EMS location-allocation problem under uncertainties

Abstract

Emergencies, especially those considered routine (i.e., occurring on a daily basis), pose great threats to health, life, and property. Immediate response and treatment can greatly mitigate these threats. This research is conducted to optimize the locations of ambulance stations, deployment of ambulances, and dispatch of vehicles under demand and traffic uncertainty, which are the main factors that influence emergency response time. The research problem is formulated as a dynamic scenario-based two-stage stochastic programming model, aiming to minimize the total cost while responding to as much demand as possible. The Sample Average Approximation is proposed to approximate the original problem using a limited number of scenarios, and a two-phase Benders Decomposition solution scheme is proposed to accelerate computation, especially when solving a large-sized problem. Numerical experiments using real-world emergency data are conducted to validate the performance of the solution method. The results demonstrate the effectiveness and efficiency of the proposed algorithm. We additionally conduct a sensitivity analysis to evaluate the influences of crucial parameters, including the response time standard, facility capacity, service capacity, and facility heterogeneity. The managerial insights derived from sensitivity analysis will provide valuable guidance for the design of an emergency response system in practice.

Keywords: Emergency medical services, Location-allocation problem, Stochastic program, Sample average approximation, Benders Decomposition

1. Introduction

An emergency situation can happen anywhere and anytime, and such situations pose great risks to people's health, lives, and property. A routine emergency, such as a heart attack, road accident, or residential fire, is a type of emergency that can happen on a daily basis and has a small scope of influence. According to research by Vos et al. (2020), this type of emergency is one of the leading causes of death worldwide. Therefore, the immediate response and treatment to a routine emergency are of great importance. In this context, treatment can be categorized as in-hospital treatment or pre-hospital treatment. In-hospital treatment depends heavily on the process design and operations

management of patient flows, which is not considered in this research. Readers who are interested in efficiency and effectiveness in the context of in-hospital treatment are referred to Kuo (2014) and Kuo et al. (2016). Pre-hospital treatment usually involves an emergency medical services (EMS) department. When an emergency call is received, the call handler determines the severity of the situation and then dispatches response vehicles accordingly. As the patient survival rate is highly dependent on the response time, defined as the time interval between the reception of a call and the arrival of the response vehicle at the emergency site, the dispatched vehicles must arrive at the emergency site within certain time limits (Bürger et al., 2018; Erkut, Ingolfsson, & Erdoğan, 2008; Knight, Harper, & Smith, 2012). The response time is directly affected by the locations of ambulance stations, number of available vehicles, and dispatching decisions. Therefore, optimizing the factors that directly influence the response time is essential to guarantee that efficient and high-quality EMS services are available for the public.

The optimization problem mentioned above is called a location-allocation problem. Such problems are challenging because they usually are formulated as mixed-integer programming models containing many decision variables and constraints. The inclusion of uncertainty further complicates this type of problem. This research addresses a location-allocation problem with system congestion under demand and traffic uncertainty. System congestion refers to a situation where the available response vehicles are insufficient to respond to demand. In this article, system congestion is captured by vehicle availability, i.e., the number of vehicles that can respond to demand, and is influenced by the dispatch of vehicles to demands that overlap in time with the current demand. The uncertainty in demand includes the number of emergencies, locations, occurrence times, numbers of ambulances needed, and the service time needed, which directly influence optimization decisions. The service time is the time interval between the arrival of a vehicle at the emergency site and its return to the station. Traffic uncertainty is mainly represented by travel time, which influences the coverage set of each demand, i.e., the subset of facility sites that can cover demand within time standard. These uncertainties are captured by scenarios, and the problem is formulated as a dynamic two-stage stochastic model. In the first stage, the model determines the optimal locations of ambulance stations and deployment of ambulances without considering realized uncertainties. In the second stage, resource decisions on vehicle

dispatching are made according to the realized uncertainties, first-stage decisions, and status of available vehicles. The objective of the model is to minimize the total cost while responding to as much demand as possible. The total cost comprises the station set-up cost, vehicle purchasing cost, demand fulfillment cost, and penalties for failing to respond to demand within the required time standard and failing to satisfy demand. We first apply Sample Average Approximation (SAA) to approximate the original problem using samples. We next propose a two-phase Benders Decomposition solution scheme (TPBD) to accelerate the computation. To evaluate the performance of our proposed methods, we conduct numerical experiments using real-world emergency data. We also conduct a sensitivity analysis to evaluate the influences of crucial parameters, including the response time standard, facility capacity, service capacity, and facility heterogeneity. We obtain managerial insights that will provide valuable guidance for the design of an emergency response system in practice.

The rest of this paper is organized as follows. Section 2 reviews relevant literature and illustrates contribution. The problem description is presented in Section 3 where the deterministic model is first introduced and then extended to consider random demands and travel time. Section 4 introduces solution approach. We conduct numerical experiments in Section 5. Section 6 concludes this research.

2. Literature Review

The research in routine EMS can be divided into many branches, such as dispatching, location, deployment, assignment, and relocation. Dispatching problem decides which ambulances are dispatched to serve each demand. Location problem determines where to set up stations that can host ambulances. Deployment problem optimizes the number of ambulances hosted at each station. Assignment problem assigns serving stations to demand. Relocation problem identifies where ambulances are moved to. Readers who are interested in these topics can refer to Aringhieri, Bruni, Khodaparasti, and van Essen (2017) and Bélanger, Ruiz, and Soriano (2019). Due to complexity of dispatching and location problem, a lot of research investigates these two problems separately. However, as all of the branches are essential components for an efficient EMS system, research that makes a combination of these branches gains popularity in recent decades. The combination further complicates the problem, especially when uncertainty is considered. Therefore, in this section, we first introduce dispatching and location

problem separately. Then we comprehensively review research that integrates several types of decisions. Finally, we analyze research gaps in existing literature and illustrate how this research fills the gaps.

2.1 Dispatching Problem

According to Lee (2012), dispatching can be divided into two types: call-initiated and server-initiated. Call-initiated dispatching is to choose an appropriate ambulance to respond to an emergency call, while server-initiated dispatching is to decide the demand in the waiting list to be served when a server is available. Regardless of the type of dispatching, decision makers usually adopt certain dispatching policy when making dispatching decisions. The most commonly used dispatching strategy is nearest available policy, which means that an emergency call is most likely to be served by available vehicles that are closest to it (Dean, 2008; Lee, 2011; Zarkeshzadeh, Zare, Heshmati, and Teimouri, 2016). In addition to nearest available policy, several other rules are proposed based on characteristics of the problem. Lee (2011) adopts the concept of preparedness, a quantitative function of the number of available ambulances and call rate, when designing dispatching algorithm for ambulance services. Lee (2012, 2013) define centrality that reflects the density of emergency calls and propose a dispatching policy based on this definition. McLay and Mayorga (2013a) proposes Markov decision process that dispatches distinguished ambulances (i.e., ambulances with different response and service time) to prioritized demand and at the same time considers estimation error of patient priority to calculate the optimal dispatching policies. The purpose is to maximize the expected coverage of true high-risk calls. It is extended by McLay and Mayorga (2013b) to consider both efficiency and equity. Efficiency is represented by expected coverage of high-priority demand. Equity is captured by four types of equity constraints, two of which reflect customer equity and the remaining two reflect server equity. Sudtachat, Mayorga, and McLay (2014) further extends the problem by dispatching two types of ambulances to three-priority-level demand. Zarkeshzadeh, Zare, Heshmati, and Teimouri (2016) develops a weighted hybrid method that combines centrality, nearest neighbor, and first-in-first-out into one model to take advantage of each method. After developing different strategies, performance evaluation is also important. Haghani, Tian, and Hu (2004) uses simulation to evaluate three response strategies, namely the first called first served

strategy, the nearest origin assignment strategy, and the flexible assignment strategy that uses real-time traffic information. Bandara, Mayorga, and McLay (2014) also tests different response strategies to find the optimal dispatching strategy for EMS systems considering demand priority. The above models are formulated under deterministic environment. Jenkins, Robbins, and Lunday (2021) optimizes dispatch of military medical evacuation assets considering uncertain demand where the uncertainty is represented by scenarios. The problem is formulated as a discounted, infinite-horizon Markov decision process model and solved by two approximate dynamic programming methods.

2.2 Location Problem

Research on EMS location problem has a long history which can date back to 1970s. Most of the studies are extensions of two classic coverage models: location set covering problem (LSCP) and maximal covering location problem (MCLP). LSCP is first proposed by Toregas, Swain, ReVelle, and Bergman (1971), which minimizes the number of facilities to cover all demands. The requirement of a mandatory coverage of all demand points may be impossible to implement under some situations, such as budget shortage. Church and ReVelle (1974) proposes MCLP to maximize the demand coverage under facility number constraint. When designing fire protection system where two different types of equipment have to be considered, Schilling, Elzinga, Cohon, Church, and ReVelle (1979) changes definition of coverage in MCLP and proposes FLEET model that requires demand be covered only when it is simultaneously within distance standard of two types of equipment. Daskin and Stern (1981) extends location problem to consider system congestion, which allows demands to be covered by multiple locations so that even the nearest vehicles are engaged, other vehicles within coverage radius can serve demands. Gendreau, Laporte, and Semet (1997) considers redundant coverage and proposes double coverage model (DCM), in which two response distances are considered. All demands are required to be covered within the larger response distance and at the same time a proportion of demands must be covered within smaller response distance.

The models in above articles are deterministic, which can obtain optimality or near-optimality in simplified assumptions of the real-life practices. However, as the operational environment keeps changing, the deterministic inputs may cause biased

results. Therefore, probability is added into the model to represent system instability. Daskin (1983) is one of the early research that adopts busy fraction, i.e., the probability that a server (i.e., ambulance) cannot respond to the demand within time requirements, and proposes a model called MEXCLP, which maximizes the expected coverage. ReVelle and Hogan (1988, 1989a, 1989b) embed busy fraction into chance constraints that require that the probability of the demand being responded is no less than a reliability level, resulting in a problem called maximum availability location problem (MALP). Sorensen and Church (2010) combines MALP with MEXCLP and compares this new model with the two original models in a range of test problems. Liu, Li, Liu, and Patel (2016) combines MALP with DCM to maximize coverage of demand at guaranteed service reliability in a primary distance standard and at the same time to ensure a full coverage in a secondary distance standard. When traffic situation is uncertain, Goldberg and Paz (1991) considers the distribution of the travel time when locating ambulance stations to maximize the expected coverage. The distribution determines the probability that the demand could be responded by the vehicle at certain station within the threshold time. Schmid and Doerner (2010) develops multi-period model to take into account time dependent speed. Berman, Hajizadeh, and Krass (2013) uses MEXCLP where travel time uncertainty is represented by scenarios with certain probability to maximize expected coverage. El Itani, Abdelaziz, and Masri (2019) considers the combination of MEXCLP and MALP and proposes a bi-objective model that simultaneously maximizes expected coverage and minimizes expected cost when paying for external ambulances is allowed. In addition to system efficiency, some research considers equity of the system. Chanta, Mayorga, Kurz, and McLay (2011) defines the concept of envy to model equity when location ambulance stations. Customer envy is calculated based on the distance between demand area and stations on a pre-determined preference list. The objective is to minimize weighted envy where demand density and vehicle availability obtained through queuing theory are two weights used in objective function. Chanta, Mayorga, and McLay (2014) simultaneously considers efficiency and equity by developing a bi-objective covering location model where the former is captured by the first objective and the latter is represented by one of three second objectives at a time.

2.3 Hybrid Problem

The combination of several problems will complicate research. In early days, the combination is usually done at the same level, e.g., strategic location, deployment and assignment, or operational dispatching, deployment, and relocation are combined in one research. Ingolfsson, Budge, and Erkut (2008) optimizes the deployment of ambulances to stations and the assignment to demand under random pre-travel delay, travel time, and vehicle availability. The randomness in the pre-travel delay and travel time is captured by the deviation from the mean time. Beraldi and Bruni (2009) innovatively incorporates joint probabilistic chance constraints into the traditional two-stage stochastic programming model to explore base station location, fleet size, and ambulance assignment problem for EMS under demand uncertainty. van den Berg and Aardal (2015) extends the MEXCLP into a multi-period version with the goal to maximize the expected coverage, minimize start-up cost, and minimize penalty for relocation throughout the day. Degel, Wiesche, Rachuba, and Werners (2015) also uses time-dependent parameters to obtain the more precise and practical solutions for maximum demand coverage. To improve the coverage level and system efficiency, the relocation and additional flexible stations are considered in the model. Liu, Li, and Zhang (2019) uses two-stage distributionally robust model with joint chance constraints to optimize location and deployment considering two types of uncertainties related with demand. Boutilier and Chan (2020) uses two-stage robust optimization model, which makes sure that the worst-case solution is also optimal, to determine location and routing of emergency response vehicles in low- and middle-income countries under demand and travel time uncertainty.

Vehicles are dispatched to satisfy certain demand. To make sure that enough demands are served, the number of ambulances deployed at each station is often jointly optimized with dispatching. Bertsimas and Ng (2019) uses both stochastic and robust two-stage models to solve ambulance dispatching and deployment problem under demand uncertainty. It requires that the number of vehicles dispatched do not exceed the total number deployed. The uncertainty set of robust optimization is calculated based on data-driven approach. When a vehicle is dispatched, the location of the vehicle is empty, reducing protection for surrounding areas. This phenomenon is especially severe in areas with high demand density. An effective method to improve the situation is to relocate vehicles from other less busy stations. Nasrollahzadeh, Khademi, and Mayorga (2018) optimizes real-time ambulance dispatching and relocation, which is formulated

as an infinite-horizon Markov decision process. The model is solved by approximate dynamic programming. Park, Waddell, and Haghani (2019) optimizes dispatch of emergency vehicles in freeway under randomness of requests. Different from research that only looks at past and current demand information, this article further looks ahead a short-term future demand based on incident distribution to dispatch and relocate vehicles. A dynamic programming based method is proposed to solve the problem.

As researchers have deeper understanding about hybrid problem, more complicated combinations (e.g., location and dispatching) are taken into account. Toro-Díaz, Mayorga, Chanta, and McLay (2013) integrates mixed-integer programming model for location and dispatching and hypercube queuing model for system congestion considering fixed priority list for each demand area. Nickel, Reuter-Oppermann, and Saldanha-da-Gama (2016) uses two-stage stochastic programming model to optimize ambulance location, fleet deployment, and the number of vehicles dispatched from stations to serve demands under demand uncertainty. Boujemaa et al. (2018) extends the problem to consider two types of vehicles. Nelas and Dias (2020) proposes a new integer linear programming model that allows vehicle substitution and considers system congestion. Bélanger et al. (2020) proposes a recursive simulation-optimization framework that iterates between an integer programming model and a discrete event simulation model. The integer programming model determines optimal ambulance location and dispatching list for each demand area under given response probability. The discrete event simulation model dispatches vehicles and updates response probability under solution obtained from integer programming model. Peng, Delage, and Li (2020) extends the problem to multiperiod and proposes envelop constraints to guarantee coverage under extreme scenarios. Yoon, Albert, and White (2021) improves solution technique for two-stage stochastic programming model.

2.4 Research Gap and Contribution

As location and dispatch belong to different decision levels, the combination of two-level problems is computationally intensive. Consequently, little relevant hybrid research is available. This article addresses location, dynamic real-time dispatching, and fleet deployment, which are rarely explored in combination in the literature. Location and fleet deployment problems are usually formulated as mixed-integer programming models, while dispatching problems are usually solved by drawing on

queuing theory, which considers dispatch policies and preference lists. However, we use an innovative mixed-integer programming model to formulate the hybrid problem. We further incorporate system congestion into the model when selecting vehicles to be dispatched. In the literature, busy fraction and queuing theory are the most commonly used methods to model system congestion. Busy fraction is generally assumed to be fixed, independent, and exogenous and thus cannot reflect the dynamic and endogenous characteristics of system congestion. When dispatch is modeled by Markov decision process, to simplify the computation, it usually adopts certain assumptions (e.g., arrival and service process) and dispatch policies (e.g., nearest available, preparedness, or centrality). In our research, we aim to establish a dynamic and endogenous dispatch strategy without excessive assumptions or preset dispatching policies. Rather than using busy fraction or queuing theory, we represent system congestion by the functions of a parameter indicating the overlap between demands, deployment decisions, and previous dispatch decisions. The dispatch policy is not pre-defined but rather determined by an objective function. We also incorporate both demand and travel time uncertainties into the model, whereas most research considers neither or only one of these variables. Demand and travel time uncertainties are usually represented by scenarios and the time-dependent travel time, respectively. We adopt the scenario method to represent demand uncertainty, as accurate demand information in each scenario is helpful to model vehicle availability. We represent travel time uncertainty by combining scenario and multi-period methods. We first generate scenarios, each of which represents traffic information from one day. We then divide each scenario into 24 equal 1-hour segments and calculate the segment-dependent travel time. This approach to travel time uncertainty is helpful for calculating the actual coverage set of each demand, which is essential for calculating the total cost. Finally, we propose a new algorithm to accelerate computation, especially for large-size problems.

3. Problem Description and Formulation

This research deals with ambulance location-allocation problem with vehicle availability under demand and traffic uncertainty. We first present a deterministic model in Section 3.1 to make it easier to understand the reasoning and logic behind the model. Then in Section 3.2, a scenario-based two-stage stochastic model is proposed to deal with uncertainties.

3.1 Deterministic Model

In this section, we assume that the information about demand and traffic situation is known a priori. We denote all demands by an ordered set $I = \{1, \dots, |I|\}$ where demand $i-1$ occurs before demand i, $i \in I \backslash \{1\}$. The whole research region is divided into several zones, each of which is represented by its centroid. The location of demand i is the centroid of the zone where i occurs. As the model is deterministic, the occurrence time t_i, the number of required vehicles d_i, and the required service time l_i (time interval between the arrival of a vehicle at the location of demand i and its return to station) are also given.

The set of candidate facilities is denoted by J. Note that in this research, facility, site, and station refer to the same thing and are used interchangeably. The location of each facility j, $j \in J$ is given. One decision of this research is to determine which sites to open, represented by a binary variable x_j, which equals 1 if facility j is open and 0 otherwise. Once facility j is open, it will incur a set-up cost f_j. The number of vehicles deployed at open station j is denoted by y_j which is a decision variable. Each vehicle is purchased at a cost h and can serve any demand, but will be penalized if it is located outside the coverage set N_i of demand i. The coverage set of demand i is the set of facilities that can cover demand i within response time standard R; $N_i \subseteq J$. Each vehicle has the workload limit, the maximum number of demands the vehicle can serve. The purpose of this limit is to balance the workload between stations.

We adopt first-come-first-serve policy for demand service and allow demands to be partially served. The first-come-first-serve policy is widely used in the EMS response, which means if there is an overlap between the service time of two demands requiring the vehicle from the same station, the one that comes first will be served by the vehicle at this station, while the one that comes later will experience one of the three situations: 1) It will be served by vehicles from the same station if there are enough available vehicles left; 2) It will be served by available vehicles from another station; 3) It will be partly served by vehicles from the same station and the rest is served by available vehicles from another station. This can be illustrated by an example in Figure 1, which shows 5 demands. The left side, right side, and length of rectangle represent the occurrence time, finish time, and time interval from occurrence to completion, respectively. The number of vehicles needed is also shown in the figure. These demands

will be served by vehicles from two stations, each of which hosts two vehicles. We assume that station 1 is preferred to all the demands than station 2. The number of available vehicles to each demand and the dispatching decisions are shown in Table 1. Demand 1 happens first, so two vehicles from station 1 are dispatched to demand 1. When demand 2 occurs, these two vehicles are still engaged in the last service, so the vehicle from the less preferred station 2 is dispatched. The same rule is applicable to demand 3 and 4. When demand 5 occurs, only two vehicles are available, but it requires three vehicles. Thus, two vehicles are dispatched and demand 5 is partially served. To calculate the vehicle availability, we need to define a binary parameter $\delta_{ii'j}$, which indicates whether there is overlapping time between demand i and i' for vehicles at station j to serve them.

$$\delta_{ii'j} = \begin{cases} 0, \text{if } t_i \geq t_{i'} + T_{i'j} + l_{i'} \\ 1, \text{otherwise} \end{cases}, \ \forall i \in I, j \in J, i' \in \{1, 2, \dots, i-1\}. \quad (1)$$

Notice that when $\delta_{ii'j} = 0$, demand i and i' are disjoint. The service to i' will not influence the available vehicles to i. However, when $\delta_{ii'j} = 1$, the allocation decision of i' has influence on vehicle availability to i.

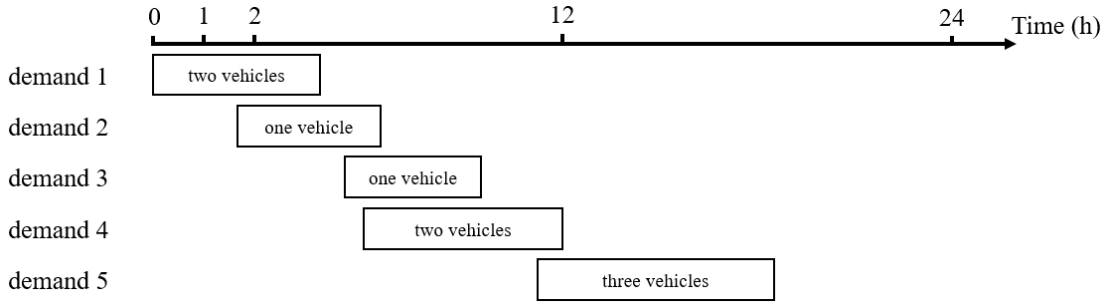


Figure 1. One example of demand in one day period

Table 1. One example of vehicle allocation

	Station 1		Station 2	
	Available	Allocated	Available	Allocated
d1	2	2	2	0
d2	0	0	2	1
d3	2	1	1	0
d4	1	1	1	1
d5	1	1	1	1

The goal of the research is to identify the optimal location of stations, the number and deployment of vehicles, and the allocation of demand at a minimum total cost. All the notations for the deterministic model are introduced in Table 2 and the model is given as follows:

Table 2. Notations for deterministic model

Sets	
I	The set of a sequence of demand, where the demand $i-1$ happens before i
J	The set of candidate facility sites
N_i	The coverage set of demand i, i.e., the set of facility sites that can cover demand i within time standard ($N_i = \{j \in J, T_{ij} \leq R\}$)
Parameters	
f_j	The fixed cost of opening station j
h	The purchasing cost of a vehicle
c_{ij}	The unit transportation cost for vehicle at station j to serve demand i
c^{lost}	The unit penalty of demand unsatisfaction
γ	The unit penalty for violating the response time standard
T_{ij}	The travel time for vehicle at station j to serve demand i
R	The response time standard
Q_j	The maximal number of vehicles that can be hosted at station j
d_i	The number of vehicles needed for demand i
μ_j	The maximum number of demands each vehicle at station j could serve
t_i	The occurrence time of demand i
l_i	The service time of demand i, the time needed after vehicle arriving at the emergency site until the vehicle goes back to station again
$\delta_{ii'j}$	1 if there is overlapping time between demand i and i' for vehicles at station j to serve them, 0 otherwise
Decision variables	
x_j	1 if a station is set up at site j, 0 otherwise
y_j	The number of vehicles hosted at location j
z_{ij}	The number of vehicles at location j that are dispatched to demand i

[M1] Min $\sum_{j\in J} f_j x_j + \sum_{j\in J} h y_j + \sum_{i\in I}\sum_{j\in J} c_{ij} T_{ij} z_{ij} + \sum_{i\in I}\sum_{j\in J\setminus N_i} \gamma(T_{ij} - R) z_{ij} +$
$c^{lost}(\sum_{i\in I} d_i - \sum_{i\in I}\sum_{j\in J} z_{ij})$ (2)

subject to

$y_j \le Q_j x_j, \ \forall j \in J$ (3)

$\sum_{j\in J} z_{ij} \le d_i, \ \forall i \in I$ (4)

$z_{ij} \le y_j - \sum_{i'=1}^{i-1} \delta_{ii'j} z_{i'j}, \ \forall i \in I, j \in J$ (5)

$\sum_{i\in I} z_{ij} \le \mu_j y_j, \ \forall j \in J$ (6)

$x_j \in \{0,1\}, \ \forall j \in J$ (7)

$y_j \in \mathbb{Z}_0^+, \ \forall j \in J$ (8)

$z_{ij} \in \mathbb{Z}_0^+, \ \forall i \in I, j \in J.$ (9)

The objective function (2) minimizes total cost, which consists of station set-up cost, vehicle purchasing cost, demand fulfillment cost, the penalty for not responding the demand in required time standard, and penalty for demand unsatisfaction. For the third term, we only account for the vehicles located outside the coverage set because only these vehicles cannot respond to demand within time standard. Constraints (3) require that vehicles can only be located at open station and the number cannot exceed the station capacity. Constraints (4) state that demand can be partially satisfied. Constraints (5) require that only vehicles available at station when demand occurs are available to serve it. The second term on the right-hand side is the number of unavailable vehicles when demand i occurs, which is determined by $\delta_{ii'j}$ and $z_{i'j}$. $\delta_{ii'j}$ indicates whether different demands would have overlap if they are allocated to the same station, which is illustrated by Equation (1). If when demand i occurs, any service to demand i' that occurs before demand i has already been completed, there is no overlap between i and i' and $\delta_{ii'j} = 0$. If when demand i occurs, there are demands being served, the overlap exists and $\delta_{ii'j} = 1$. $z_{i'j}$ is the number of vehicles at location j that are dispatched to demand i' that occurs before demand i. When $\delta_{ii'j} = 0$, i.e., there is no overlap between demand i and i', whatever the decision $z_{i'j}$ is, the available vehicles at station j will not be influenced. When $\delta_{ii'j} = 1$, i.e., there is overlap between demand i and i', if $z_{i'j} = 0$, the available vehicles at station j will not be influenced. If $z_{i'j} > 0$, $z_{i'j}$number of the vehicles are engaged in demand i' when i occurs, thus the available vehicles at station j will be reduced by $z_{i'j}$. Then the total number of unavailable vehicles because of the preexisting demands is $\sum_{i'=1}^{i-1} \delta_{ii'j} z_{i'j}$. Constraints (6) state that the demands served by each station cannot

exceed the workload of the vehicle at this station. Constraints (7) to (9) set the domain of decision variables.

3.2 Stochastic Model

One of the important limitations of deterministic model is the assumption that all the parameters are known in advance. However, in practice, it is hard to know what will happen in the future, especially the demand and traffic condition. For this reason, we have to develop a model that could take the uncertainty into consideration. The model developed in this section is a scenario-based two-stage stochastic programming model. In the first stage, the model determines the optimal location of ambulance stations, fleet size, and the deployment of ambulances without considering the realization of uncertainties. In the second stage, recourse decisions on ambulances dispatching are made based on scenarios, first-stage decisions, and state of available vehicles. We denote by S the set of scenarios. Each scenario $s \in S$ contains the information of demand and traffic situation during a one-day period and is associated with a probability of occurrence p_s. The information includes a sequence of demands that happen during the day, their location and occurrence time, the travel time between candidate stations and demand sites when the demand occurs, and the service time. Under the changing scenarios, the value of some parameters and variables may change accordingly. The demand coverage set N_i is different because the demand location and the travel time between demand and candidate locations change. Parameter $\delta_{ii'j}$ and decision variables z_{ij} and σ_i also change. The definition of $\delta_{ii'j}$ under scenario s is given as follows:

$$\delta_{ii'j}^{s} = \begin{cases} 0, \text{ if } t_i^s \geq t_{i'}^s + T_{i'j}^s + l_{i'}^s \\ 1, \text{ otherwise} \end{cases}, \ \forall i \in I, j \in J, i' \in \{1, 2, \dots, i-1\}, s \in S. \quad (10)$$

The objective function turns to calculate the minimum expected total cost. The additional parameters and variables for stochastic model are listed in Table 3 and the scenario-based two-stage stochastic programming model is as follows

Table 3. Additional notations in stochastic model

Sets	
S	The set of scenarios
I^s	The set of a sequence of demand under scenario s
N_i^s	The coverage set of demand i under scenario s
Parameters	
T_{ij}^s	The travel time for vehicle at station j to serve demand i under scenario s
d_i^s	The number of vehicles needed for demand i under scenario s
t_i^s	The occurrence time of demand i under scenario s
l_i^s	The service time of demand i under scenario s
$\delta_{ii'j}^s$	1 if there is overlapping time between demand i and i' for vehicles at station j to serve them under scenario s, 0 otherwise
Decision variables	
z_{ij}^s	The number of vehicles at location j that are dispatched to demand i under scenario s

$$\text{Min } \sum_{j\in J} f_j x_j + \sum_{j\in J} h y_j + \sum_{s\in S} p_s \Big(\sum_{i\in I^s}\sum_{j\in J} c_{ij} T_{ij}^s z_{ij}^s + \sum_{i\in I^s}\sum_{j\in J\setminus N_i^s} \gamma (T_{ij}^s - R) z_{ij}^s + c^{lost}\big(\sum_{i\in I^s} d_i^s - \sum_{i\in I^s}\sum_{j\in J} z_{ij}^s\big)\Big) \quad (11)$$

subject to (3), (7), (8)

$$\sum_{j\in J} z_{ij}^s \le d_i^s,\ \forall i \in I^s, s \in S \quad (12)$$

$$z_{ij}^s \le y_j - \sum_{i'=1}^{i-1} \delta_{ii'j}^s z_{i'j}^s,\ \forall i \in I^s, j \in J, s \in S \quad (13)$$

$$\sum_{i\in I^s} z_{ij}^s \le \mu_j y_j,\ \forall j \in J, s \in S \quad (14)$$

$$z_{ij}^s \in \mathbb{Z}_0^+,\ \forall i \in I^s, j \in J, s \in S. \quad (15)$$

4. Solution Approach

The scenario-based two-stage stochastic programming model is challenging to solve because in real life the demand and traffic situation are changing all the time, resulting in a large number of scenarios, which makes the problem computationally intractable. In this section, we will first introduce SAA, which is used to approximate the original problem with samples. Then a Benders Decomposition based approach is proposed to speed up the computation.

4.1 Sample Average Approximation

The SAA method is a Monte Carlo simulation-based approach to solve stochastic optimization problems. The basic idea of this approach is to approximate the true distribution by empirical distribution obtained from samples. The sample is represented by S', which is a finite set of scenarios sampled from S with the same probability of occurrence, i.e., $S' \subseteq S$, $|S'|$ is the sample size, and each scenario in S' has the same probability $1/|S'|$. The SAA formulation is given as follows:

$$[SAA]\ \text{Min}\ \sum_{j\in J} f_j x_j + \sum_{j\in J} h y_j + \sum_{s\in S'} \frac{1}{|S'|}\Big(\sum_{i\in I^s}\sum_{j\in J} c_{ij} T_{ij}^s z_{ij}^s + \sum_{i\in I^s}\sum_{j\in J\setminus N_i^s} \gamma\left(T_{ij}^s - R\right) z_{ij}^s + c^{lost}\Big(\sum_{i\in I^s} d_i^s - \sum_{i\in I^s}\sum_{j\in J} z_{ij}^s\Big)\Big) \quad (16)$$

subject to (3), (7), (8), (12)–(15) in which S is replaced by S'.

When using SAA to solve the problem, one essential procedure is to determine the number of scenarios in S'. The solution quality will be improved with the increase of sample size, while the model will become computationally intractable. We need to strike a balance between precision and computational tractability. Algorithm 1 describes the procedure to evaluate the solution quality of SAA under given sample size, which includes the calculation of confidence intervals (CIs) for lower bound, upper bound, and optimality gap under given sample size. We will discuss the choice of $|S'|$ in Section 5 using real-world emergency data.

Algorithm 1 Estimate $(1-\tau)$-CI for lower bound, upper bound, and optimality gap of two-stage stochastic program

---

1. Generate a set of scenarios S'.
2. Solve the SAA with S' and obtain the optimal first-stage solution x^*, y^*.
3. **for** $m = 1, 2, \dots, M$ **do**
4. Generate a set of new independent scenarios S_m, $|S_m| = |S'|$.
5. Solve SAA with S_m and obtain the objective value v_m.
6. Generate a set of new independent scenarios S'_m, $|S'_m| \gg |S_m|$.
7. Evaluate the quality of the first-stage solution x^*, y^* on scenarios in S'_m. Input x^*, y^* into SAA, obtaining cost $v^m_{x^*,y^*}$.
8. Let $g_m := v^m_{x^*,y^*} - v_m$.
9. **end for**

10. Estimate $(1-\tau)$-CI for lower bound

11. Let $L:=\frac{1}{M}\sum_{m=1}^{M} v_m$ and $S_L:=\frac{1}{M-1}\sum_{m=1}^{M}(v_m-L)^2$.

12. The $(1-\tau)$-CI for lower bound is $\left[L-\frac{t_{M-1,\frac{\tau}{2}}\sqrt{S_L}}{\sqrt{M}}, L+\frac{t_{M-1,\frac{\tau}{2}}\sqrt{S_L}}{\sqrt{M}}\right]$, $t_{M-1,\frac{\tau}{2}}$ is the t-value obtained from t-distribution with degrees of freedom $M-1$ and confidence level $1-\tau$.

13. Estimate $(1-\tau)$-CI for upper bound

14. Let $U:=\frac{1}{M}\sum_{m=1}^{M} v_{x^*,y^*}^m$ and $S_U:=\frac{1}{M-1}\sum_{m=1}^{M}(v_{x^*,y^*}^m-U)^2$.

15. The $(1-\tau)$-CI for upper bound is $\left[U-\frac{t_{M-1,\frac{\tau}{2}}\sqrt{S_U}}{\sqrt{M}}, U+\frac{t_{M-1,\frac{\tau}{2}}\sqrt{S_U}}{\sqrt{M}}\right]$.

16. Estimate $(1-\tau)$-CI for optimality gap

17. Let $G:=\frac{1}{M}\sum_{m=1}^{M} g_m$ and $S_G:=\frac{1}{M-1}\sum_{m=1}^{M}(g_m-G)^2$.

18. The $(1-\tau)$-CI for optimality gap is $\left[0, G+\frac{t_{M-1,\tau}\sqrt{S_G}}{\sqrt{M}}\right]$.

4.2 Benders Decomposition Based Solution Scheme

Benders decomposition is efficient in dealing with complicated large-scale two-stage stochastic programming problem. The basic idea is to decompose the scenario-based two-stage stochastic programming model into a master problem and a number of subproblems. Then these two types of problems are solved iteratively, adding additional constraints to the master problem. Usually the second-stage model under each scenario is regarded as an independent subproblem, see Peng, Delage, and Li (2020). In this case, many Benders cuts, one for each subproblem, are added at each iteration to accelerate the convergence. According to Adulyasak, Cordeau, and Jans (2015), adding too many cuts at each iteration can lead to a worse performance because of the time taken to solve the master problem. Therefore, in this research, we first separate first- and second-stage models and then aggregate all the scenarios in the SAA for the second-stage model, resulting in a master problem MP and a subproblem $SP(x,y)$ that depend on the first-stage solutions as follows:

$$[MP]\ \text{Min}\sum_{j\in J} f_j x_j + \sum_{j\in J} h y_j + \theta \tag{17}$$

subject to (3), (7), (8)

$$\theta \geq Q(x,y) \tag{18}$$

$$[\ SP(x,y)\ ]\quad Q(x,y) = \min \sum_{s\in S'} \frac{1}{|S'|}\Big(\sum_{i\in I^s}\sum_{j\in J} c_{ij} T_{ij}^s z_{ij}^s + \sum_{i\in I^s}\sum_{j\in J\setminus N_i^s} \gamma\big(T_{ij}^s - R\big) z_{ij}^s + c^{lost}\Big(\sum_{i\in I^s} d_i^s - \sum_{i\in I^s}\sum_{j\in J} z_{ij}^s\Big)\Big) \tag{19}$$

subject to

$$\sum_{j\in J} z_{ij}^s \le d_i^s,\ \forall i \in I^s, s \in S' \quad (20)$$

$$z_{ij}^s \le y_j - \sum_{i'=1}^{i-1} \delta_{ii'j}^s z_{i'j}^s,\ \forall i \in I^s, j \in J, s \in S' \quad (21)$$

$$\sum_{i\in I^s} z_{ij}^s \le \mu_j y_j,\ \forall j \in J, s \in S' \quad (22)$$

$$z_{ij}^s \in \mathbb{Z}_0^+,\ \forall i \in I^s, j \in J, s \in S' \quad (23)$$

where S' represents the sample used in SAA.

Both master problem and subproblem are integer optimization problems. Repeatedly solving these two types of problems is time-consuming. What is more, the coefficient matrix in constraints is not necessarily totally unimodular, making the problem more difficult to solve. Thus, we do not apply the classical Bender decomposition method (Benders, 1962) to solve this problem. As an alternative, we propose a two-phase Benders Decomposition based method. In phase 1, we relax all the integer constraints in MP and $SP(x,y)$, obtaining relaxed version called RMP and $RSSP(x,y)$ respectively, where θ in Eq. (35) is an underestimator of $Q_r(x,y)$ (the subscript "r" means "relaxed") and successively approximates the shape of $Q_r(x,y)$ by adding cuts derived from dual problem $DRSP(x,y)$. We iterate between RMP and $RSSP(x,y)$, adding Benders cuts (35) into RMP until no Benders cut is added, the gap between upper bound UB and lower bound LB is below a predetermined tolerance ϵ, or a preset time is reached.

$$[RMP]\ \text{Min} \sum_{j\in J} f_j x_j + \sum_{j\in J} h y_j + \theta \quad (24)$$

subject to (3)

$$\theta \ge Q_r(x,y) \quad (25)$$

$$0 \le x_j \le 1,\ \forall j \in J \quad (26)$$

$$y_j \ge 0,\ \forall j \in J \quad (27)$$

$$[\ RSP(x,y)\ ]\quad Q_r(x,y) = \min \sum_{s\in S'} \frac{1}{|S'|} \Big(\sum_{i\in I^s}\sum_{j\in J} c_{ij} T_{ij}^s z_{ij}^s + \sum_{i\in I^s}\sum_{j\in J\setminus N_i^s} \gamma\big(T_{ij}^s - R\big) z_{ij}^s + c^{lost}\Big(\sum_{i\in I^s} d_i^s - \sum_{i\in I^s}\sum_{j\in J} z_{ij}^s\Big)\Big) \quad (28)$$

subject to (20)–(22)

$$z_{ij}^s \ge 0,\ \forall i \in I^s, j \in J, s \in S'. \quad (29)$$

$$[\,DRSP(x,y)\,]\ \ D_r(x,y) = \max \sum_{s\in S'} \Big(\sum_{i\in I^s} \lambda_{is}^1 d_i^s + \sum_{i\in I^s}\sum_{j\in J} \lambda_{ijs}^2 y_j + \sum_{j\in J} \lambda_{js}^3 \mu_j y_j + \frac{c^{lost}}{|S'|} \sum_{i\in I^s} d_i^s\Big) \quad (30)$$

subject to

$$\lambda_{is}^1 + \lambda_{ijs}^2 + \sum_{i'=i+1}^{|I^s|} \delta_{i'ij}^s \lambda_{ijs}^2 + \lambda_{js}^3 \le \frac{1}{|S'|}\left(c_{ij}T_{ij}^s + \mathbb{I}_{j\in J\setminus N_i^s}\gamma\left(T_{ij}^s - R\right) - c^{lost}\right), \quad \forall i \in I^s, j \in J, s \in S' \tag{31}$$

$$\lambda_{is}^1 \le 0, \ \forall i \in I^s, s \in S' \tag{32}$$

$$\lambda_{ijs}^2 \le 0, \ \forall i \in I^s, j \in J, s \in S' \tag{33}$$

$$\lambda_{js}^3 \le 0, \ \forall j \in J, s \in S' \tag{34}$$

where $\mathbb{I}_{j\in J\setminus N_i^s}$ is an indicator function. If $j \in J\setminus N_i^s$, the value is 1, otherwise, it is 0.

$$\theta \ge \sum_{s\in S'}\left(\sum_{i\in I^s}\bar{\lambda}_{is}^1 d_i^s + \sum_{i\in I^s}\sum_{j\in J}\bar{\lambda}_{ijs}^2 y_j + \sum_{j\in J}\bar{\lambda}_{js}^3 \mu_j y_j + \frac{c^{lost}}{|S'|}\sum_{i\in I^s} d_i^s\right) \tag{35}$$

where $(\bar{\lambda}_{is}^1, \bar{\lambda}_{ijs}^2, \bar{\lambda}_{js}^3)$ are optimal solutions of $DRSP(x,y)$ given optimal master problem solution $(\bar{x},\bar{y})$. These expressions are added to RMP to tighten the lower bound.

In phase 2, we restore the integrality constraints in master problem and keep all the Benders cuts generated in phase 1. We input the optimal solution of MP to $RSSP(x,y)$ to evaluate whether additional cuts are needed. If no Benders cut is added or a preset time is reached, solve $SP(x,y)$ under given optimal solution of MP. The whole algorithm is presented as follows:

Algorithm 2 Benders Decomposition based solution approach (TPBD)

Phase 1 procedure:
Input A tolerance $\epsilon_1 \ge 0$ and maximum phase 1 run time TL_1.
Initialize $UB_1 = \infty$, $LB_1 = 0$.
for $i = 1, 2, \ldots$ **do**
    Solve RMP to obtain optimal solution $(\bar{x}_r, \bar{y}_r)$ and the optimal objective value $lobj$.
    Update $LB := max(LB, lobj)$.
    Solve $RSP(\bar{x}_r, \bar{y}_r)$ to obtain $Q_r(\bar{x}_r, \bar{y}_r)$.
    if $\theta < Q_r(\bar{x}_r, \bar{y}_r)$ **then**
        Add optimality cut (35) to RMP.
    end if
    Calculate $uobj = \sum_{j\in J} f_j \bar{x}_{rj} + \sum_{j\in J} h\bar{y}_{rj} + Q_r(\bar{x}_r, \bar{y}_r)$.
    Update $UB := min(UB, uobj)$
    if No Benders cut is added or $UB - LB \le \epsilon_1$ or $runtime \ge TL_1$ **then**
        Break.

33. **end if**

34. **end for**

35. **Phase 2 procedure:**

36. **Input** Maximum phase 2 run time TL_2.

37. **Initialize** $UB_2 = \infty$, $LB_2 = LB_1$.

38. Keeping all Benders cuts generated in phase 1 procedure.

39. **for** $i = 1, 2, \ldots$ **do**

40. Solve MP to obtain optimal solution $(\bar{x}, \bar{y})$ and the optimal objective value $lobj^*$.

41. Update $LB := max(LB, lobj^*)$.

42. Solve $RSP(\bar{x}, \bar{y})$ to obtain $Q_r(\bar{x}, \bar{y})$.

43. **if** $\theta < Q_r(\bar{x}_r, \bar{y}_r)$ **then**

44. Add optimality cut (35) to MP.

45. **end if**

46. **if** no Benders cut is added or $runtime \geq TL_2$ **then**

47. Solve $SP(\bar{x}, \bar{y})$ to obtain $Q(\bar{x}, \bar{y})$.

48. **end if**

49. Calculate $uobj^* = \sum_{j \in J} f_j \bar{x}_j + \sum_{j \in J} h \bar{y}_j + Q(\bar{x}, \bar{y})$.

50. Update $UB := min(UB, uobj)$.

51. Break.

52. **end for**

53. **Return** UB and corresponding optimal solution $(\bar{x}, \bar{y})$.

5. Numerical Experiments

To evaluate the performance of our method, we conducted numerical experiments using real-world emergency data. We first determined how many scenarios are needed to obtain a high level of approximation precision while at the same time make the model computationally tractable in SAA. Then we compared the computational performances of SAA with TPBD under the same sample. To show the robustness of solution method, we conducted out-of-sample analysis. Next, we evaluated the benefit of using stochastic programming approach over deterministic model. Finally, we conducted sensitive analysis to show how the value of some crucial parameters will influence the optimal objective value of our model, which yields some valuable managerial insights. All the experiments were carried out on a computer with i9-12900K CPU, 3.20 GHz processing

speed and 32 GB of memory. The model and the algorithm were implemented in C++ programming and solved by CPLEX 12.10.

5.1 Parameter Setting

We used the emergency incident data of Manhattan, which is provided by Fire Department of New York City (FDNY). The data spans from the time the incident is created to the time the incident is closed in the system, including the incident datetime, incident location, response time, response police precinct, etc. We finally chose the data of year 2011, which includes 225634 emergency call logs occurring at 22 police precincts, as the data of the other years has a large number of missing values. The 22 police precincts were regarded as the demand areas. The candidate facility site was obtained from the website of New York City Fire Department Bureau of Emergency Medical Services (FDNY EMS), which is divided into four sectors, but nearly all the incidents are served by two sectors: FDNY EMS municipal (FEM) and Voluntary Hospital EMS (VHE). The former controls 70% of the ambulances in the New York City 911 System and serves 63% of ambulance tours while the latter controls the rest of the ambulances and serves 37% of the ambulance tours. FEM now operates 6 stations in Manhattan and VHE provides emergency services through 10 stations. Totally, 16 stations were used as the candidate facility sites in this section. The demand area and candidate stations are shown in Figure 2, where the numbers and letters are the indices of the police precinct and candidate station, respectively. The red and blue circles represent the location of stations operated by FEM and VHE, respectively. In order to reflect the fact that the 6 stations operated by FEM work as the main emergency facilities, the fixed cost was set lower and facility and service capacity were higher than those of stations operated by VHE. The vehicle travel speed data was calculated using the 2011 Yellow Taxi Trip Data, which includes trip distance and trip duration. We divided each day into 24 equal segments, i.e., each segment is one hour, and we calculated the average speed of each segment as the speed at which vehicles responded to emergency calls occurring during that segment. The response time standard was set to 9 minutes according to National Fire Protection Association benchmark. The fixed cost to set up a station, including land cost, construction cost, material cost, etc., is amortized according to service life of buildings to $1500 and $4500 per day for stations belong to FEM and VHE, respectively. The reason for cost difference between FEM

and VHE stations is that FEM stations are more convenient and flexible for vehicle deployment and dispatching. When an emergency occurs, FEM stations are preferred for service. Therefore, facility capacity and vehicle service capacity for FEM stations are also greater than that of VHE stations. Vehicle purchase cost, including ambulance cost, equipment cost, maintenance cost, etc., is amortized to $300 per vehicle per day according to service life of ambulance. Unit transportation cost, including labor cost, fuel cost, etc., is $30 per minute. The penalty cost is set to 5 times the transportation cost because survivability is highly dependent on response time and the response that is later than response time standard may have serious consequences to life and property. The values of parameters used in numerical experiments are listed in Table 4.

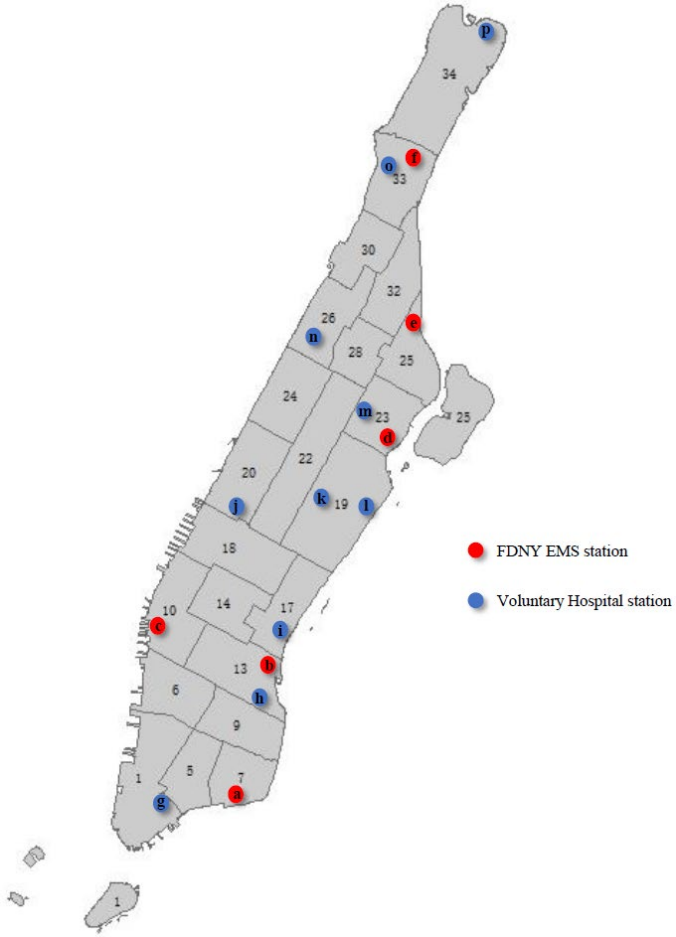


Figure 2. The demand area and candidate stations

583 Table 4. The values of parameters used in numerical experiments

Parameter	Value
f_j	\$1500 per day if j belongs to FEM
	\$4500 per day if j belongs to VHE
h	\$300 per vehicle per day
c_{ij}	\$30 per minute
γ	\$150 per minute
Q_j	10 vehicles if j belongs to FEM
	5 vehicles if j belongs to VHE
α	90%
μ_j	20 demands per day if j belongs to FEM
	5 demands per day if j belongs to VHE

584

585 **5.2 Determination of Sample Size**

586 In this section, we will determine the sample size for the following numerical
587 experiments. We first introduce how a sample is generated. Then, we illustrate the
588 procedures and results of solution quality evaluation. Accordingly, we finally determine
589 the sample size.

590

591 5.2.1 Sample Generation

592 A sample is made up of a specific number of scenarios. A scenario is real emergency
593 rescue data for a day, including demand and travel speed data. The specific number of
594 scenarios is called sample size. We have emergency call logs and yellow taxi trip data
595 of year 2011, which spans over 365 days. We randomly generated a number between 1
596 and 365 and extracted all emergency call logs and yellow taxi trip data for the day
597 corresponding to this number as scenario data. We repeated this procedure in the
598 remaining data set until the number of scenarios equals the sample size.

599

600 5.2.2 Solution Quality Evaluation

601 We used Algorithm 1 to evaluate the performance of different sample sizes, which
602 were set to 1, 3, 5, 8, 10, 15, 20, 30, 50, 80, and 100. We first ran line 1 and 2 of
603 Algorithm 1 to obtain the optimal first-stage solution. Then we iterated line 4 to 8 for
604 10 times where $|S'_m|$ was set to 150. Finally, we calculated solution quality, which is

shown in Table 5. NS, PE, CI, CI-L, CI-U, and Time represent sample size, point estimate, 95% confidence interval, ratio between 95% confidence interval for optimality gap and point estimate of lower bound, ratio between 95% confidence interval for optimality gap and point estimate of upper bound, and running time, respectively. When sample size was set to 80 and 100, the computer ran out of memory due to the large number of variables and constraints. Therefore, we do not show the results of sample size 80 and 100. It can be seen from the table that when sample size is 20, the error rate is below 1% with a high probability. Considering the trade-off between solution quality and computational tractability, we set the sample size to 20 in the following calculation.

Table 5. Solution quality of different sample sizes (95% CIs)

NS	Lower bound		Upper bound		Optimality gap		CI-L	CI-U	Time
	PE	CI	PE	CI	PE	CI			
1	188700	(182300, 195100)	191700	(177200, 206200)	3063	(0, 18340)	9.72%	9.57%	164
3	178900	(173500, 184300)	181000	(177000, 185000)	2141	(0, 9291)	5.19%	5.13%	674
5	186400	(184300, 188500)	189300	(184000, 194600)	2904	(0, 8149)	4.37%	4.30%	1380
8	183400	(181600, 185200)	185900	(182000, 189800)	2466	(0, 5760)	3.14%	3.10%	2679
10	180800	(175600, 186000)	182000	(178100, 185900)	1194	(0, 4330)	2.39%	2.38%	2648
15	180400	(175900, 184900)	181600	(176200, 187000)	1252	(0, 2197)	1.22%	1.21%	3835
20	180300	(176400, 184200)	181500	(178900, 184100)	1220	(0, 1404)	0.78%	0.77%	6587
30	183700	(180700, 186700)	184700	(180300, 189100)	952	(0, 1072)	0.58%	0.58%	11960
50	179400	(176500, 182300)	180200	(177700, 182700)	786	(0, 803)	0.45%	0.45%	35360

5.3 Numerical Performance of Our TPBD Algorithm

Since each scenario contains around 800 demands, with the increase of sample size, SAA will become computationally inefficient or even run out of memory. In this section, we compared the computational performances of SAA and TPBD under different samples, which are shown in Table 6. The first column represents the sample size. *Total Cost* is the objective value. *Time* is the computation time. It can be seen that TPBD obtains the same results as SAA. When the sample size is small, SAA gets the optimal solution in shorter time. As the sample size increases, TPBD becomes more efficient. If we set the time limit to one hour, TPBD can get the optimal solutions under all the

sample sizes. However, SAA can only finish the computation for sample sizes less than 50. The results suggest that TPBD is an efficient algorithm, especially for large-size problem. Therefore, in the following numerical experiments, we will use TPBD to solve the problem unless otherwise specified.

Table 6. Comparison of computational performances between SAA and TPBD

	Total Cost		Time	
NS	SAA	TPBD	SAA	TPBD
1	157800	157800	1.21	6.27
3	180100	180100	9.19	23.41
5	179700	179700	27.75	64.05
8	188600	188600	40.05	111.80
10	172500	172500	51.41	154.50
15	170400	170400	127.80	257.10
20	183000	183000	842.40	550.80
30	183800	183800	1037.00	620.50
50	186900	186900	4942.00	1355.00
80	178900	178900	9207.00	2907.00
100	180200	180200	11650.00	3588.00

5.4 Robustness Evaluation

Optimal solutions obtained through TPBD are based on generated scenarios. It is highly possible that in practice the realized demand is not a member of used samples, resulting in poor performance of solution approach. To test the robustness of TPBD method, we conduct out-of-sample analysis, which evaluates performance of optimal solutions using out-of-sample data. The procedures are shown in Algorithm 3. We divide all scenarios into two sets: in-sample scenario S^{in} and out-of-sample scenario S^{out}, where $S^{in} \cup S^{out} = S$. All scenarios in S^{in} are input into TPBD to obtain optimal first-stage solution x^* and y^*. Then N scenarios are randomly selected from S^{out}, each of which is input into deterministic model [M1] to calculate optimal dispatching decisions. There are two possible outcomes: optimal solution exists and optimal solution does not exist. In the first case, calculate coverage level, i.e., the percentage of demand covered by opening stations within response time standard, response level, i.e., the proportion of demand that is responded within response time standard, and demand loss rate, i.e., the percentage of demand that is unserved. Robustness is measured by robustness level, the proportion of scenarios where optimal

649 solution can be found, $(1-\tau)$-CI of coverage level, response level, and demand loss
650 rate.
651

Algorithm 3 Robustness evaluation

1. Generate a set of scenarios S^{in}.
2. Solve the TPBD problem with S^{in} and obtain the optimal first-stage solution x^* and y^*.
3. **for** $n = 1, 2, \dots, N$ **do**
4. Generate a new independent scenario s_n from S^{out}, $s_n \in S^{out}$.
5. Solve deterministic model [M1] with s_n, x^* and y^*.
6. If optimal solution can be found, calculate coverage level CL_n, response level RL_n, and demand loss rate DL_n.
7. **end for**
8. Calculate the number of iterations N_o where optimal solution can be found.
9. Calculate robustness level $BL = \frac{N_0}{N}$.
10. Let $ACL := \frac{\sum_{n=1}^{N_0} CL_n}{N_0}$ and $S_{ACL} := \frac{1}{N_0-1}\sum_{n=1}^{N_0}(CL_n - ACL)^2$.
11. Let $ARL := \frac{\sum_{n=1}^{N_0} RL_n}{N_0}$ and $S_{ARL} := \frac{1}{N_0-1}\sum_{n=1}^{N_0}(RL_n - ARL)^2$.
12. Let $ADL := \frac{\sum_{n=1}^{N_0} DL_n}{N_0}$ and $S_{ADL} := \frac{1}{N_0-1}\sum_{n=1}^{N_0}(DL_n - ADL)^2$.
13. The $(1-\tau)$-CI of coverage level is $\left[ACL - \frac{t_{N_0-1,\frac{\tau}{2}}\sqrt{S_{ACL}}}{\sqrt{N_0}}, ACL + \frac{t_{N_0-1,\frac{\tau}{2}}\sqrt{S_{ACL}}}{\sqrt{N_0}}\right]$.
14. The $(1-\tau)$-CI of response level is $\left[ARL - \frac{t_{N_0-1,\frac{\tau}{2}}\sqrt{S_{ARL}}}{\sqrt{N_0}}, ARL + \frac{t_{N_0-1,\frac{\tau}{2}}\sqrt{S_{ARL}}}{\sqrt{N_0}}\right]$.
15. The $(1-\tau)$-CI of demand loss rate is $\left[ADL - \frac{t_{N_0-1,\frac{\tau}{2}}\sqrt{S_{ADL}}}{\sqrt{N_0}}, ADL + \frac{t_{N_0-1,\frac{\tau}{2}}\sqrt{S_{ADL}}}{\sqrt{N_0}}\right]$

652

653 We ran TPBD and deterministic model [M1] 10 times separately. There are two types
654 of solutions for TPBD according to the number of opening stations: one with 12 opening
655 stations and the other with 13 opening stations. For the solution with 13 opening stations,
656 the number of deployed vehicles is the same. We denoted this type of solution by
657 TPBD-13. For the solution with 12 opening stations, the number of deployed vehicles
658 varies. Therefore, we selected two solutions with the fewest and most vehicles, denoting
659 them by TPBD-12-F and TPBD-12-M, respectively. For deterministic model, there are
660 two types of solutions according to the number of opening stations: one with 11 opening

stations and the other with 12 opening stations. The number of deployed vehicles under both solution types varies. We took the same measure as TPBD with 12 opening stations and finally we got four solution types, namely MV-12-F, MV-12-M, MV-11-F, and MV-11-M. Totally, we got seven different optimal first-stage solutions, which were put into Algorithm 3 to calculate robustness level and 95% CI of coverage level, response level, and demand loss rate with N set to 150. Figure 3 and Figure 4 show 95% CI of coverage level and response level, respectively. Results show that compared with deterministic model, stochastic model can obtain solutions that cover and respond to more emergency calls within response time standard with high probability. Even though these solutions are tested in out-of-sample data, at least 98.68% and 95.55% of the emergency demands can be covered and responded to in time with high probability. While solutions of deterministic model can provide a maximum 98.69% and 94.52% of coverage and responses respectively within response time standard with high probability. Most of the time, the response level is below 90%. All solution types do not have demand unsatisfied, except MV-12-F and MV-11-F having demand loss rate 0.2% and 0.6% respectively with a high probability. Robustness level is 100% for all solution types, indicating that solutions obtained by TPBD can find optimal value under every scenario in out-of-sample analysis. All these results suggest that TPBD is a robust solution approach.

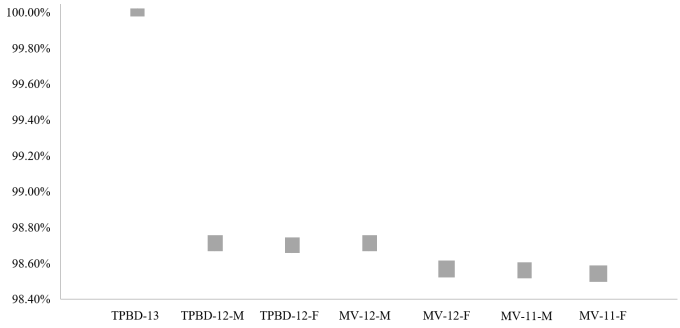


Figure 3. 95% CI of coverage level

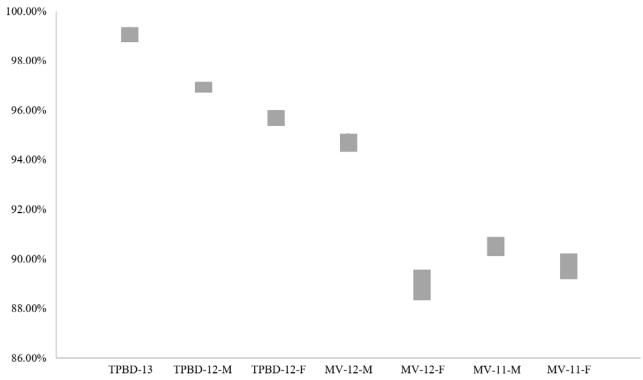


Figure 4. 95% CI of response level

5.5 The Benefit of Stochastic Programming

We evaluate the benefit of stochastic programming by comparing the performance of the mean value solution with the stochastic solution under the same sample. The mean value solution was obtained by taking the mean value of the sample into the deterministic model [M1]. For a sample with 20 scenarios, we first ignored the date of the emergency calls and summed up all the emergency calls that happened at the same demand area and during the same time segment. There were 24 time segments, each representing one hour of the 24 hours. Therefore, we obtained demand distribution over 22 demand areas × 24 time segments. Then we divided the distribution by 20, i.e., the number of scenarios, and obtained the average number of emergency calls occurred at each demand area during each time segment, which is the mean value of a sample. We put the average data into the deterministic model [M1] and obtained the first-stage solution. Then we solved the second-stage problem by fixing the first-stage solution in TPBD under the same sample, the result of which was compared with that of the TPBD using the same scenarios. The results are shown in the Table 7. The first and second row are performance of mean value solution and stochastic solution, respectively. Improvement indicates the percentage cost saving and coverage improvement of stochastic model. The second column is the total cost. FC, PC, TRC, and Penalty represent fixed cost of opening stations, purchasing cost of ambulances, transportation cost of demand fulfillment, and penalty for overtime, respectively. CL is the coverage level, i.e., the percentage of demand covered by opening stations within response time standard. RL is the response level, i.e., the percentage of demand being served within response time standard. DL is the demand loss rate, i.e., the percentage of demand that is unserved. The last column is the computation time.

The results show that stochastic programming could reduce the total cost by 7.93%, which is achieved by opening more facilities and purchasing more vehicles to reduce the demand fulfillment cost and delay penalty. As there are more facilities and vehicles, the coverage level and response level are improved by 4.67% and 5.25%, respectively. Therefore, stochastic programming could achieve a better coverage with less cost compared with deterministic one.

Table 7. The performance of the mean value solution and the stochastic solution under the same sample

	TC	FC	PC	TRC	Penalty	CL	RL	DL	Time
MV	192900	31500	19500	120400	21500	94.30%	88.98%	0.00%	15.26
TPBD	177600	36000	24000	110800	6800	98.70%	93.65%	0.00%	280
Improvement	7.93%	-14.29%	-23.08%	7.97%	68.37%	-4.67%	-5.25%	0.00%	-1736.83%

5.6 Sensitive analysis

We conducted the sensitive analysis to evaluate the influence of the value of crucial parameters on the optimal value of the stochastic programming. The following parameters are considered: response time standard, facility capacity, service capacity, and facility heterogeneity.

5.6.1 Impact of response time standard

Table 8 reports the effect of response time standard on cost, the number of facilities opened, the number of vehicles purchased, coverage level, and response level. RT means the value of response time standard. We can see that with the release of response time standard, total cost will reduce to certain value and then stay unchanged. Because when response time standard is relaxed, more demands can be served within time requirement, greatly reducing penalty. At the same time, with the enlarge of coverage radius, stations and vehicles can cover further demands without bearing penalty, reducing the number of stations and vehicles needed and thus reducing set-up and purchasing cost. When response time standard is raised to certain level, demands are fully covered and can be responded without any penalty under the required service level. The best combination of costs components is achieved so that further increasing

response time standard will not influence the system performance.

The results suggest that increasing response time standard under threshold can reduce total cost but does not have any effect when response time standard is already beyond the threshold. Besides, response time standard highly determines emergency survivability. When authorities set up response time standard, it is important to make a balance between cost and survivability.

Table 8. The influence of response time standard

RT	Total Cost	Fixed Cost	Purchase Cost	Transportation Cost	Penalty	Coverage Level	Response Level
1	644500	45000	27600	115500	456400	1.85%	1.71%
3	424300	45000	27600	115500	236200	34.48%	24.38%
5	281100	45000	27600	115600	92900	73.24%	57.34%
7	214600	45000	27000	116300	26300	90.43%	81.96%
9	188400	36000	24900	120700	6800	98.79%	94.29%
11	178400	27000	22500	127100	1800	100.00%	98.12%
13	174000	18000	20100	135900	0	100.00%	100.00%
15	174000	18000	20100	135900	0	100.00%	100.00%
17	174000	18000	20100	135900	0	100.00%	100.00%

5.6.2 Impact of facility capacity and vehicle service capacity

The impact of facility capacity is reflected by Table 9. The first and second column are the capacity of stations operated by FEM and VHE, respectively. We assume that fixed cost of station does not change with capacity. We can see that increasing the capacity of stations operated by FEM does not change the model performance while increasing the capacity of stations operated by VHE first decreases total cost and when the capacity increases to 20 vehicles, the optimal value does not change anymore. Because current station capacity of FEM is enough to achieve the required service level, making increased capacity redundant. However, increasing station capacity of VHE, which is originally set to a low value, allows more demands to be served by closer vehicles, and therefore reducing the number of stations, transportation cost, and penalty. It is worth noting that even though the number of stations is reduced, the coverage level stays the same and what's more, response level is improved, which means that restricting station capacity of VHE will waste money building redundant stations and

763 forcing demands to be served by further vehicles.

764 Results in Table 9 suggest that there is a threshold for station capacity where when
765 the value is below the threshold, increasing capacity can reduce total cost because more
766 vehicles can be used to serve demands, thus reducing station set-up cost, transportation
767 cost, and penalty, while when the value is above the threshold, increasing capacity does
768 not influence the performance of the system because existing vehicles can already
769 achieve the required service level at the minimal cost and additional capacity dose not
770 contribute to better service, thus being redundant.

771

772 Table 9. The impact of facility capacity on optimal value

CM	CV	TC	FC	NM	NV	PC	VM	VV	TRC	Penalty	CL	RL
10	5	182400	36000	6	6	24600	52	30	115800	6000	98.53%	94.13%
15	5	182400	36000	6	6	24600	52	30	115800	6000	98.53%	94.13%
20	5	182400	36000	6	6	24600	52	30	115800	6000	98.53%	94.13%
25	5	182400	36000	6	6	24600	52	30	115800	6000	98.53%	94.13%
30	5	182400	36000	6	6	24600	52	30	115800	6000	98.53%	94.13%
10	5	182400	36000	6	6	24600	52	30	115800	6000	98.53%	94.13%
10	10	165800	27000	6	4	25800	47	39	109600	3400	98.53%	96.31%
10	15	160500	27000	6	4	28500	43	52	104100	900	98.53%	98.11%
10	20	160400	27000	6	4	28800	43	53	103900	700	98.53%	98.16%
10	25	160400	27000	6	4	28800	43	53	103900	700	98.53%	98.16%
10	30	160400	27000	6	4	28800	43	53	103900	700	98.53%	98.16%

773

774 Table 10 shows impact of service capacity, where the first and second column are
775 the vehicle service capacity of stations operated by FEM and VHE, respectively. Table
776 10 has the same results as Table 9, because increasing vehicle service capacity under
777 fixed facility capacity has a similar effect to increasing facility capacity under fixed
778 vehicle service capacity, both of which increase demands served by each station.

779 Results of Table 9 and 10 inform us that there are thresholds for facility and vehicle
780 service capacity. Setting both capacities to their thresholds could achieve maximal
781 coverage with minimal cost.

782 Table 10. The impact of service capacity on optimal value

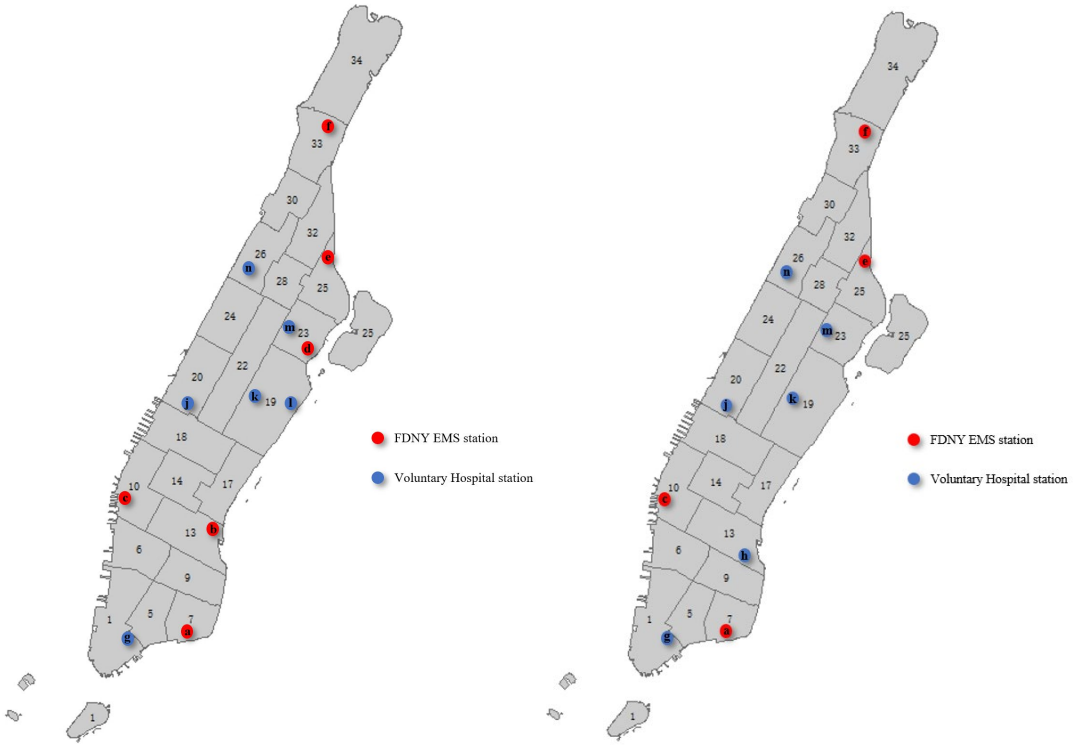
SM	SV	TC	FC	NM	NV	PC	VM	VV	TRC	Penalty	CL	RL

20	5	180900	36000	6	6	24300	51	30	114400	6200	98.23%	93.88%
20	10	160800	31500	6	5	21300	46	25	105400	2600	98.23%	96.74%
20	15	153200	31500	6	5	19800	41	25	98400	3500	98.23%	97.23%
20	20	153200	31500	6	5	19800	41	25	98400	3500	98.23%	97.23%
20	25	153200	31500	6	5	19800	41	25	98400	3500	98.23%	97.23%
20	5	180900	36000	6	6	24300	51	30	114400	6200	98.23%	93.88%
25	5	180900	36000	6	6	24300	51	30	114400	6200	98.23%	93.88%
30	5	180900	36000	6	6	24300	51	30	114400	6200	98.23%	93.88%
35	5	180900	36000	6	6	24300	51	30	114400	6200	98.23%	93.88%
40	5	180900	36000	6	6	24300	51	30	114400	6200	98.23%	93.88%

5.6.3 Impact of facility heterogeneity

To reflect the fact that stations of FEM control 70% of the ambulances in the New York City 911 System and serve 63% of ambulance tours while stations of VHE control the rest of the ambulances and serve 37% of the ambulance tours, we differentiated these two types of stations in terms of cost and capacity. In this section, we will compare results of problem setting with facility heterogeneity to those of problem setting where all facilities are assumed homogenous in term of cost and capacity.

We ran 10 times of each problem setting. The optimal location solutions are the same for 10 iterations, which are shown in Figure 5. Figure 5(a) and 5(b) show optimal location solutions for heterogenous and homogenous facilities, respectively. We can see that when heterogeneity is considered, all the 6 stations operated by FEM, i.e., the red circles, are selected for the following reasons: first, these stations are cheaper than the rest of the stations; second, they can serve more demands; third, they are sparsely distributed throughout the whole region. When the 6 stations cannot meet the requirements, stations in the place not covered by FEM or with high-density population, i.e., stations *g*, *j*, *k*, *l*, *m*, and *n*, are selected to fill the vacancy. This suggests that with heterogeneity, municipal stations are the best choice to serve emergency calls and when demands exceed their service capacity, stations that can fill the vacancy are preferred.



(a) Heterogenous facilities (b) Homogenous facilities

Figure 5. Comparison of optimal location solutions for heterogenous and homogenous facilities

If we assume that all facilities are the same, results will change. The number of selected stations of FEM and VHE are 4 and 6, respectively. Stations *b*, *d*, and *l* are replaced by station *h*. The coverage level and response level are shown in Table 11, which indicate better performances of location solutions for homogenous facilities

Table 11. The coverage level and response level of optimal location solutions for heterogenous and homogenous facilities

	CL	RL
Homogenous	98.72%	98.13%
Heterogenous	98.70%	93.65%

Results of comparison indicate that even though stations of FEM are preferred by New York City 911 System for various reasons, increasing utilization of VHE stations will improve system performances.

6. Conclusion

In this research, we propose a dynamic scenario-based two-stage stochastic programming model to optimize location of ambulance stations, deployment of ambulances, and vehicle dispatching under uncertain demand and traffic situation. The objective is to minimize total cost composed of station set-up cost, vehicle purchasing cost, demand fulfillment cost, penalty for overtime under service level requirements, and penalty for demand unsatisfaction. We apply Sample Average Approximation to approximate the original problem and propose a two-phase Benders Decomposition algorithm to accelerate computation. Numerical experiments are conducted to evaluate the performance of the solution method. Results show that proposed solution method is effective and efficient. Besides, it is proved to be robust in out-of-sample analysis. The comparison between stochastic model and deterministic model reveals that stochastic program could achieve a better coverage with less cost compared with deterministic one. We also conduct sensitive analysis to evaluate the influence of the value of crucial parameters on the optimal value of the stochastic programming model. Results suggest that there exist thresholds for response time standard, facility capacity, and vehicle service capacity. Increasing each of them under threshold can reduce total cost but does not have any effect when the threshold is reached. Response time standard highly determines emergency survivability. Authorities should make a balance between cost and survivability when setting up response time standard. For facility capacity and vehicle service capacity, it is recommended to set both to their thresholds in order to achieve maximal coverage with minimal cost. Facility heterogeneity will also influence problem solutions and performances. When considering heterogeneity, municipal stations are the best choice to serve emergency calls and when demands exceed their service capacity, stations that can fill the vacancy are preferred. When all facilities are homogeneous, better performances can be achieved. Therefore, even though stations of FEM are preferred by New York City 911 System for various reasons, increasing utilization of VHE stations will improve system performances.

Future research can be extended in four directions. First, emergency demands can be prioritized with the level of priority representing severity of emergency. We can require that high priority calls be served first. Second, we can use multi-type vehicles to serve demands, such basic life support ambulance and advanced life support ambulance. Each type of vehicle can only serve specified demands. Third, we allow vehicles not to

852 respond immediately to a demand because a more severe emergency may happen in the
853 near future, resulting in no vehicle response. Fourth, as vehicles may not be able to
854 serve all demands, we can consider demand queuing.

855 **References**

856 Aringhieri, R., Bruni, M. E., Khodaparasti, S., & van Essen, J. T. (2017). Emergency
857 medical services and beyond: Addressing new challenges through a wide
858 literature review. *Computers & Operations Research*, 78, 349-368.

859 Adulyasak, Y., Cordeau, J. F., & Jans, R. (2015). Benders decomposition for production
860 routing under demand uncertainty. *Operations Research*, 63(4), 851-867.

861 Bandara, D., Mayorga, M. E., & McLay, L. A. (2014). Priority dispatching strategies
862 for EMS systems. *Journal of the Operational Research Society*, 65(4), 572-587.

863 Bélanger, V., Lanzarone, E., Nicoletta, V., Ruiz, A., & Soriano, P. (2020). A recursive
864 simulation-optimization framework for the ambulance location and dispatching
865 problem. *European Journal of Operational Research*, 286(2), 713-725.

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

869 Benders, J. F. (1962). Partitioning procedures for solving mixed-variables programming
870 problems. *Numerische Mathematik*, 4(1), 238-252.

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

873 Beraldi, P., Bruni, M. E., & Conforti, D. (2004). Designing robust emergency medical
874 service via stochastic programming. *European Journal of Operational Research*,
875 158(1), 183-193.

876 Berman, O., Hajizadeh, I., & Krass, D. (2013). The maximum covering problem with
877 travel time uncertainty. *IIE Transactions*, 45(1), 81-96.

878 Bertsimas, D., & Ng, Y. (2019). Robust and stochastic formulations for ambulance
879 deployment and dispatch. *European Journal of Operational Research*, 279(2),
880 557-571.

881 Boujemaa, R., Jebali, A., Hammami, S., Ruiz, A., & Bouchriha, H. (2018). A stochastic
882 approach for designing two-tiered emergency medical service systems. *Flexible*
883 *Services and Manufacturing Journal*, 30(1), 123-152.

884 Boutilier, J. J., & Chan, T. C. (2020). Ambulance emergency response optimization in
885 developing countries. *Operations Research*, 68(5), 1315-1334.

886 Bürger, A., et al. (2018). The effect of ambulance response time on survival following
887 out-of-hospital cardiac arrest: an analysis from the German resuscitation

registry. *Deutsches Ärzteblatt International*, 115(33-34), 541.

Chanta, S., Mayorga, M. E., Kurz, M. E., & McLay, L. A. (2011). The minimum p-envy location problem: a new model for equitable distribution of emergency resources. *IIE Transactions on Healthcare Systems Engineering*, 1(2), 101-115.

Chanta, S., Mayorga, M. E., & McLay, L. A. (2014). Improving emergency service in rural areas: a bi-objective covering location model for EMS systems. *Annals of Operations Research*, 221(1), 133-159.

Church, R., & ReVelle, C. (1974). The maximal covering location problem. Paper Presented at the Papers of the Regional Science Association.

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

Daskin, M. S., & Stern, E. H. (1981). A hierarchical objective set covering model for emergency medical service vehicle deployment. *Transportation Science*, 15(2), 137-152.

Dean, S. F. (2008). Why the closest ambulance cannot be dispatched in an urban emergency medical services system. *Prehospital and Disaster Medicine*, 23(2), 161-165.

Degel, D., Wiesche, L., Rachuba, S., & Werners, B. (2015). Time-dependent ambulance allocation considering data-driven empirically required coverage. *Health Care Management Science*, 18(4), 444-458.

El Itani, B., Abdelaziz, F. B., & Masri, H. (2019). A bi-objective covering location problem: case of ambulance location in the Beirut area, Lebanon. *Management Decision*, 57(2), 432-444.

Erkut, E., Ingolfsson, A., & Erdoğan, G. (2008). Ambulance location for maximum survival. *Naval Research Logistics*, 55(1), 42-58.

Fire Department of New York City. EMS Incident Dispatch Data. https://data.cityofnewyork.us/Public-Safety/EMS-Incident-Dispatch-Data/76xm-jjuj.

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

Goldberg, J., & Paz, L. (1991). Locating emergency vehicle bases when service time depends on call location. *Transportation Science*, 25(4), 264-280.

Haghani, A., Tian, Q., & Hu, H. (2004). Simulation model for real-time emergency

vehicle dispatching and routing. *Transportation Research Record*, 1882(1), 176-183.

Ingolfsson, A., Budge, S., & Erkut, E. (2008). Optimal ambulance location with random delays and travel times. *Health Care Management Science*, 11(3), 262-274.

Jenkins, P. R., Robbins, M. J., & Lunday, B. J. (2021). Approximate dynamic programming for military medical evacuation dispatching policies. *INFORMS Journal on Computing*, 33(1), 2-26.

Knight, V. A., Harper, P. R., & Smith, L. (2012). Ambulance allocation for maximal survival with heterogeneous outcome measures. *Omega*, 40(6), 918-926.

Kuo, Y. H. (2014). Integrating simulation with simulated annealing for scheduling physicians in an understaffed emergency department. *HKIE Transactions*, 21(4), 253-261.

Kuo, Y. H., Rado, O., Lupia, B., Leung, J. M., & Graham, C. A. (2016). Improving the efficiency of a hospital emergency department: a simulation study with indirectly imputed service-time distributions. *Flexible Services and Manufacturing Journal*, 28(1), 120-147.

Laporte, G., & Louveaux, F. V. (1993). The integer L-shaped method for stochastic integer programs with complete recourse. *Operations Research Letters*, 13(3), 133-142.

Lee, S. (2011). The role of preparedness in ambulance dispatching. *Journal of the Operational Research Society*, 62(10), 1888-1897.

Lee, S. (2012). The role of centrality in ambulance dispatching. *Decision Support Systems*, 54(1), 282-291.

Lee, S. (2013). Centrality-based ambulance dispatching for demanding emergency situations. *Journal of the Operational Research Society*, 64(4), 611-618.

Liu, K., Li, Q., & Zhang, Z. H. (2019). Distributionally robust optimization of an emergency medical service station location and sizing problem with joint chance constraints. *Transportation Research Part B: Methodological*, 119, 79-101.

Liu, Y., Li, Z., Liu, J., & Patel, H. (2016). A double standard model for allocating limited emergency medical service vehicle resources ensuring service reliability. *Transportation Research Part C: Emerging Technologies*, 69, 120-133.

McLay, L. A., & Mayorga, M. E. (2013a). A model for optimally dispatching

954 ambulances to emergency calls with classification errors in patient priorities.
955 *IIE Transactions*, 45(1), 1-24.
956 McLay, L. A., & Mayorga, M. E. (2013b). A dispatching model for server-to-customer
957 systems that balances efficiency and equity. *Manufacturing & Service*
958 *Operations Management*, 15(2), 205-220.
959 Nasrollahzadeh, A. A., Khademi, A., & Mayorga, M. E. (2018). Real-time ambulance
960 dispatching and relocation. *Manufacturing & Service Operations Management*,
961 20(3), 467-480.
962 Nelas, J., & Dias, J. (2020). Optimal emergency vehicles location: An approach
963 considering the hierarchy and substitutability of resources. *European Journal of*
964 *Operational Research*, 287(2), 583-599.
965 New York City Fire Department Bureau of Emergency Medical Services.
966 https://emspac.org/sectors/municipal/.
967 Nickel, S., Reuter-Oppermann, M., & Saldanha-da-Gama, F. (2016). Ambulance
968 location under stochastic demand: A sampling approach. *Operations Research*
969 *for Health Care*, 8, 24-32.
970 Park, H., Waddell, D., & Haghani, A. (2019). Online optimization with look-ahead for
971 freeway emergency vehicle dispatching considering availability. *Transportation*
972 *Research Part C: Emerging Technologies*, 109, 95-116.
973 Peng, C., Delage, E., & Li, J. (2020). Probabilistic envelope constrained multiperiod
974 stochastic emergency medical services location model and decomposition
975 scheme. *Transportation Science*, 54(6), 1471-1494.
976 ReVelle, C., & Hogan, K. (1988). A reliability-constrained siting model with local
977 estimates of busy fractions. *Environment and Planning B: Planning and Design*,
978 15(2), 143-152.
979 ReVelle, C., & Hogan, K. (1989a). The maximum availability location problem.
980 *Transportation Science*, 23(3), 192-200.
981 ReVelle, C., & Hogan, K. (1989b). The maximum reliability location problem and α-
982 reliable p-center problem: derivatives of the probabilistic location set covering
983 problem. *Annals of Operations Research*, 18(1), 155-173.
984 Schilling, D., Elzinga, D. J., Cohon, J., Church, R., & ReVelle, C. (1979). The
985 TEAM/FLEET models for simultaneous facility and equipment siting.
986 *Transportation Science*, 13(2), 163-175.

Schmid, V., & Doerner, K. F. (2010). Ambulance location and relocation problems with time-dependent travel times. *European Journal of Operational Research*, 207(3), 1293-1303.

Sorensen, P., & Church, R. (2010). Integrating expected coverage and local reliability for emergency medical services location problems. *Socio-Economic Planning Sciences*, 44(1), 8-18.

Sudtachat, K., Mayorga, M. E., & McLay, L. A. (2014). Recommendations for dispatching emergency vehicles under multitiered response via simulation. *International Transactions in Operational Research*, 21(4), 581-617.

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

Toro-Díaz, H., Mayorga, M. E., Chanta, S., & McLay, L. A. (2013). Joint location and dispatching decisions for emergency medical services. *Computers & Industrial Engineering*, 64(4), 917-928.

van den Berg, P. L., & Aardal, K. (2015). Time-dependent MEXCLP with start-up and relocation cost. *European Journal of Operational Research*, 242(2), 383-389.

Vos, T., et al. (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, 396(10258), 1204-1222.

Yellow Taxi Trip Data (2011). https://data.cityofnewyork.us/Transportation/2011-Yellow-Taxi-Trip-Data/jr6k-xwua.

Yoon, S., Albert, L. A., & White, V. M. (2021). A stochastic programming approach for locating and dispatching two types of ambulances. *Transportation Science*, 55(2), 275-296.

Zarkeshzadeh, M., Zare, H., Heshmati, Z., & Teimouri, M. (2016). A novel hybrid method for improving ambulance dispatching response time through a simulation study. *Simulation Modelling Practice and Theory*, 60, 170-184.