EMS LOCATION-ALLOCATION PROBLEM UNDER UNCERTAINTIES

WANG WEI

MPhil

The Hong Kong Polytechnic University

2022

The Hong Kong Polytechnic University

Department of Logistics and Maritime Studies

EMS location-allocation problem under uncertainties

Wang Wei

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Philosophy

April 2022

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

________________________(Signed)

_______WANG Wei_______(Name of student)

Abstract

Emergency especially routine emergency that happens on a daily basis poses great threat to our health, life, and property. Immediate response and treatment can greatly mitigate these threats. This research optimizes location of ambulance stations, deployment of ambulances, and vehicle dispatching under demand and traffic uncertainty, which are the main factors that influence response time. The problem is formulated as a dynamic scenario-based two-stage stochastic programming model with the aim of minimizing total cost under service level requirements and is solved by Sample Average Approximation. Finally, we conduct numerical experiments using real-world emergency data to evaluate the performance of our methods, which yields valuable managerial insights for the design of EMS response system.

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

Acknowledgements

During the entire MPhil period, I would like to express my most sincere thanks to my supervisor Prof. Shuaian (Hans) WANG. The first time I knew the professor's name was when I was preparing to apply for PhD, I read many articles published by professor in journals that are relevant to my research field. I was so interested in many of these articles that I plucked up courage to submit my PhD application to the professor. During the communication, professor showed good personality and great patience, which surprised and warmed me, because as a world-renowned researcher, it is admirable to be kind to an ordinary student like me. Thanks to professor's great help, I finally got the offer. After becoming a student, I learnt a lot from Prof. Wang, which reshaped my attitude towards academics, teaching, and life. As a researcher, professor works hard, has critical thinking, has rigorous academic attitude, and is open-minded. He started to work early in the morning. He was still thinking and absorbing new knowledge in his spare time, such as walking and waiting for public transport. He encouraged us to raise questions and to be critical to articles, news, and even those that are published in top platforms. When he revised our articles, he even noticed details that others would easily overlook. He welcomed opinions and ideas that differ from his own. As a teacher, professor is patient, conscientious, careful, and kind. He was very careful and dedicated to prepare teaching materials. He updated materials according to students' feedback and new trends in the field. He responded to students' questions quickly and detailly. This is just to name a few. All good qualities I learn from professor will help me not only in daily academic study but also in preparation to be a teacher in the future. Besides, Prof. Wang provided me precious opportunities to participate in various conferences and seminars and gave me valuable suggestions in my academic study. He demonstrates noble personality and great capability. He is a good role model whom I can learn from in my whole life.

Then I would express my heartful thanks to my parents and two elder cousins. My parents gave me unconditional love and meticulous care when I was growing up. They always supported my decisions, even if I decided to quit my job and pursue PhD. After quitting the job, I spent one year preparing for language tests and application materials. My parents took care of me well and encouraged me during this period. When I was in Hong Kong, they often sent me messages of concern. Because of supporting and love from my parents, I am who I am. I am very grateful to them. However, I feel very sorry for them because I cannot accompany them

as they do to me and I cannot come back to them when they are ill. Gratefully, my two elder cousins helped me to take care of my parents and help me out when I am in trouble. Without them, I might not have been able to move on during those difficult times.

Third, I am thankful to my friends. I stayed with friends in the same office nearly every day during one and a half years of MPhil study. We had lunch together, chatted with each other, and tackled problems together. Thank you for your sincere help and company. The friends who live outside Hong Kong always comforted me when I was down. Thank you for listening to me and comforting me.

Finally, I am grateful to the academic staff of PolyU for their excellent and useful courses and the administrative staff for their help in all aspects of my life at PolyU.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

- Confidence intervals (CIs)
- Double coverage model (DCM)
- Emergency medical services (EMS)
- Fire Department of New York City (FDNY)
- FDNY EMS municipal (FEM)
- Location set covering problem (LSCP)
- Maximum availability location problem (MALP)
- Maximal covering location problem (MCLP)
- New York City Fire Department Bureau of Emergency Medical Services (FDNY EMS)
- Sample Average Approximation (SAA)
- Voluntary Hospital EMS (VHE)

Chapter 1: Introduction

1.1 BACKGROUND

Emergency is a kind of situation that can happen anywhere and anytime, posing great risks to people's health, life, and properties. Routine emergency, such as heart attack, road accident, and residential fire, is a type of emergency that happens on a daily basis and has a small scope of influence. According to the research by Vos et al. (2020), this type of emergency is one of the leading causes of death globally. Therefore, immediate response and treatment to such emergency are of great importance. The treatment can be divided into pre-hospital treatment and in-hospital treatment. The former usually involves emergency medical services (EMS) department while the latter highly depends on process design and operations management of patient flows, which is not considered in this research. For readers who are interested in efficiency and effectiveness of in-hospital treatment, please refer to Kuo (2014) and Kuo et al. (2016). For pre-hospital treatment, when an emergency call comes, the receptionist determines the severity of the situation and then dispatches responding vehicles accordingly. Considering that survival rate highly depends on response time, i.e., the time interval between reception of call and arrival of vehicle at the emergency site, the dispatched vehicles must arrive at the emergency site as soon as possible or within certain time limits (Bürger et al., 2018; Erkut, Ingolfsson, & Erdoğan, 2008; Knight, Harper, & Smith, 2012). Response time is directly affected by location of ambulance stations, the number of available vehicles, and dispatching decisions. Therefore, optimizing factors that directly influence response time is essential to guarantee efficient and high-quality EMS services for the public.

The optimization problem mentioned above is called location-allocation problem. It is quite challenging, because it is usually formulated as a mixed-integer programming model with a large number of decision variables and constraints. When uncertainty is incorporated into problem, it becomes more complicated. This research addresses location-allocation problem with system congestion under demand and traffic uncertainty. System congestion refers to a situation where there are not enough vehicles available to respond to demand, which in this article is captured by vehicle availability, i.e., the number of vehicles that could respond to demand. It is influenced by vehicles dispatched to demands that overlap in time with current demand. The uncertainty in demand includes the number of emergencies, location, occurrence time, and the number of ambulances and the service time needed, which directly influence optimization decisions. Service time is the time interval between the arrival of a vehicle at the emergency site and its return to 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 location of ambulance stations and the deployment of ambulances without considering the realization of uncertainties. In the second stage, recourse decisions on vehicle dispatching are made based on realization of uncertainties, first-stage decisions, and state of available vehicles. The objective of the model is to achieve the required service level at minimal cost, which is comprised of station set-up cost, vehicle purchasing cost, demand fulfillment cost, and the penalty for failing to respond to the demand within required time standard. The problem is solved by Sample Average Approximation (SAA) where a balance between precision and computational tractability has to be achieved. To evaluate the performance of the method, we conduct numerical experiments using real-world emergency data. Through comparison between stochastic and deterministic method and results of sensitive analysis, we obtain several valuable managerial insights.

1.2 THESIS OUTLINE

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

Chapter 2: Literature Review[1]

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 chapter, 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

[1] Wang, W., Wu, S., Wang, S., Zhen, L., & Qu, X. (2021). Emergency facility location problems in logistics: Status and perspectives. Transportation research part E: logistics and transportation review, 154, 102465.

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 dispatching belong to different decision levels, the combination of these two-level problems is computationally intensive, resulting in not much hybrid research. This article combines location, dynamic real-time dispatching, and fleet deployment, which is rarely explored in combination in existing literature. Location and fleet deployment problem are usually formulated as a mixed-integer programming model, while dispatching problem is usually solved by queuing theory considering dispatching policies and preference lists. However, this research innovatively uses a mixed-integer programming model to formulate the hybrid problem. When selecting vehicles to be dispatched, we further incorporate system congestion into the model. The most commonly used methods to model system congestion in literature are busy fraction and queuing theory. Busy fraction is generally assumed to be fixed, independent and exogeneous, which cannot reflect the dynamic and endogenous characteristics of system congestion. When dispatching is modeled by Markov decision process, it usually adopts certain assumptions (e.g., assumption of arrival and service process) and dispatching policies (e.g., nearest available, preparedness, or centrality policy). Each assumption and policy have their suitable applications. If they are applied in an inappropriate problem setting, the results may be suboptimal. Therefore, in this research, we do not use these two methods. Instead, system congestion is represented by functions of a parameter indicating the overlap between demands, deployment decision, and previous dispatching decisions, while dispatching policy is not pre-defined but determined by objective function. Besides, we incorporate both uncertain demand and uncertain travel time into the model, whereas most literature considers at most one of them. Demand and travel time uncertainties are usually represented by scenarios and time-dependent travel time, respectively. We adopt the scenario method for demand uncertainty as the accurate demand information in each scenario is helpful to model vehicle availability, while for travel time uncertainty, we combine scenario and multi-period methods. We first generate scenarios, each of which represents one day traffic information. Then we divide each scenario into 24 equal segments, each of which is one hour, and calculate segment-dependent travel time. The method to deal with travel time uncertainty is helpful to calculate actual coverage set of each demand, which is essential in calculating total cost.

Chapter 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 Chapter 3.1 to make it easier to understand the reasoning and logic behind the model. Then in Chapter 3.2, a scenario-based two-stage stochastic model is proposed to deal with uncertainties.

3.1 DETERMINISTIC MODEL

In this chapter, 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 3-1, 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 3-1. 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.

$$\delta_{ii'j} = \begin{cases} 0, \text{ if } t_i \geq t_{i'} + T_{i'j} + l_{i'} \\ 1, \text{ otherwise} \end{cases}, \forall i \in I, j \in J, i' \in \{1, 2, \dots, i-1\}. \tag{3.1}$$

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.

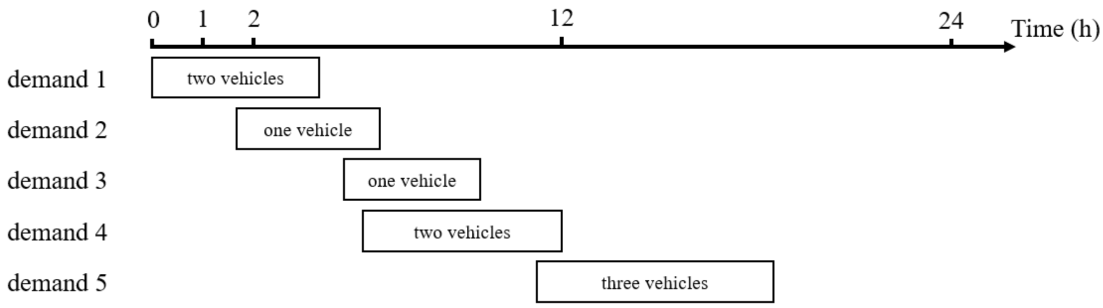


Figure 3-1: One example of demand in one day period

Table 3-1: One example of vehicle allocation

	Station 1		Station 2	
	Available	Allocated	Available	Allocated
d_1	2	2	2	0
d_2	0	0	2	1
d_3	2	1	1	0
d_4	1	1	1	1
d_5	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 cost while maintaining the required service level. All the notations for the deterministic model are introduced in Table 3-2 and the model is given as follows:

Table 3-2: 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
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
α	The minimum service level ($\alpha \in [0,1]$)
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]

$$\text{Min} \sum_{j\in J} f_j x_j + \sum_{j\in J} h y_j + \sum_{i\in I}\sum_{j\in J} c_{ij}T_{ij}z_{ij} + \sum_{i\in I}\sum_{j\in J\setminus N_i} \gamma(T_{ij}-R)z_{ij} \quad (3.2)$$

subject to

$$\sum_{j\in J} z_{ij} \leq d_i, \forall i \in I \quad (3.3)$$

$$\sum_{i\in I}\sum_{j\in J} z_{ij} \geq \alpha \sum_{i\in I} d_i \quad (3.4)$$

$$z_{ij} \leq y_j - \sum_{i'=1}^{i-1} \delta_{ii'j} z_{i'j}, \forall i \in I, j \in J \quad (3.5)$$

$$\sum_{i\in I} z_{ij} \leq \mu_j y_j, \forall j \in J \quad (3.6)$$

$$y_j \leq Q_j x_j, \forall j \in J \quad (3.7)$$

$x_j \in \{0,1\}, \forall j \in J$ (3.8)

$y_j \in \mathbb{Z}_0^+, \forall j \in J$ (3.9)

$\sigma_i \in \{0,1\}, \forall i \in I$ (3.10)

$z_{ij} \in \mathbb{Z}_0^+, \forall i \in I, j \in J.$ (3.11)

The objective function (3.2) minimizes total cost, which consists of station set-up cost, vehicle purchasing cost, demand fulfillment cost, and the penalty for not responding the demand in required time standard. For the last term, we only account for the vehicles located outside the coverage set because only these vehicles cannot respond to demand within time standard. Constraints (3.3) state that demand can be partially satisfied. Constraint (3.4) ensures a minimum service level, which sets a lower bound for the demands to be served. Constraints (3.5) 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 (3.1). 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 (3.6) state that the demands served by each station cannot exceed the workload of the vehicle at this station. Constraints (3.7) require that vehicles can only be located at open station and the number cannot exceed the station capacity. Constraints (3.8) to (3.11) 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 chapter 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:

$$\delta_{ii'j}^{s} = \begin{cases} 0, & \text{if } t_i^s \geq t_{i'}^s + T_{i'j}^s + l_{i'}^s \\ 1, & \text{otherwise} \end{cases}, \forall i \in I, j \in J, i' \in \{1, 2, \ldots, i-1\}, s \in S. \quad (3.12)$$

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

Table 3-3: 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

$$\text{Min} \sum_{j \in J} f_j x_j + \sum_{j \in J} h y_j + \sum_{s \in S} p_s \left(\sum_{i \in I^s} \sum_{j \in J} c_{ij} T_{ij}^s z_{ij}^s + \sum_{i \in I^s} \sum_{j \in J \setminus N_i^s} \gamma (T_{ij}^s - R) z_{ij}^s \right) \tag{3.13}$$

subject to

$$\sum_{j \in J} z_{ij}^s \le d_i^s, \forall i \in I^s, s \in S \tag{3.14}$$

$$\sum_{i \in I^s} \sum_{j \in J} z_{ij}^s \ge \alpha \sum_{i \in I^s} d_i^s, \forall s \in S \tag{3.15}$$

$$z_{ij}^s \le y_j - \sum_{i'=1}^{i-1} \delta_{ii'j}^s z_{i'j}^s, \forall i \in I^s, j \in J, s \in S \tag{3.16}$$

$$\sum_{i \in I^s} z_{ij}^s \le \mu_j y_j, \forall j \in J, s \in S \tag{3.17}$$

$$y_j \le Q_j x_j, \forall j \in J \tag{3.18}$$

$$x_j \in \{0,1\}, \forall j \in J \tag{3.19}$$

$$y_j \in \mathbb{Z}_0^+, \forall j \in J \tag{3.20}$$

$$\sigma_i^s \in \{0,1\}, \forall i \in I^s, , s \in S \tag{3.21}$$

$$z_{ij}^s \in \mathbb{Z}_0^+, \forall i \in I^s, j \in J, s \in S. \tag{3.22}$$

Chapter 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. To solve this challenge, we use SAA, which 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:

$$\text{Min} \sum_{j \in J} f_j x_j + \sum_{j \in J} h y_j + \frac{1}{|S'|}\left(\sum_{s \in S'} \sum_{i \in I^s} \sum_{j \in J} c_{ij} T_{ij}^s z_{ij}^s + \sum_{s \in S'} \sum_{i \in I^s} \sum_{j \in J \setminus N_i^s} \gamma\, (T_{ij}^s - R) z_{ij}^s\right) \tag{4.1}$$

subject to (3.14) – (3.22) 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 Chapter 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 problem 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 the SAA problem 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. The resulting cost is $v^m_{x^*,y^*}$, $v^m_{x^*,y^*} = \min \sum_{j\in J} f_j x^*_j + \sum_{j\in J} h y^*_j + \frac{1}{|S'_m|}\left(\sum_{s\in S'_m}\sum_{i\in I^s}\sum_{j\in J} c_{ij} T^s_{ij} z^s_{ij} + \sum_{s\in S'_m}\sum_{i\in I^s}\sum_{j\in J\setminus N^s_i} \gamma(T^s_{ij} - R) z^s_{ij}\right)$ subject to Eq. (14)–(22) in which S is replaced by S'_m.
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 $U := \frac{1}{M}\sum_{m=1}^{M} v^m_{x^*,y^*}$ and $S_U := \frac{1}{M-1}\sum_{m=1}^{M}(v^m_{x^*,y^*} - U)^2$.
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 $\left[0, G + \frac{t_{M-1,\tau}\sqrt{S_G}}{\sqrt{M}}\right]$.

Chapter 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. Then we evaluated the benefit of using stochastic programming approach over deterministic model. Next, we show the robustness of solution method. 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 Dell XPS 15 9500 laptop with i7-10750H CPU, 2.60 GHz processing speed and 16 GB of memory. The model and the algorithm were implemented in C++ programming and both SAA and deterministic model were 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 5-1, 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 5-1.

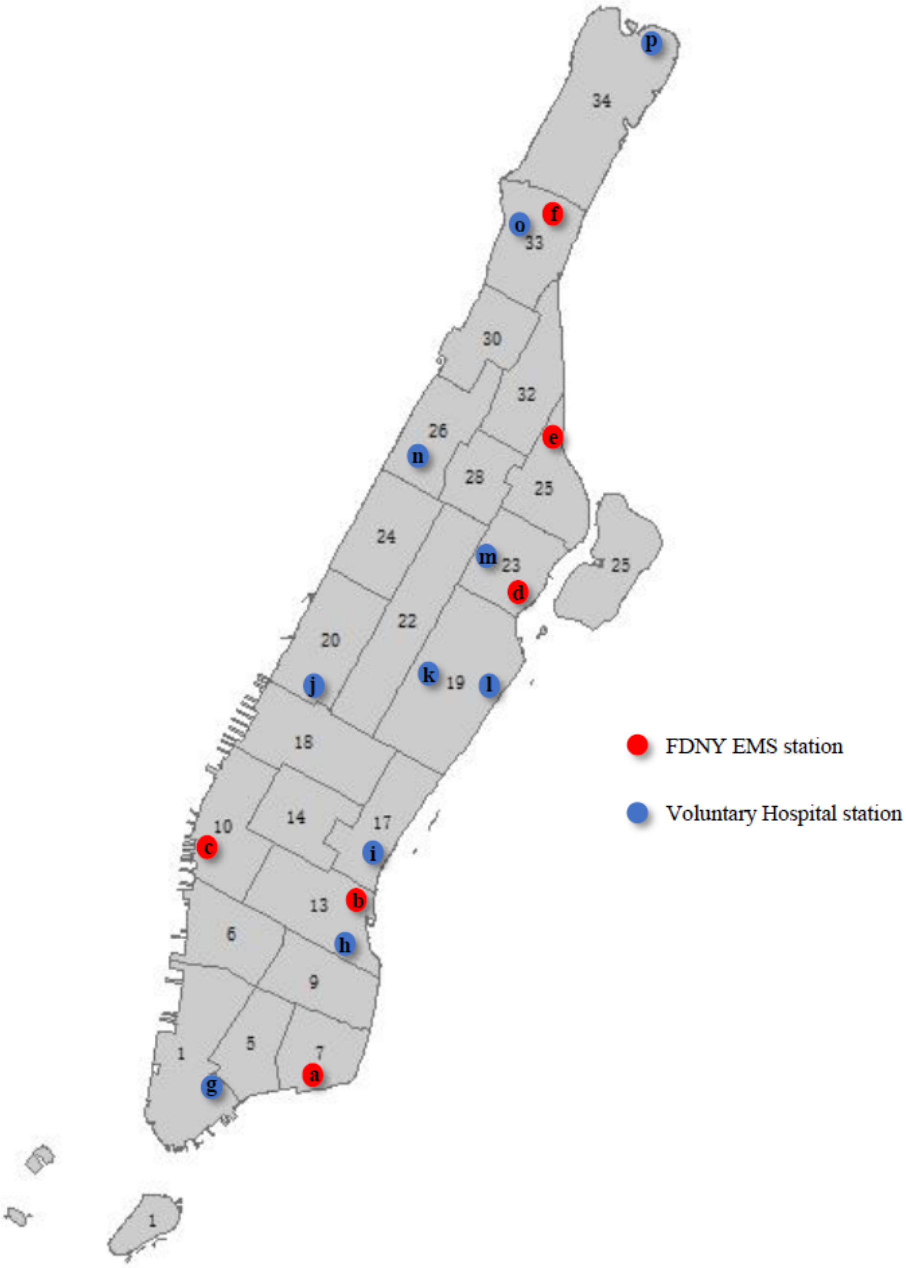


Figure 5-1: The demand area and candidate stations

Table 5-1: 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

5.2 DETERMINATION OF SAMPLE SIZE

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

5.2.1 SAMPLE GENERATION

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

5.2.2 SOLUTION QUALITY EVALUATION

We used Algorithm 1 to evaluate the performance of different sample sizes, which were set to 10, 20, 30, 40, and 50. We first ran line 1 and 2 of Algorithm 1 to obtain the optimal first-stage solution. Then we iterated line 4 to 8 for 10 times where $|S'_m|$ was set to 100. Finally, we calculated solution quality, which is shown in Table

5-2. 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 is 40, 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 40 in the following calculation.

Table 5-2: 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			
10	146000	(142350, 149650)	144700	(143500, 145900)	2157	(0, 4604)	3.15%	3.18%	3801
20	142700	(139700, 145700)	143900	(142850, 144950)	1898	(0, 3795)	2.66%	2.64%	11420
30	143500	(140750, 146250)	143800	(142800, 144800)	1428	(0, 2614)	1.82%	1.82%	22210
40	143300	(141050, 145550)	143800	(142950, 144650)	643	(0, 1370)	0.96%	0.95%	32800
50	143800	(142900, 144700)	143700	(143050, 144350)	501	(0, 1018)	0.71%	0.71%	43610

Note: NS represents sample size. PE represents point estimate. CI represents 95% confidence interval. CI-L represents ratio between 95% confidence interval for optimality gap and point estimate of lower bound. CI-U represents ratio between 95% confidence interval for optimality gap and point estimate of upper bound. Time represents running time.

5.3 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 40 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 40, 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 SAA under the same sample, the result of which was compared with that of the SAA using the same scenarios. The results are shown in the Table 5-3. 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. NM and NV represent the number of opening stations run by FEM and VHE, respectively. VM and VV represent the number of ambulances purchased in the opening stations operated by FEM and VHE, respectively. CL is the coverage level, i.e., the percentage of demand covered by opening stations. RL is the response level, i.e., the percentage of demand being served within response time standard. The last column is the computation time.

The results show that stochastic programming could reduce the total cost by 6.27%, 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 4.17% and 9.21%, respectively. Therefore, stochastic programming could achieve a better coverage with less cost compared with deterministic one.

Table 5-3: The performance of the mean value solution and the stochastic solution under the same sample

	TC	FC	NM	NV	PC	VM	VV	TRC	Penalty	CL	RL	Time
MV	156400	22500	6	3	15900	38	15	107300	10700	92.09%	81.78%	38.41
SAA	146600	27000	6	4	20100	47	20	98230	1270	95.93%	89.31%	2381.66
Improvement	6.27%	−20%	-	-	−26.42%	-	-	8.45%	88.13%	4.17%	9.21%	-

Note: TC means total cost. FC means fixed cost of opening stations. NM means the number of opening stations run by FEM. NV means the number of opening stations run by VHE. PC means purchasing cost of ambulances. VM means the number of ambulances purchased in the opening stations operated by FEM. VV means the number of ambulances purchased in the opening stations operated by VHE. TRC means transportation cost of demand fulfillment. Penalty means penalty for overtime. CL means the coverage level. RL means the response level. Time means computation time.

5.4 ROBUSTNESS EVALUATION

Optimal solutions obtained through SAA 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 SAA method, we conduct out-of-sample analysis, which evaluates performance of optimal solutions using out-of-sample data. The procedures are shown in Algorithm 2. 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 SAA 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 response level, i.e., the proportion of demand that is responded within response time standard. Robustness is measured by robustness level, the proportion of scenarios where optimal solution can be found, and $(1-\tau)$-CI of response level.

Algorithm 2 Robustness evaluation

1. Generate a set of scenarios S^{in}.
2. Solve the SAA 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 response level RL_n, i.e., the proportion of demand that is responded within response time standard.
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 $A := \frac{\sum_{n=1}^{N_0} RL_n}{N_0}$ and $S_A := \frac{1}{N_0 - 1}\sum_{n=1}^{N_0}(RL_n - A)^2$
11. The $(1-\tau)$-CI of response level is $\left[A - \frac{t_{N_0-1,\frac{\tau}{2}}\sqrt{S_A}}{\sqrt{N_0}}, A + \frac{t_{N_0-1,\frac{\tau}{2}}\sqrt{S_A}}{\sqrt{N_0}}\right]$.

We ran SAA 10 times. Only one of the optimal first-stage location solutions is different. The rest have the same location solution but different deployment solutions. Therefore, we selected two solutions with the fewest and most vehicles from the remaining iterations. Totally, we got three different optimal first-stage solutions. Then we conducted procedures described in Section 5.3 to get one optimal first-stage solution obtained from the mean value model. The four solutions were put into Algorithm 2 to calculate robustness level and 95% CI of response level where N was set to 150. The 95% CI of response level and robustness level for four solutions are shown in Figure 5-2 and Table 5-4, respectively.

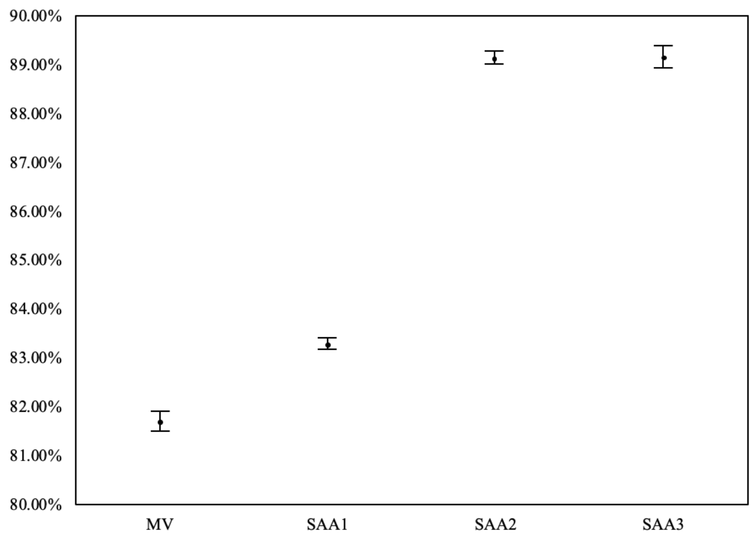


Figure 5-2: 95% CI of response level for one deterministic and three stochastic solutions

In Figure 5-2, MV and SAA1–3 represent optimal first-stage solutions for mean value model and stochastic models, respectively. The results show that in comparison with deterministic model, stochastic model can obtain solutions that respond more emergency calls within response time standard with high probability. Even though these solutions are tested in out-of-sample data, at least 83% of the emergency demands can be responded in time with high probability. Table 5-4 shows that solutions obtained by SAA can find optimal value under nearly every scenario in out-of-sample analysis. Therefore, we can conclude that results of SAA method are robust.

Table 5-4: Robustness level for one deterministic and three stochastic solutions

	Robustness level
SAA1	100.00%
SAA2	99.33%
SAA3	100.00%

5.5 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, service level, facility capacity, service capacity, and facility heterogeneity.

5.5.1 IMPACT OF RESPONSE TIME STANDARD

Table 5-5 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 5-5: The influence of response time standard

RT	TC	FC	NM	NV	PC	VM	VV	TRC	Penalty	CL	RL
1	493100	40500	6	7	24600	47	35	88730	339270	1.80%	1.68%
3	305900	40500	6	7	24600	47	35	88750	152050	33.21%	24.42%
5	194000	40500	6	7	24300	46	35	88920	40280	72.27%	58.13%
7	154900	36000	6	6	23100	47	30	90360	5440	89.09%	80.93%
9	143500	22500	6	3	19200	49	15	99340	2460	95.72%	86.97%
11	139100	18000	6	2	16800	46	10	104100	200	97.70%	89.58%
13	138900	18000	6	2	16500	45	10	104400	0	100.00%	90.07%
15	138900	18000	6	2	16500	45	10	104400	0	100.00%	90.07%
17	138900	18000	6	2	16500	45	10	104400	0	100.00%	90.07%

Note: RT means response time standard. TC means total cost. FC means fixed cost of opening stations. NM means the number of opening stations run by FEM. NV means the number of opening stations run by VHE. PC means purchasing cost of ambulances. VM means the number of ambulances purchased in the opening stations operated by FEM. VV means the number of ambulances purchased in the opening stations operated by VHE. TRC means transportation cost of demand fulfillment. Penalty means penalty for overtime. CL means the coverage level. RL means the response level.

5.5.2 IMPACT OF SERVICE LEVEL

Table 5-6 shows the influence of service level on optimal value. The first column is the value of service level. When service level is reduced, all cost components decrease. As less demands are required to be served, redundant stations and vehicles are no longer needed, saving a large amount of cost. At the same time, coverage level and response level decrease accordingly.

The results indicate a positive relationship between total cost and service level. If authorities want to provide emergency service with high service level, costs are bound to be high. Therefore, authorities must make a trade-off between cost and service level when making decisions.

Table 5-6: The influence of service level on optimal value

SL	TC	FC	NM	NV	PC	VM	VV	TRC	Penalty	CL	RL
0.9	139900	22500	6	3	18900	48	15	96460	2040	96.06%	87.31%
0.8	112000	13500	6	1	14700	44	5	83350	450	89.02%	79.17%
0.7	90280	9000	6	0	10500	35	0	70750	30	78.67%	69.96%
0.6	72440	9000	6	0	9600	32	0	53840	0	78.67%	60.05%
0.5	56970	9000	6	0	8100	27	0	39870	0	78.67%	50.03%
0.4	43890	9000	6	0	6000	20	0	28890	0	78.67%	40.05%
0.3	31610	6000	4	0	5100	17	0	20510	0	59.26%	30.06%
0.2	19850	4500	3	0	3300	11	0	12050	0	40.56%	20.04%
0.1	9248	3000	2	0	1800	6	0	4448	0	33.61%	10.05%

Note: SL means service level. TC means total cost. FC means fixed cost of opening stations. NM means the number of opening stations run by FEM. NV means the number of opening stations run by VHE. PC means purchasing cost of ambulances. VM means the number of ambulances purchased in the opening stations operated by FEM. VV means the number of ambulances purchased in the opening stations operated by VHE. TRC means transportation cost of demand fulfillment. Penalty means penalty for overtime. CL means the coverage level. RL means the response level.

5.5.3 IMPACT OF FACILITY CAPACITY AND VEHICLE SERVICE CAPACITY

The impact of facility capacity is reflected by Table 5-7. 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 15 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 forcing demands to be served by further vehicles.

Results in Table 5-7 suggest that there is a threshold for station capacity where when the value is below the threshold, increasing capacity can reduce total cost because more vehicles can be used to serve demands, thus reducing station set-up cost, transportation cost, and penalty, while when the value is above the threshold, increasing capacity does not influence the performance of the system because existing vehicles can already achieve the required service level at the minimal cost and additional capacity dose not contribute to better service, thus being redundant.

Table 5-7: The impact of facility capacity on optimal value

CM	CV	TC	FC	NM	NV	PC	VM	VV	TRC	Penalty	CL	RL
10	5	142600	27000	6	4	19800	46	20	94940	860	95.69%	88.57%
15	5	142600	27000	6	4	19800	46	20	94940	860	95.69%	88.57%
20	5	142600	27000	6	4	19800	46	20	94940	860	95.69%	88.57%
25	5	142600	27000	6	4	19800	46	20	94940	860	95.69%	88.57%
30	5	142600	27000	6	4	19800	46	20	94940	860	95.69%	88.57%
10	5	142600	27000	6	4	19800	46	20	94940	860	95.69%	88.57%
10	10	132500	22500	6	3	21000	41	29	88840	160	95.69%	89.76%
10	15	131500	22500	6	3	22800	39	37	86150	50	95.69%	89.99%
10	20	131500	22500	6	3	22800	39	37	86150	50	95.69%	89.99%
10	25	131500	22500	6	3	22800	39	37	86150	50	95.69%	89.99%
10	30	131500	22500	6	3	22800	39	37	86150	50	95.69%	89.99%

Note: CM means the capacity of stations operated by FEM. CV means the capacity of stations operated by VHE. TC means total cost. FC means fixed cost of opening stations. NM means the number of opening stations run by FEM. NV means the number of opening stations run by VHE. PC means purchasing cost of ambulances. VM means the number of ambulances purchased in the opening stations operated by FEM. VV means the number of ambulances purchased in the opening stations operated by VHE. TRC means transportation cost of demand fulfillment. Penalty means penalty for overtime. CL means the coverage level. RL means the response level.

Table 5-8 shows impact of service capacity, where the first and second column are the vehicle service capacity of stations operated by FEM and VHE, respectively. Table 5-8 has the same results as Table 5-7, because increasing vehicle service capacity under fixed facility capacity has a similar effect to increasing facility capacity under fixed vehicle service capacity, both of which increase demands served by each station.

Table 5-8: The impact of service capacity on optimal value

SM	SV	TC	FC	NM	NV	PC	VM	VV	TRC	Penalty	CL	RL
20	5	147000	27000	6	4	21600	52	20	96640	1760	96.02%	88.33%
20	10	132900	27000	6	4	19500	45	20	85680	720	96.02%	89.53%
20	15	127700	27000	6	4	18000	40	20	82320	380	96.02%	89.83%
20	20	126700	25500	5	4	18300	41	20	82560	340	96.02%	89.85%
20	25	126700	25500	5	4	18300	41	20	82560	340	96.02%	89.85%
20	5	147000	27000	6	4	21600	52	20	96640	1760	96.02%	88.33%
25	5	146700	27000	6	4	21000	50	20	96750	1950	96.02%	88.30%
30	5	146700	27000	6	4	21000	50	20	96750	1950	96.02%	88.30%
35	5	146700	27000	6	4	21000	50	20	96750	1950	96.02%	88.30%
40	5	146700	27000	6	4	21000	50	20	96750	1950	96.02%	88.30%

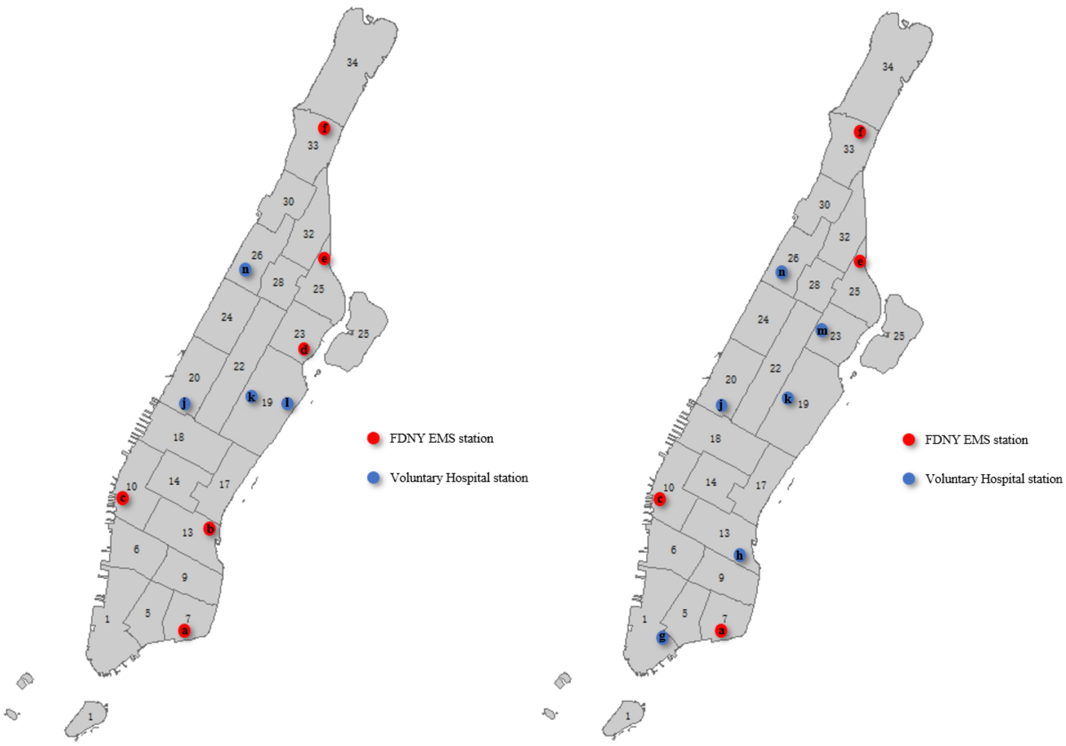
Note: SM means the vehicle service capacity of stations operated by FEM. SV means the vehicle service capacity of stations operated by VHE. TC means total cost. FC means fixed cost of opening stations. NM means the number of opening stations run by FEM. NV means the number of opening stations run by VHE. PC means purchasing cost of ambulances. VM means the number of ambulances purchased in the opening stations operated by FEM. VV means the number of ambulances purchased in the opening stations operated by VHE. TRC means transportation cost of demand fulfillment. Penalty means penalty for overtime. CL means the coverage level. RL means the response level.

Results of Table 5-7 and Table 5-8 inform us that there are thresholds for facility and vehicle service capacity. Setting both capacities to their thresholds could achieve maximal coverage with minimal cost.

5.5.4 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 5-3. Figure 5-3(a) and Figure 5-3(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 middle and east part, i.e., stations j, k, l, and n, are selected because there are no FEM stations in these places and the selected 4 stations can 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 5-3: 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 g, h, and m. The coverage level and response level are shown in Table 5-9, which indicate better performances of location solutions for homogenous facilities.

Table 5-9: The coverage level and response level of optimal location solutions for heterogenous and homogenous facilities

	CL	RL
Homogenous	98.61%	90.07%
Heterogenous	95.99%	88.60%

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.

Chapter 6: Conclusions and Future Research

6.1 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, and the penalty for overtime under service level requirements. We apply Sample Average Approximation to solve the problem and conduct numerical experiments to evaluate the performance of the solution method. Results show that stochastic program could achieve a better coverage with less cost compared with deterministic one and that solutions obtained by SAA are proved to be robust. 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. Second, service level is positively related to total cost. Improving service level will inevitably increase total cost. Besides, 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.

6.2 FUTURE RESEARCH

Future research can be extended in two ways: problem-related extension and solution-related extension.

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

Solution-related extension means we can apply heuristics or exact algorithms that solve problems in shorter time, such as L-shaped method and Branch-and-Cut.

References

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

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

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

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

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

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

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

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

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

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

Bürger, A., et al. (2018). The effect of ambulance response time on survival following 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.

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.

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 ambulances to emergency calls with classification errors in patient priorities. *IIE Transactions*, 45(1), 1-24.

McLay, L. A., & Mayorga, M. E. (2013b). A dispatching model for server-to-customer systems that balances efficiency and equity. *Manufacturing & Service Operations Management*, 15(2), 205-220.

Nasrollahzadeh, A. A., Khademi, A., & Mayorga, M. E. (2018). Real-time ambulance dispatching and relocation. *Manufacturing & Service Operations Management*, 20(3), 467-480.

Nelas, J., & Dias, J. (2020). Optimal emergency vehicles location: An approach considering the hierarchy and substitutability of resources. *European Journal of Operational Research, 287*(2), 583-599.

Nickel, S., Reuter-Oppermann, M., & Saldanha-da-Gama, F. (2016). Ambulance location under stochastic demand: A sampling approach. *Operations Research for Health Care, 8*, 24-32.

Park, H., Waddell, D., & Haghani, A. (2019). Online optimization with look-ahead for freeway emergency vehicle dispatching considering availability. *Transportation Research Part C: Emerging Technologies*, 109, 95-116.

Peng, C., Delage, E., & Li, J. (2020). Probabilistic envelope constrained multiperiod stochastic emergency medical services location model and decomposition scheme. *Transportation Science, 54*(6), 1471-1494.

ReVelle, C., & Hogan, K. (1988). A reliability-constrained siting model with local estimates of busy fractions. *Environment and Planning B: Planning and Design, 15*(2), 143-152.

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

ReVelle, C., & Hogan, K. (1989b). The maximum reliability location problem and α-reliable p-center problem: derivatives of the probabilistic location set covering problem. *Annals of Operations Research, 18*(1), 155-173.

Schilling, D., Elzinga, D. J., Cohon, J., Church, R., & ReVelle, C. (1979). The TEAM/FLEET models for simultaneous facility and equipment siting. *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.

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.