{y}_i = P.$$

For the second constraint, consider $r \in J$. By our construction, $\tilde{x}_{i^r r} = 1$ and $\tilde{x}_{ir} = 0, \forall i \in N \setminus \{i^r\}$. Therefore, $\sum_{i \in N} \tilde{x}_{ir} = \tilde{x}_{i^r r} + \sum_{i \in N \setminus \{i^r\}} \tilde{x}_{ir} = 1$. Consider, $w \in I$. By our construction, $\tilde{x}_{ww} = 1$ and $\tilde{x}_{iw} = 0, \forall i \in N \setminus \{w\}$. Therefore, $\sum_{i \in N} \tilde{x}_{iw} = \tilde{x}_{ww} + \sum_{i \in N \setminus \{w\}} \tilde{x}_{iw} = 1$. Combining these implies $\sum_{i \in N} \tilde{x}_{ij} = 1, \forall j \in N$.

For the third constraint, consider $r \in J$. By our construction $\tilde{x}_{rr} = 0$ and $\tilde{x}_{rk} = 0, \forall k \in N \setminus \{r\}$. Therefore, $\tilde{x}_{rr} \geq \tilde{x}_{rk}, \forall r \in J, k \in N \setminus \{r\}$. Consider $w \in I$. By our construction, $\tilde{x}_{ww} = 1$ and $\tilde{\mathbf{x}} \in \{0,1\}$ (i.e., $\tilde{\mathbf{x}} \leq 1$), therefore we have $\tilde{x}_{ww} \geq \tilde{x}_{wk}, \forall w \in I, k \in N \setminus \{w\}$. Combining these implies $\tilde{x}_{ii} \geq \tilde{x}_{ij}, \forall i \in N, j \in N \setminus \{i\}$.

Lastly, we show that the objective function values are equal. First we must derive some intermediate information. Consider the following optimization problem with $\hat{\mathbf{y}}$ fixed,

$$\begin{aligned} \underset{\mathbf{f}}{\text{minimize}} \quad & \sum_{(i,j) \in E} f_{ij} c_{ij} \\ \text{subject to} \quad & \sum_{j \in O(i)} f_{ij} - \sum_{j \in I(i)} f_{ji} \leq \alpha \hat{y}_i - d_i, \forall i \in N, \\ & f_{ij} \geq 0, \forall (i,j) \in E. \end{aligned} \tag{EC.4}$$

Denote the optimal solution of (EC.4) by $\hat{\mathbf{f}}$. The dual of (EC.4) is given by,

$$\begin{aligned} \underset{\mathbf{p}}{\text{maximize}} \quad & \sum_{i \in N} p_i (\alpha \hat{y}_i - d_i) \\ \text{subject to} \quad & p_i - p_j \leq c_{ij}, \forall (i,j) \in E, \\ & p_i \leq 0, \forall i \in N. \end{aligned} \tag{EC.5}$$

Denote the optimal solution to (EC.5) by $\hat{\mathbf{p}}$. The dual variable $\hat{p}_k$ represents the change in optimal cost due to increasing d_k by one unit. If we increase d_k by one unit, the optimal solution will increase by the length of the shortest path from $i^k \in I$ to k. Therefore, at optimality, the value of -p_k (because p_k is negative in EC.5) is equal to the length of the shortest path from $i^k \in I$ to $k \in J$. Mathematically, $-p_k = t_{i^k,k}$. Note that this implies that $p_k = 0, \forall k \in I$ because the shortest path from a facility to itself, has length zero.

Now we show that optimal costs are equal:

$$\begin{aligned}
\sum_{w\in N}\sum_{r\in N}\tilde{x}_{rw}d_w t_{rw} &= \sum_{r\in J}\sum_{w\in N}\tilde{x}_{rw}d_w t_{rw} + \sum_{r\in I}\sum_{w\in J}\tilde{x}_{rw}d_w t_{rw} + \sum_{r\in I}\sum_{w\in I\setminus\{r\}}\tilde{x}_{rw}d_w t_{rw} + \tilde{x}_{ww}d_w t_{ww} \\
&= \sum_{w\in J}\sum_{r\in I}\tilde{x}_{rw}d_w t_{rw} && (t_{ww}=0,\ \tilde{x}_{rw}=0, \forall r\in J, w\in N, \\
& && \text{and } x_{wr}=0, \forall w\in I, r\in I\setminus\{w\}) \\
&= \sum_{w\in J} d_w \sum_{r\in I}\tilde{x}_{rw}t_{rw} \\
&= \sum_{w\in J} d_w t_{r^w w} && \text{(By our construction)} \\
&= -\sum_{w\in J} d_w \hat{p}_w && \text{(From duality)} \\
&= -\sum_{w\in N} d_w \hat{p}_w && (\hat{p}_w=0, \forall w\in I) \\
&= \alpha\sum_{w\in N} y_w\hat{p}_w - \sum_{w\in N} d_w\hat{p}_w && (\hat{p}_w=0, \forall w\in I \text{ and } y_w=0, \forall w\in J) \\
&= \sum_{w\in N}\hat{p}_w(\alpha y_w - d_w) \\
&= \sum_{(r,w)\in E}\hat{f}_{rw}c_{rw} && \text{(By strong duality)}
\end{aligned}$$

Combining both directions, we have that

$$\sum_{(i,j)\in E}\bar{f}_{ij}c_{ij} = \sum_{i\in N}\sum_{j\in N}\hat{x}_{ij}d_j t_{ij} \le \sum_{i\in N}\sum_{j\in N}\tilde{x}_{ij}d_j t_{ij}, \forall\tilde{\mathbf{x}},$$

and that

$$\sum_{i\in N}\sum_{j\in N}\tilde{x}_{ij}d_j t_{ij} = \sum_{(i,j)\in E}\hat{f}_{ij}c_{ij} \le \sum_{(i,j)\in E}\bar{f}_{ij}c_{ij}, \forall\bar{\mathbf{f}}.$$

Therefore,

$$\sum_{i\in N}\sum_{j\in N}\hat{x}_{ij}d_j t_{ij} = \sum_{(i,j)\in E}\hat{f}_{ij}c_{ij}. \quad \square$$

EC.4. Proof of Theorem 2.

Proof. Let $\mathbb{Y} = \{\mathbf{y} \mid \mathbf{e}'\mathbf{y} = P, \mathbf{y}\ge\mathbf{0}\}$ and $\mathbb{F}(\mathbf{y},\mathbf{d}) = \{\mathbf{f} \mid \mathbf{A}\mathbf{f} \le \boldsymbol{\alpha}\mathbf{I}\mathbf{y} - \mathbf{d}, \mathbf{f}\ge\mathbf{0}\}$. Then, **R-NFF** can be written as

$$\min_{\mathbf{y}\in\mathbb{Y}}\ \max_{\mathbf{c}\in\mathcal{C},\mathbf{d}\in\mathcal{D}}\ \min_{\mathbf{f}\in\mathbb{F}(\mathbf{y},\mathbf{d})}\ \mathbf{c}'\mathbf{f},$$

or in epigraph form

$$\begin{aligned}
\underset{\mathbf{y}\in\mathbb{Y},\mathrm{t}}{\text{minimize}} \quad & t \\
\text{subject to} \quad & t \ge \max_{\mathbf{c}\in\mathcal{C},\mathbf{d}\in\mathcal{D}}\ \min_{\mathbf{f}\in\mathbb{F}(\mathbf{y},\mathbf{d})}\ \mathbf{c}'\mathbf{f}.
\end{aligned}$$

Enumerating the elements of $\mathcal{D}$, the model becomes

$$\begin{aligned} \underset{\mathbf{y}\in\mathbb{Y},\mathrm{t}}{\text{minmize}} \quad & t \\ \text{subject to} \quad & t \geq \max_{\mathbf{c}\in\mathcal{C}} \min_{\mathbf{f}\in\mathbb{F}(\mathbf{y},\mathbf{d}^k)} \mathbf{c}'\mathbf{f}, \quad k=1,...,N. \end{aligned}$$

Since $\mathcal{C}$ and $\mathbb{F}(\mathbf{y},\mathbf{d}^k)$ are disjoint, we can swap the min and max using the min-max theorem (Neumann 1928):

$$\begin{aligned} \underset{\mathbf{y}\in\mathbb{Y},\mathrm{t}}{\text{minimize}} \quad & t \\ \text{subject to} \quad & t \geq \min_{\mathbf{f}\in\mathbb{F}(\mathbf{y},\mathbf{d}^k)} \max_{\mathbf{c}\in\mathcal{C}} \mathbf{c}'\mathbf{f}, \quad k=1,...,N. \end{aligned}$$

We then replace $\mathbf{f}$ by $\mathbf{f}^k$ for each scenario k, which yields:

$$\begin{aligned} \underset{\mathbf{y}\in\mathbb{Y},\mathrm{t}}{\text{minimize}} \quad & t \\ \text{subject to} \quad & t \geq \min_{\mathbf{f}^k\in\mathbb{F}(\mathbf{y},\mathbf{d}^k)} \max_{\mathbf{c}\in\mathcal{C}} \mathbf{c}'\mathbf{f}^k, \quad k=1,...,N. \end{aligned}$$

We can now move $\mathbf{f}^k$ to the outer minimization problem:

$$\begin{aligned} \underset{\mathbf{y}\in\mathbb{Y},\mathrm{t},\mathbf{f}^\mathrm{k}}{\text{minimize}} \quad & t \\ \text{subject to} \quad & t \geq \underset{\mathbf{c}\in\mathcal{C}}{\text{maximize}} \ \mathbf{c}'\mathbf{f}^k, \quad k=1,...,N, \\ & \mathbf{f}^k \in \mathbb{F}(\mathbf{y},\mathbf{d}^k), \quad k=1,...,N. \end{aligned}$$

Finally, for each k, we take the dual of the inner maximization problem to obtain the required result. □

EC.5. Comparison of solution approaches

In this section, we present results from a set of computational experiments that compare the effectiveness of our exact and heuristic scenario generation algorithms. To do so, we use smaller randomly-generated problem instances that can be solved to optimality.

EC.5.1. Experimental setup.

We use three random network instances to conduct our experiments. The first network has 30 nodes and 90 edges, the second has 50 nodes and 150 edges, and the third network has 75 nodes and 226 edges. For each graph, we vary the number of scenarios ($|\mathcal{S}|$) in $\mathcal{D}$, the interdiction budget (B) in $\mathcal{C}$, and the number of vehicle outposts (P). Specifically, we consider: $|\mathcal{S}| \in \{1, 10, 100, 1000, 10000\}$, $P \in \{1, 2, 5, 10, 25\}$, and $B \in \{0, 10, 50, 100, 250, 500, 1000\}$. Hence, we solve 175 problem instances for each random network, for a total of 525 problem instances. We solve each instance using 1) a commercial solver (Gurobi), 2) our exact scenario generation algorithm (SGen), and 3) our heuristic scenario generation algorithm (HSGen) with 10 random starts and 10 interchanges. We chose this number of random starts and interchanges after testing our heuristic with different values (see

Figure EC.13). We set a maximum time limit of 36,000 seconds for each instance. All experiments were programmed using `MATLAB2016a` and run on a desktop computer with an Intel Core i7-4790K 4.0 GHz processor and 32 GB of RAM.

To estimate realistic edge-lengths for these instances, we randomly sample from the edge data distribution introduced in Figure EC.6. To generate node-weights and demand scenarios, we used a modified version of the methodology outlined in Section 3.2.1. We estimate the population and ξ as in Section 3.2.1. Without underlying ward features, our logistic regression model is not applicable so we use the naive approach from EC.1.2.3 instead.

EC.5.2. Scenario generation algorithm performance.

The scenario generation algorithm was able to solve all 525 problem instances to optimality, while Gurobi struggled with larger instances. Table EC.7 compares solution times as a function of uncertainty set size. Gurobi was not able to solve any of the instances that had 100 or more demand scenarios, except the one with no travel time uncertainty. Table EC.8 compares the solution times for instances that vary in the size of the underlying network and the number of outposts located. The scenario generation approach enjoys the largest speed up for intermediate values of P.

EC.5.3. Heuristic algorithm performance.

Table EC.7 also compares the optimal cost and solution time between SGen and HSGen as a function of the number of scenarios and the interdiction budget. The objective function value is displayed as mean response time, in seconds. To determine this value, we divide the actual objective function value by the total number of trips. The performance of the heuristic algorithm remains relatively stable as the number of scenarios increases with solutions times that are an order or magnitude less than SGen. Table EC.8 compares the optimal cost and solution time between SGen and HSGen as a function of the size of the graph and the number of outposts, while holding both the interdiction budget and the number of scenarios constant at 100. Across all instances, the heuristic algorithm was able to obtain the optimal solution when the number of outposts was small. The performance also remains relatively stable as the size of the graph grows. However, the performance degrades as P grows. This degradation in performance is balanced by up to an order of magnitude speed-up in certain cases. While HSGen does not close the optimality gap as the size of the problem increases, for the large-scale, real-world instances of the robust problems that we solve in Section 5, it is the only method capable of generating a solution in a reasonable time limit.

Figure EC.13 displays the performance of HSGen as a function of the number of random starts and random interchanges for different numbers of scenarios and different numbers of outposts. Figure 13(a) shows that HSGen improves significantly from one to ten random starts, but does

Table EC.7 Comparison of objective function values and solution times between Gurobi, SGen, and HSGen as a function of the number of scenarios and uncertainty budget. The number of vehicles and the size of the graph are held constant at 5 and, 75 nodes and 226 edges, respectively.

		Objective function value			Solution time		
$\|S\|$	Budget	SGen	HSGen	Optimality gap (%)	Gurobi (s)	SGen (s)	HSGen (s)
1	1	133.4	142.8	6.6	3.4	3.4	3.7
	10	134.8	141.0	4.4	4.3	4.3	3.6
	100	145.9	171.7	15.0	5.6	5.7	3.7
	1000	205.1	214.7	4.4	10.7	11.1	4.3
10	1	135.1	156.7	13.7	179.5	37.3	9.4
	10	136.6	147.8	7.5	192.5	31.2	9.2
	100	148.4	169.0	12.2	625.3	109.7	10.7
	1000	208.3	209.8	0.7	4581.7	72.0	12.1
100	1	143.2	164.9	13.1	26857.0	137.9	19.1
	10	144.5	163.1	11.4	-	170.1	24.1
	100	153.7	167.1	8.0	-	197.5	16.9
	1000	218.4	227.2	3.9	-	2974.8	24.4
1000	1	140.1	153.8	8.9	-	585.6	84.2
	10	141.4	161.8	12.6	-	945.3	63.0
	100	152.5	158.4	3.8	-	1160.0	93.7
	1000	212.1	219.0	3.1	-	3721.8	34.8
10000	1	142.5	162.4	12.2	-	3856.3	581.2
	10	143.8	161.6	11.0	-	3264.3	462.7
	100	154.7	164.0	5.7	-	9306.5	581.5
	1000	217.8	227.0	4.0	-	5475.0	350.7

not appear to improve much beyond ten. By contrast, Figure 13(b) displays a small improvement from one to ten random starts and only marginal improvements thereafter. Thus, we use ten random starts to conduct our real experiments on the Dhaka road network. Figures 13(c) and 13(d) show that there does not appear to be a correlation between increasing the number of random interchanges and the overall solution quality. However, in all cases, there is a small improvement from one to 10 random interchanges. Thus, we use we use ten random interchanges to conduct our real experiments on the Dhaka road network.

EC.6. Tactical Simulation Model

Algorithm 1 displays high-level pseudo-code for our simulation framework. Note that t_C denotes the current time, t_W denotes the waiting time, t_D denotes the drive time to the emergency location, and t_S denotes the scene time sampled from an exponential distribution with mean of 15 min. The DISPATCH function uses a greedy dispatching policy that assigns the closest ambulance and we determine the closest ambulance by solving the robust shortest path problem with $B = 1000$ for each available ambulance. After reaching the scene and picking up the patient, ambulances

Table EC.8 Comparison of objective function values and solution times between SGen and HSGen as a function of graph size and P. The number of scenarios and the interdiction budget are held constant at 100 and 100, respectively.

			Objective function value			Solution time		
Nodes	Edges	P	SGen	HSGen	Optimality gap (%)	Gurobi (s)	SGen (s)	HSGen (s)
30	90	1	213.9	213.9	0	87.8	12.7	2.0
		2	155.8	155.8	0	82.2	25.5	9.2
		5	93.0	102.8	9.6	473.9	25.1	7.7
		10	46.1	60.3	23.5	210.4	94.8	42.9
		25	7.3	11.3	35.4	10.6	142.1	40.1
50	150	1	287.0	287.0	0	525.1	38.5	9.6
		2	191.0	191.0	0	1213.5	25.9	5.9
		5	109.9	130.0	15.5	3534.0	89.6	25.4
		10	62.8	88.7	29.2	3410.6	100.0	86.7
		25	17.2	31.8	45.7	211.6	197.4	99.4
75	226	1	279.5	279.5	0	2671.8	31.7	6.9
		2	232.2	232.2	0	-	88.8	13.1
		5	153.7	167.1	8.0	-	197.5	16.9
		10	99.4	137.3	27.6	-	400.9	32.0
		25	40.4	65.9	38.7	2591.1	176.9	113.9

are routed to the closest hospital determined by solving the robust shortest path problem with $B = 1000$ and t_H denotes the drive time to the hospital. The ROUTEHOME function determines the time until an ambulance has returned to its home base location (t_B) by solving the robust shortest path problem with $B = 100$.

EC.7. Dhaka Policy Experiments

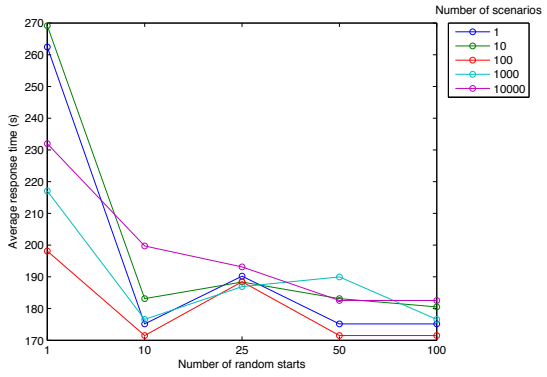
Figure EC.14 depicts the locations of all 67 current outposts.

EC.7.1. What is the impact of the number of ambulances per outpost?

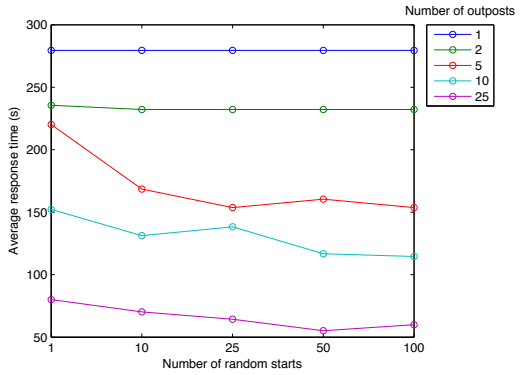
Figure EC.15 displays the two major components of response time (drive time and waiting time) as a function of the number of ambulances per outpost for 20 outpost locations and the current baseline scenario (67 hospital-based outposts). The current baseline scenario includes a total of 269 ambulances spread across 67 outposts, while the 20 outpost solution with nine ambulances per outpost includes only 180 total ambulances. In other words, similar response time performance can be achieved with far fewer resources, if the resources are utilized more effectively. We find that diminishing returns are reached with seven ambulances per outpost (a total of 140) and we use seven per outpost for our policy experiments. Note that three or fewer ambulances per outpost result in a system with waiting times over 24 hours.

Algorithm 1 Tactical ambulance simulator

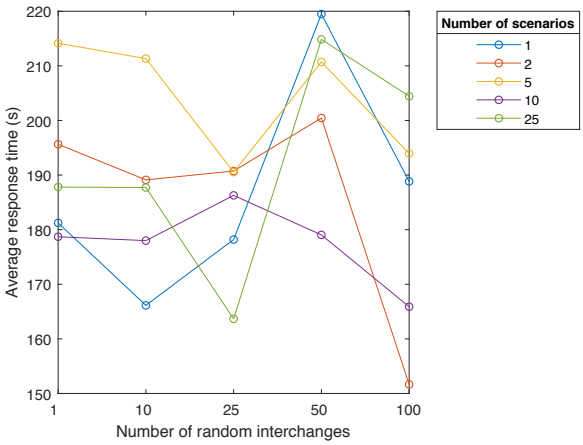
function SIMULATE(C,R,Y)
  $C \leftarrow$ Load the simulated call times and locations
  $R \leftarrow$ Load the road network with hospital locations and travel times for each day/hour
  $Y \leftarrow$ Load the optimized outpost locations and the number of ambulances per outpost
  $Q \leftarrow \emptyset$ ▷ initialize empty call queue
  $\mathcal{E} \leftarrow C$ ▷ initialize event queue with calls
  $A \leftarrow Y$ ▷ initialize ambulance availability list
  while $|\mathcal{E}| > 0$ **do**
    Remove next event e from $\mathcal{E}$
    Update current time t_C
    if e is a new call **then**
      if $|A| = 0$ **then** ▷ No available ambulances
        $Q \leftarrow Q + e$ ▷ Queue the call
      else
        $V(c) \leftarrow$ DISPATCH(c,A,R) ▷ Dispatch closest ambulance
        $A = A - V(c)$ ▷ Remove dispatched vehicle from available list
        e_{new} ▷ Create new event for when ambulance is free at hospital
        $\mathcal{E} \leftarrow e_{new}$ ▷ Insert new event at time $t = t_C + t_W + t_D + t_S + t_H$
    else if e is ambulance becomes available **then**
      if $|Q| > 0$ **then**
        $V(c) \leftarrow$ DISPATCH(c,A,R) ▷ Dispatch newly available ambulance
        e_{new} ▷ Create new event for when ambulance is free at hospital
        $\mathcal{E} \leftarrow e_{new}$ ▷ Insert new event at time $t = t_C + t_W + t_D + t_S + t_H$
      else
        ROUTEHOME(c,Y,R) ▷ Route ambulance home
        e_{new} ▷ Create new event for when ambulance is free at its base
        $\mathcal{E} \leftarrow e_{new}$ ▷ Insert new event at time $t = t_C + t_B$
    else if e is ambulance returned to base **then**
      if $|Q| > 0$ **then**
        $V(c) \leftarrow$ DISPATCH(c,A,R) ▷ Dispatch newly available ambulance
        e_{new} ▷ Create new event for when ambulance is free at hospital
        $\mathcal{E} \leftarrow e_{new}$ ▷ Insert new event at time $t = t_C + t_W + t_D + t_S + t_H$
      else
        $A = A + V(c)$ ▷ add ambulance to available list



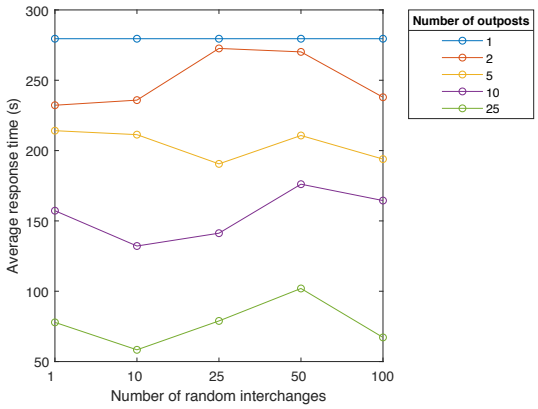
(a) Number of scenarios



(b) Number of outposts



(c) Number of scenarios



(d) Number of outposts

Figure EC.13 Objective function value as a function of the number of random starts and interchanges used for HSGen.

EC.7.2. What is the impact of the travel time budget?

In this section, we conduct a sensitivity analysis to determine the impact of the travel time budget. We use the HSGen algorithm with 10 random starts and 10 random interchanges to solve (3) with $P = 20$ and $B = \{0, 100, 1000, 2500, 5000, 7500, 10000\}$. We then apply the outpost locations resulting from a budget of 1000 seconds to all seven budget instances. We compare these results with the response time of outpost locations optimized and evaluated on the same budget. We conduct separate experiments for each of the three temporal combinations.

Figure EC.16 compares the response time performance between a fixed travel time budget and a problem-specific travel time budget for each of the three temporal combinations. For rush hour, the fixed budget outpost locations perform better when evaluated on networks with budgets of 2500 and 10000 with an average improvement of 1.45 minutes (4.5%). For all other instances, the

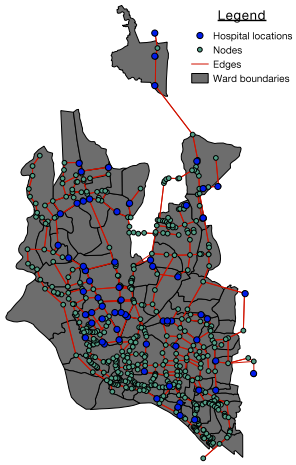


Figure EC.14 Map of all 67 current outpost locations.

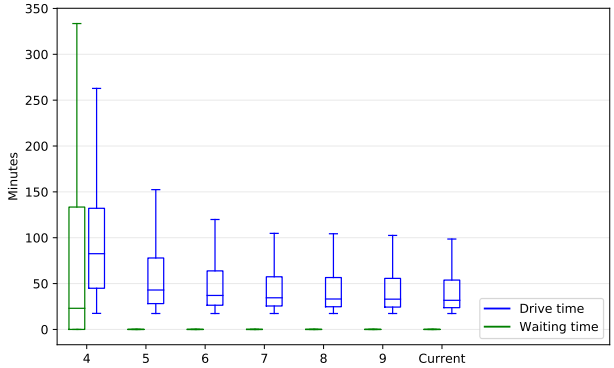
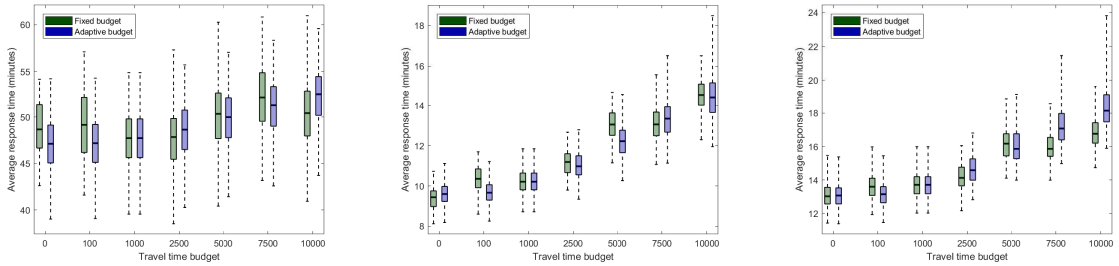


Figure EC.15 Drive time and waiting time as a function of the number of ambulances per outpost for 20 outpost locations. "Current" is the baseline scenario with 67 hospital-based outposts.



(a) Rush hour (b) Overnight (c) Weekend

Figure EC.16 Response time performance of outpost locations determined using a fixed budget of 1000 seconds versus outpost locations determined using an adaptive travel time budget.

fixed budget performed worse with an average degradation of 1.29 minutes (4%). Similar results are observed for overnight and weekend baseline traffic scenarios.

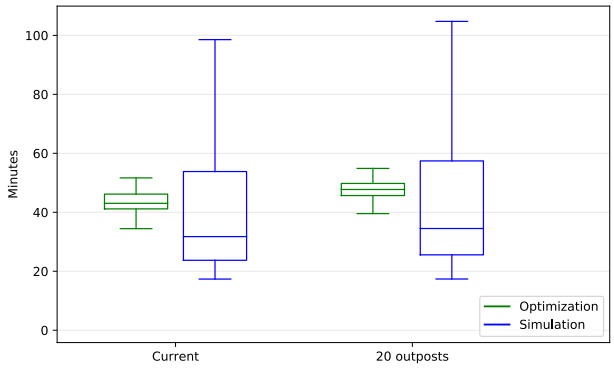


Figure EC.17 A comparison of optimization and simulation estimated response times for the current outpost locations and a network with 20 outpost locations.

EC.7.2.1. Discussion and policy implications. Our results suggest that the outpost locations determined using a travel time budget of 1000 seconds are relatively insensitive to changes the travel time budget. This is an important result that implies that ambulance providers in Dhaka can use the optimal outpost locations from a budget of 1000 seconds without concern that these locations will perform significantly worse for other travel time budgets.

EC.7.3. How do the optimization-estimated response times compare to the simulation-estimated response times?

Figure EC.17 compares the response times estimated by the optimization and simulation models for the current and 20 outpost solutions. The median response time estimated via optimization overestimate the median response time estimated via simulation by 11.3 min (26.2%) and 13.3 min (27.7%) for the current and 20 outpost solutions, respectively. Although the optimization results fall within the interquartile range of the simulation results, the optimization model underestimates the total range of response times by 208.5 min and 311.6 min for the current and 20 outpost solutions, respectively. In summary, we find that the response times estimated by the optimization model provide a conservative estimate on the median response time, but significantly underestimate the total range of response times.