ABSTRACT

Title of Dissertation:	REAL-TIME DISPATCHING AND REDEPLOYMENT OF HETEROGENEOUS EMERGENCY VEHICLES FLEET WITH A BALANCED WORKLOAD
	Chenyang Fang
Dissertation directed by:	Professor Ali Haghani Department of Civil and Environmental Engineering

The emergency management service (EMS) system is a complicated system that tries to coordinate each system component to provide a quick response to emergencies. Different types of vehicles cooperate to finish the tasks under unified command. The EMS system tries to respond quickly to emergency calls and communicate with each department to balance the resources and provide maximal coverage for the whole system.

This work aims to develop a highly efficient model for the EMS system to assist the coordinator in making the dispatching and relocation decisions simultaneously. Meanwhile, the model will make a route decision to provide the vehicle drivers with route guidance. In the model, heterogenous emergency vehicle fleets consisting of police vehicles, Basic Life Support (BLS) vehicles, Advanced Life Support (ALS) vehicles, Fire Engines, Fire Trucks, and Fire Quants are considered. Moreover, a coverage strategy is proposed, and different coverage types are considered according to the division of vehicle function. The model tries to provide maximal coverage by advanced vehicles under the premise of ensuring full coverage by basic vehicles. The workload

balance of the vehicle crews is considered in the model to ensure fairness. A mathematical model is proposed, then a numerical study is conducted to test the model's performance. The results show that the proposed model can perform well in large-scale problems with significant demands. A comprehensive analysis is conducted on the real-case historical medical data. Then a discrete event simulation system is built. The framework of a discrete event simulation model can mimic the evolution of the entire operation of an emergency response system over time. Finally, the proposed model and discrete event simulation system are applied to the real-case historical medical data. Three different categories of performance measurements are collected, analyzed, and compared with the real-case data. A comprehensive sensitivity analysis is conducted to test the ability of the model to handle different situations. The final results illustrate that the proposed model can improve overall performance in various evaluation metrics.

REAL-TIME DISPATCHING AND REDEPLOYMENT OF HETEROGENEOUS EMERGENCY VEHICLES FLEET WITH A BALANCED WORKLOAD

by

Chenyang Fang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Ali Haghani, Chair
Professor Cinzia Cirillo
Professor Martin Dresner
Professor Paul Schonfeld
Professor Chenfeng Xiong

Acknowledgment

First and foremost, I want to express my heartfelt gratitude and appreciation to my advisor, Dr. Ali Haghani. It has always been my great honor to be his student. When I transferred from structure to transportation engineering, Dr. Haghani kindly agreed to adopt me as his new student. I will never forget that moment. He generously mentored me through my doctoral studies and led me to a fulfilling journey of intellectual adventure.

I want to thank the members of my dissertation committee, Dr. Cinzia Cirillo, Dr. Martin Dresner, Dr. Paul Schonfeld, and Dr. Chenfeng Xiong, for their insightful comments and invaluable suggestions throughout this research.

Words cannot express my gratitude to my parents for their unwavering support, patience, and understanding during my academic life and for every significant decision which could change my life. Their constant encouragement, trust, and love have been a source of motivation on my journey. Furthermore, I am expressing my sincere love and gratitude to my beloved girlfriend, Siyu Xie, for accompanying, encouraging, and inspiring me to pursue a better self and chase my dream bravely.

Finally, I would like to acknowledge the support provided by the University of Maryland Medical Center (UMMC). Their support enabled me to conduct my research effectively and accomplish the goals of my dissertation.

Table of Contents

List of Figures

List of Tables

Chapter 1 : Introduction

1.1 Emergency management service

Millions of Americans suffer from accidental injuries yearly, resulting in billions of dollars of direct or indirect economic losses. According to the Institute of Medicine report, more than 600,000 Americans die of cardiac arrest strikes each year. Approximately 395,000 cases of cardiac arrest occur outside hospitals, where less than 6% survive, while 200,000 cases happen inside hospitals, where about 24% survive (ScienceDaily, n.d.).

The cardiac arrest survival rate is extremely low and greatly depends on where the cardiac arrest happens and if the patients can get professional treatment. So, immediate effective treatment is critical to the survival rate, and every minute with professional treatment will increase the likelihood of survival without disability. It is the same for other types of injuries, such as car accidents. Unfortunately, based on the statistics, approximately thousands of deaths annually result from inadequate emergency medical care (Roemer, Kramer, & Frink, 1977).

The term 'emergency management service system' means a system that provides for the arrangement of personnel, facilities, and equipment for the corrective and coordinated delivery of health care services in an appropriate geographical area under emergency conditions (occurring either as a result of the patient's conditions or natural disasters or similar simulation), and which is administered by a public or nonprofit private entity which has the authority and the resources to provide effective administration of the system.

EMS system act of 1973 and its subsequent amendment in 1976 have been proposed to support building an adequate, standard pre-hospital emergency system. As a result, more than 300 EMS systems across the nation have been set up, and funding has been set aside for future planning and

growth ("A Brief History of Emergency Medical Services in the United States EMRA," n.d.). By 1975, more than 30 EMS residencies had developed nationwide, preparing physicians who would interface with EMS at all levels: from responders and educators to medical directors.

Within the last ten years, EMS has become a focus of intense research on pre-hospital interventions into many commonly encountered acute care issues observed in emergency medicine ("A Brief History of Emergency Medical Services in the United States EMRA," n.d.). With increasingly integrated technology used in pre-hospital care and hospital emergency rooms, the emergency response crew implements more and more services. Hospital emergency rooms are responsible for providing medical and surgical care to patients arriving at the hospital needing immediate care. Their work is limited to providing quick responses and transporting the patients to the assigned hospitals. In addition, earlier determination of patient condition severity, preparation before the arrival at the hospital, and coordination with the trauma center also can be done to shorten the time.

The EMS system is not an isolated system. It needs multi-party collaboration between health care, public health, and public safety. For example, pre-hospital medical services are usually provided by a fire department, a hospital, an independent government agency, a nonprofit corporation, or commercial, for-profit companies. The EMS system's responsibility is to coordinate each system component to play essential roles and quickly respond to emergencies. The EMS system consists of the following components ("EMS.gov | What is EMS?," n.d.):

- All kinds of private or public agencies and organizations
- Communication and transportation networks
- Trauma systems, hospitals, trauma centers, and specialty care centers
- Rehabilitation facilities

- Highly trained professionals
    - Volunteer and career pre-hospital personnel
    - Physicians, nurses, and therapists
    - Administrators and government officials
- An informed public that knows what to do in a medical emergency

Figure 1.1 illustrates the structure of the EMS system. The large circle represents each system element activated in response to an incident in the diagram. The arrows within the circle represent the specialty care areas within EMS. The list within the circle represents the elements acting behind the scenes to support the system.

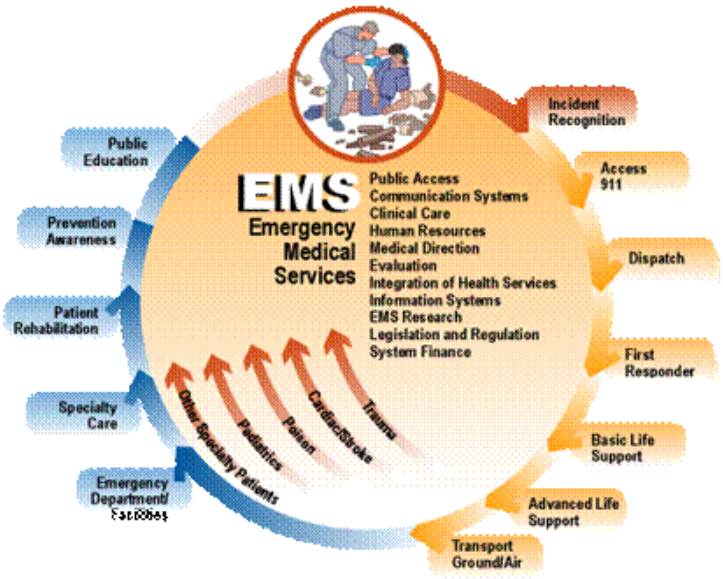


Figure 1.1 Illustration of the EMS system

1.2 Emergency call center (911)

The emergency call center is responsible for prioritizing and dispatching the required number and type of units to the calls to provide corresponding assistance. All calls to 911 are answered by

trained universal call-takers at the Office of Unified Communications. For example, calls requiring fire vehicles or medical units will be immediately transferred to dispatchers within the call center. While calls that need police cars will be forwarded to the police dispatcher within the call center. The center's Computer-Aided Dispatch (CAD) system helps with the dispatch of first responders. In the meantime, the dispatcher needs to monitor the activities of the responding personnel.

1.3 Maryland institute for emergency medical service systems (MIEMSS)

The Maryland Emergency Medical Services (EMS) System is a coordinated statewide network that includes volunteer and career EMS clinicians, medical and nursing personnel, communications, transportation systems, trauma and specialty care centers, and emergency departments ("Who We Are," n.d.).

The role of the Maryland Institute for Emergency Medical Service System (MIEMSS) is to provide oversight and coordinate all components of the Maryland Emergency Medical Service System in obedience to statute and regulation, including planning, operations, evaluation, and research.

In addition, MIEMSS provides leadership and medical direction, conducts, and supports EMS educational programs, operates, and maintains a statewide communications system, designates trauma and specialty centers, licenses and regulates commercial ambulance services, and participates in EMS-related public education and prevention programs.

In Maryland, more than half of the pre-hospital clinicians are volunteers operating in public service EMS agencies, while others are employed as career EMS clinicians by public service agencies and/or commercial EMS services (ground and air). Both the volunteers and employed pre-hospital clinicians and personnel receive professional training. They are certified by the state and provide medical care and assistance in accordance with the Maryland Medical Protocols.

In Maryland, pre-hospital clinicians are divided into two categories based on professional skills: basic life support (BLS) and advanced life support (ALS). BLS is provided by state-certified emergency medical dispatchers (EMD), first responders, and emergency medical technicians-basic (EMT-B). ALS, which is available in all jurisdictions, is provided by state-licensed cardiac rescue technicians (CRT), cardiac rescue technicians-intermediate (CRT-I), and emergency medical technicians-paramedic (EMT-P).

The MIEMSS uses the statewide EMS communication system to integrate and coordinate the entire EMS system in Maryland. By taking advantage of the advanced radio and microwave technology, the EMS communication system strengthens the links between the various departments of the system and enhances the connection between ambulances, helicopters, and hospitals.

The Emergency Medical Resources Center (EMRC), a communications center at MIEMSS, can assist with the heavy demand for medical communication and coordinate medical consultations between medical units and hospital physicians. After the medical units pick up the patients and request medical assistance, operators at the EMRC will guide the medical vehicles to transfer the patients to the available hospitals based on the severity of the patients and the availability and priority of hospitals. Pre-hospital clinicians transmit patient information to a hospital physician online when the medical unit is heading toward the assigned hospital. Physicians may direct the pre-hospital clinician to follow specific medical protocols and approve additional treatment.

The System Communication Center (SYSCOM), also located within the communication center, works with a Maryland State Police Duty Officer to dispatch and coordinate all Maryland State Police med-evac missions. In addition, the Maryland State Police Aviation Division, with a fleet

of 11 helicopters based in seven (7) sections across the state, transports over 5,000 critically injured or ill patients each year.

MIEMSS is not only committed to balancing medical resources but also providing fast, effective, and accurate medical rescue for patients. According to the MIEMSS, about 85% of patients are taken to the nearest hospital emergency department. There are 48 hospital emergency departments in Maryland equipped with professional medical teams working around the clock to ensure timely medical assistance for patients. For patients who need a higher level of care, Maryland's Trauma and EMS System can ensure that the patients get to the proper facilities to receive the right care through the use of statewide medical protocols for EMS clinicians.

1.4 Emergency response time

Many ambulance service providers use the percentage of highest urgency incidents reached by an ambulance within this maximum allowed response time as their performance criterion (Jagtenberg, Bhulai, & van der Mei, 2015). Three basic factors have been used to measure fire departments' emergency response performance: availability, capability, and operational effectiveness ("Understanding and Measuring Fire Department Response Times," n.d.)

For precise quantification and evaluation of the dispatching, the total dispatching time has been divided into several parts, shown in Figure 1.2.

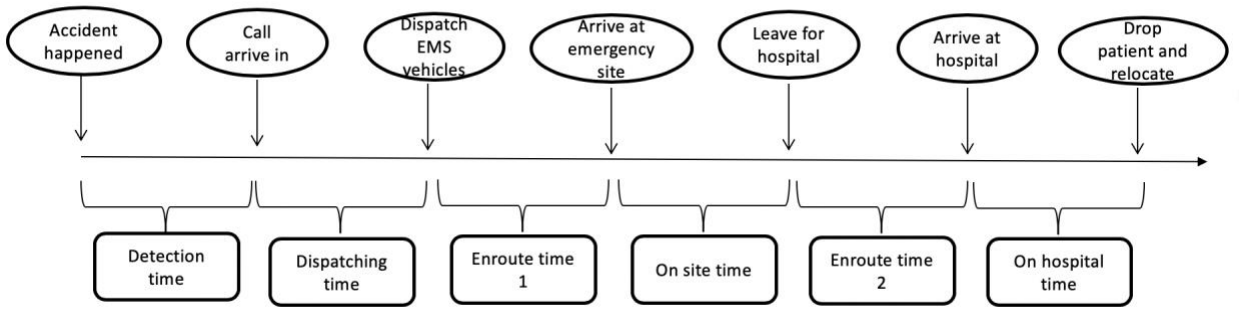


Figure 1.2 Emergency response time

One of the key EMS benchmarks for municipal and career fire departments is the National Fire Protection Association's (NFPA) 1710 (Standard for the Organization and Deployment of Fire Suppression Operations, Emergency Medical Operations, and Special Operations to the Public by Career Fire Departments). According to a combination of practices and more than 30 years of study, research, testing, and validation, NFPA committee members establish the standard, including representatives from various fire agencies. The NFPA establishes a one-minute turnout time, the elapsed time from when a unit is dispatched until the time its status changes to responding, and 240 seconds travel time benchmark for not less than 90% of dispatched incidents for EMS first response. Moreover, NFPA Standard 1710 establishes a 60-second "turnout time" and 480-second "travel time" benchmark goal for "the arrival of an advanced life support (ALS) unit at an emergency medical incident" for not less than 90% of dispatched incidents ("EMS Response Time | fems," n.d.).

1.5 Objectives of the research

Empirical studies suggest that after a cardiac and circulatory arrest, the chances for resuscitation to be successful decrease dramatically. Typically, chances decrease by 10% per minute if the patient is not treated accordingly. Providing a quick response to emergency requests

is crucial for the patients' health. In practice, simple rules for dispatching and relocation are in use. In Austria, the closest ambulance will always be dispatched in case of an emergency because of regulatory rules. After serving a request, ambulances are supposed to return to their home base. The majority of EMS systems in North America have implemented a redeployment strategy. A highly efficient dispatching strategy can quickly respond to the emergency call to protect property and people's safety and allocate resources reasonably to avoid waste.

This work aims to develop a highly efficient and comprehensive integrated model of dispatching and relocating emergency service vehicles to assist the coordination center in making dispatching decisions. The model determines the required number and types of vehicles that need to be dispatched in response to emergency calls. Moreover, redeployment decisions will be made simultaneously to ensure the system can have maximal coverage in preparation for future incidents. Different types of vehicles can cooperate under the unified command to finish the tasks. Historical accident data, real-time traffic data, and the status of vehicles and the system will also be incorporated into the model to help with decision-making.

1.6 Contributions of the research

This work proposes a modified integrated model to assist in making dispatching and relocation decisions. The proposed optimization model can increase the system's performance compared with previous work. A heterogeneous vehicle fleet, which consists of three types of emergency vehicles, is considered in the proposed model. Police vehicles, ambulances, and fire vehicles will cooperate to handle the tasks under the unified scheduling. According to the different functions, two types of ambulances (Advanced Life Support and Basic Life Support) and three types of fire vehicles (Fire Engine, Fire Truck, and Fire Quint) are considered in the model.

Moreover, according to the different types of vehicles arriving at the scene, different coverage types are considered to measure the performance of the dispatching and relocation. ALS vehicles can work as BLS and provide advanced medical care. Fire engines and fire trucks need to cooperate to finish all the jobs. It is assumed that when two types of vehicles arrive at the scene simultaneously, the task is handled fully functionally. Fire quint is equivalent to a combination of a fire engine and a fire truck. Fully functional coverage and partial functional coverage are considered in the model, and the corresponding benefits are added to the system.

Different types of vehicles will take different responsibilities in performing the tasks at hand. The BLS vehicles try to provide quick response, and ALS vehicles provide more professional and border treatment for the patients. Therefore, different standards are proposed for different types of vehicles. Different types of travel limits are considered in the model for different vehicles.

Moreover, more parameters are introduced into the model to make it more realistic. The priority of the hospital will be used to help the coordinator to assign the proper hospital for the patients.

Besides the decision-making constraints, some reasonable operational constraints are considered in the model. Workload balance is a symbol of fairness and affects the efficiency of relocation. Since the proposed approach is implemented in real-time, the accumulated workload will be calculated to help the system improve efficiency and avoid excessive moves. Work shifts are also considered in the model.

1.7 Organization of the dissertation

In Chapter 1, an introduction was made, and related concepts were introduced to have a better and more comprehensive understanding of the problem. In Chapter 2, previous work and literature

will be reviewed, and the evolution of existing models and algorithms will be sorted out. In Chapter 3, the problem statement and the mathematical model will be introduced. In Chapter 4, a numerical study is presented to test the performance of the proposed model. Chapter 5 presents a comprehensive numerical analysis of the historical medical data. The underlying physical network and the operational data are illustrated for Baltimore City. In Chapter 6, a discrete event system simulation model that is built to test the system is explained in detail. The framework of this discrete event simulation model can mimic the evolution of the entire operation of an emergency response system over time. The sophisticated discrete event simulation model is designed and coded in Python programming language. Chapter 7 presents the results of applying the proposed model in the discrete event simulation system in the real-world case study and compares those results with the actual operation data. Then an extensive sensitivity analysis is performed on the parameters in the model to test how the model will perform under various conditions. Chapter 8 summarizes this research and presents some avenues for future research.

Chapter 2 : Literature Review

This chapter comprehensively reviews the emergency vehicle dispatching problem, deployment, and redeployment problem.

2.1 Deterministic models

The location problem usually arises from the planning phase. The choice of the location of facilities will affect the system's performance.

Totegas et al. (Constantine Toregas, 1970) proposed the first emergency facilities location model as a location set covering problem (LSCP). This model aims to find the minimum number of facilities to cover all the demand nodes within a specific response time or distance. The LSCP tries to provide full coverage for all demand nodes, which is unrealistic, especially when the resources are limited.

Another basic model (Church & Davis, 1992) maximizes coverage by using a fixed number of facilities, formulated as a maximal covering location problem (MCLP) by Church and ReVell in 1974. However, these two models assume the minimum distance or response time between the demand nodes and the facilities is given. They overlook the stochastic feature of the problem.

Once a facility is in service in these two models, then the demand points within the coverage previously will be unattended. They omit the feature of unavailability of the facilities and randomness of the accidents. One strategy to deal with the stochastic feature of demand and spatial and temporal availability is multi-objective approaches, consisting of hierarchical and true multi-objective methods. Daskin and Stern have formulated a hierarchical objective set covering (HOSC) problem by considering the hierarchical programming and MCLP. The single objective model can

be modified to incorporate two objectives. The first objective is to minimize the number of facilities to cover all demands, while the second objective is to measure the degree of multiple coverages in the system. (Mark S. Daskin, 1981)

Hogan and ReVelle (Hogan & ReVelle, 1986) proposed two backup coverage models to ensure that areas with high demand can maintain a more uniform level of service, BACOP1 and BACOP2. The BACOP1 model provided backup coverage to as much of the population as possible with a given number of facilities. It incorporates both aspects of LSCP and MCLP to protect coverage from varying with time or with unit availability.

The BACOP2 model extended the Maximal Covering Location Problem as a multi-objective problem, allowing simultaneous optimization of both backup coverages. The model can trade off the population with single coverage against those with double coverage.

Daskin/Stern's formulation divides the whole region into N zones while the model equally weights the zones. So, Eaton et al. (Eaton, Ml Sánchez, & Morgan, 1986) extended Daskin/Stern's formulation to take into account weighted demand. The model tries to maximize the multiple coverages of demand within a given critical response time with a minimum number of ambulances. A multi-objective heuristic method is proposed to overcome the issue of insufficient random-access memory. The algorithm is tested on the Santo Domingo EMS system in the Dominican Republic.

One of the first models used to handle multiple types of vehicles was the tandem equipment allocation model (TEAM) in 1979 (Schilling, Elzinga, Cohon, Church, & Schilling, 1979). The model provides multiple types of services and some special capabilities to meet particular demands. The model was originally carried out for fire companies where protection is mainly provided by two types of fire vehicles: trucks (ladders) and engines (pumpers). This model can be applied in

the medical vehicle location model, in which ALS and BLS units provide full protection. The TEAM model assumes that demand is fully covered only if it has both primary and specialty equipment coverage with the pre-determined standard. Moreover, the specialty equipment can only be located in tandem with primary equipment. The constraint of required ordering has been relaxed in the facility-location, equipment-emplacement technique (FLEET) model (Schilling et al., 1979).

2.2 Probabilistic models

The early papers on the location problem focus on the location problem under deterministic conditions. Because the deterministic models overlook the fact that the facilities are not always available and the ambulance work in a queueing system, ignoring these features may result in inaccuracy or inefficiencies. The following papers concentrated on the location problem in which stochastic information has been considered. The probabilistic models have been put out to mimic this feature.

Chapman and White (Chapman & White, 1974) formulated the first probabilistic version of the location set covering problem under the assumption of a uniform system-wide busy fraction for servers, which is denoted as the maximum expected covering location problem (MEXCLP).

Daskin (Daskin, 1983) reformulated the maximum covering location model and extended it to incorporate the busy possibility factor for each facility. He assumed that not all the facilities are always available, and some facilities may be in service and cannot be accessed by the new request. The author assumes all the facilities are independent and have the same busy fraction q.

In 1989, Batta et al. (Batta, Dolan, & Krishnamurthy, 1989) thought that even if the system could be modeled as a queueing system, the workloads may not be equal due to the geospatial distribution of the demand nodes. So, they denoted the adjusted MEXCLP model as AMEXCLP

and tried to relax three assumptions to incorporate more stochastic information. Their model assumes that the busy probabilities are not identical between different facilities and are variant with respect to their locations. An interactive version of the hypercube queueing model is used to provide statistics about the system. Moreover, they assume all the servers operate dependently and use the correction factors for the hypercube queueing model. The final comparison is conducted between different models.

In Daskin's MEXCLP model, the travel time is assumed to be deterministic. Ignoring the features of stochastic travel time and the location position results in low system operation efficiency. Goldberg et al. (Goldberg et al., 1990) extended the model to a nonlinear integer programming model for finding the optimal base location. The model tries to maximize the expected system success rate by considering stochastic travel times. Unequal vehicle utilization and multiple call classes are also considered in the model. The model is validated by using the Tuscan Emergency Medical System in Arizona.

In 1994, Bernardo and Repede (Repede & Bernardo, 1994) extended the MEXCLP to incorporate temporal variation in the daily demand process in addition to spatial variation and multiple states of vehicle availability. Then the TIMEXCLP model was incorporated into a decision support system to determine the vehicles' location initially. This approach can help EMS planners analyze the EMS system characteristics under alternative scenarios.

Revelle and Hogan (Revelle & Hogan, 1989) introduced two maximum availability location problems derived from maximal covering location problems (MCLP). These two models try to maximize the population within the desirable travel time with stated reliability. The randomness only comes from the server availability in the paper, while the travel time is deterministic. The first MALP model assumed all the busy fractions for the servers in the system were the same.

While the second MALP model relaxed this assumption and allowed the busy fractions to vary in different city zones. Even the author noticed that a site-specified busy fraction is preferable to an area-specified busy fraction. However, the area-specified busy fraction was used in the paper because of some operation limitations. The site-specified busy fraction can be obtained by using the hypercube model.

In 1994, ReVelle and Marianov (Marianov & Revelle, 1994) extended the PLSCP model to formulate the queueing probabilistic location set covering problem (Q-PLSCP). The main difference between these two models is the minimum number of servers that must be located within the time or required distance of the node.

All of these models (MEXCLP, PLSCP, MALP (PROFLEET)) make the simplifying assumption that the probabilities of two vehicles being busy within the same region are independent. In the Q-PLSCP model, the dependence between servers with different busy fractions within the same neighborhood is considered.

Ball and Lin (Ball & Lin, 1993) formulated a new version of probabilistic LSCP and extended the LSCP model to a reliability probabilistic model denoted as Rel-P. In addition, they tried to prove that the upper bound of the uncovered probability for the demand nodes is limited to a pre-defined value.

The previous probabilistic models only consider one type of vehicle. ReVelle and Marianov (Marianov & Revelle, 1994) extended the FLEET model to a probabilistic facility location equipment emplacement technique, PROFLEET, applied to a fire protection system. The model tried to locate the depots or stations to maximize the population (number of calls), which is covered by an engine company and a truck company within the given distance with the same reliability α.

Their model tried to locate p^S fire stations and p^{E+T} vehicles to maximize the population which is covered by three engines and two tucks with the same reliability α.

In these models, fire trucks and fire engines are two distinct types of vehicles, and they take different parts of the job. Therefore, based on the fire department's request, the demand node is fully covered only if the demand node is simultaneously covered by these two types of vehicles. In 1998, Mandell (Mandell, 1998) extended the model to a two-tiered emergency medical service system, denoted as TTM. Emergency medical service system usually consists of two types of vehicles with different capabilities: basic life support (BLS) units and advanced life support (ALS) units. In contrast to the previous models, this model allows one type of server to substitute for the other. Moreover, two different types of servers may have different time standards.

Iannoni et al. adapted the hypercube queueing model to analyze the emergency medical system (EMS) on highways, which operates according to specific policies (Iannoni & Morabito, 2007). Different types of emergency calls and servers, partial backup of the servers, and multiple dispatching are considered in the model. Then those models were applied to a case study of an EMS operating on Brazilian highways.

In 2009, a modified method was proposed by Iannoni et al. to optimize the location of ambulances based along the highway (Iannoni, Morabito, & Saydam, 2009). The location GA/hypercube algorithm includes a local search procedure to evaluate the local neighborhood of each solution generated by the GA operators, which can achieve better performance than the districting GA/hypercube algorithm.

Expanding the system to incorporate multiple servers and represent each server individually can bring more complex and flexible dispatching policies. Morabito et al. (Morabito, Chiyoshi, & Galvão, 2008) proposed a model which assumed that the servers are non-homogeneous within

applications of the hypercube model. Experiments based on the proposed illustrative examples showed that model output measures could differ substantially depending on the degree of servers' non-homogeneity. The results imply that the non-homogeneous hypercube models should be used to generate the appropriate dispatching decision in practice.

By considering the spatial and temporal demand characteristics and server busy probabilities, Geroliminis et al. (Geroliminis, Karlaftis, & Skabardonis, 2009) developed a model for the congested service system, which is a spatial queueing model (SQM). The model also considered that the service rates are not identical and may vary between servers. Districting and dispatching problems are also integrated with the location model to optimize emergency vehicle deployment rather than based on the dispatching preference. A heuristic algorithm is proposed to find near-optimal solutions.

In 2009, a novel approach was proposed by Beraldi et al. (Beraldi & Bruni, 2009) to address uncertainty based on the probabilistic constraints in the traditional two-stage framework. In effect, the location and the definition of the corresponding capacities are the first-stage strategic decisions. In contrast, the tactical decisions concerning the allocation of customers to facilities can be taken in the second stage. Unlike the previous stochastic programming problem, replacing the stochastic constraints with probabilistic constraints allows the decision-makers to evaluate different solutions by varying the reliability level. The model assumes that the main source of uncertainty comes from the emergency calls process. Dependent assumptions among servers and demand points are modeled and tackled with joint probabilistic constraints, which is more general than the separate chance-constrained formulation. A Branch and Bound scheme is proposed to deal with the continuous variables to overcome the computationally overwhelming problem.

The local-reliability estimates and the α-reliability construct, two methods to measure the coverage level of the system, have been introduced and extended in the past decades. However, these issues are still subject to little empirical analysis and may not be modeled appropriately. In 2010 a hybrid model designated the local reliability-based maximum expected covering location problem (LR-MEXCLP) was proposed to examine these features (Sorensen & Church, 2010). In the model, the local-reliability estimates are incorporated into the original maximum expected coverage goal of MEXCLP. Then, different scenarios were generated, and solutions for MEXCLP, MALP, and LR-MEXCLP were generated using a mixed-integer program solver.

2.3 Dynamic models

Static models are usually applied at the strategic level instead of operational due to a lack of flexibility (Schmid & Doerner, 2010). Once the location decision is made at the beginning of the planning horizon, certain assumptions and constraints will be given, and stochastic features of the systems will be ignored. With the development of powerful computer systems and global positioning system (GPS) technology, the dynamic relocation approach has attracted more attention and is applied in the real world. With demand varying geospatially and temporally, dynamic relocation approaches need to be used to relocate the vehicles to cover for the busy ambulances.

Ingolfsson (2013) conducted a literature review on simulation models applied to emergency medical service operations. In the first part, typical processes and specific features related to the EMS system are described to reflect the system's complexity. Different categories of performance measures associated with EMS operations are introduced. Then, a comprehensive literature review is conducted to analyze the main features of the problem, such as data collection, model design,

verification, and validation. Finally, an overview of different scenarios in the literature is proposed aiming at improving EMS system performance and conducting sensitivity analysis.

Unlike previous research, which paid more attention to mathematical models, methodologies, and simulations, Ingolfsson (Ingolfsson, 2013) surveyed research on planning and management for emergency medical service emphasizing four different topics. First, three components as input information are investigated in detail: forecasting demand, response times, and workload. Second, appropriate EMS performance measures are discussed, and stochastic models which can be analytically solved are used to predict the EMS performance evaluation. Third, the optimization models for choosing station locations are interpreted. Finally, the optimization models designed for ambulance allocation are summarized based on predictable and unpredictable demand and travel time changes.

In 2018, Belanger et al. (V. Bélanger, Ruiz, & Soriano, 2019) summarized and discussed modern modeling approaches to address problems related to ambulance fleet management, particularly those related to vehicle location and relocation, as well as dispatching decisions. First, the early static ambulance location problem has been reviewed as the introduction. Then, the authors traced the evolution of both multi-period and dynamic approaches to show various methods and factors have been incorporated into the model to make the model more realistic and easier to implement. By completing existing reviews, this work provided a precise and up-to-date picture of research on optimization models and presented all the models and variants in a summarized table.

The most recent review paper was done by Joseph (Tassone & Choudhury, 2020), focusing on the developments of the ambulance routing problem (ARP) and ambulance location problem (ALP). First, multiple versions of ARP and ALP are structured with integer programming and

solved subject to a set of constraints. Then, a comprehensive review of simulation and mathematical programming is summarized in a table. Finally, various heuristic methods, which are: ant colony optimization (ACO), genetic algorithm (GA), local search-based solutions, neighborhood search description, particle swarm optimization (PSO), and clustering method, are analyzed according to the previous works and papers.

2.3.1 Offline redeployment approach

2.3.1.1 Planned multi-period deployments

In 1994, Repede and Bernardo (Repede & Bernardo, 1994) formulated the first multi-period location model (TIMEXCLP). For a fleet vehicle, the model divides the day into multiple time periods and randomly selects the busy factors for the vehicles to consider the variation in demand pattern and the number of available ambulances. The model tries to maximize the expected coverage of the system.

In the TIMEXCLP model, each time period is independent, and there may be a huge difference in the configuration of vehicles' locations between the periods. Van den Berg and Aardal (Van Den Berg & Aardal, 2015) extended the TIMEXCLP model by incorporating the start-up and relocation costs.

In 2008, Rajagopalan et al. (Rajagopalan, Saydam, & Xiao, 2008) extended Marianov and ReVelle's Q-PLSCP model for multiple periods, which is a dynamic available coverage location (DACL) model. The model tries to minimize the required number of ambulances to satisfy pre-determined coverage requirement with certain reliability and try to relocate the ambulances to satisfy the demand requests for each cluster of time with varying demand pattern. The model also uses Jarvis's hypercube approximate algorithm to calculate the ambulance busy fractions. Finally, a tabu search heuristic algorithm is given to solve the model.

Although redeployment can improve the performance of the emergency medical service system and reduce rescue time, frequent redeployment will cause an increase in operational costs and crew fatigue. Saydam et al. (Saydam, Rajagopalan, Sharer, & Lawrimore-Belanger, 2013) tried to reduce the number of redeployments without sacrificing coverage requirements. They considered the number of ambulance deployments in the objective function and extended the DACL model into a dynamic redeployment model (DRCL). Moreover, a new heuristic search algorithm is given to solve the problem.

In 2010, Schmid and Doerner (Schmid & Doerner, 2010) extended the double standard model (DSM), which was introduced by Gendreau et al. (Michel Gendreau, 1997), to formulate a multi-period mixed-integer model (mDSM) by considering the time-dependent variation in travel speed and resulting time-dependent variation in coverage areas. The double standard model (DSM) requires multiple coverages, such that demand points are supposed to be covered by more than one vehicle. The mDSM model tries to maximize the coverage throughout the entire planning horizon and considers the penalty to eliminate the unnecessary relocations. A variable neighborhood search method is proposed to solve the problem.

Basar et al. (Başar, Çatay, & Ünlüyurt, 2011) extended the BACOP 1 and DSM model to formulate a multi-period backup double coverage model, which is denoted as MPBDCM. The model tries to maximize the total population double covered in all the periods within two distinct time standards for two different EMS vehicles. And some operational requests have been added to the stations. Moreover, a tabu search approach is proposed and demonstrated on the generated data.

In 2013, Naoum-Sawaya and Elhedhli (Naoum-Sawaya & Elhedhli, 2013) proposed a two-stage stochastic optimization model for the redeployment problem. The model's objective is to minimize the number of relocations in the first stage and the number of emergency calls not served

within the pre-determined time. The historical data of the Waterloo region is used to test the algorithm's performance.

The DSM and its extensions are good approaches to dealing with the location and relocation problem, while the stochastic features of the demand and travel time are ignored in the model. Not considering these features may cause the model to be inaccurate and inefficient. Degel et al. (Degel, Wiesche, Rachuba, & Werners, 2015) extended the DSM model to a data-driven optimization model by incorporating empirical data, dynamic variations, and economic aspects. The model tries to force adequate coverage on the tactical level by using historical data.

In 2016, Bélanger et al. (V. Bélanger, Kergosien, Ruiz, & Soriano, 2016) proposed a comparison analysis study of location and relocation strategies to test the performance of different location and relocation strategies. The paper tried to quantify the benefits and drawbacks of the different relocation strategies compared with the static strategies. Therefore, four different relocation strategies were proposed. Strategy 1 and Strategy 2 are apriori models, and all decisions are made ahead regardless of the system's current state. Strategy 1 allows no relocations between periods, while relocation between periods is permitted in Strategy 2. In this model, the new idle ambulance can be relocated to the assigned station after the service. Strategies 3 and 4 are real-time dynamic relocation strategies.

2.3.1.2 Offline-dynamic redeployment

Because the relocation strategies need to be executed each time a vehicle is dispatched to a call, it makes the dynamic relocation algorithm time-consuming and infeasible when the calls arrive quickly. Since parallel computing was not always an option for dynamic relocation in the early days, a series of pre-planned scenarios can be applied to deal with the case in real life. In 2006, Gendreau et al. (M. Gendreau, Laporte, & Semet, 2006) proposed a dynamic model, a

maximum expected coverage relocation problem (MECRP). The model tries to maximize the expected coverage and consider the upper bounds of the relocation times.

The hypercube queueing model (HQM) proposed by Larson in 1974 was a Markov chain model of a queueing system with distinguishable servers and assumed a static deployment strategy where the idle ambulances are assigned to the fixed home station. Alanis et al. (Alanis, Ingolfsson, & Kolfal, 2013) extended the HQM by considering repositioning policies and used the same data requirements and output as the HQM model. The model is a two-dimensional Markov chain model of an emergency medical service system with redeployment using a compliance table policy. The paper demonstrated the Markov chain model is near-optimal, and a well-designed compliance table is vital to the final result.

In 2004, Yang (Yang, Hamedi, & Haghani, 2004) developed one of the most complete simulation models for the EMS vehicle dispatching system. The model is integrated with a Genetic algorithm to solve an EMS dispatching and redeployment problem.

In 2007 Andersson and Värbrand (Andersson & Värbrand, 2007) suggested preparedness as a way to evaluate the ability of ambulances serving the regions. They proposed a new quantifiable measure for preparedness and new decision support tools for vehicle dispatching and relocation, which is a dynamic ambulance relocation model (DYNAROC). The model considers the prioritization of an ambulance call, and the data in Sweden was used. Moreover, the redirection of vehicles between calls with different severity is considered. The relocation occurs when the preparedness of one or more regions is lower than the pre-determined level. The relocation model aims to minimize the maximal travel time for any relocation vehicles to reach the pre-defined preparedness in all zones. Then a tree-search heuristic method is used to solve the DYNAROC problem.

In 2015, Valerie et al. (Valérie Bélanger, Lanzarone, Ruiz, & Soriano, 2015a) addressed the ambulance dispatching and relocation decision problem simultaneously as the ambulance relocation and dispatching problem (ARDP), which tries to determine the location of available ambulances as well as the best-dispatching policy defined as a set of pre-assignment lists, one for each demand zone. ARDP suggests a joint dispatching and relocation strategy to minimize the total expected response time considering relocation efforts. In the model, the capacity of ambulances and maximal workload are considered. A metaheuristic decomposition approach consisting of heuristic algorithms that include the interoperation of metaheuristics and mathematical programming techniques is developed to solve the problem.

Liu et al. (Liu, Li, Liu, & Patel, 2016) extended the DYNAROC model by considering other random elements consisting of travel time and emergency demands. They proposed a chance-constrained programming model with probabilistic constraints to achieve a reliable level of services. The preparedness is modified using an approximate hypercube approach in the model. The model's objective is to maximize the profit with a given level of reliability within the limitation of the number of relocations. Finally, a genetic algorithm is proposed to solve the problem.

As mentioned above, the compliance table is optional for relocating ambulances in real life. Sudtachat et al. (Sudtachat, Mayorga, & Mclay, 2016) extended the model to a nested-compliance table model and restricted the number of relocations that can occur simultaneously. The advantage of nested policies is that only a pre-defined number of vehicles are relocated, and unnecessary movements can be avoided. The nested-compliance table model is formulated as an integer programming model, and the objective is to maximize the expected coverage using the compliance table. This paper considers only one type of vehicle and one type of call priority, and no patient queue is formed. The model uses the Markov chain method to approximate the EMS system's

steady-state probability and other parameter measures. Finally, the real-world data is used to validate the model, and a comparative analysis study is proposed to prove the benefit of the relocation with respect to the adjusted maximum expected covering location problem.

By extending previous work, two types of medical units are considered: Rapid Responder Ambulances (RRAs) and Regular Transport Ambulances (RTAs). The key difference between these two types of vehicles is that RRAs are faster, but they lack the ability to transport a patient to the hospital. In the model (van Barneveld, van der Mei, & Bhulai, 2017), the number of ambulances per vehicle type and traveling time that is instructed to relocate are constrained. If there is no EMS vehicle available for the response at the moment, the call enters a first-come-first-serve queue and waits for the next available EMS vehicle to be dispatched. Then an integer linear programming is formulated to compute the compliance tables for the EMS system, which uses the outcomes of a hypercube model as input parameters. Finally, a discrete-event simulation is conducted to generate the two-dimensional compliance tables.

The dynamic redeployment problem provides adequate coverage with reliability and relocates the idle vehicle to cover for the busy vehicles. In contrast to the urban areas, the rural areas have different features in the following aspects: rural areas have a limited number of ambulances, the demand per area in the rural region fluctuates, and the events occurring in the rural area are sparse. In 2017, van Barneveld et al. proposed a dynamic ambulance management system for a rural area with a limited number of ambulances (van Barneveld, Bhulai, & van der Mei, 2017). They proposed a discrete-time Markov decision process (MDP) to find a good configuration of ambulances that can respond quickly to the coming request. MDP provides a general framework for modeling sequential decision-making under uncertainty. However, the state space of the corresponding dynamic program may be extremely large, referred to as the curse of dimensionality,

which makes the computation intractable. Then a one-step look ahead heuristic method is proposed to address this problem, and a numerical comparative analysis with compliance table policy is conducted to test the performance.

Maxwell divided the relocation models into three categories in the paper “Approximate Dynamic Programming for Ambulance Redeployment” (Maxwell, Restrepo, Henderson, & Topaloglu, 2010). The first class of redeployment models involves solving integer programs in real time whenever an ambulance redeployment decision needs to be made. The second class of models is based on solving integer programs in a preparatory phase. The third class of models attempts to explicitly capture the system's randomness, either through a dynamic programming formulation or through heuristic approaches. The paper also proposed an approximate dynamic programming (ADP) approach. Compared with these three classes of redeployment models, the ADP approach can capture the stochastic features of the emergency medical system and quickly obtain the relocation decision. Compared with the lookup table policy, the ADP approach can fully automate the decision-making process. In the lookup table approach, the model is solved in a preparatory phase and provides a lookup table describing the deployment of all vehicles. The ADP approach can work with a variety of objective functions and can handle large-size dimensions of state-space problems. The model minimizes the total expected discounted number of calls by relocating the idle ambulances within a given delay threshold. However, due to the size of the problem, only new idle ambulances can be relocated sometimes. Finally, the performance of the ADP approach is validated by using realistic problems, and the parameters in the model are tuned.

In Maxwell’s model, the travel time is deterministic. Only one type of ambulance and call priority is considered. Schmid (Schmid, 2012) extended the ADP model to incorporate time-dependent travel times and demand requests based on the proposed work. The objective of the

approximate dynamic programming model is to minimize the resulting response times for all demands that occur in a day. In the model, only the new idle ambulances are relocated to the stations because Austria's law does not allow the empty ambulances to be repositioned from one station to another one. Despite this, the model has proved useful based on real data from the city of Vienna. Moreover, the stochastic travel time and demand volume are proved to be vital to the final dispatching result.

As mentioned in the previous work, constrained relocation time was important for implementing the relocation model in real-world systems. Partitioning the whole area into districts can impose a restriction on relocation time and reduce the complexity of the problem. In 2020, Sudtachat et al. (Sudtachat, Mayorga, Chanta, & Albert, 2020) partitioned the service area into small sub-areas. Each sub-area operated as a distinguishable sub-system. The nested-compliance model is formulated as an integer programming model (MRCRDP) to maximize the expected coverage for each district. Then the algorithm of the nested-compliance model embedded into a Tabu search heuristic for the MECRDP is proposed. Finally, the performance comparison is conducted using real-world data between combined relocation and districting policies and static policy based on the adjusted maximum expected covering location problem (AMEXCLP).

2.3.2 Online redeployment approach

For offline redeployment approaches, deployment decisions are made using pre-determined computations, and they do not consider the current state of the system and vehicles. These models are computationally efficient and can improve the performance of the EMS system compared with the static models. However, they cannot capture the stochastic features of the system and sometimes cannot reflect the real situation of the real world. Therefore, they cannot achieve the

best result when considering the real status of the system. As mentioned above, stochastic information is vital to the computational results.

With the development of advanced computational tools and powerful global positioning systems, the emergency medical system can be more precise, and the vehicles are much easier to track. Therefore, a higher level of requirement can be put forward in designing the emergency medical management system. In the new types of models, dispatching and redeployment of the vehicles should consider the system's current status in real time.

The first real-time redeployment model was proposed by Gendreau et al. (Michel Gendreau, Laporte, & Semet, 2001) and formulated as an integer programming model defined as RP^t. Several practical constraints and factors are considered in the model. The objective of the model is to maximize the zones covered by at least two vehicles to make sure they have extra coverage. Meanwhile, the corresponding penalties of relocations have been considered in the objective function to keep the location plan stable throughout the day, and the round trips and long trips will be avoided. Moreover, the model takes into account the workload and avoids redeploying the ambulance repeatedly. In 1997, Gendreau et al. used tabu search, a local search method, to solve the static ambulance location model. This method has been developed into a parallel tabu search heuristic method and applied to the dynamic problem.

Yang (Yang, 2006) and Yang et al. (Yang, Hamedi, & Haghani, 2005) were the first to develop a model which considers different categories of vehicles (police vehicles, ambulance vehicles, and fire vehicles). According to the emergency severity, the model will dispatch the appropriate types and numbers of vehicles to the incident site. Predetermined dispatching time for different types of vehicles is considered. Moreover, diversion of vehicles and redeployment of vehicles are also considered. For this problem and model, a Genetic algorithm heuristic method is used to find good

solutions in a reasonable time. Finally, a simulation model is conducted to check the performance of the optimization model on real-case data. In this research, the functional division of vehicles is not considered. For example, ALS vehicles can work as BLS vehicles.

More and more dynamic redeployment models are applied in the real-time decision-making process. However, some limitations make redeployment difficult to be implemented in the real world. The efficiency and accuracy of the algorithms cannot be guaranteed for some large problems due to incomplete system information. In 2015, Jagtenberg et al. (Jagtenberg et al., 2015) proposed a polynomial-time heuristic method to solve the real-time dynamic ambulance relocation problem. In the model formulation, the maximum expected covering location problem (MEXCLP) (Daskin, 1983), which uses an integer programming model to search static policy, is used to calculate the marginal coverage contribution to obtain a dynamic redeployment strategy. The heuristic method tries to relocate the newly idle ambulances to the base, resulting in the largest marginal coverage according to the MEXCLP model. The model ignores various information details because of the state space definition, resulting in guaranteed optimality. However, the small region and the realistic case study proved that the dynamic MEXCLP methods can achieve better results than the static model in most conditions.

System status management was first introduced to describe the process of dynamically reconfiguring EMS ambulance deployments to balance ambulance availability and demands across time and space (Stout, 1989). By matching the limited system resources with gradually increasing demands, system status management tries to derive dispatching and redeployment algorithms that dynamically assign ambulances to incidents. System status is the total number of ambulances available to respond to calls. In 2015, an ambulance allocation dynamic model (Lam et al., 2015) was proposed based on the system status management (SSM) strategy. The geographical

information system-based analysis and mathematical programming were used to develop the dynamic ambulance deployment plans for SSM based on real-world data. Discrete event simulation (DES) was used to compare the performance of the SSM strategy derived using the GIS-based analysis and MP approach. Three measures, response times, system utilization, and coverage proportions, were used to compare static allocation and SSM strategy performance under various demands and travel time uncertainties.

As an extension of the previous study, an ADP modeling framework (Lam, Ng, Nguyen, Ng, & Ong, 2017) was proposed in 2017 to derive the optimal dynamic ambulance allocation policies by leveraging the DES model for the Singapore EMS system. The ADP approach based on the DES model can overcome the problem that the MDP model can quickly become intractable by considering more parameters and system configurations. This is the first study on applying the ADP approach for a national EMS system based on the actual ambulance demands over a continuous two-year study horizon. The study used the temporal difference (TD) and the least-squares temporal difference (LSTD) learning algorithm to train the parameters of the ADP model. According to the organization of Singapore's EMS system, various scenarios and deployment policies were evaluated.

Due to the difficulty of achieving the optimality mentioned above, Jagtenberg et al. (Jagtenberg, van den Berg, & van der Mei, 2017) proposed a heuristic method instead of an optimal one to calculate the redeployment of the ambulances. This paper makes two contributions: first, some quantified standards have been set up to test the necessity of relocations in the heuristic method. The model tries to balance the benefits of relocation versus the disadvantages regarding the number of additional ambulance relocations to achieve this gain. Second, any idle ambulances can be relocated instead of dispatching the ambulances from base to destination. This strategy intends to

decrease the redeployment time to achieve the new allocation configuration. The model examines the relationship between relocations and performance by solving a linear bottleneck assignment problem (LBAP). Finally, some distinct scenarios are used to validate the performance of the heuristic method.

The literature mentioned above shows that most of the dispatching and relocation of the ambulances are computed independently. And the joint dispatching and location policy have not been studied extensively. In 2015, Bélanger et al. (Valérie Bélanger, Lanzarone, Ruiz, & Soriano, 2015b) proposed a joint dispatching and relocation model (ARDP) to determine the location of the ambulances as well as the best dispatching policy. Three main characteristics make the ARDP model different from others. First, the ARDP model will simultaneously consider dispatching and relocation to maintain an adequate service level and minimize the response time and relocation time. Second, the model takes into account the capacities and workload of ambulances. Finally, the expected response time is used to measure the performance of the policy rather than the number of coverages. The authors claim that the model that minimizes expected response time is much easier to understand and execute for the decision-maker for large-size problems, even when the expected coverage is not as precise as queueing theory-based model to estimate system performances under uncertainty. Then a metaheuristic decomposition approach is proposed to handle the large-size problem. The approach lies between the decomposition approaches and improvement heuristics. The problem is divided into serval small-size subproblems, and a specific model is applied to each subproblem. Finally, a real-world case is used to validate the performance of the heuristic method.

In 2016, Bélanger et al. proposed two dynamic repositioning problems: the third and fourth models in the paper. The first dynamic model corresponding to the double standard repositioning

model maximizes overall coverage within different pre-defined standards. The model only allowed redeploying the new idle ambulances, which had just finished the service. While in the second dynamic model, the dynamic relocation of available vehicles in real-time is permitted. Both newly idle vehicles that just finished service and idle vehicles located in the stations can be relocated based on the system's current status. The model is formulated as a double standard dynamic relocation model defined as DSDRM. The multi-objective model tries to maximize the double coverage while minimizing the vehicle relocating penalty.

In Sharifi's paper (Sharifi, 2014), the author proposed an integrated model to do the dispatching and relocation as well as considering the workload balance, work shift, and vehicle home stations. Based on the assumptions, the influence of these three factors is considered in the travel cost. A function is proposed to calculate the adjusted travel cost matrix because it is assumed that the assignment and relocation costs are different under different conditions.

In 2018, a two-stage redeployment optimization approach was proposed (Enayati, Mayorga, Rajagopalan, & Saydam, 2018). In the first step, the model seeks to maximize the coverage of the demand zone, which is weighted by historical call volume by making redeployment decisions. Hence, the relocation decision can only be made when the accumulated busy time plus the time related to the redeployment decision is smaller than a specified amount. In the second step, the minimum total travel time location problem with the workload restriction model is formulated to minimize total travel time with respect to redeployment moves.

Shakiba et al. (Enayati, Özaltın, Mayorga, & Saydam, 2018) proposed a two-stage stochastic programming model to redeploy and dispatch ambulances to maximize the expected coverage. The model incorporates multiple call priority levels and considers the balance of personnel workload in a shift. In the first stage, the idle ambulances are deployed, then dispatching decisions are made

in the second stage. To overcome the gradually increasing scale of the problem and increase the computational efficiency, the authors proposed a dual decomposition algorithm based on a split-variable reformulation to take advantage of the block separability of the underlying problem structure. By evaluating the model performance based on average coverage and ambulance workload during a shift, the proposed Lagrangian branch-and-bound algorithm can perform better than solving the integer programming model, especially for large-size problems.

In 2018, Nasrollahzadeh et al. (Ali Nasrollahzadeh, Khademi, & Mayorga, 2018) developed a flexible optimization framework to consider dispatching, redeployment, and reallocation for real-time dynamic EMS systems. Moreover, an approximate dynamic programming model is proposed to generate the solutions, and a lower bound on the expected response time of relocation policy is calculated to assess the quality of solutions. In addition, different basis functions and static and dynamic benchmarks are used to test the performance of the ADP model by using real-world data.

In 2020, Carvalho et al. (Carvalho, Captivo, & Marques, 2020) focused on the operational level to solve the ambulance dispatching and relocation problems. Their work proposed two generic approaches based on real-life features: a mathematical model that integrates dispatching and relocation decisions and aims to maximize system coverage and a pilot method heuristic. Both approaches use a time-preparedness measure to evaluate the system's capability to handle new emergencies.

The brief literature review shows that joint dispatching and redeployment models are formulated to address the problem, and various parameters are added to the models to reflect the complexity of the EMS system. Effective algorithms are then proposed to solve the problems. However, except from the works of Sharifi (Sharifi, 2014), it seems that joint dispatching and redeployment of heterogeneous emergency vehicle fleets under a balanced workload have not been

studied extensively. Therefore, this work aims to develop a joint dispatching and redeployment model to compute the configuration of a heterogeneous emergency vehicle fleet consisting of police, ambulance, and fire vehicles. In addition, workload balance among vehicles is also considered to ensure fairness and effectiveness.

Moreover, some operational constraints are considered in the model to reflect the actual situations in the real world. Different scenarios will be generated from real-world data to test the model's performance. Finally, an efficient heuristic approach will be proposed to solve the large-size problem and deal with complex situations in the real world.

Chapter 3 : Problem Statement and Mathematical Formulation

This chapter will elaborate on the characteristics and properties of the problem. The model's notations, parameters, and variables will be given and explained. Reasonable assumptions and simplifications about the practical problem will be provided, and then the mathematical formulation will be presented. Some details about the mathematical model will also be explained thoroughly.

3.1 Problem statement

In the real world, where medical resources and budget are limited, the Emergency Management Service (EMS) system is critical to coordinate the medical units and provide quick responses for the patients. Every second counts in life-threatening emergencies, which means the ability to provide fast and effective rescue can make the difference between survival and death. The main tasks of the EMS system are to coordinate all parts of the system to provide a quick response under the uniform command and shorten the response and rescue time. Therefore, ambulance service providers must meet strict requirements regarding response time. Response time is the core parameter to estimate the performance of the system. Response time is related to the dispatching system and largely depends on the system coverage by the remaining idle vehicle units. This research aims to develop an integrated model for the EMS system to assist dispatching and relocation decision-making. The integrated model will dispatch the required number and types of vehicles to the incident site and relocate the idle vehicles to other stations to maintain the maximal system coverage. Redeployment tries to make sure all regions are attended to and makes preparation for the accidents that may happen in the future. Both dispatching and relocation will

follow the guidelines and regulations. Some assumptions based on the actual conditions will be considered in the model to make it more realistic.

In summary, the integrated model reflects the actual need and considers many uncertainties that can be used in real life to assist the emergency management center in making decisions. The following section will discuss the characteristics and details of the problem.

3.1.1 Excessive workload conflict

Surveys of North American EMS operators have shown that operators who use a dynamic strategy increased from 23% in 2001 (Cady, 2002) to 37% in 2009 (Williams, 2009). Relocation of idle ambulances can achieve good use of medical resources through proper and strategic scheduling. By relocating the vehicle to unattended areas, the system can prepare for the accidents that may happen in the future and improve coverage with the same or even fewer resources. With increased population density and motor vehicles, the EMS system must handle steady and rapidly growing demands even under budget cuts or no increase in funds. So, although repositioning strategies can increase coverage and shorten first response time under some conditions, there is no doubt that these strategies impose extra workload on the personnel and increase the cost of system operation and maintenance. Therefore, good strategies need to be considered to reduce unnecessary moves under the premise of ensuring the improvement of system coverage. A lot of research proposed various methods to limit the number of moves or consider the relocation cost in the decision-making process to increase the useability in real life.

Sofianopoulos et al. (Sofianopoulos, Williams, Archer, & Thompson, 2011) conducted a study to investigate the impact of shift work on physical fatigue, sleep, and psychological factors among paramedics in Australia. Ambulance paramedics play an important, indispensable, and unique role as members of a component that forms a unique part of emergency services. As pre-hospital

clinicians, they are constantly and increasingly faced with heavy workloads that are physically, mentally, and emotionally tiring. These factors can compromise the effectiveness of these workers.

Fatigue and back problem are associated with EMS technicians due to the extended work hours and lack of fitness.

The system workload is better shared among the stations based on their cooperation. Stations are spatially distributed and operate independently, and the workloads may vary with the location of the stations (Aringhieri, Bruni, Khodaparasti, & van Essen, 2017). In some previous works, the total workload is evenly distributed among the available stations. Some others have mentioned that it would be desirable to have individual response time (the mean response time for each demand zone) that does not vary too much among the demand zones. In 2013, Toro-Diaz et al. (Toro-Díaz, Mayorga, Chanta, & McLay, 2013) illustrated two alternative criteria, in particular, variability in individual response time as well as variability in ambulance workloads which can be seen as fairness performance indicators from the perspective of internal and external customers.

3.1.2 Preference for home stations

An ambulance station is a structure or other area for storing ambulances and their medical equipment, as well as working and living space for their staff. Ambulance stations have the basic facilities to maintain the ambulance vehicles and provide the supplies for the medical units, such as a charger for the vehicles' batteries and medical supplies. The stations have specific areas for storing medical equipment and supplies. Stations are also equipped with office and living areas for clinicians and technicians. In addition, the stations may have an alerting system and training rooms, which depend on whether crews routinely wait at the station. In some conditions, ambulance stations may be co-located with or integral to other emergency service facilities, such as fire stations or police stations, especially where the fire department runs the ambulance service.

After providing the basic medical treatment at the scene or dropping the patients off at the hospital, an ambulance can be on standby at the hospital or relocated to the station to allow staff to rest and replenish medical resources. Based on the configuration of stations, there is no doubt that crew members prefer to locate at their home stations. Compared with ambulances, the cost of relocating the fire vehicles to other stations is much higher than relocation to the home station. EMS workers are commonly deployed in teams and work 12 or 24-hour shifts. At the end of the shift, vehicles located at other locations need to return to their home station for crews' preparation. Also, a vehicle at the end of the shift is not suitable for relocation.

3.2 Characteristics of the problem

3.2.1 Emergency vehicle fleet

Due to unexpected events' complexity, different departments sometimes need to cooperate to handle the accident. After the incident happens, the patrolling police vehicle will be dispatched to the accident site. At the incident scene, the police officer must ensure everyone is safe and manage logistics like automobile removal, traffic control, recording accident, and finishing documents. After the accident, the police officer may also need to determine fault.

Many EMS calls present situations that cannot be adequately addressed by a two-person ambulance crew. That is why a fire engine, fire truck, or fire quint is needed to help with the scene. The staff also needs to ensure the patient's care needs are met promptly and safely if the patients are trapped, critically ill, or unable to walk. In addition, personnel must be prepared to address any hazards found and provide standby fire protection as needed. For example, suppose a fire were to break out. In that case, crews must be prepared to extinguish the flame and ensure personnel safety. Ambulances must reach the incident site to provide medical rescue and pre-hospital care. Then

communicate with the hospital, if necessary, to ensure the hospital finishes the necessary preparation and saves rescue time.

So, three categories of emergency vehicles are considered in the model:

- Police vehicles: only one type of police vehicle is considered
- Ambulance vehicles:
    - Basic Life Support (or BLS) is a specific level of pre-hospital medical care provided by trained responders. Basic Life Support consists of a number of life-saving techniques focused on the "ABCs" of pre-hospital emergency care.
    - ALS (Advance Life Support) provider may perform advanced procedures and skills on a patient involving invasive and non-invasive procedures
- Fire vehicles:
    - Fire Engines are equipped with hoses and water so that the personnel can fight with fire.
    - Fire trucks carry ladders, rescue equipment, and other tools to support firefighting activities.
    - Fire quint is a firefighting apparatus that serves the dual purpose of an engine and a ladder truck.

According to different functions, the vehicles have a more detailed division which can make dispatching and scheduling quicker and more efficient. Therefore, considering three different types of emergency fleets and six different vehicle types can make the model more realistic.

3.3 Assumption related to the model

This section elaborates on the assumption behind the model. Key characteristics of the problem need to be extracted, and reasonable assumptions must be made to handle the uncertainties in real life.

3.3.1 Spatial assumptions

Unlike Sharifi's research (Sharifi, 2014), which divides all nodes into two categories: critical nodes and ordinary nodes, demand zones will be used to do the relocation. Here are several advantages:

- Because some nodes have similar properties, such as similar dispatching times and accident rates, using demand zones instead of nodes can decrease problem complexity significantly and save computational resources.
- The importance of points is not fixed. The accident rate for a given location is different during different time periods. For example, the regions with significant office space need to be paid more attention during working hours. While during the night, after the staff leaves the office, the residential areas may need more coverage. Fixing the importance of the locations cannot reflect the geospatial and temporal distribution of accidents which may cause a huge difference in the results.
- The importance of locations is a binary value. Even for two nodes that are both critical nodes, their importance may not be the same. For example, larger hospitals may need more equipment among hospitals of different sizes.
- Fixed demand nodes do not fully use the historical data, which is precious in making dispatching and relocation decisions.

An EMS system serves a large region and coordinates various departments. The continuous space of potential accident locations will be divided into several smaller subregions to reduce the system's complexity and computational difficulties. It is assumed that the subregions are small enough that all nodes in the subregion share common properties—moreover, each subregion's demand aggregates to the zone's centroid. Then the accident rate of each zone during different time periods can be calculated according to the historical data.

3.3.2 Temporal assumptions

The location of the accident site and stations is known for dispatching, and the travel time can be computed based on the road network. For relocation, however, the travel time between the stations and each zone can be represented as the travel time between stations and centroids of zones. Accurate travel time estimation will affect the dispatching decision and change the redeployment decision and workload balance. In the real-world case study, the Open Street Map will be used in Python to generate the road network of Baltimore City. And the travel distance matrix is calculated based on the network using Dijkstra's algorithm. When calculating travel time, travel condition is not considered for several reasons. First, sirens and lights are used to clear the path when the vehicles are en route. So, traffic conditions may not affect travel speed. Second, in the historical data, the actual travel conditions at the time of the incidents are missing. Real-time travel information can be obtained from Google if travel conditions are necessary. The real-world case study's purpose is to ensure the model works.

For relocation, long-distance dispatching is not an option in real life. A constraint can be imposed on the relocation travel time. If the relocation travel time is greater than a specific value, relocation is not an option.

The occurrence of the relocation has strict restrictions. When an emergency vehicle unit is dispatched to respond to a call, personnel must provide basic rescue at the scene. The medical clinicians may provide pre-hospital treatment and then transfer the patient to the assigned hospital. After finishing a series of tasks, the ambulance will become idle and can be dispatched to a new accident site or relocated to a station. The real-world case study will assume an average working time for each dispatching. If the interarrival time between two incidents is smaller than a threshold, it usually means that the relocated vehicle cannot get to the assigned station to prepare for a new incident. On the other hand, if the interarrival time of two incidents is larger than the average working time, the dispatched vehicle has finished its job and can return to its home station.

To account for the workload balance in the model, after dispatching the vehicles to the incident sites and relocating the vehicles to the stations, the travel time of each vehicle needs to be added to the accumulated busy time.

3.3.3 System dynamic assumptions

It is assumed that the current status of vehicles and the system are known in real time. Therefore, the online redeployment approach is based on the event. Each time a call comes into the system, information about the vehicle and the system will be updated. Then the decision-making will be based on the updated information.

3.3.4 Assumptions related to crew

The proposed model tries to give out efficient dispatching and relocation decisions to guide the operation of the EMS system as well as consider the workload balance of vehicle units. This measure tries to maintain fairness and avoid excessive workload. First, according to the configuration of the stations and vehicles, it is assumed that the crews prefer their home station when relocation decisions are being made. Second, after their shift, the vehicles and crews must

return to their home stations and rest. Third, the accumulated work hour for each vehicle is calculated. If a vehicle's accumulated work hours exceeds a given value, it is assumed that it is unavailable for relocation, but it can still be sent to a new emergency scene. This is reasonable in real life. When an accident happens, sending the nearest available vehicle to the accident site will result in the best solution in most cases. Dispatching the proper vehicle to take care of the patients is highly recommended, rather than leaving them at the incident site waiting. While the relocation is not mandatory, the workload should be considered in the decision-making for relocation. Excessive workload should be avoided to help avoid crew fatigue.

3.4 Mathematical formulations

3.4.1 Notations

The real-time dispatching and redeployment problem is formulated as an integer-programming problem subject to various constraints. Table 3.1 illustrates the variables, parameters, and notations used in the mathematical model.

3.4.2 The integer programming model

Traditionally, the dispatching and relocation decisions are considered separately. Some studies have shown that the response times and busy fractions largely depend on the station locations, frequency of the incidents, and the policy of the EMS vehicles dispatching. Various dispatching policies and many important factors can affect the dispatching of EMS vehicles. There is no way to guarantee that one method is helpful in all situations. The most widely used dispatching policy for EMS vehicles is to dispatch the closest vehicles to the emergency site. It is the most straightforward method to reduce the system response time.

Table 3.1 Summary of notations

GLOBAL PARAMETERS AND SETS	
N	Initial fleet size at the beginning of a shift
S	Set of ambulance locations
A	Set of incident sites
H	Set of hospitals
D	Set of demand zones
T	Total hours of a shift
$\boldsymbol{\tau_{ss'}}$	Shortest travel time from station $s \in S$ to station $s' \in S$
$\boldsymbol{\tau_{sa}}$	Shortest travel time from station $s \in S$ to incident site $a \in A$
$\boldsymbol{\tau_{as}}$	Shortest travel time from accident site $a \in A$ to station $s \in S$
$\boldsymbol{\tau_{ah}}$	Shortest travel time from accident site $a \in A$ to hospital h $\in H$
$\boldsymbol{z_d}$	Importance of demand zone d $\in D$
$\boldsymbol{PD_{kd}}$	The penalty for deficiency of type k vehicles at demand zone $d \in D$
$\boldsymbol{PT_{ka}}$	The penalty for the shortage of type k vehicles at accident site $a \in A$
$\boldsymbol{\zeta_d}$	Set of locations that cover zone d $\in D$
$\boldsymbol{\gamma}$	Maximum allowed workload of each ambulance in a shift (hours)
A	Set of accidents
$\boldsymbol{T_{ka}}$	The pre-defined travel time limit for type k vehicles to the accident site a
DYNAMIC PARAMETERS AND SETS	
t	Current time in the shift, t $\in \{0, T\}$
$\boldsymbol{L_t}$	Set of available ambulances at time t $\in T$, $L_t \leq N$
$\boldsymbol{\alpha_t}$	Maximum allowed accumulated busy time for each ambulance at time t $\in T$ after the redeployment decision at time t $\in T$
$\boldsymbol{\beta_{kis}^t}$	Accumulated busy time of type k vehicle I located at station s at time t
$\boldsymbol{I_{kis}}$	Binary parameter, =1 if type k ambulance I is located at station $s \in S$ at time t, =0 otherwise.
$\boldsymbol{Req_{ka}}$	Integer parameter, number of type k vehicles required by accident site a at time t
$\boldsymbol{tt_{limit}}$	Integer parameter, travel time limit for the relocation

DECISION VARIABLES

$\boldsymbol{R_{ss'}}$	Binary variable, =1 if a vehicle located at station s (regardless of type) moves from $s \in S$ to $s' \in S$, =0 otherwise
$\boldsymbol{DIS_{ksa}}$	Binary variable, =1 if type k vehicle is dispatched from station s to accident site a, =0 otherwise
$\boldsymbol{DIS_{kas}}$	Binary variable, =1 if type k vehicle is dispatched from accident site a to station s, =0 otherwise
$\boldsymbol{DIS_{kah}}$	Binary variable, =1 if type k vehicle is dispatched from accident site a to hospital h, =0 otherwise
$\boldsymbol{Y_{kd}}$	Integer variable, =n if d $\in D$ is covered n times by type k ambulance after redeployment decision, =0 otherwise
$\boldsymbol{X_{kis}}$	Binary variable, =1 if a type k ambulance i is located at station $s \in S$ after deployment decision, =0 otherwise
$\boldsymbol{XAB_{is}^{a}}$	Binary variable, =1 if ALS ambulance i located at station $s \in S$ works as a BLS ambulance for dispatching at time t $\in T$, =0 otherwise
$\boldsymbol{XQE_{is}^{a}}$	Binary variable, =1 if Fire Quint i located at station $s \in S$ works as a Fire Engine for dispatching at time t $\in T$, =0 otherwise
$\boldsymbol{XQF_{is}^{a}}$	Binary variable, =1 if Fire Quint i located at station $s \in S$ works as a Fire Truck for dispatching at time t $\in T$, =0 otherwise
$\boldsymbol{XAB_{is}^{r}}$	Binary variable, =1 if ALS ambulance i located at station $s \in S$ works as a BLS ambulance for redeployment at time t $\in T$, =0 otherwise
$\boldsymbol{XQE_{is}^{r}}$	Binary variable, =1 if Fire Quint i located at station $s \in S$ works as a Fire Engine for redeployment at time t $\in T$, =0 otherwise
$\boldsymbol{XQF_{is}^{r}}$	Binary variable, =1 if Fire Quint i located at station $s \in S$ works as a Fire Truck for redeployment at time t $\in T$, =0 otherwise
$\boldsymbol{EXT_{ksi}}$	Binary variable, =1 if type k vehicle i dispatched from station s exceeds the required travel time limit, =0 otherwise
$\boldsymbol{x_{kis}^{a}}$	Binary parameter, =1 if a type k ambulance i located at location $s \in S$ at time t is dispatched to an emergency site, =0 otherwise.

x^r_{kis}	Binary parameter, =1 if a type k ambulance i located at location $s \in S$ is relocated to another station, =0 otherwise.
x_{kia}	Binary parameter, =1 if a type k ambulance i is located at accident site a$\in A$, =0 otherwise.
x_{kih}	Binary parameter, =1 if a type k ambulance i is located at hospital h$\in H$, =0 otherwise.

In 2017, Jagtenberg et al. (Jagtenberg, Bhulai, & van der Mei, 2017) proposed that the closest-idle dispatching policy is not always optimal. They showed that under light traffic conditions, the model using the myopic dispatching policy, which dispatches the closest vehicles, can lead to an optimal solution, while this might not be the case in heavy traffic conditions. However, the final result is close to the optimal state in the latter condition.

Toro-Diaz et al. (Toro-Díaz et al., 2013) proposed a joint location and dispatching decisions model for the EMS system that integrated the location and dispatching decisions. Their study found that the closest vehicle dispatching policy can lead to the optimal solution when the objective is to minimize the average response time. While if the objective is to maximize the coverage, the policy may result in a suboptimal result. It is also noticed that the joint location and dispatching model may bring little benefit to the system if the two most common criteria, which are response time and expected coverage, are used. So, to test the actual performance of the integrated model, another two criteria, in particular variability on individual response and ambulance workloads, are considered to measure the potential benefits.

This section will present the formulation of an integrated integer programming model for dispatching and redeployment. The model seeks to minimize the total travel time and considers the deficiency of each type of EMS vehicle. On the other hand, the model tries to maximize the total coverage weighted by the historical demand in each zone at different time periods T.

3.4.2.1 Objective function

Minimize

$$Z_D = \sum_{a \in A}\sum_{k}\sum_{s} DIS_{ksa} x_{ksi} \tau_{sa} + \sum_{k}\sum_{s}\sum_{s'} R_{ss'} x_{ksi} \tau_{ss'} + \sum_{a\in A}\sum_{k}\sum_{s} DIS_{kas} x_{kai} \tau_{as} + \sum_{a\in A}\sum_{k}\sum_{h} DIS_{kah} x_{kai} \tau_{ah} + \sum_{h\in H}\sum_{k}\sum_{s} DIS_{khs} x_{khi} \tau_{hs} + \sum_{k}\sum_{a\in A} PD_{ka} \cdot D_{ka} + \sum_{k}\sum_{d\in D} PD_{kd} \cdot D_{kd} + \sum_{k}\sum_{s}\sum_{a} PT_{ka} \cdot |\tau_{sa} - T_{ka}| \cdot EXT_{ksi} - \sum_{k}\sum_{d} z_{kd} \cdot Y_{kd} \quad (3\text{-}1)$$

Equation (3-1) is the objective function that minimizes the total travel cost of the problem. The first two terms of the equation minimize the travel cost of dispatching the vehicles to the incident site and relocating the idle vehicles to other stations.

After pre-hospital treatment, some patients may need to be sent for higher-level medical care, and the ambulances need to transfer the patients to the assigned hospital while other patients are good to go. Then the ambulance needs to return to the station. The third and fourth terms are set to do the relocation from the incident site and escort the patient to a hospital. Like above, the fifth term minimizes the travel time of vehicles relocating to stations from hospitals.

Under ideal circumstances, all calls can be answered immediately, and the nearest vehicles can be dispatched to the incident site to rescue the patients. Moreover, there are enough numbers and types of idle vehicles that can be relocated to cover all the regions in the system. With limited budgets and resources, existing vehicles cannot meet the needs of dispatching and relocation. Therefore, limited resources are used to satisfy higher priority demand, which is the purpose of

the sixth and seventh terms in the objective function. If there are not enough numbers and types of vehicles to do the dispatching or relocation, penalties will be added to the model.

Since response time is critical to pre-hospital treatment, different standards have been set up to regulate rescue operations and estimate the performance of the dispatching algorithm. If the vehicles cannot reach the incident sites within the specified time, penalties represented by the eighth term will be added to the model.

For the relocation part, the last term in the model provides maximal coverage for the whole system based on the importance of different zones, calculated from the historical distribution of incidents.

3.4.2.2 Constraints

- Dispatching constraints

$$\sum_{a} DIS_{sa} x_{kis}^{a} + \sum_{s'} R_{ss'} x_{kis}^{r} \leq 1, \forall k \in K, s \in S, i \in L_s \tag{3-2}$$

$$x_{kis}^{a} + x_{kis}^{r} \leq I_{kis}, \forall k \in K, s \in S, i \in L_s \tag{3-3}$$

Constraints (3-2) require each vehicle to have at most one destination at each time step. Constraints (3-3) ensure that dispatching or relocation can only happen if there is a vehicle on duty. I_{kis} is an index variable used to indicate if there is a type k vehicle i located at station s.

For the police vehicle,

$$\sum_{s} \sum_{i \in L_s} DIS_{sa} x_{kis}^{a} + D_{ka} \geq Req_{ka}, \forall k = 1 \tag{3-4}$$

Constraints (3-4) ensure enough emergency vehicles are dispatched to the incident sites. D_{ka} are variables used to record the deficiency of emergency vehicle type k at the incident site a. If there is a shortage of required types of vehicles, penalties will be added to the model to ensure all demands can be appropriately satisfied.

For the ambulances,

$$\sum_{s}\sum_{i\in L_s} DIS_{sa}x_{kis}^{a} + XAB_{is}^{a} + D_{ka} \geq Req_{ka}, \forall k = 2 \tag{3-5}$$

$$x_{kis}^{a} \geq XAB_{is}^{a}, \forall\, k = 3, s \in S, i \in L_s \tag{3-6}$$

$$\sum_{s}\sum_{i\in L_s} DIS_{sa}x_{kis}^{a} - XAB_{is}^{a} + D_{ka} \geq Req_{ka}, \forall k = 3 \tag{3-7}$$

$$x_{kis}^{r} \geq XAB_{is}^{r}, \forall\, k = 3, s \in S, i \in L_s \tag{3-8}$$

Constraints (3-5) to (3-8) define the number of deficiencies of ALS and BLS in an emergency. According to the functional division, ALS is a higher-level and advanced version of BLS. So, if ALS can reach the incident site before BLS, it is assumed that ALS can perform the job of a BLS. This is the reason that the definition of deficiency of BLS is different. XAB_{is}^{a} is the variable indicating whether an ALS works as a BLS. The superscript a indicates that ALS is dispatched to an incident site to provide medical care, while r means that relocation action happens.

$$\sum_{s}\sum_{i\in L_s} DIS_{sa}x_{kis}^{a} + \sum_{s}\sum_{i\in L_s} XQE_{is}^{a} + D_{ka} \geq Req_{ka}, \forall k = 4\,, a \in A \tag{3-9}$$

$$\sum_{s}\sum_{i\in L_s} DIS_{sa}x_{kis}^{a} + \sum_{s}\sum_{i\in L_s} XQT_{is}^{a} + D_{ka} \geq Req_{ka}, \forall k = 5\,, a \in A \tag{3-10}$$

$$x_{ksi}^{a} \geq XQE_{si}^{a} + XQT_{is}^{a}, \forall\, k = 6, s \in S, i \in L_s \tag{3-11}$$

$$x_{ksi}^{r} \geq XQE_{si}^{r} + XQT_{is}^{r}, \forall\, k = 6, s \in S, i \in L_s \tag{3-12}$$

$$\begin{aligned}\sum_{s}\sum_{i\in L_s} DIS_{sa}x_{kis}^{a} - \sum_{s}\sum_{i\in L_s} XQE_{is}^{a} - \sum_{s}\sum_{i\in L_s} XQT_{is}^{a} - \sum_{s}\sum_{i\in L_s} XQE_{is}^{r} \\ - \sum_{s}\sum_{i\in L_s} XQT_{is}^{r} \qquad + D_{ka} \geq Req_{ka}, \forall k = 6\,, a \in A\end{aligned} \tag{3-13}$$

The condition of fire vehicles is similar to the ambulances but with a bit of difference. Because the fire engine and truck can only finish a part of the job, it is assumed that both the fire engine

and fire truck must arrive at the incident site so that the task can be considered fully completed. Since a fire quint is functionally a combination of a fire engine and a fire truck, a fire quint can be dispatched to finish all the work. Constraints (3-9), (3-10), and (3-13) ensure that the required number of fire vehicles are dispatched to the incident site. If there are not enough vehicles, the deficiencies of each type of vehicle will be calculated. Constraints (3-11) and (3-12) ensure that there are enough fire quints to be dispatched as a basic vehicle: fire engine and fire truck.

The dispatching and relocation need to satisfy the flow conservation constraints. Constraints (3-14) ensure that the number of dispatched vehicles should be less than the number of available vehicles located in the station.

$$\sum_{s'}\sum_{i} R_{ss'}x^{r}_{kis} + \sum_{a}\sum_{i} DIS_{sa}x^{a}_{kis} \le n_{ks}, \forall s \in S \tag{3-14}$$

$$\begin{aligned} X_{kis} = I_{kis} - (R_{ss'}x^{r}_{kis} + DIS_{sa}x^{a}_{kis}) + R_{s's}x_{ks'i} + DIS_{as}x_{kih} \\ + DIS_{hs}x_{kih}, \qquad \forall k \in K, i \in I_s, s \in S \end{aligned} \tag{3-15}$$

$$x_{kih} \le I_{kih}, \forall k \in K, h \in H, i \in L_H \tag{3-16}$$

$$x_{kia} \le I_{kia}, \forall k \in K, a \in A, i \in L_H \tag{3-17}$$

Constraints (3-15) define the status of vehicles and the system after the dispatching and relocation. Constraints (3-16) and (3-17) define the existence of the vehicles that need to be redeployed.

$$\sum X_{ksi} \le C_{ks}, \forall k \; in \; K, \forall s \in S \tag{3-18}$$

Constraints (3-18) are about the capacity of the stations. Therefore, when relocating, the stations' capacities need to be considered. If the number of vehicles exceeds the stations' capacities, it is assumed that relocation is not an option.

$$\tau_{sa} \cdot DIS_{sa} \cdot x_{kis} - M \cdot EXT_{ksi} \leq T_{ka}, \forall k \in K, s \in S, i \in L_s, a \in A \quad (3\text{-}19)$$

Constraints (3-19) ensure that emergency vehicles can reach the incident sites within the required time. If they cannot respond to the incidents in time, variable EXT_{ksi} will be equal to 1, and the penalty will be added to the objective function.

- Required coverage constraints

For ambulance

$$\sum_{s \in \zeta_d} \sum_{i \in L_s} (X_{kis} + \sum_{s\prime} R_{s\prime s} \cdot XAB^r_{is\prime}) + D_{kd} \geq Y_{kd}, \forall d \in D, k = 2 \quad (3\text{-}20)$$

$$Y_{kd} \geq 1, \forall\, d \in D, k = 2 \quad (3\text{-}21)$$

$$\sum_{s \in \zeta_d} \sum_{i \in L_s} (X_{kis} - \sum_{s'} R_{s's} \cdot XAB^r_{is'}) \geq Y_{kd}, \forall d \in D, k = 3 \quad (3\text{-}22)$$

For fire vehicle

$$\sum_{s \in \zeta_d} \sum_{i \in I_s} (X_{kis} + \sum_{s'} R_{s's} \cdot XQE^r_{is'}) + D_{kd} \geq Y_{kd}, \forall d \in D, k = 4 \quad (3\text{-}23)$$

$$Y_{kd} \geq 1, \forall\, d \in D, k = 4 \quad (3\text{-}24)$$

$$\sum_{s \in \zeta_d} \sum_{i \in I_s} (X_{kis} + \sum_{s\prime} R_{s\prime s} \cdot XQT^r_{is\prime}) + D_{ks} \geq Y_{kd}, \forall d \in D, k = 5 \quad (3\text{-}25)$$

$$Y_{kd} \geq 1, \forall\, d \in D, k = 5 \quad (3\text{-}26)$$

$$\sum_{s \in \zeta_d} \sum_{i \in I_s} (X_{kis} - \sum_{s'} R_{s's} \cdot XQE^r_{is'} - \sum_{s'} R_{s's} \cdot XQT^r_{is'}) \geq Y_{kd}, \forall d \in D, k = 6 \quad (3\text{-}27)$$

Constraints (3-20), (3-22), (3-23), (3-25), and (3-27) define the times that each zone is covered by different types of vehicles. Constraints (3-21), (3-24), and (3-26) ensure that each demand zone is covered by a basic vehicle at least once.

- Priority constraints

$$DIS_{ah} \cdot P_h \geq N_p, \forall a \in A, \forall h \in H \quad (3\text{-}28)$$

The patients are not always assigned to the nearest hospitals in the real world. The pre-hospital clinicians need to coordinate with the medical center to escort the patients to the proper hospitals according to various factors, including the severity of patients' conditions, the type of treatment and cure needed, and the availability of hospitals. The coordination operation is a complex process that needs multi-professional decisions. Constraints (3-28) ensure that the patient can be escorted to the corresponding level hospital based on the different priorities.

- Workload balance constraints

$$\beta_{kis} + R_{ss\prime}\tau_{ss\prime}x^r_{kis} \leq \alpha_t, \forall k \in K, s \in S, i \in L_s \quad (3\text{-}29)$$

$$R_{ss\prime}\tau_{ss\prime}x^r_{kis} \leq tt_{limit}, \forall k \in K, s \in S, i \in L_s \quad (3\text{-}30)$$

$$\alpha_t = \frac{t \cdot \gamma}{T} \quad (3\text{-}31)$$

Workload balance is the system's fairness and affects the performance of decision-making. Not considering the workload may cause crew fatigue and decrease the operations' efficiency. Constraints (3-29) ensure the accumulated work time of vehicle units is shorter than a pre-defined value in a shift. If the total work time of a vehicle unit exceeds the threshold value, it is assumed that this vehicle unit is not suitable for relocation. This constraint only works for relocation.

Excessive and frequent relocation may cause crew fatigue and increase the unnecessary waste of workload and budget. On the other hand, efficient relocation can help the system get maximal coverage and shorten the response time. Therefore, there are strict restrictions on the occurrence of relocation action. Constraints (3-30) ensure that long-distance relocation is not allowed. Constraints (3-31) are used to calculate the pre-defined value.

- Operational constraints

$$I_{kis}^{T} = 0, \forall s \in S_p \tag{3-32}$$

$$I_{kis}^{0} = I_{kis}^{T}, \forall s \in S \backslash S_p \tag{3-33}$$

Besides constraints required to make the system perform better, operational constraints are considered to make the model more realistic and in accordance with real situations. Constraints (3-32) to (3-33) ensure the vehicle units return to their home stations at the end of the shift. The vehicles on standby in temporary locations also need to return to home stations.

- Variable domain constraints

$$DIS_{sa} \in \{0,1\}, \forall s \in S, a \in A \tag{3-34}$$

$$R_{ss'} \in \{0,1\}, \forall s, s' \neq s \in S \tag{3-35}$$

$$x_{kis}^{r} \in \{0,1\}, \forall s \in S \tag{3-36}$$

$$x_{kis}^{a} \in \{0,1\}, \forall s \in S \tag{3-37}$$

$$X_{kis} \in \{0,1\}, \forall s \in S \tag{3-38}$$

$$Y_{kd} \in integer, \forall d \in D \tag{3-39}$$

$$XAB_{is}^{a} \in \{0,1\}, \forall s \in S, i \in L_s \tag{3-40}$$

$$XQE_{is}^{a} \in \{0,1\}, \forall s \in S, i \in L_s \tag{3-41}$$

$$XQT_{is}^{a} \in \{0,1\}, \forall s \in S, i \in L_s \tag{3-42}$$

$$XAB_{is}^{r} \in \{0,1\}, \forall s \in S, i \in L_s \tag{3-43}$$

$$XQE_{is}^{r} \in \{0,1\}, \forall s \in S, i \in L_s \tag{3-44}$$

$$XQT_{is}^{r} \in \{0,1\}, \forall s \in S, i \in L_s \tag{3-45}$$

Constraints (3-34) to (3-45) define the variable domains.

Chapter 4 : Numerical Study

In this chapter, a small-size problem will be solved, and different scenarios will be tested to evaluate the performance of the integrated model. Gurobi optimizer, a commercial optimization solver for linear programming, quadratic programming, quadratically constrained programming, mixed-integer linear programming, mixed-integer quadratic programming, and mixed-integer quadratically constrained programming in Python, is used to solve numerical problems and to find the optimal solution. This numerical study first proves that the proposed mathematical model can handle complex situations and assist the coordinator in making dispatching and relocation decisions. Second, by using Sharifi's model as a baseline, a comparison is made between these two models to evaluate the performance of the proposed model. Third, by increasing the problem's size, the model's characteristics and properties are explored, and the model's performance under large demands is analyzed. The numerical study aims to prove that the proposed model can handle large-scale problems and perform better than other models. In real life, incidents always come into the system one by one in a small region. Therefore, to better compare and display the performance of models, several generated large-scale and complex scenarios are used to make the comparison.

4.1 Some clarifications about the numerical study

The numerical studies are conducted on a randomly generated graph to eliminate the influence of human intervention.

First, the Erdos-Renyi model (Erdös & Rėnyi, 1959) is used to generate the random graph. In the model, a graph is constructed by connecting labeled nodes randomly. Then, each edge is included in the graph with probability p, independently from the other edges. After the graph is generated, random weights will be generated using the Python built-in function 'random.'

Second, after the randomly generated graph is ready, NetworkX, a Python package for creating, manipulating, and studying complex networks' structure, dynamics, and functions, is used to prepare the data. The shortest path lengths are calculated using Dijkstra's algorithm. If two nodes are not connected, a large number represents the lack of connectivity.

Third, the travel distance matrices for the two models are calculated respectively based on the same graph. This measurement ensures the fairness of the comparison. As a reminder, Sharifi's model considers two types of nodes according to their importance. In contrast, the proposed model in this research divides the whole region into some small zones with the same properties. This move can vastly decrease the size and complexity of the problem, which will shorten the computational time and the data preprocessing time. Besides the size of the problem, to compare the performance of the two models fairly, the same size of the demand zones will be used to do the computation.

Fourth, based on the above procedure, different sizes of problems are used to test the model's performance in different situations. The characteristics of these randomly generated cases are shown in Table 4.1.

The cases are solved with the latest version of Gurobi in Python on a computer with the same configuration. The number of constraints and variables and running time for each scenario are shown in Table 4.2.

Table 4.2 summarizes the properties and characteristics of both models and the number of constraints and variables for each case. Both models are dedicated to achieving two functions: dispatching and relocation. Both models try to dispatch the required number and type of vehicles to the incident sites within the pre-defined travel time for dispatching.

Table 4.1 Characteristics of 6 cases

Case	Number of demand zones	Number of ordinary nodes	Number of critical modes	Number of vehicles for each type	Number of stations for each type	Number of accidents in the system
Case 1	55	50	5	5	5	1
Case 2	55	50	5	5	5	5
Case 3	550	500	50	50	50	5
Case 4	550	500	50	50	50	10
Case 5	550	500	50	50	50	20
Case 6	550	500	50	50	50	40

Table 4.2 The number of variables and constraints and the running time for 6 cases

Case	Number of constraints	Number of variables	Running time (Dispatching)	Running time (Dispatching and relocation)
Case 1 - Sharifi's Model	2733	2261	0.015	0.21
Case 1 - This research	593	976	0.0145	0.049
Case 2 - Sharifi's Model	3257	2725	0.0233	0.255
Case 2 - This research	617	1312	0.01	0.064
Case 3 - Sharifi's Model	32276	40480	0.098	178.97
Case 3 - This research	5792	14182	0.032	2.66
Case 4 - Sharifi's Model	38556	46010	0.153	264.85
Case 4 - This research	5822	17302	0.032	2.724
Case 5 - Sharifi's Model	51116	57070	0.245	289.65
Case 5- This research	5882	23542	0.04	2.9
Case 6- Sharifi's Model	76236	79190	0.436	247.73
Case 6 - This research	6002	36002	0.044	10.41

If there are vehicle deficiencies or travel time exceeds the pre-defined travel limit, penalties will be added to the objective function. Both models use the same algorithm and have the same objective. It is easy to quantify the performance of the final results. Therefore, the running time of dispatching has been measured independently. In the dispatching model, all constraints have been commented out. It is illustrated from the table that the proposed model in this work can get the same results by using less time. The modified structure of the model can shorten the computational time.

Two different covering strategies have been chosen for the two models for the relocation part. Sharifi's model attempts to provide double coverage for critical nodes within T_1 minutes and provide double coverage for ordinary nodes within T_2 minutes, where T_1 and T_2 are pre-defined time limits. Several constraints have been set to define how often each node is covered. While the proposed model in this work tries to use the BLS vehicles to cover all the demand zones to shorten the first response time and use ALS vehicles to get maximal coverage. Both models have their advantages and suitable applicable scenarios. Due to the complexities of the problem, it is hard to quantify the coverage performance. It is shown that the proposed model in this research can finish the relocation decision in a few seconds, while for Sharifi's model, the computational time increases a lot with the increase in problem size.

According to Table 4.2, it is shown that for the integrated model, the relocation part consumes the most computational resources. Both models can finish the decision process without considering relocation within a few seconds. When considering relocation, the computational time increases significantly, especially when the sizes of the problems increase. Fire departments nationwide seek to decrease their response times to emergencies. When the problem sizes get larger, the running time of relocation will exceed the allowable running time.

The performance of relocation is affected by many factors, such as the graph's connectivity, the available number of vehicles, and the location of critical and ordinary nodes.

In summary, the proposed model can handle large-scale and complex situations and produce dispatching and relocation decisions in a reasonable time. Furthermore, the modified structure of the model can decrease not only the dispatching time but also the relocation time. With the increase in problems' sizes, the running time of the integrated model will also increase, but the increase in the running time is within an acceptable range.

Chapter 5：A Large-Scale Case Study

This chapter presents a comprehensive numerical analysis of the historical medical data. The underlying physical network and the operational data are from Baltimore City.

5.1 Case study operational data

The data used in the analysis comes from the University of Maryland Medical Center (UMMC) and the Maryland Institute for Emergency Medical Service System (MIEMSS). The data are generated from the real-world operations of the ambulances and medical units for the incidents that happened in Baltimore City. The study data comprised all ambulance calls over a continuous one-year period from 1 January 2019 to 31 December 2019 and was analyzed across the time of the day.

According to medical data, 61,233 incidents occurred in Maryland in January 2019. For each incident, a variety of main vital signs are routinely monitored by medical professionals and healthcare providers. Multiple medical records might correspond to one incident. Duplicate records and unnecessary data were dropped according to the PSAP (public safety answering point) call time to help with the emergency vehicle dispatching study.

The data consisted of approximately 99,100 calls in Maryland state per year. 15,162 of these incidents happened in Baltimore City. 7,501 incidents occurred in January, and 1,140 of these incidents happened in Baltimore City.

The medical data records detailed and comprehensive information about the incidents, including the time and location of the incidents and detailed dispatching information. The medical data provides valuable data for future research and study. Moreover, it also provides scientific theoretical support for model and solution optimization. However, some data are not fully recorded

due to special regulations and limitations. For example, only physical address information has been recorded for each incident that happened in Baltimore City, while exact GPS location information is missing. Moreover, only the dispatched vehicle unit number is kept in records, and corresponding station information is missing, which makes it difficult to find the original location for each unit.

Several processing methods and reasonable assumptions are devised to make the data accessible and to handle these problems. The programming language cannot identify physical addresses. It is necessary to convert the data type through a geocoding process. Geocoding is the computational process of transforming a physical address description to a location on the Earth's surface (spatial representation in numerical coordinates). In this study, geocoding in Python with the help of Geopy and Geopandas libraries is conducted to convert the physical addresses to their corresponding latitude and longitude. Due to several reasons, such as missing zip code, misspelled street name, misspelled city name, and different kinds of abbreviations in the data, only 797 records of incident data can be used to find the geospatial information.

5.2 Case study network

The network consists of 12,643 nodes and 32,264 directed links. According to the Baltimore City fire department website, in this region, there are 24 fire stations strategically located throughout the city, shown as green icons in Figure 5.1. These fire stations are equipped with various types and numbers of ambulances, medical units, and fire vehicles. According to the records, 17 hospitals in this region were used to transfer patients as destinations.

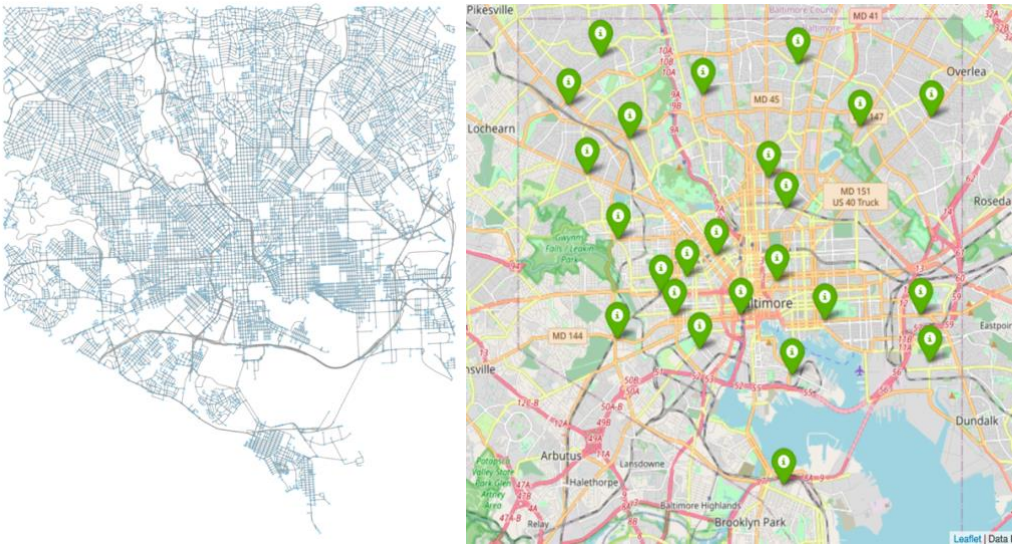


Figure 5.1 The case study network (left) and the geospatial distribution of fire stations of Baltimore City (right)

This study will use the demand zones to generate the calls and calculate the performance measurements. There are several advantages:

- Using the demand zones instead of nodes can vastly decrease the size of the problems and shorten the computational time.
- Second, sufficient data does not exist to support the model and generate the necessary distribution function for each node, which may cause inaccurate final results. In some scenarios, some nodes are close enough to be assumed to have the same properties, such as incident rates and dispatching time.

A geohash is a unique identifier of a specific region on the earth. This study uses geohash to generate the city's demand zones. It is assumed that each demand zone is small enough that all nodes in the same demand zone will have unified properties, such as the same call arrival rates and traveling time from stations. Python package Geohash is used to generate the spatial representations, and different precisions can be chosen depending on the accuracy demand. In this study, precision 6 is used, which means the service region of Baltimore City, MD, is divided into

356 demand zones, where each demand zone is a 0.7625 by 0.38125 miles square region. The city is about 92.05 square miles with 356 demand zones and has a population of about 585,000. Demands are defined to be the calls requesting paramedic units from an EMS system. Some zones had no demand (no calls for service) during the observed period and, therefore, were removed from consideration. The total number of demand zones with positive demand is 356 during the day and 329 during the night period.

5.2.1 Emergency Vehicles

One hundred fifty-five vehicle identification numbers are reported in the data, and 114 of these vehicles took care of fewer than 30 incidents in 2019. Based on the Baltimore City Fire Station's official website, 41 registered ambulance and medical units are considered in the study. According to the vehicle types on the records and official website, 16 basic life support (BLS) ambulances and 31 advanced life support (ALS) ambulances are selected for the dispatching and redeployment operations.

In the medical record data, only dispatched vehicle unit number is recorded, while the original station information is missing. All EMS vehicles' original stations are assumed to be identical to the information on Baltimore City Fire Station's official website. Next, for the missing station information for some vehicles, the K-means clustering method is used to aggregate the dispatching trajectories. The cluster is generated, and the nearest station is selected as their home station.

Figure 5.2 The job assignments for different ALS ambulance vehiclespresents the detailed job assignment for different ALS ambulance vehicles, and

Figure 5.3 shows the same for BLS ambulance vehicles.

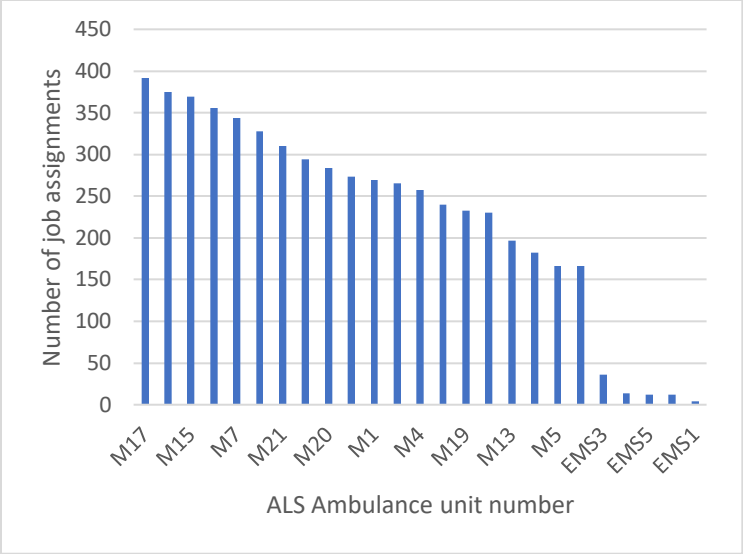


Figure 5.2 The job assignments for different ALS ambulance vehicles

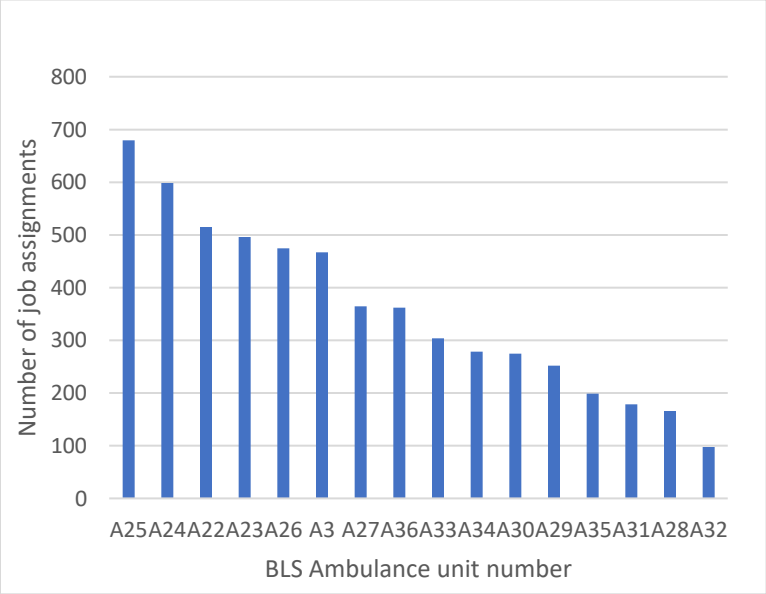


Figure 5.3 The job assignments for different BLS ambulance vehicles

The probability distribution analysis on the input data was conducted using Python and some libraries. Some packages were used to identify the best-fitted distribution for the generated database and determine the quality of fitting of probability distribution functions to the input data. First, histograms were plotted to show the general type of the input data. Then numerous

distributions and fitting functions were tested, and a summary of the best distribution was obtained. The five best distributions were kept and sorted from the best to the worst according to the respective error terms.

In the probability distribution analysis, the whole dispatching process was divided into different periods. For each period, various distribution functions were tested at different times of the day. In this study, the whole day was divided into two 12-hour periods to represent day and night to reflect the actual demands and operation proprieties. The daytime is from 8 am to 8 pm, while the rest is nighttime.

5.2.2 Preparation time

The preparation time starts with the call arriving at the 911 center and ends with the emergency vehicles dispatched from the stations. During this period, the incident information is collected, dispatching and rescue decisions must be made, and the crew members must prepare to finish the task. A total of 15,162 records were used to calculate the preparation time for each incident. According to the test results shown in **Error! Reference source not found.** and Table 5.3, the best distribution is Lognormal Distribution with a squared error equal to 1.253 for nighttime and 0.071 for the daytime. The comparison of histograms of the real-world data and the fitted data is shown in

Figure 5.4.

Table 5.1 The results of the estimation of the distribution of emergency incidents preparation time during the night

Fitted Function	Squared Error
Exponential	1.31121726
Normal	1.37892533
Gamma	1.70083921

Chi-squared	1.35897134
Lognormal	1.25345486
Beta	1.31910642
Burr	1.38015642

Table 5.2 The results of the estimation of the distribution of emergency incidents preparation time during the day

Fitted Function	Squared Error
Exponential	0.10154709
Normal	0.13182412
Gamma	0.25853525
Chi-squared	0.18216477
Lognormal	0.07091998
Beta	0.1189358
Burr	0.09237762

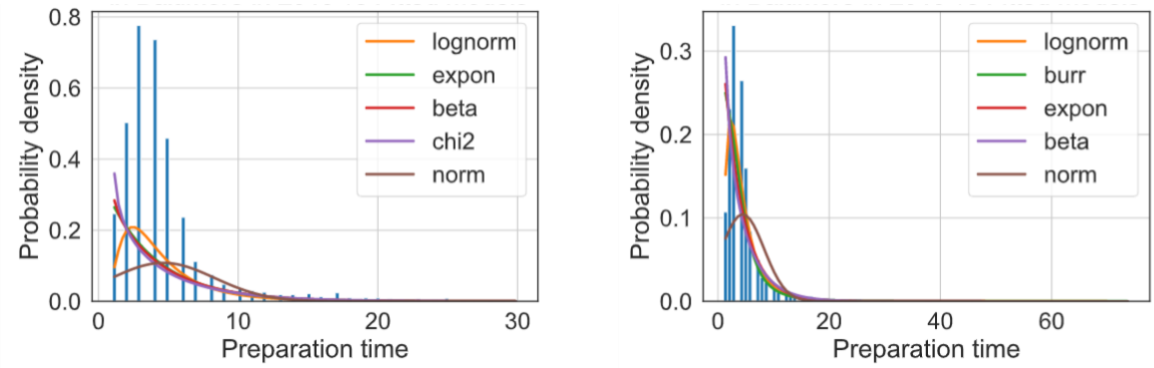


Figure 5.4 The histogram of preparation time for incidents during the night (left) and the day (right) in Baltimore City in 2019 vs. fitted models

5.2.3 Interarrival time

A total of 15,162 medical data records were used to calculate the interarrival time between each consecutive pair of incidents. According to the distribution fitting analysis, the top six best-fitted distribution functions are shown in Table 5.3 and Table 5.4, corresponding to nighttime and daytime. The results in Table 5.3 and Table 5.4 indicate that the exponential distribution with a squared error equal to 0.00023 for nighttime and 0.00021 for daytime is the best.

Figure 5.5 shows the histograms of the real-world data and the fitted data.

Table 5.3 The results of the estimation of the distribution of emergency incidents interarrival time during the night

Fitted Function	Squared Error
Normal	0.00176799
Exponential	0.00023617
Gamma	0.00025205
Lognormal	0.00030228
Beta	0.00034868
Burr	0.00025675

Table 5.4 The results of the estimation of the distribution of emergency incidents interarrival time during the day

Fitted Function	Squared Error
Normal	0.00210502
Exponential	0.00021134
Gamma	0.00020898
Lognormal	0.00802299
Beta	0.0002089
Burr	0.00052968

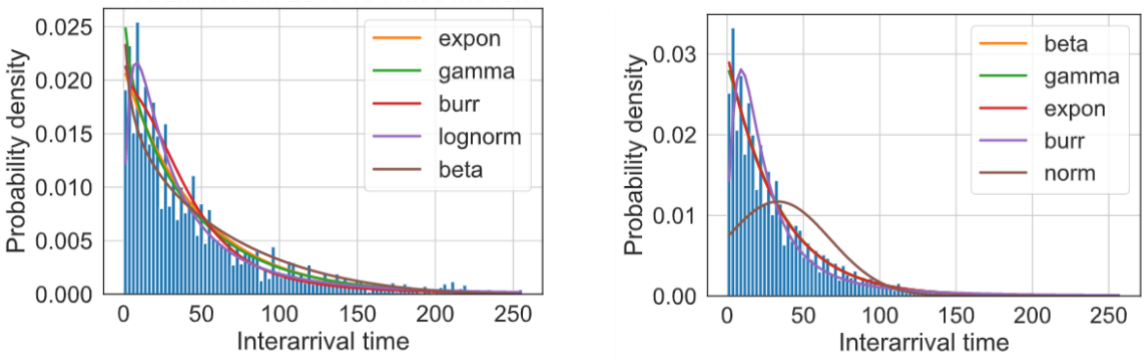


Figure 5.5 The histogram of the interarrival time for incidents during the night (left) and the day (right) in Baltimore City in 2019 vs. fitted models

The average call arrival rate across the time of the day is summarized in the box plot in

Figure 5.6. The vertical axis represents the total number of incidents that occurred at that time of the day for each month and the variation between different months. Twenty-four box plots represent the different hours of the day. The figure indicates the highest accident rate of the day is around 6 pm, and the lowest is around 5 to 6 am.

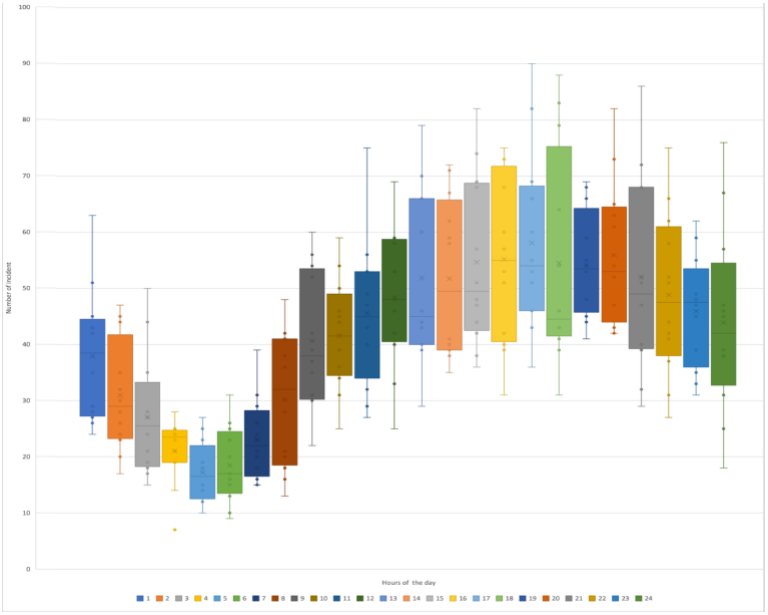


Figure 5.6 Incident arrival rates in Baltimore City in 2019

Figure 5.7 shows a pie chart that represents different categories of inter-arrival time of incidents. All data in the graph are in minutes. The figure shows that the inter-arrival time between 5 to 30 minutes and 30 to 60 minutes accounts for 44% and 23% of all data. The inter-arrival time is critical to the dispatching and redeployment model. Too short an inter-arrival time will make the redeployment decision model inaccessible. The vehicle cannot be dispatched to the new incident site while en route to another station. So, the redeployment process can sometimes even decrease the total coverage and negatively impact the system.

On the other hand, too long an inter-arrival time will cause a waste of resources because the dispatched vehicles can finish their jobs and return to their home locations to prepare for future

incidents. A redeployment decision is not necessary for the system. An inter-arrival time between 20-60 minutes is assumed to be good enough.

Figure 5.7 indicates the incident data meets the expectations. Good dispatching and redeployment decisions may benefit the system by finding the best configuration for the deployment of emergency vehicles.

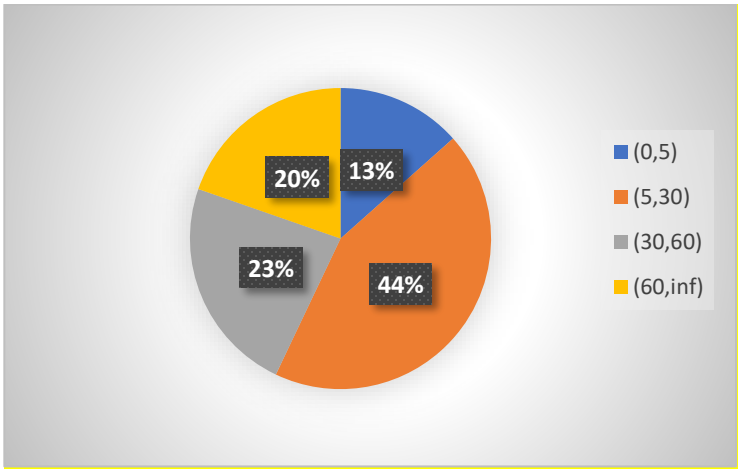


Figure 5.7 Categories of interarrival time of incidents in Baltimore City in January 2019

5.2.4 En-Route time

Vehicle En Route time represents the time period from the vehicle leaving the station for the incident site to the arrival at the incident site. A total of 15,162 incident records were used to generate the En Route time for each dispatching. According to analysis results shown in **Error! Reference source not found.** and **Error! Reference source not found.**, the best distribution is Lognormal Distribution with a squared error equal to 0.0869 for nighttime and 0.2715 for the daytime. The comparison of histograms of the real-world data and the fitted data is shown in

Figure 5.8.

5.2.5 Service time

The service time is calculated as the difference between when the vehicle arrived at the site and when the same vehicle departed from the site. A total of 15,162 records were used to generate

the service time for each rescue. According to the test result shown in **Error! Reference source not found.** and **Error! Reference source not found.**, the best distribution is Lognormal Distribution with a squared error equal to 0.0258 for the nighttime and 0.0323 for the daytime. The comparison of histograms of the real-world data and the fitted data is shown in

Figure 5.9.

Table 5.5 The results of the estimation of the distribution of emergency incidents En Route time during the night

Fitted Function	Square Error
Normal	0.09849947
Exponential	0.12528977
Gamma	0.08706707
Lognormal	0.08690945
Beta	0.08708189
Burr	0.08676187

Table 5.6 The results of the estimation of the distribution of emergency incidents En Route time during the day

Fitted Function	Square Error
Normal	0.28418342
Exponential	0.31233214
Gamma	0.27086472
Erlang	0.27086462
Lognormal	0.27156524
Beta	0.27087528
Burr	0.27180689

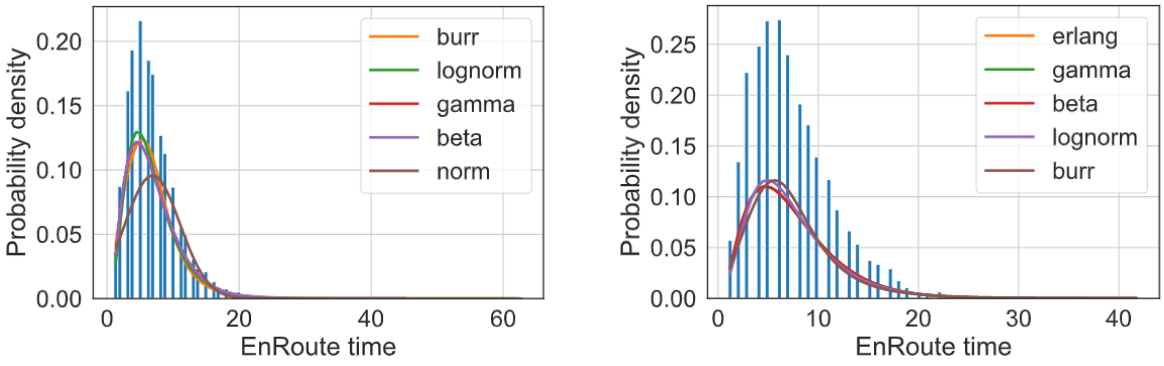


Figure 5.8 The histogram of the En Route time for incidents during the night (left) and the day (right) in Baltimore City in 2019 vs. fitted models

Table 5.7 The results of the estimation of the distribution of emergency incidents service time during the night

Fitted Function	Squared Error
Normal	0.028858237
Exponential	0.048426183
Gamma	0.025827384
Lognormal	0.025833742
Beta	0.025831267
Burr	0.02603148

Table 5.8 The results of the estimation of the distribution of emergency incidents service time during the day

Fitted Function	Squared Error
Normal	0.03584363
Exponential	0.058591606
Gamma	0.032502182
Lognormal	0.032375977
Beta	0.032500347
Burr	0.032720868

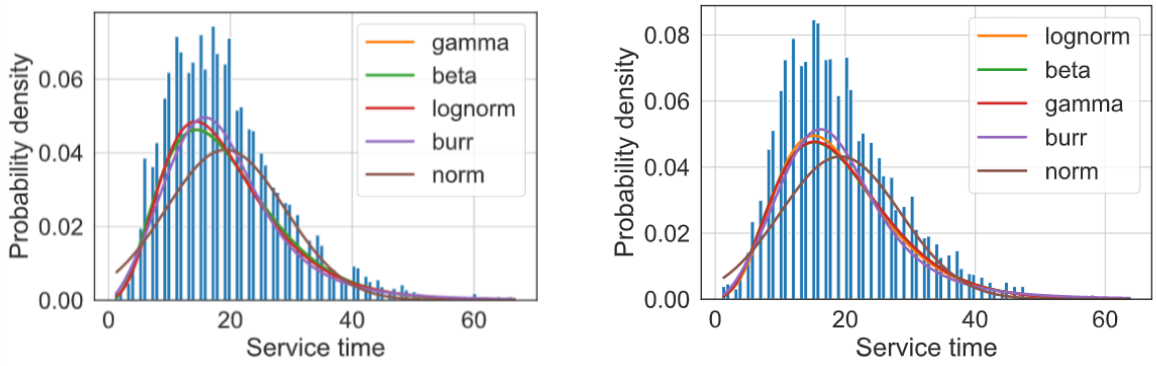


Figure 5.9 The histogram of the service time for incidents during the night (left) and the day (right) in Baltimore City in 2019 vs. fitted models

The call characteristics for the emergency medical demand data in the study are shown in Table 5.9 to Table 5.14. Table 5.9 represents the number and ratio of choosing the destination of incidents. It indicates that 73% of patients have been sent to the closest facility. Nearest dispatching can make sure the patients get access to immediate hospital treatment. Sometimes, the closest facility is not always an optimal solution. Choosing a destination highly depends on the capacity of hospitals, what treatments the patients need, and what treatments the hospitals can provide. Among the incidents, 10% of patients have been escorted to the designed hospital according to the protocols.

Table 5.9 Comparison of the number and ratio of incidents managed through dispatching policies

Closest Facility	8609	0.73966836
Protocol - When NOT closest Hospital/Specialty Center	1194	0.10258613
State Specialty Center	723	0.06211874
Patient's Choice	534	0.04588023
On-Line/On-Scene Medical Direction	271	0.02328379
Other	234	0.02010482
Diversion	74	0.00635793

According to the study data, of the 15,162 incidents, 11,618 incidents indicate the patients requiring continued hospital care, which is 76% of the incidents. Based on the patients' hospital

records, 22 hospitals took on the main task of treatment, and some other patients, smaller than 1.2%, were transferred to other hospitals. The distribution of these hospitals as final destinations is shown in

Table 5.10.

In rescue missions, ALS and BLS ambulances perform the main dispatch and escort tasks. For some specific or urgent rescue tasks, other means of rescue, such as an aircraft, will also be included. A variety of rescue methods can provide maximum coverage in different ways to provide more timely assistance to patients. Table 5.11 shows detailed information on the number and ratio of incidents handled by different rescue transportation modes.

For ambulance rescue mode, different types of ambulances undertake different treatment tasks. Table 5.12 describes the detailed dispatching records. It presents that ALS and BLS ambulances are each responsible for nearly half of the incident's responses. A better dispatching and redeployment decision strategy can vastly enhance the role of ALS vehicles in rescue and provide more professional pre-hospital treatment for the patients.

Table 5.10 Distribution of incidents between public hospitals

Public Hospital 1	1784	0.15355483
Public Hospital 2	1482	0.12756068
Public Hospital 3	1158	0.09967292
Public Hospital 4	1098	0.09450852
Public Hospital 5	1062	0.09140988
Public Hospital 6	927	0.07978998
Public Hospital 7	803	0.06911689
Public Hospital 8	633	0.05448442
Public Hospital 9	496	0.04269237
Public Hospital 10	462	0.03976588
Public Hospital 11	452	0.03890515
Public Hospital 12	445	0.03830263
Public Hospital 13	391	0.03365467

Public Hospital 14	124	0.01067309
Public Hospital 15	58	0.00499225
Public Hospital 16	55	0.00473403
Public Hospital 17	51	0.00438974
Public Hospital 18	36	0.00309864
Public Hospital 19	31	0.00266827
Public Hospital 20	25	0.00215183
Public Hospital 21	20	0.00172147
Public Hospital 22	9	0.00077466
Other Hospital	16	0.00137717

Table 5.11 Comparison of the number and ratio of incidents handled by different rescue transportation modes

Ground - Ambulance	11337	0.97113243
Ground - ATV or Rescue Vehicle	253	0.02167209
Air Medical - Rotor Craft	66	0.00565359
Ground - Other Not Listed	18	0.00154189

Table 5.12 Comparison of the number and ratio of incidents serviced by different ambulances

ALS - Paramedic	6460	0.42606516
BLS - EMT-IV	5832	0.38464583
BLS - EMT	1781	0.11746471
ALS - CRT	1067	0.0703733
BLS - First Responder/EMR	14	0.00092336
ALS - Community Paramedicine	8	0.00052763

Levels of care refer to the comprehensiveness of services hospitals provide to patients. Hospitals are certified based on services available in their emergency departments, inpatient units, and on-campus outpatient clinics. Table 5.13 compares the number and ratio of incidents handled by different level hospitals.

Table 5.13 Comparison of the number and ratio of incidents handled by different hospitals

Hospital (General)	3787	0.64002028
MD SAFE Hospital	877	0.148217
Level II Trauma Center	621	0.10495183
Level I Trauma Center	559	0.09447355
Burn Center	47	0.00794321
Others	26	0.00439412

The priority level is a selection made during incident creation that conveys the severity of an incident so that responders can react accordingly. According to the Maryland Institute for Emergency Medical Service System, priority 1 is a critically ill or injured person requiring immediate attention. From priority 1 to priority 4, the degree of urgency of the accident decreases in order, as does the degree of injury to the patients. Table 5.14 keeps a record of the incidents' priority. The data can help the decision model to predict the level of injury.

Table 5.14 Distribution of the number and ratio of incidents priority

Priority 3	10505	0.69606414
Priority 2	3346	0.22170686
Priority 1	1038	0.06877816
Priority 4	114	0.00755367
Dead without Resuscitation Efforts	89	0.00589716

5.3 Case study zone importance

5.3.1 Baltimore city original incidents data

The geospatial and temporal distribution of historical incidents is shown in

Figure 5.10 and

Figure 5.11. The height of the bars represents the number of incidents, which also means the incidents rate.

Figure 5.10 illustrates the incidents that occurred in Baltimore City in 2019 during the day, while

Figure 5.11 shows the incidents during the night. In the figure, areas are marked with different colors based on the number of incidents that occurred in this zone. They are red, orange, yellow, and green from high to low. Different colors are used to distinguish the area and compare different incident rates.

Figure 5.10 illustrates incidents concentrated in the downtown area. The number of incidents in the surrounding areas is significantly lower than in the downtown area. The number of incidents in certain surrounding regions is also in the top 25 percent. During the night, the number of incidents is significantly less. The areas where incidents are concentrated are reduced accordingly. However, they are still concentrated in the middle of the downtown area.

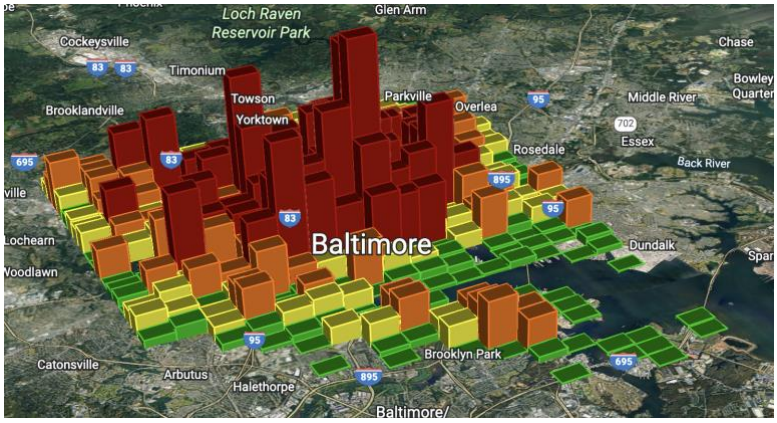


Figure 5.10 Geospatial and temporal distribution of zone importance in Baltimore City in 2019 during the day

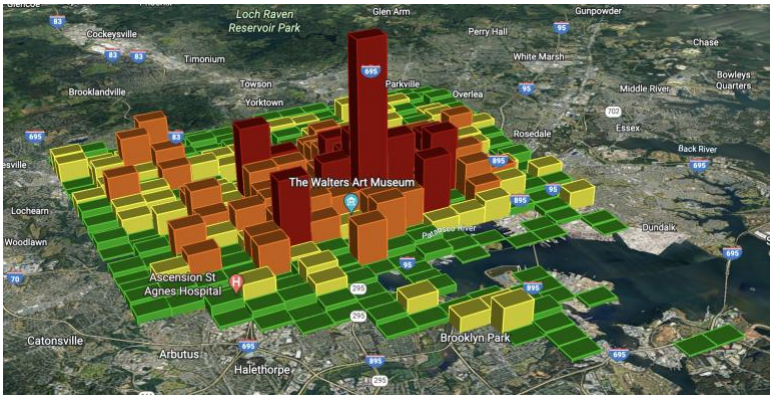


Figure 5.11 Geospatial and temporal distribution of zone importance in Baltimore City in 2019 during the night

5.3.2 Baltimore City incidents data adjusted by priority

The number of incidents is a critical metric to evaluate the importance of zones, while the severity of incidents also impacts the importance of zones. If the severities of incidents in a specific area are very high, the area needs more protection and coverage. In the case study, if there is no specific regulation of coverage for some specific areas, such as hospitals, schools, or other important places, then the incident rate and severity of incidents can be used to evaluate the importance of zones. According to the Maryland Institute for Emergency Medical Services Systems (MIEMSS), clinical priority is divided into four levels. The strategy for generating zones' importance is to use the number of incidents that occurred in the zone multiplied by the priority of each incident divided by four.

Figure 5.12 and

Figure 5.13 show the geospatial and temporal distribution of the importance of zones adjusted by priority.

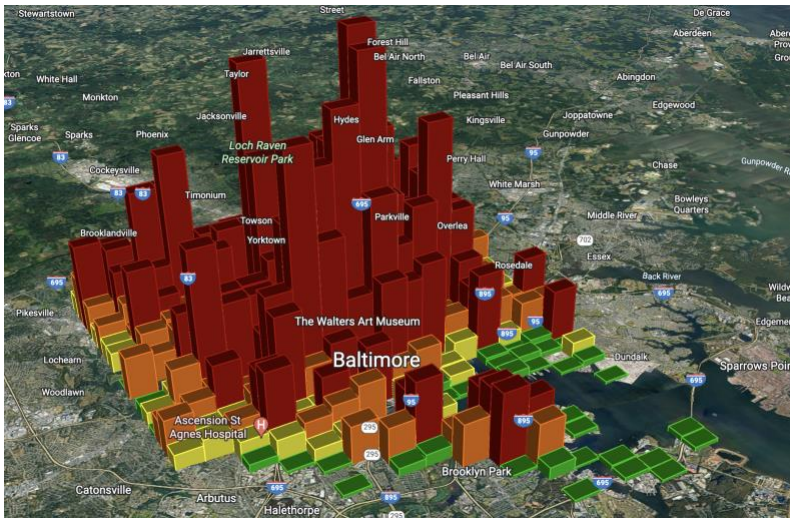


Figure 5.12 Geospatial and temporal distribution of zone importance adjusted by priority in Baltimore City in 2019 during the day

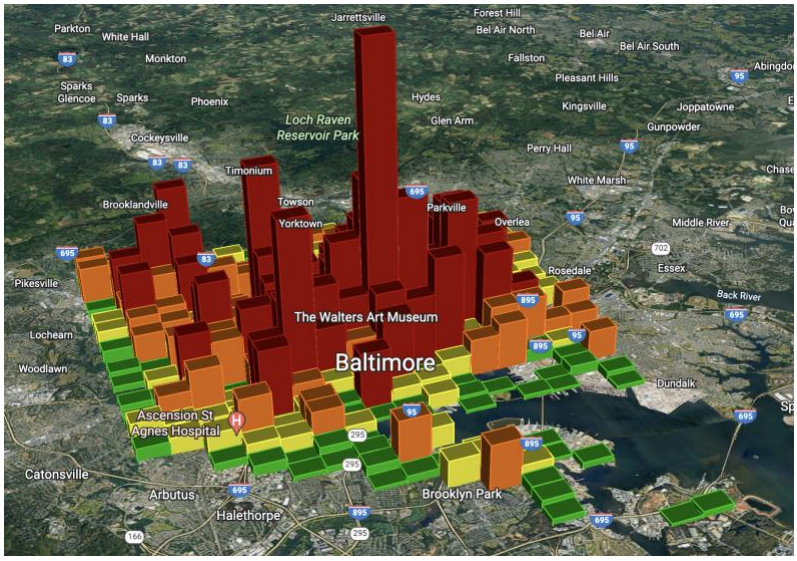


Figure 5.13 Geospatial and temporal distribution of zone importance adjusted by priority in Baltimore City in 2019 during the night

Compared with
Figure 5.10 and
Figure 5.11 without considering the priority, in
Figure 5.12 and
Figure 5.13, important areas extend from the central area to the surrounding areas. More areas need to be focused.

5.3.3 Baltimore city categorical incidents data

The above graphs show a huge difference in the height of different areas. According to the records of incidents, the number of incidents in some specific areas may exceed one hundred in a year, while in some surrounding suburban areas, the number of incidents is less than ten. If the importance of zones is expressed in terms of the real number of incidents, there is a probability that emergency vehicles would abandon the surrounding suburban areas and converge on the central area of the city. This phenomenon is not acceptable. Thus, instead of using ordinal numbers to express the zones' importance, using categories to express the zones' importance can avoid this problem.

Figure 5.14 and

Figure 5.15 illustrate the importance of zones using categories. According to the incident rate, the whole city is divided into four categories to indicate the importance of zones.

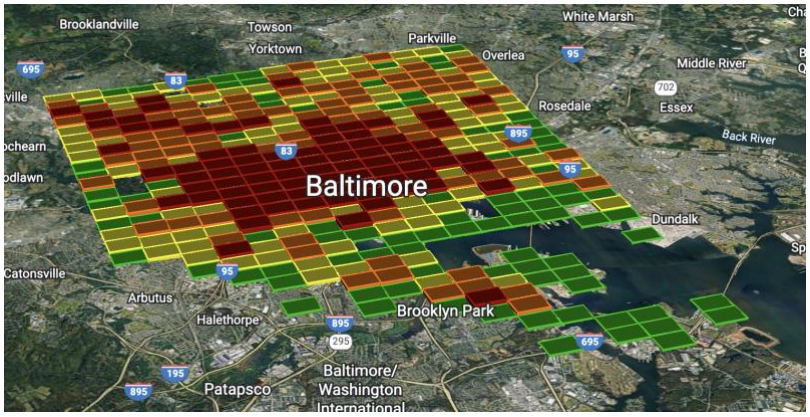


Figure 5.14 Geospatial and temporal distribution of categories of zone importance in Baltimore City in 2019 during the day

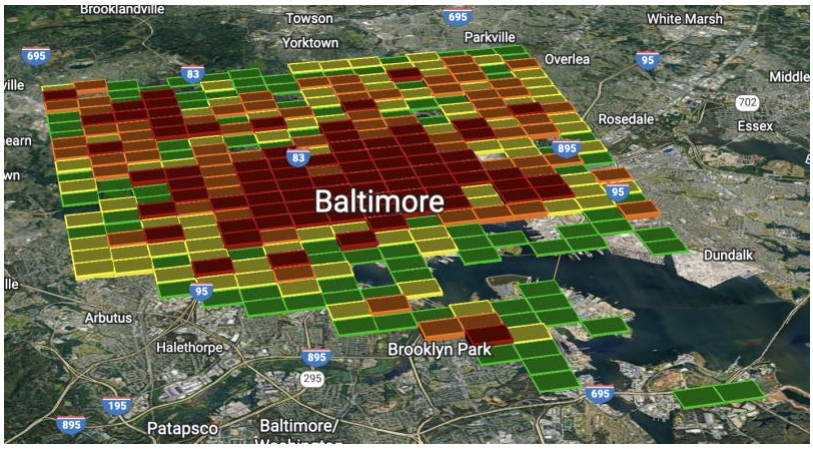


Figure 5.15 Geospatial and temporal distribution of categories of zone importance in Baltimore City in 2019 during the night

5.3.4 Baltimore city categorical incidents data adjusted by priority

Based on the above arguments,

Figure 5.16 and

Figure 5.17 represent the categorical zone importance adjusted by priority. When the priority is considered, the important areas change a little.

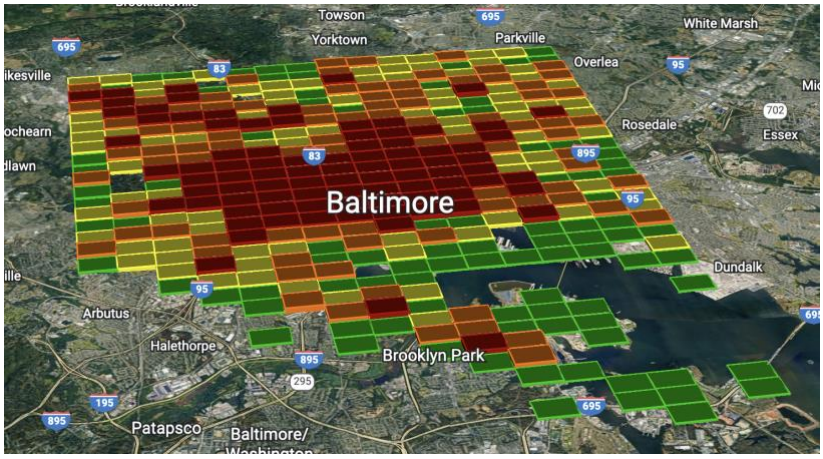


Figure 5.16 Geospatial and temporal distribution of categories of zone importance adjusted by priority in Baltimore City in 2019 during the day

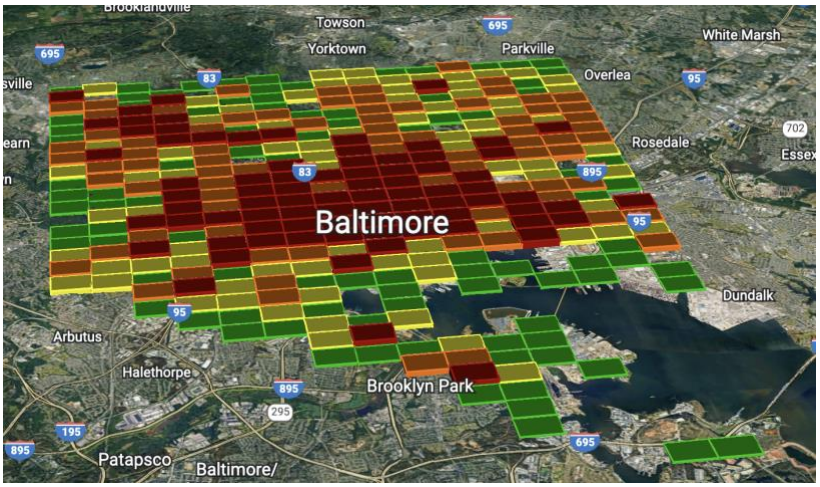


Figure 5.17 Geospatial and temporal distribution of categories of zone importance adjusted by priority in Baltimore City in 2019 during the night

Different strategies were proposed above, and corresponding figures were plotted to illustrate the geospatial and temporal distribution of zone importance. The performance of different strategies on the final results varies from case to case. In this study, models with different strategies will be tested and compared in the following chapter to determine the performance of strategies. The strategy with the best performance will be selected for further analysis.

Chapter 6 : Discrete Event Simulation

In Chapter 3, an integrated dispatching and redeployment decision-making model was proposed. At any given moment, when the exact status of the system and demand requests are known, the model can provide the best dispatching and redeployment decisions. Chapter 5 provided a comprehensive data analysis. The data generated from the real-world operations for the ambulances and medical units for the incidents that occurred in Baltimore City has a record of vehicle dispatching trajectory and real operational times. So, a simulation procedure is critical and necessary to see how the proposed model performs in a real-world case study. In this chapter, a discrete event system simulation model is built and explained in detail. The framework of this discrete event simulation model can mimic the evolution of the entire operation of an emergency response system over time. The sophisticated discrete event simulation model is designed and coded in Python programming language.

6.1 Introduction to the discrete event system

The system simulation model is event-based and is evolved whenever there is an event in the system. Discrete event-based methods are applicable to systems that can be interpreted as a set of interrelated entities which can change their status at discrete time points and, as a result, can cause a change in the system's state (Ullrich & Lückerath, 2017).

The potential behavior of an entity that may cause the change of system can be defined as an event. In this study, the events can be:

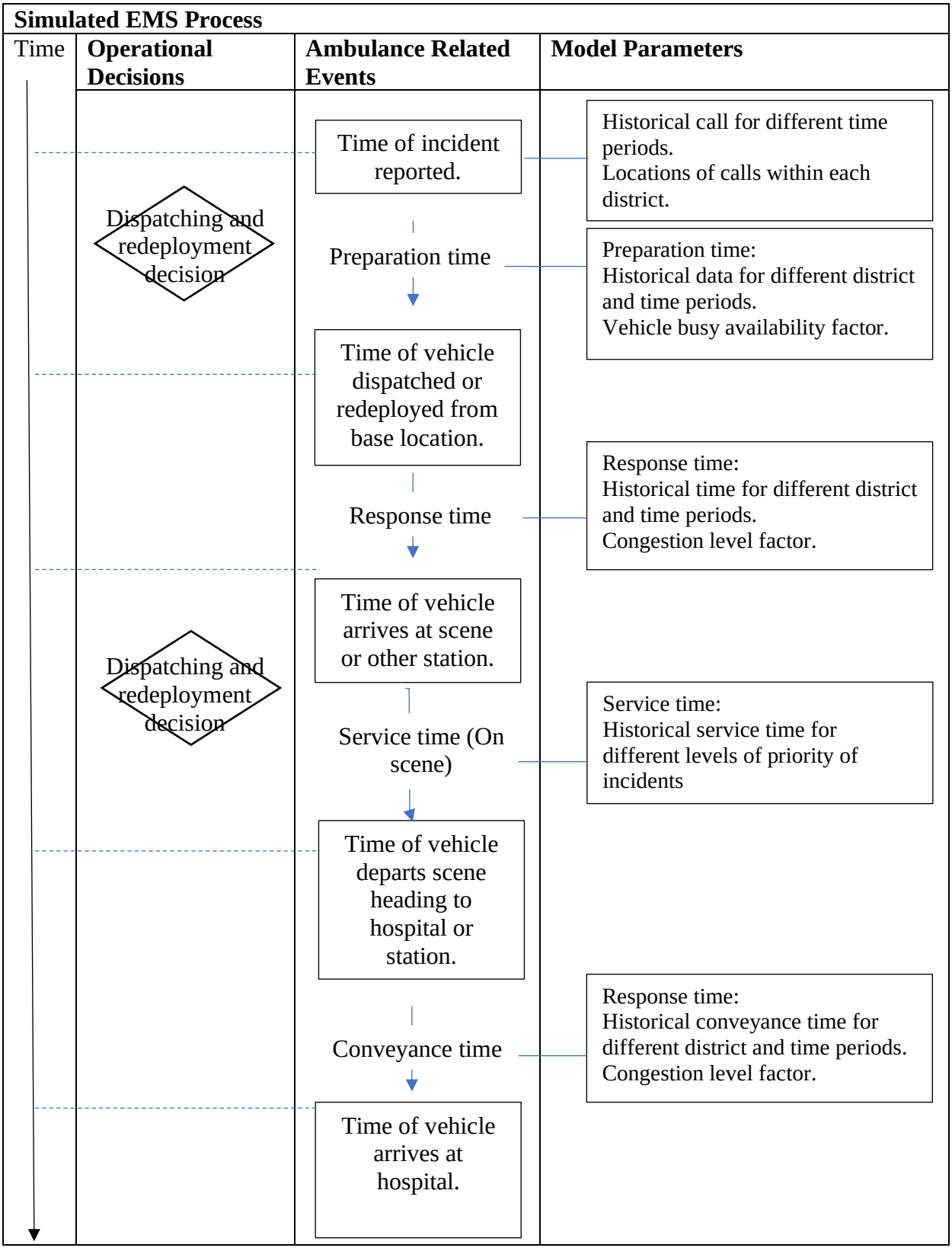
- Occurrence of an emergency: when an emergency call comes into the system, the dispatching and redeployment decisions need to be made for the available emergency

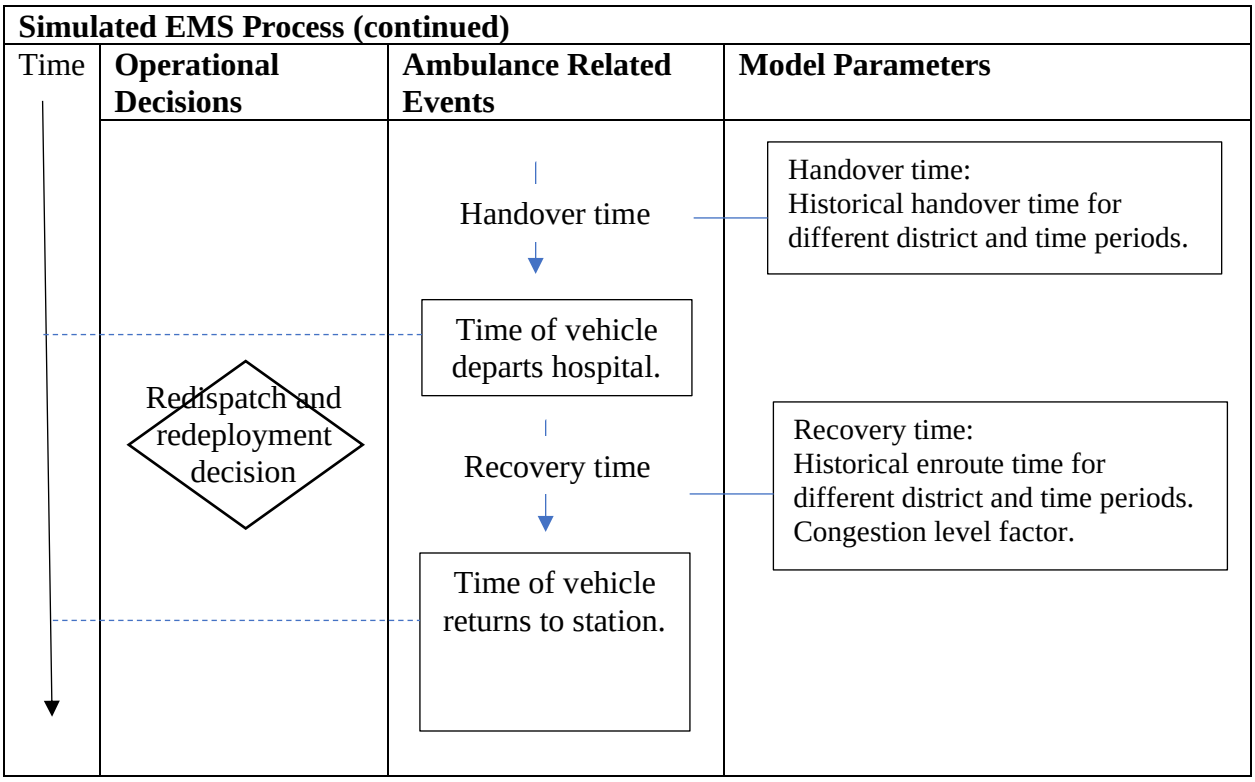
vehicles. If there are not enough vehicles to be dispatched to deal with the tasks, the call will be put into a queue system.

- Change in the status of vehicles: When the status of emergency vehicles changes, the total expected coverage for the system may vary. Moreover, a subsequent decision must be made for the vehicles that have completed their tasks. So, the change in the status of vehicles can be defined as an event that may cause a change in the system's status.
- Change in the traffic data: The integer model primarily relies on the input data to make decisions. Any change in the input data may result in a different decision. For example, when the traffic conditions vary, the model needs to be called to calculate the shortest path to escort the patients to the designed hospitals.
- Change in the likelihood of an emergency in demand zones: For the integrated integer model, both dispatching and redeployment decisions must be made simultaneously to reduce the rescue time and maintain the maximal coverage for the whole area. The coverage calculation largely relies on each zone's importance, which is the likelihood of an emergency happening in the demand zone. If the likelihood changes, the deployment configuration may be changed accordingly.

A detailed process flowchart for the simulated EMS process is shown in Table 6.1. The table defines the operational decisions that evolve with time, and the ambulance-related events in each period are also defined. When the events come into the system, the data and parameters required for the proposed model to make decisions are also described. The detailed discrete event simulation system is built according to the flowchart for the simulated EMS process.

Table 6.1 Process flowchart for simulated EMS process





6.2 Conceptual framework of the simulation model

Figure 6.1 illustrates the conceptual framework of the discrete event simulation model, developed for this research and implemented in Python. During the simulation process, the simulation engine recursively retrieves and removes the event according to the minimum time stamp.

Several modules and components are identified in the simulation, including their attributes, relations, activities, and event. These modules are:

- Travel time module
- Vehicle module
- Emergency module
- Emergency call module

- Statistics module
- Traffic condition module
- Busy factor module
- Optimization module

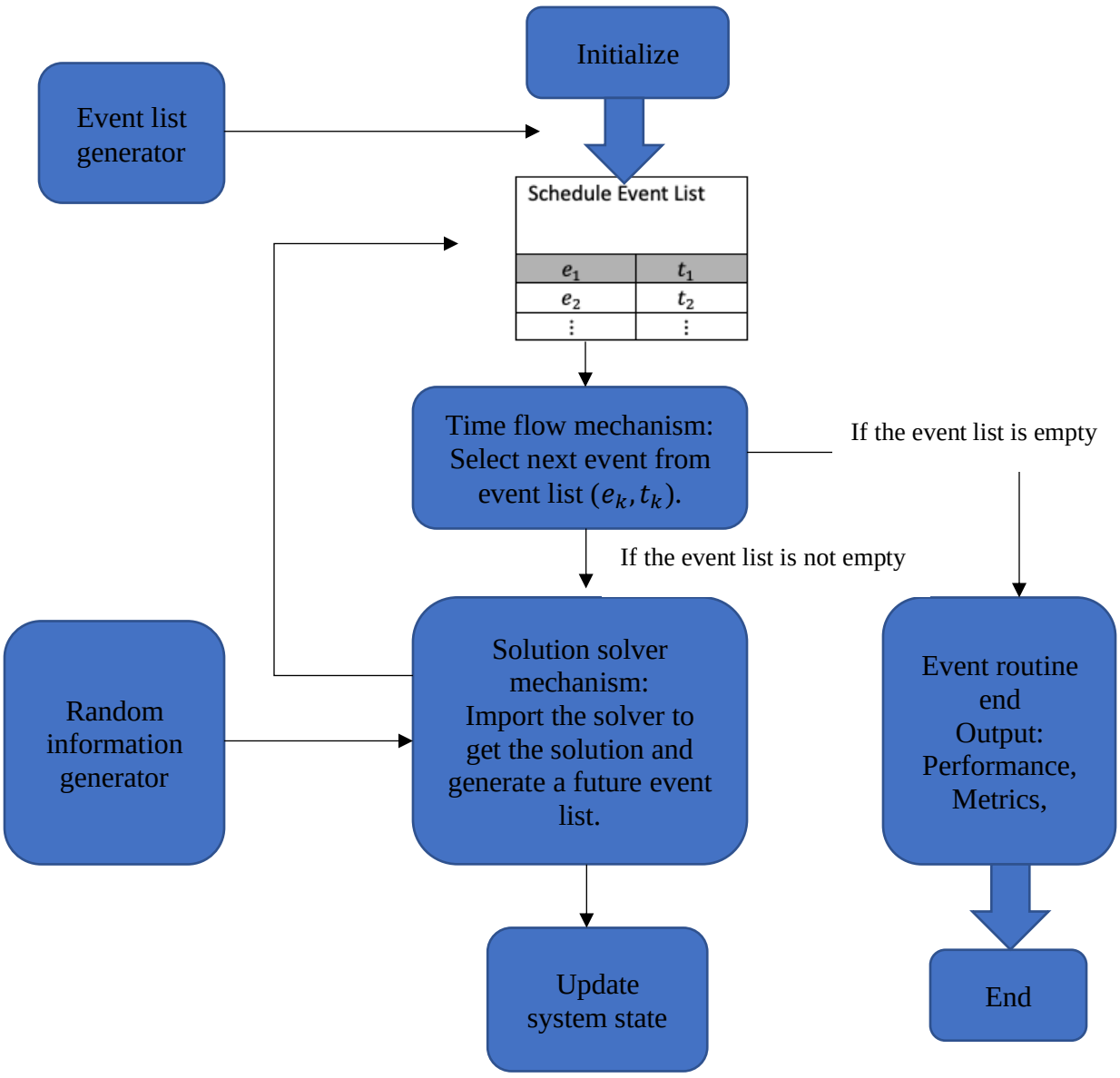


Figure 6.1 The event scheduling scheme in EMS simulation

6.2.1 Travel time module

The calculation of expected travel time is based on the existing street network of the designed area. Accurate and preprocessed spatiotemporal historical travel time is critical to the decision model and final result.

Google Maps is a web mapping service that leverages GPS crowdsource to retrieve accurate traffic data and provide access to the traffic data publicly, but with limited features and requires further pre-processing. Google API can only provide historical travel data for public transportation, and historical traffic data and congestion levels are not accessible from Google Maps.

The Open Source Routing Machine (OSRM) is an open-source router designed for use with data from the OpenStreetMap project. It combines sophisticated routing algorithms with the open and free road network data of the OpenStreetMap (OSM) project. Moreover, OSRM can compute and output the shortest path between any origin and destination pair. In this research, The OSRM will be used in Python to compute the travel time between any pair of nodes with the free flow speed. Each time an incident is identified, and the GPS location of the incident is given, the OSRM is imported to calculate the distance matrix between the incident site and stations and the distance matrix between the incident site and all hospitals.

6.2.2 Vehicle module

The vehicle module is designed to track and update the status and locations of the vehicles. At a specific time stamp, the optimization solver module is imported to make dispatching and redeployment decisions to satisfy the demand requests and find the best configuration for the deployment. The job assignment and route choice for each vehicle are provided. Meanwhile, each vehicle's location and destination are updated after the decision-making. The vehicle module is called to update the status and location of each vehicle.

6.2.3 Emergency module

The emergency module is used to track and update the status of the emergencies in the system. When new emergency calls come into the system, based on the arrival time, location, severity, priority, and request information, the emergency module is used to check if the emergencies are fully serviced or if they are in the queue system waiting for the response. The emergency module keeps tracking the status of the emergency request until it is fully serviced.

6.2.4 Emergency call module

This module carefully defines the emergency call arrival time, location information, request information, status severity, and priority. According to the historical medical data, the spatial distribution and geographical distribution are known in advance. The historical emergency calls are imported by using the module.

6.2.5 Statistics module

The statistics module, also called the performance module, is designed to collect the metrics and performance measure parameters. Several performance measures will be collected during the simulation to fairly and comprehensively reflect the system's performance and compare different methods under uncertainty. These include:

- Performance related to the dispatching efficiency, including the response time, en route time, the number of calls that can be reached within the required time, and the total coverage for the whole system.
- Performance related to the system operating cost, including the number of relocations, total relocation distances, etc.
- Performance related to the workload, including the average working time for the crew members and the average relocation times.

6.2.6 Traffic condition module

When travel times are needed, the model cannot access the historical travel times due to data source issues. The OSRM package generates the static, optimal travel time matrix for the calculation. Without considering the traffic condition, the computational result may be inaccurate and inapplicable to the real world. Exogenous information is needed to evaluate the model's reaction to specific disruptions or disturbances. A traffic condition module is developed to generate disruptions that are systematically injected into the model as exogenous information. The traffic condition module generates random traffic jams for each timestamp based on a pre-defined congestion level for each road segment. Different congestion levels will be systematically generated and used in the corresponding simulation. Its effectiveness can be examined systematically.

6.2.7 Busy factor module

Similar to the traffic condition module, the busy factor module is designed to generate exogenous information for the vehicles. In the actual dispatching operations, the availabilities of emergency vehicles at that moment are not accessible. Ignoring this factor would overstate the optimization effect of the new model. The mechanism of the busy factor module is using a Binomial distribution to generate a Boolean-valued outcome based on a pre-defined probability factor. According to the Boolean-valued result, the status of vehicles at each time stamp is available. The available vehicles will be selected to fulfill the dispatching or redeployment tasks.

6.2.8 Optimization module

This module imports the integrated integer programming model, which is the core module in the simulation system. The purpose of the optimization module is to calculate the best decisions based on the current status of the system. This module will be called whenever an event comes

into the system. After the dispatching or redeployment decisions have been made, the subsequent tasks for the corresponding vehicles will be added to the future event list.

Once the components of the discrete event simulation system are identified, including their attributes, relations, activities, and events, the model is implemented using a programming language. The simulation events are managed by the Future Event List (FEL). This data structure is managed as a priority queue, ordered by a key parameter, such as a time stamp.

Figure 6.1 depicts the model's operational logic in detail. During a simulation run, the system will repeatedly retrieve the next event from the Future Event List (FEL) according to the time stamp, import the corresponding module to execute it, and remove it after finishing. The simulation time will be set to the currently executed event's time stamp in these processes. The generated future events will be added to the FEL if applicable, and the system's state will be updated. The simulation will run until the stop conditions (such as an empty FEL) are fulfilled.

6.3 Summary

In this chapter, a sophisticated discrete event simulation system was defined. Each component in the system was identified, and the logic behind the simulation was built up. A large-scale case study was conducted using the abovementioned discrete event simulation framework. The results and analysis will be presented and discussed in the next chapter.

Chapter 7 : Case Study Results

In previous chapters, the proposed integer programming model was tested in different numerical cases to prove the capability to handle multiple tasks in different scenarios. A discrete event simulation (DES) system is built to test the performance of the integer programming model in actual operations during a long time horizon. This chapter compares the results of applying the proposed model in the discrete event simulation system in the real-world case study with the actual operation data. Then an extensive sensitivity analysis is performed on the parameters in the model to test how the model will perform under various conditions.

7.1 Introduction to case study result analysis

In Chapter 6, we carefully defined the input data source at different time periods and developed the basic simulation system framework. The simulation model is applied to the real case data, and the performance measures are collected and evaluated.

In the output analysis, three different categories of statistics are analyzed, which are:

- Performance related to the dispatching efficiency, including the response time, en route time, the number of calls that can be reached within the required time, the total coverage for the whole system, and basic coverage, double coverage, and fully functional double coverage level for the entire system.
- Performance related to the system operating cost, including the number of relocations, etc.
- Performance related to the workload, including the average number of dispatches for each vehicle and the workload distribution for different types of vehicles.

Chapter 6 proposed four different strategies regarding the zones' importance. In section 7.2, the proposed discrete event simulation system is used to implement models with different importance of zones. Three categories of performance measures mentioned above are collected and compared among four strategies regarding the zones' importance. Finally, the strategy with the best performance is selected as the baseline model and used in the following output and sensitivity analyses.

7.2 Zone importance strategies analysis

Table 7.1 summarizes the detailed characteristics of the set of experiments that implemented the models with different importance of zones. For each metric, specific graphs are used to illustrate the model's performance in detail.

Table 7.1 Summary of performance measurements of four strategies regarding the zones' importance

	Strategy 1	Strategy 2	Strategy 3	Strategy 4
En Route Time	4.7910	4.7911	4.8086	4.8086
Actual Total Coverage Rate	9.1458	9.1108	6.9846	6.9618
Total Coverage Rate	11.1194	11.0113	7.8338	7.7851
Total Coverage Improvement	17.7%	17.2%	10.8%	10.57%
Basic Coverage	0.8916	0.8912	0.8896	0.8896
Double Coverage	0.6459	0.6454	0.642458	0.642470
Fully Functional Double Coverage	0.5641	0.56408	0.5563	0.5563
Number of Incidents Reached by BLS	158	158	160	160
Number of Incidents Reached by ALS	626	626	624	624
The average number of redeployments	0.7665	0.8482	0.8048	0.9145

7.2.1 En Route time

Table 7.1 illustrates the average En Route time for the four strategies. The maximum difference between the four values is less than one percent, which indicates that the models with different zone importances can achieve similar performance. This is because the system has multiple emergency medical vehicles on standby. Even when the strategies differ, the system can achieve excellent coverage in most situations. The vast majority of incidents can have the nearest vehicle dispatched. Only in a few cases, no vehicles are available for dispatch from the nearest station because of the deployment configuration. This leads to a different En Route time. Figure 7.1 compares En Route time for different zone importance strategies. The variation of En Route time among the four strategies is so minimal that it is difficult to tell the difference, which is consistent with the results in the table. Strategy 1 can achieve the shortest En Route time among the four strategies.

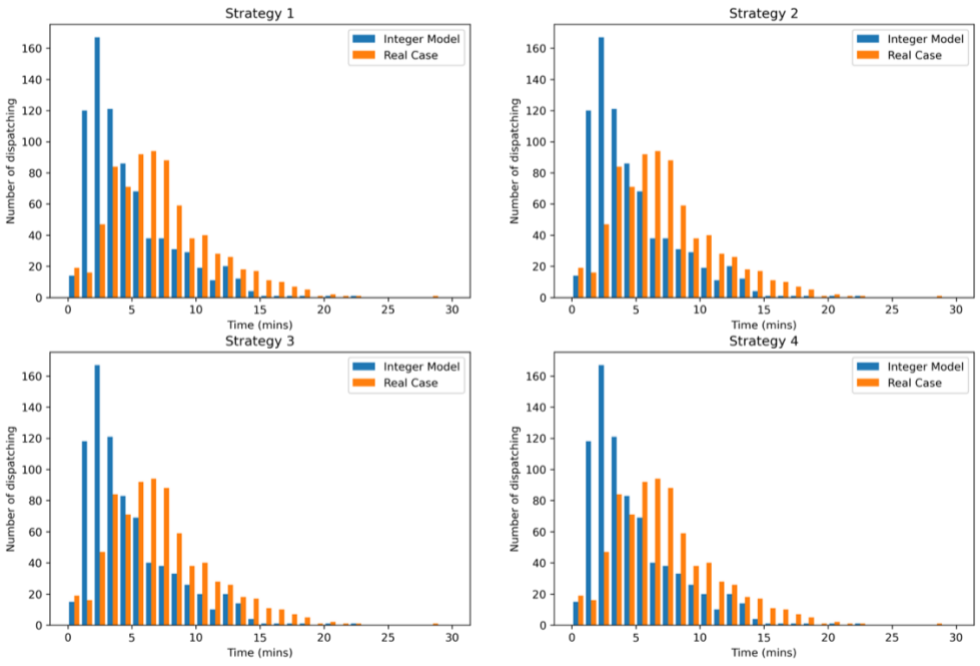


Figure 7.1 Comparison of the En Route time for different zone importance strategies

7.2.2 Coverage level

Table 7.1 illustrates the actual total coverage rate and total coverage rate based on computation. The coverage rate is the summation of the importance of all zones, which can be reached by all stations. Because the calculation of the importance of zones is different, it is impossible to compare the coverage rates among strategies directly. To overcome these issues, the total coverage improvement is calculated to compare how much the system coverage rate is improved based on the strategy. The percentage of progress can be used to evaluate the performance of strategies. According to the final results, it is concluded that Strategy 1 and Strategy 2 can maintain a higher level of total coverage.

Figure 7.2 compares total coverage rates among the real-case data and models with different strategies. Table 7.1 only reflects the increase in coverage adopting different strategies. Figure 7.2 clearly displays the change in total coverage rates as events enter the system. Compared with Strategy 2, Strategy 1 can maintain a higher level of total coverage throughout the calculation period. The total coverage rates based on the models can outperform the results according to the real-case data in almost all scenarios. While Strategy 3 and Strategy 4 can also significantly improve total coverage compared to the real-case data, the percentage of improvement is less than Strategy 1 and Strategy 2.

Moreover, according to the figure, in a few extreme cases, the total coverage rates of Strategy 3 and Strategy 4 drop to a worse level than the real-case data. This situation never happens to Strategy 1 and Strategy 2. This means Strategy 1 and Strategy 2 are more suitable than the other two strategies in real-life applications.

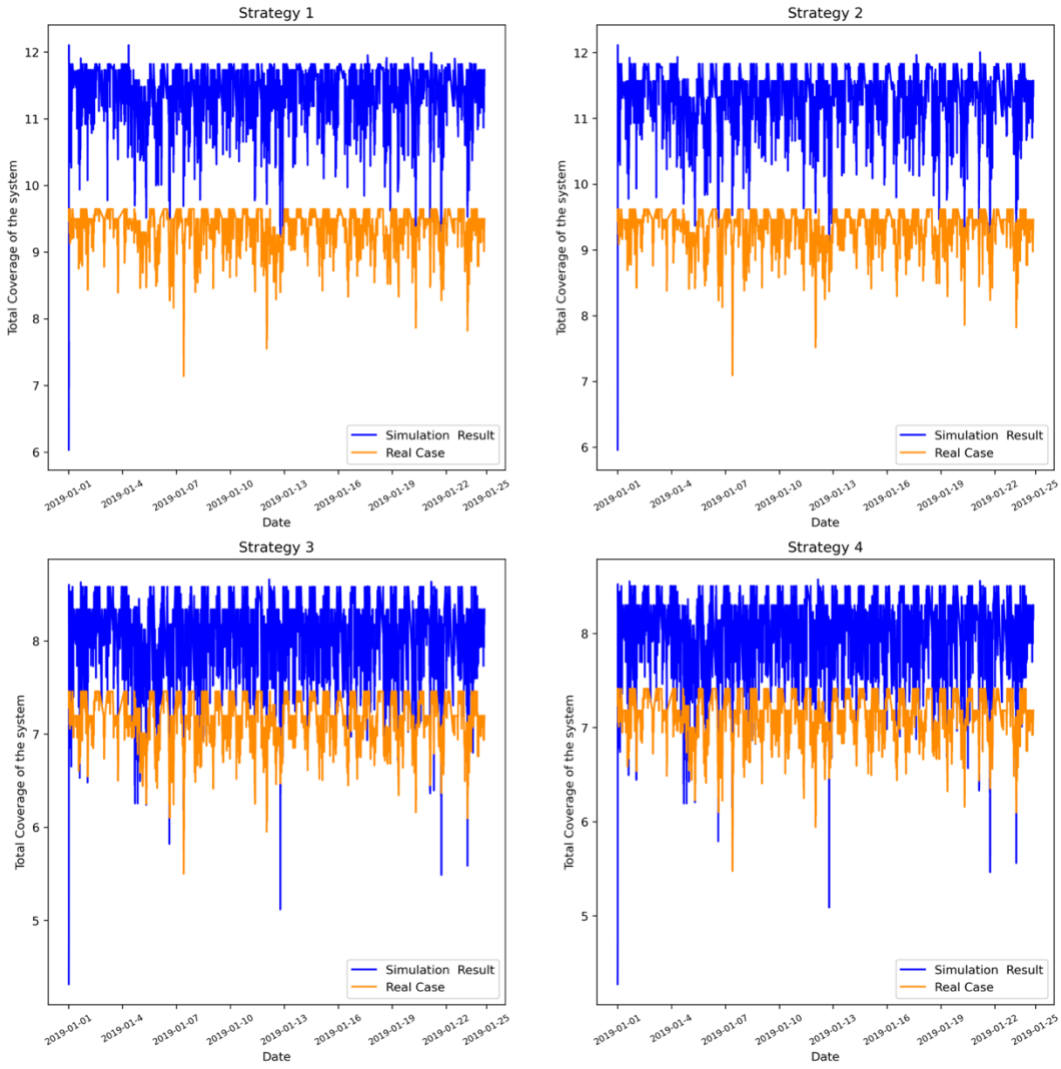


Figure 7.2 Comparison of the total coverage rate of the real case data with models with different zone importance strategies

Besides the total coverage rate, basic, double, and fully functional double coverage rates are also calculated. This is to avoid the fact that some models only emphasize the total system coverage rates and provide more protection for the incident-intensive areas, which reduces the basic coverage of surrounding areas. These three parameters, basic coverage rate, double coverage rate, and fully functional double coverage rate, are thus introduced into the system. The results indicate that four strategies can maintain a high level of similar results on these three parameters. Strategy 1 narrowly outperforms other strategies. Strategy 1 can cover a broader area while maintaining

maximum system coverage, allowing more areas to gain basic coverage rate, double coverage rate, and fully function double coverage rate.

Figure 7.3 shows the comparison of basic coverage between models with different strategies. In the figure, most of the four curves overlap, which means the basic coverage can be maintained at a similar level among the four strategies. The basic coverage level can be kept at a stable value of about 0.9. The variation of basic coverage for Strategy 1 is more stable than for Strategy 2 and Strategy 4.

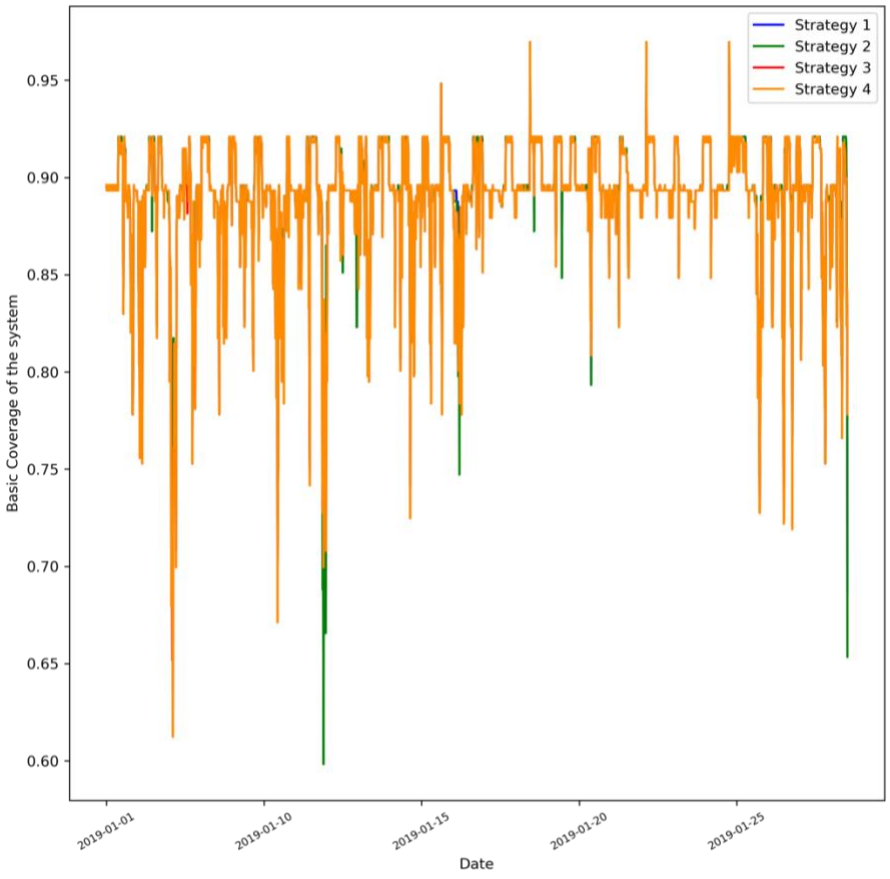


Figure 7.3 Comparison of the basic coverage for different strategies

Figure 7.4 shows the comparison of double coverage between models with different strategies. The double coverage level can be maintained around 0.65. The variation of double coverage for

Strategy 4 is larger than others. In some extreme situations, the double coverage rate can reach a level smaller than 0.5.

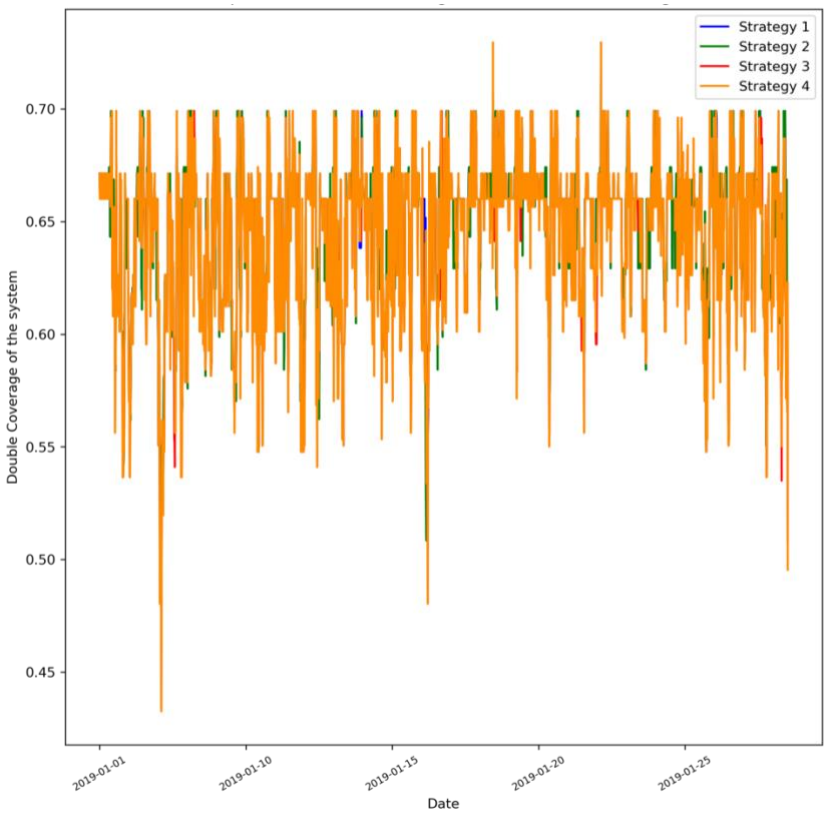


Figure 7.4 Comparison of the double coverage for different strategies

Figure 7.5 compares fully functional double coverage between models with different strategies. The fully functional double coverage level can be maintained around 0.55. Similar to the basic and double coverage rates, the variation of fully functional double coverage rates for Strategy 4 is larger than others. In some extreme situations, the fully functional double coverage rate can reach a level smaller than 0.5.

Among these three parameters, the variation of coverage level for Strategy 4 is larger than others. In some cases, the minimal coverage value for Strategy 2 is smaller than in others. In summary, Strategy 1 is the best strategy regarding these performance measurements. It can

maintain a high level of basic coverage rate, double coverage rate, and fully functional double coverage rate with fewer variations.

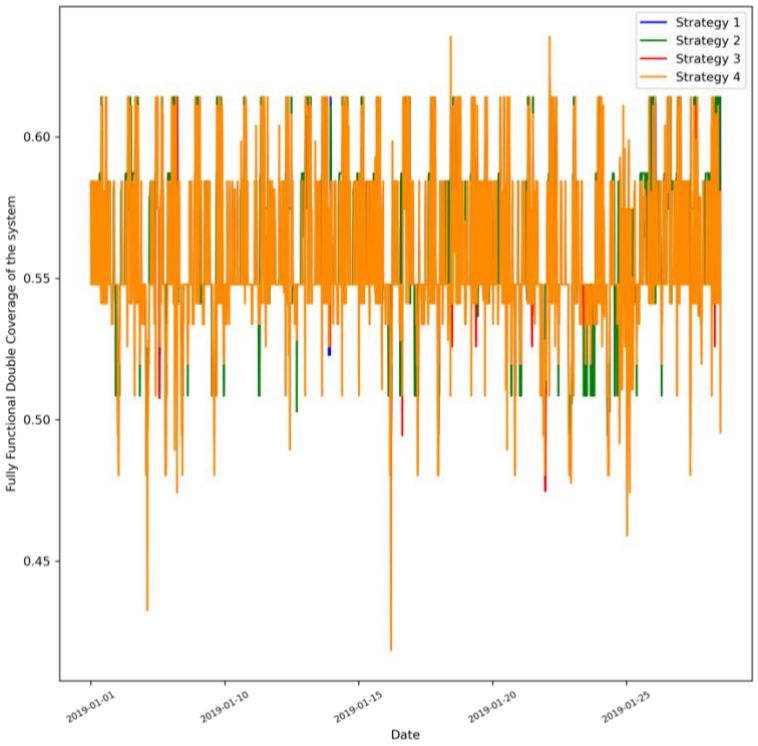


Figure 7.5 Comparison of the fully functional double coverage for different strategies

7.2.3 Workload

Meanwhile, Table 7.1 displays the workload distribution between different types of vehicles. The number of incidents reached by BLS and ALS is close among the four strategies. Strategies 3 and 4 can dispatch more advanced emergency medical vehicles to the incident sites and increase the utilization of ALS vehicles.

7.2.4 Redeployment

Table 7.1 also shows the average number of redeployments. This performance measure ensures the system can maintain the best performance with the least number of redeployments. If the average number of redeployments for each incident is too high, it means the system requires a very

high cost to maintain the optimal condition. In some cases, the high cost is unachievable. This is the reason that the parameter is introduced into the performance measures. According to the final results, it illustrates that the average number of redeployments is smaller than 1. Strategy 1 uses the least redeployments to maintain the highest level of system performance.

Figure 7.6 shows the number of redeployments when making decisions for each incident in the system. It illustrates that the number of redeployments in most conditions does not exceed one. In some conditions, the numbers of redeployments are around five. Compared with other strategies, Strategy 1 can maintain fewer redeployments.

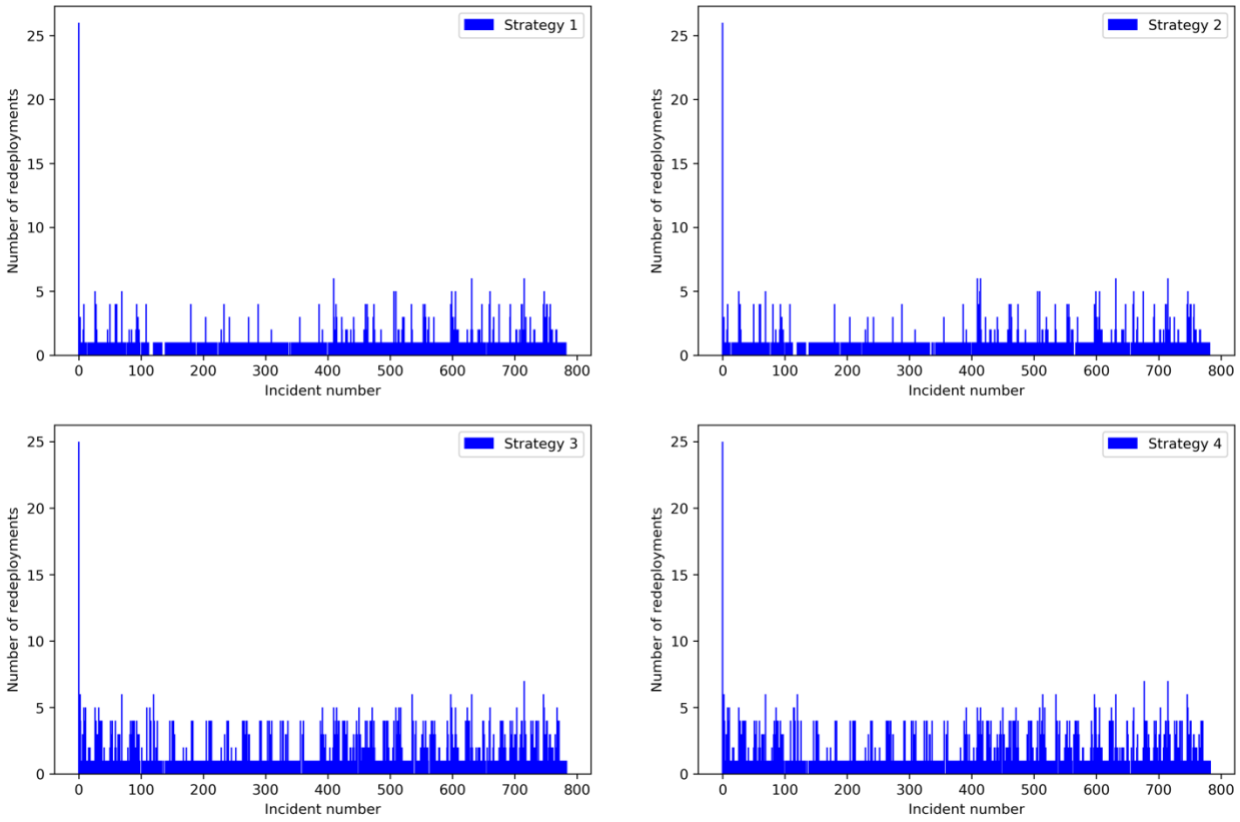


Figure 7.6 Comparison of the number of redeployments for different strategies

7.2.5 Summary

In summary, Strategy 1 for generating the zones' importance is the best among all strategies. It can not only achieve the shortest En Route time but also maintain a high level of system coverage rate. At the same time, Strategy 1 can find an optimal deployment configuration for all vehicles to maximize the basic coverage rate, double coverage rate, and fully functional double coverage rate.

Except for the total coverage improvement, all four strategies can achieve similar results in the remaining metrics. This is because the distribution pattern of the importance of zones is roughly the same across strategies. Certain areas may show large differences due to different calculation strategies, but because of the high coverage rate, most incidents can have the shortest dispatch time from the nearest stations, and the final rescue times do not differ much.

Strategy 1 will be used as a baseline in this study in the following analysis. Using the number of incidents to represent the importance of zones can not only emphasizes zones with high incident rates but also show the variation in incident distribution.

For other studies in the future, the results may be completely different. Therefore, the above analysis should be conducted to compare the performance of strategies to determine which one is more suitable for other areas.

7.3 Proposed model results

In section 7.2, the four strategies regarding the importance of zones were applied in the proposed model and discrete event simulation system. This section conducts a comprehensive output results analysis, and different categories of performance measures are collected and compared. Based on the conclusions in Section 7.2, the importance of zones generated according to Strategy 1 can help the proposed model find the best configuration for the deployment of

vehicles. Strategy 1 is used as a baseline in this section, and the simulation results generated according to the corresponding importance of zones will be compared with the real-case data. Multi-dimensional metrics are collected to estimate the performance improvement when applying the proposed model to the real-case data.

7.3.1 Total coverage

Total coverage is calculated based on the importance of zones which is calculated according to the ratio of incidents that happened in the zones. For each pair of zone and station, if the travel time is shorter than the threshold, the zone is covered by the station, and the importance of the zone will be added to the total coverage. The comparison of total coverage rates between real-case data and the simulation results is shown in Figure 7.7. It illustrates the total coverage rate provided by the emergency medical vehicles based on real-case data and computation according to the proposed model. The orange line indicates the fluctuation of the coverage level of the whole area under the actual dispatch data of emergency medical vehicles. While the blue line represents the coverage level under the calculation. Line segment fluctuations reflect emergency medical vehicles' deployment at different stations. If a vehicle is dispatched on a mission or redeployed to other stations, it is not considered available to provide coverage for the system. Therefore, the total coverage level decreases. Figure 7.7 illustrates that the average total coverage level, according to the proposed model, is around 11. While the average total coverage level, according to the actual operating data, is about 9.3. In almost all scenarios, the proposed model can increase the total coverage for the system and provide more protection than the actual operating policy.

7.3.2 Response time

Table 7.2 summarizes the statistics of En Route time between real-case data and computational data. The table records the minimal, maximal, and different percentage values for both data sets.

Except for the minimal value for the real-case data, which is 0 and better than the proposed model, in all other data, the computational results can achieve better results than real-case data.

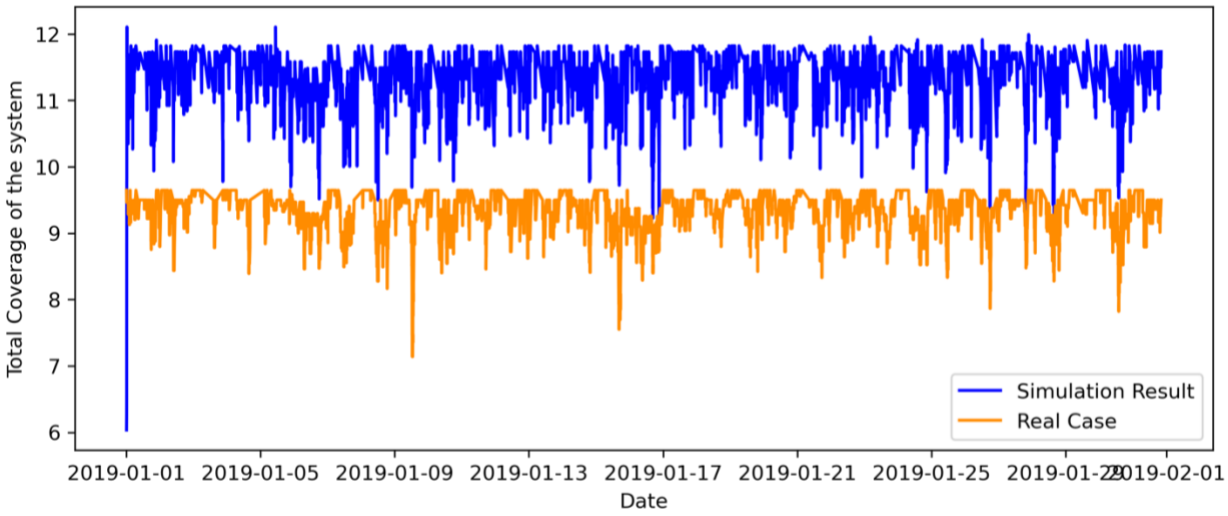


Figure 7.7 Comparison of the total coverage for real-case data and simulation results

Table 7.2 Comparison of En Route time between real-case data and simulation results

	Real Case	Integer Model
Minimal value	0	0.1
25th percentile	4	2.349
50th percentile	6	3.718
75th percentile	9	6.254
Maximal value	52	22.66

Figure 7.8 compares En Route time between real-case data and the proposed model. In this comparison, the ratio of incidents occurring in the zone is used to represent the importance of the zone. The importance of zones will affect the coverage rate. Figure 7.8 shows the histogram of two data sets and reflects the distribution of En Route time under different scenarios. It demonstrates a significant increase in the proportion of incidents that use less rescue time in the proposed model. More than half of the incidents can be reached within 5 minutes, which may take more time in the real-case date.

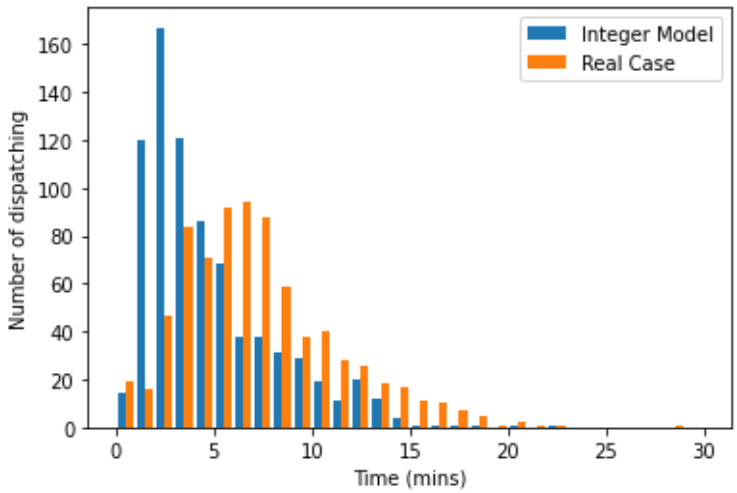


Figure 7.8 Comparison of the En Route time for real-case data and simulation results

Response time is an important measure to evaluate the performance of dispatching policies. The percentage of incidents that can be reached within a pre-defined time threshold is an important indicator of the response time. The NFPA standard clearly states the time threshold. Ensuring that all incidents can be reached within the specified time as far as possible can not only meet the requirement but also ensures that more incidents can be rescued in a timely manner, thereby reducing the rescue time. Figure 7.9 compares the percentages of incidents reached within a pre-defined time threshold. It illustrates that the ratio of incidents reached within 5 minutes will increase from 0.45 to 0.65 if the proposed mode is applied. The ratio of incidents reached within 8 minutes will increase from 0.75 to 0.83. The results illustrate that the proposed model can largely increase the proportion of incidents that can be reached within the pre-defined time threshold.

7.3.3 Workload

The proposed model aims to make the dispatching and redeployment decisions for the heterogeneous emergency vehicle fleet to shorten the rescue time and consider the workload between different units and crew members. Figure 7.10 shows the comparison of workload distributed among different types of vehicles. By applying the proposed model to the real-case

data, the utilization of ALS vehicles will increase from 50 percent to nearly 80 percent. Increased utilization of advanced level vehicles allows for more professional pre-hospital treatments for patients.

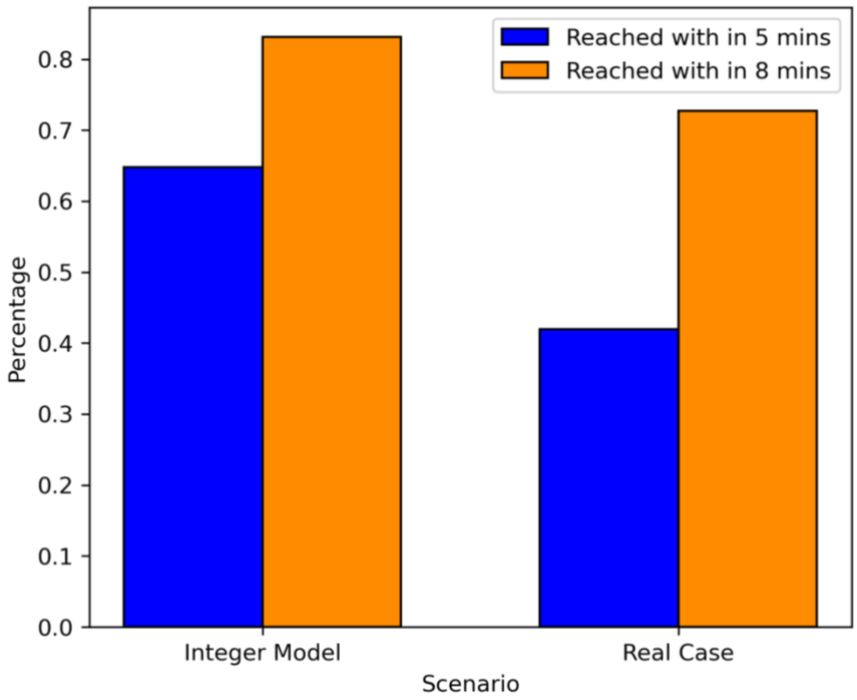


Figure 7.9 Comparison of the percentage of incidents that can be reached within a predefined time threshold

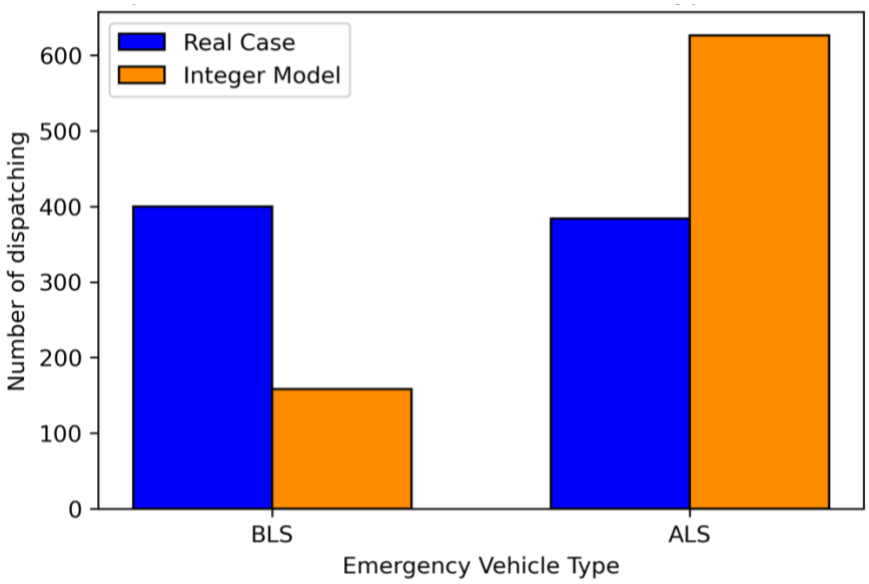


Figure 7.10 Comparison of the workload for different types of vehicles

Figure 7.11 compares workload distribution between different types of emergency vehicles in real-case data and computational results. The standard deviation for BLS vehicles in real-case data is 10.9137, while in the computation is 3.2572. The standard deviation for ALS ambulances in real-case data is 10.0012, while in the computation is 7.1861. Figure 7.11 demonstrates the proportion of incidents reached by ALS ambulances increases, and the workload distribution among ALS ambulances is more balanced than in the real-case data. The utilization of BLS ambulances is significantly reduced. The tasks are more evenly distributed in the BLS ambulances. It must be mentioned that ambulance A25 has not been dispatched to any incidents during the computation. This is not a miscalculation. This may be due to the deployment of the vehicles resulting in ambulance A25 being located consistently at the stations. This situation will not occur again in the following sensitivity analysis when small perturbations are added to the system.

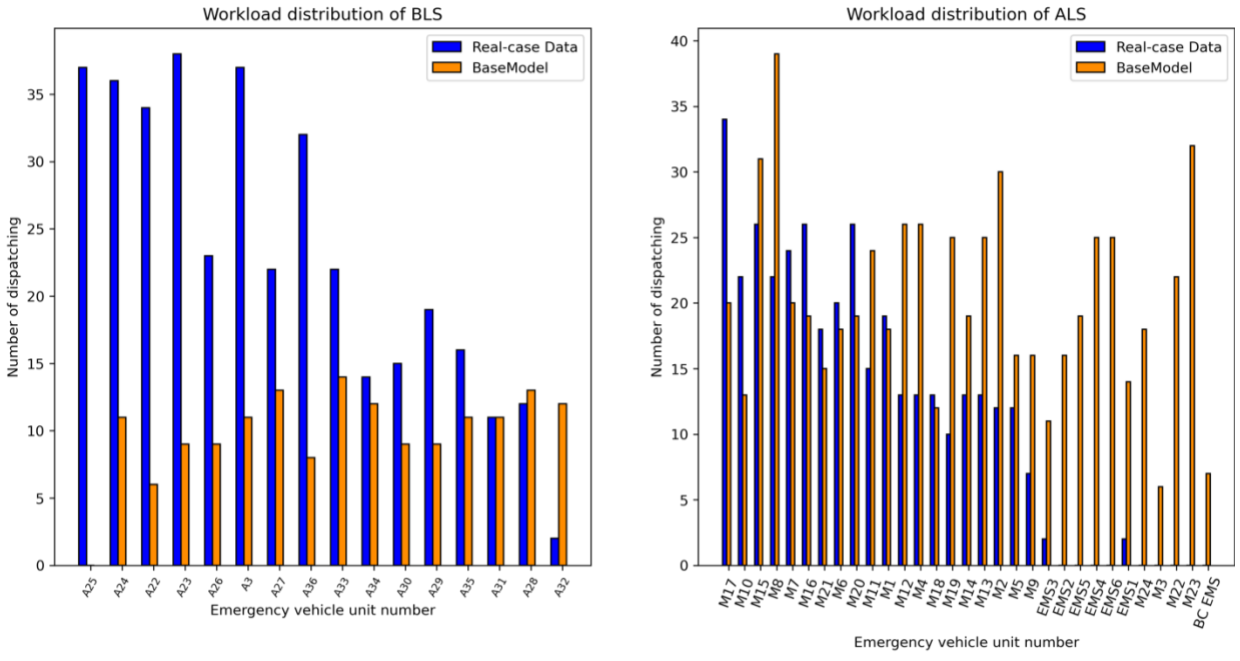


Figure 7.11 Comparison of the workload distribution of emergency vehicles for the base condition and real-case data

7.3.4 Redeployment

The previous sections demonstrate the proposed model can shorten the rescue time and find the best configuration for the deployment of vehicles to maintain a higher level of total coverage. Part of this overall performance improvement is achieved by increasing the number of redeployments. Therefore, the number of redeployments should be limited to a certain number and not increased indefinitely. Too many deployments will not only increase crew members' workload but also increase the cost. In the proposed model, the number of redeployments is not explicitly limited by adding a constraint. However, some other parameters are added to the model to simulate real situations, thus indirectly restricting the number of redeployments. Figure 7.12 shows that the number of redeployments does not exceed one for the vast majority of incidents. Table 7.1 shows the average number of redeployments is 0.7665. In a few cases, the number of redeployments is around 5. The results demonstrate that the proposed model can improve overall performance with a few redeployments, which is acceptable. In future applications, if certain limits on the number of redeployments are needed, a corresponding constraint can be added to the integer model.

7.3.5 Optimality

Chapter 4 shows that the proposed model can handle large-scale and complex situations and produce dispatching and relocation decisions in a reasonable time. In chapter 7, the proposed model is applied to the real-case data and can be solved optimally using the Gurobi optimizer solver. In real situations, Baltimore city has plenty of emergency vehicles on standby to prepare for the future, and extreme situations where a large number of incidents occur simultaneously are rare. It ensures the model can always provide a feasible optimal solution for the current situation. In the future, if the incident number is greater than the available vehicles in the system, the model might be infeasible. A queueing system can be developed and incorporated into the discrete event

system to handle infeasibility. Suppose there are not enough emergency vehicles to be dispatched to satisfy the requirements. In that case, the incidents should be sorted according to the priority or arrival time, then wait in the queue until fulfillment.

7.4 Sensitivity Analysis

7.4.1 Traffic time increment

Historical traffic congestion data is not available for the recorded medical data. Actual traffic time largely determines the efficiency of rescue. Neglecting the effect of travel time increase may result in overestimating the optimality of the model. So, testing the system's performance at different travel time conditions is critical. The traffic system is an interrelated and complex system that is integrated with multiple entities. Numerous reasons cause traffic congestion. In this study, the causes of the congestion and its attributes and properties are out of the scope of the research. The study focuses on the impacts of travel time increase on the final performance of the proposed model. Thus, the study will assess the effects of traffic congestion by increasing the congestion level, which is a predefined parameter to quantify the increase in travel time at different congestion levels.

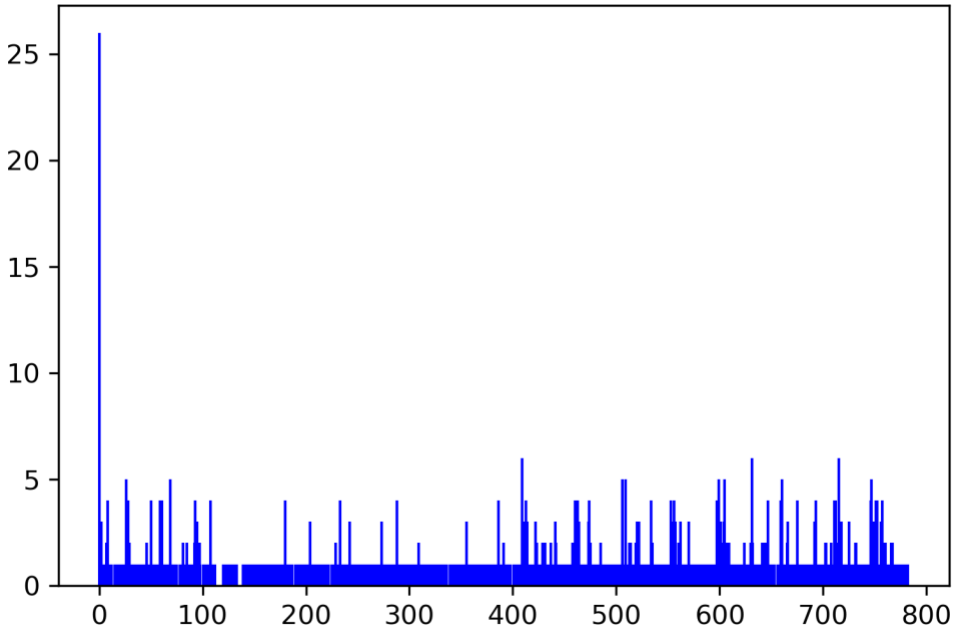


Figure 7.12 The number of redeployments for incidents

Due to the lack of sufficient historical data to generate a specific congestion probability density function for each road segment, a random generator will be used, which is implemented by using the package ‘random’ in Python. The random traffic congestion generator will generate an extra traveling time for each origin and destination pair based on a predefined congestion level. And the random congestion information will be updated at every time step. It means the whole area uses the same congestion level at each time. This assumption might be unreasonable in some real-world situations because it doesn’t distinguish the busy and non-busy areas. The results might not accurately reflect the real dispatching situations. But when important historical data is missing, this approach can illustrate the characteristics of the problem and show how the response time will change due to different congestion conditions. Moreover, in some serious incidents, the emergency vehicles will turn on the lights and sirens, and surrounding vehicles will make way for the rescue vehicle to proceed. This will also make the rescue time not heavily dependent on traffic congestion information.

In this study, eight different given congestion levels, which are: 0.1, 0.2, 0.3, 0.4, 0.5, 1.0, 1.5, and 2.0, will be tested on the historical medical data. A given congestion level is the parameter used to calculate the additional travel time between each pair of origins and destinations caused by traffic jams. The congestion level is assumed to start from 0.1 and increase gradually in increments of 0.1 to reflect the slightly congested traffic conditions. Then congestion levels 1.0, 1.5, and 2.0 are chosen to simulate the severely congested traffic conditions. These numbers have been selected to reflect travel times with traffic congestion that may be several times greater than the free-flow travel times. Five simulations will be conducted for each congestion level to reduce the randomness. And same statistics will be collected during the computation. The results and analysis are shown below.

Figure 7.13 compares the average En Route time between the base conditions and different congestion levels. Under each level, the congestion time is not identical to all the routes. A specific congestion level is randomly selected between 0 and the given congestion level for each pair of stations, zones, and incident sites. Then the actual travel time is calculated using the travel time without traffic congestion multiplied by one plus the specific congestion level. With the increase in congestion level, the En Route time will increase accordingly. When the congestion level reaches 0.5, the average En Route time is about 5.8 minutes, indicating a 21% increase. If the congestion level increase to 1.0, the average En Route time is about 6.8 minutes, indicating a 42% increase. If the congestion level reach 2.0, the average En Route time will increase by 62.5%. Such extreme situations are not common in real life because emergency vehicles will turn on lights and sirens, which can reduce the impact of congestion. However, the sensitivity analysis can roughly reflect the effect brought by congestion.

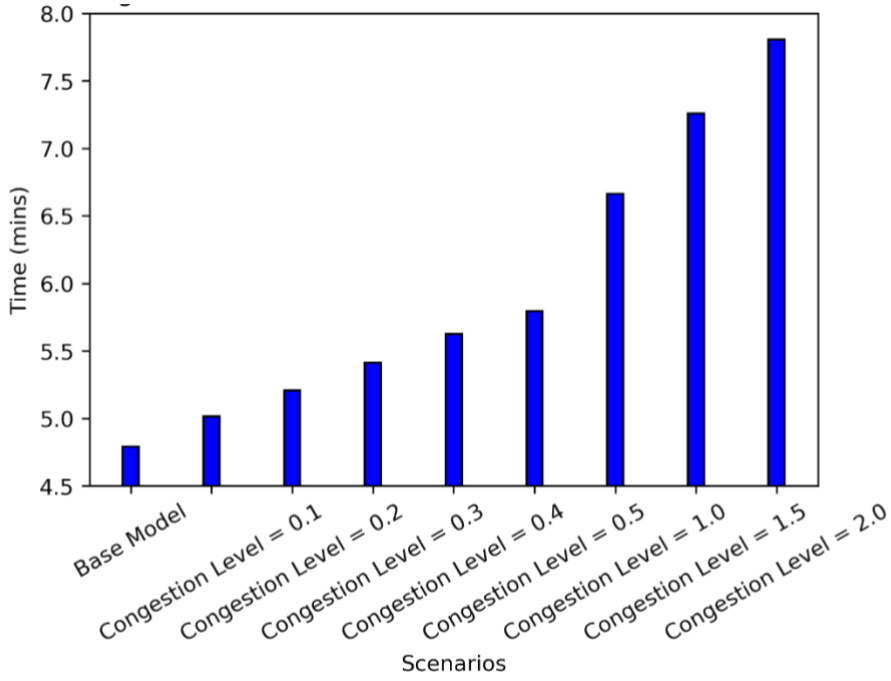


Figure 7.13 Comparison of the average En Route time for the base condition and different congestion levels

Figure 7.14 and Figure 7.15 illustrate the percentage of incidents that can be reached within 5 and 8 minutes under base conditions and conditions with different congestion levels. Only 56% of incidents can be reached within 5 minutes if the congestion level increases to 0.5, which is a 16% decrease. If the congestion level reaches 2.0, only less than 40% of incidents can be reached within 5 minutes, indicating a 45% decrease. The figures also illustrate when the congestion level reaches 0.5, 76% of incidents can be reached within 8 minutes, indicating an 8% decrease compared with the base model. If the congestion level expands to 2.0, only 63% of incidents can be reached within 8 minutes.

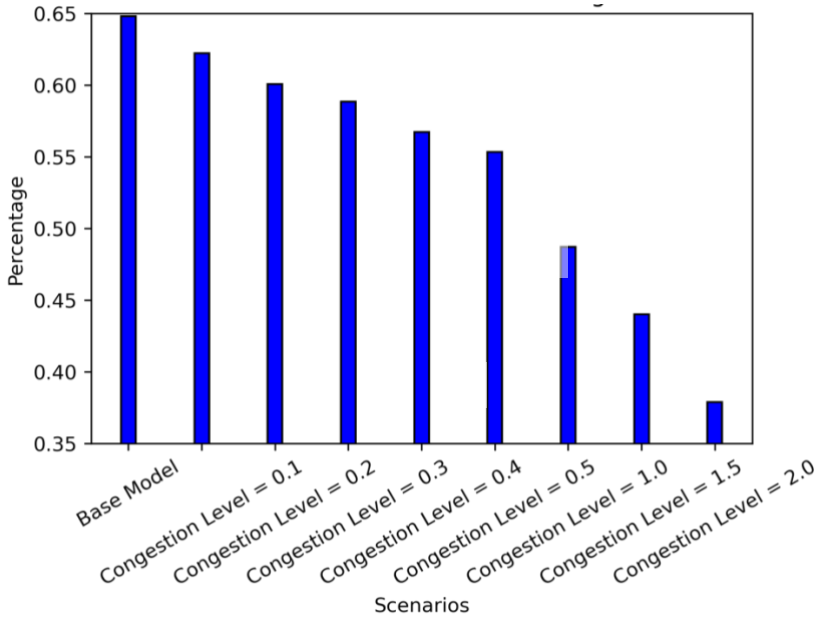


Figure 7.14 Comparison of the percentage of incidents reached within 5 minutes for the base condition and models with different congestion levels

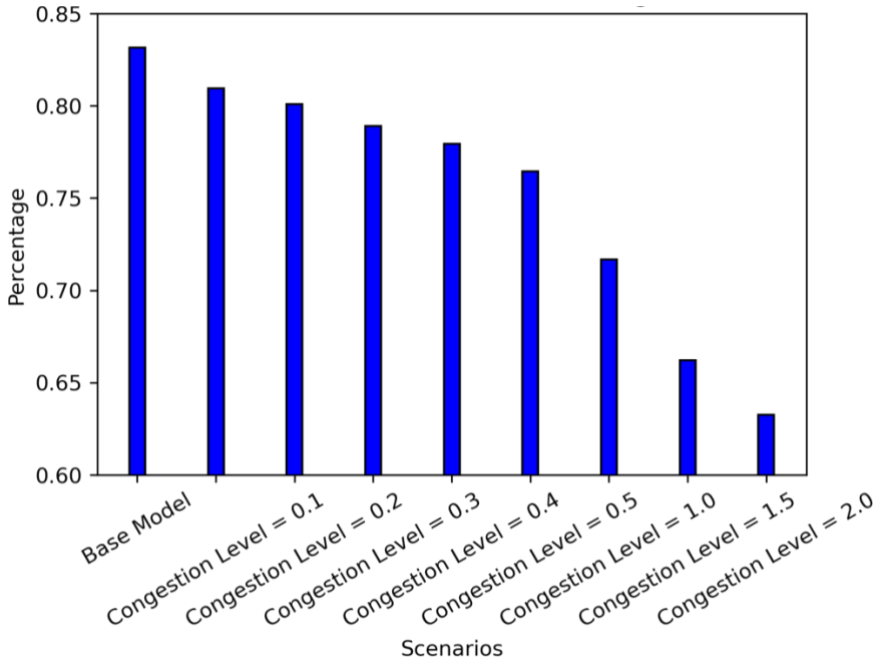


Figure 7.15 Comparison of the percentage of incidents reached within 8 minutes for the base condition and models with different congestion levels

Table 7.3 and

Table 7.4 summarize the average En Route time statistics between ideal data, which is the base model without considering the traffic condition and results with different congestion levels.

Table 7.3 Average En Route time for the base model and models with different congestion levels

	Base Model	Congestion level_0.1	Congestion level_0.2	Congestion level_0.3	Congestion level_0.4	Congestion level_0.5
Replica 1	4.79	5.021	5.213	5.437	5.635	5.822
Replica 2		5.019	5.197	5.405	5.594	5.811
Replica 3		5.019	5.213	5.422	5.637	5.773
Replica 4		4.998	5.217	5.414	5.619	5.779
Replica 5		5.015	5.189	5.386	5.633	5.783

Table 7.4 Average En Route time for the base model and models with different congestion levels (continued)

	Base Model	Congestion level_1.0	Congestion level_1.5	Congestion level_2.0
Replica 1	4.79	6.6624	7.2565	7.8068
Replica 2		6.6578	7.2967	7.8265
Replica 3		6.6639	7.3575	7.9876
Replica 4		6.6196	7.1987	7.7756
Replica 5		6.6465	7.4675	7.8376

Figure 7.16 and Figure 7.17 compare the En Route time for real-case data, the base model, and models with different congestion levels. The boxplots can clearly show that with the congestion level increase, the average En Route time will gradually increase. When the congestion level is smaller than 0.5, the models can perform better than real-case data. When the congestion level continues to increase to 1.0, 1.5, and 2.0, the actual En Route time will deteriorate or be even worse than the real-case data. The dispatch time for some extreme cases can be significantly increased compared to the base model.

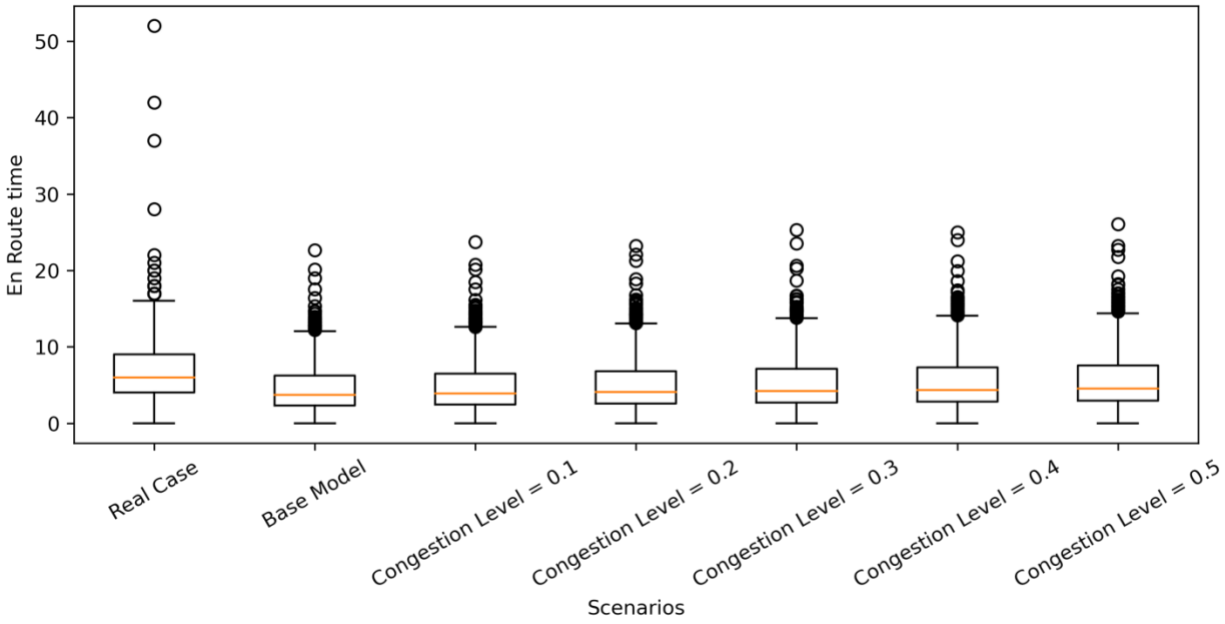


Figure 7.16 Comparison of the En Route time for real-case data, the base model, and models with different congestion levels

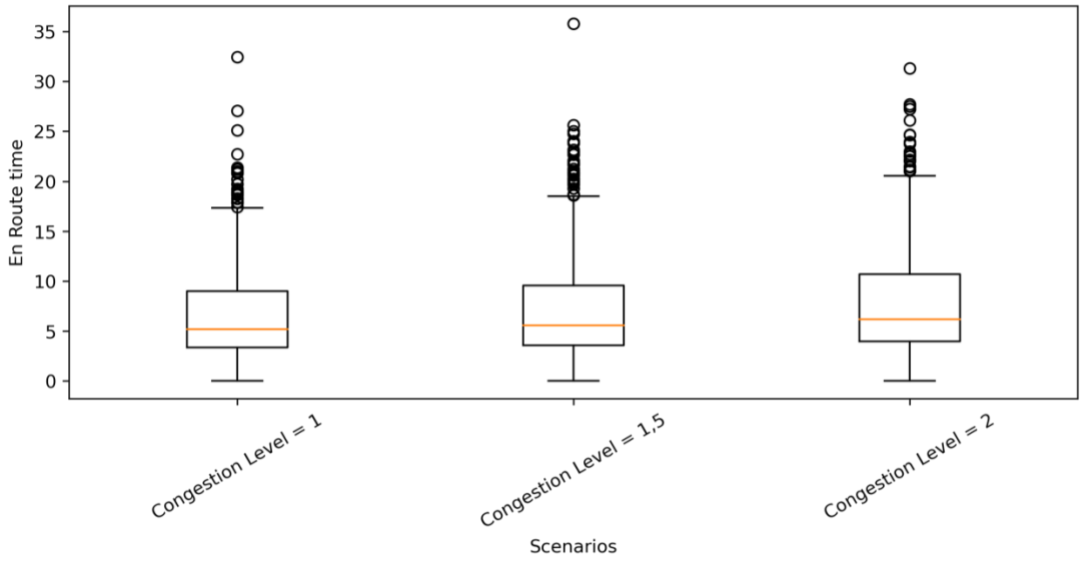


Figure 7.17 Comparison of the En Route time for real-case data, the base model, and models with different congestion levels (continued)

Table 7.5 summarizes the comparison of the percentage of basic coverage between the base model without considering traffic conditions and results with considering different congestion levels. Five different statistics are collected and computed for the base model and models

considering congestion levels. Except for the minimal value, the integer model can maintain a higher coverage level than those considering the congestion. This is consistent with common sense. More severe congestion may increase vehicle dispatching and redeployment time, thus affecting decision-making. If the redeployment duration exceeds the threshold, the redeployment decision will not be considered. This will affect the whole coverage level for the system. The maximal value is kept at 0.970, which is close to 1. The system can cover almost all zones except for very few outlying areas. This value ensures that virtually all areas in the system can get immediate access to basic medical assistance. If the congestion level is smaller or equal to 0.3, the minimal coverage is greater than the base model without considering the traffic condition. This result may be because congestion is gathering in the downtown area, and emergency vehicles are not dispatched from the nearest station but from other stations.

Table 7.5 Comparison of percentage of basic coverage between the base model and models with different congestion levels

	Integer Model	Congestion level_0.1	Congestion level_0.2	Congestion level_0.3	Congestion level_0.4	Congestion level_0.5
Minimal value	0.598	0.670	0.617	0.631	0.578	0.582
25th percentile	0.890	0.889	0.888	0.887	0.883	0.883
50th percentile	0.893	0.893	0.893	0.893	0.893	0.893
75th percentile	0.915	0.914	0.911	0.910	0.907	0.905
Maximal value	0.970	0.970	0.970	0.970	0.970	0.970

Table 7.6 and Table 7.7 compare the percentage of double coverage rate and fully functional coverage rate between the base model and models with different congestion levels. The maximal double coverage rate can be maintained at 72.9%, and the maximal fully functional double coverage rate can reach 63.5%. Double and fully functional double coverage can maintain small fluctuations in different congestion levels.

Table 7.6 Comparison of percentage of double coverage between the base model and models with different congestion levels

	Integer Model	Congestion level_0.1	Congestion level_0.2	Congestion level_0.3	Congestion level_0.4	Congestion level_0.5
Minimal value	0.483	0.490	0.487	0.495	0.495	0.493
25th percentile	0.626	0.620	0.620	0.620	0.617	0.618
50th percentile	0.660	0.660	0.656	0.653	0.650	0.649
75th percentile	0.669	0.669	0.669	0.667	0.666	0.666
Maximal value	0.729	0.729	0.729	0.729	0.729	0.729

Table 7.7 Comparison of percentage of fully functional double coverage between the base model and models with different congestion levels

	Integer Model	Congestion level_0.1	Congestion level_0.2	Congestion level_0.3	Congestion level_0.4	Congestion level_0.5
Minimal value	0.483	0.490	0.487	0.487	0.485	0.489
25th percentile	0.548	0.548	0.547	0.547	0.547	0.547
50th percentile	0.574	0.570	0.569	0.569	0.569	0.570
75th percentile	0.584	0.584	0.584	0.582	0.583	0.582
Maximal value	0.635	0.635	0.635	0.635	0.635	0.635

Figure 7.18 and Figure 7.19 compare the percentage of basic coverage between the base model and models with different congestion levels. They show that for the base model, the system can maintain around 90 percent of the basic coverage, regardless of the impact of some special cases. In some extreme conditions, when the requests in the medical system increase and multiple vehicles are on-site or at the hospital to deal with tasks, the basic cover of the system will decrease to 60 percent. For different congestion levels, there will be a reduction in the coverage of the system. But the system can still maintain a high level of base coverage. With the increase in congestion levels, the basic coverage rate will decrease. In more than half of the scenarios, the basic coverage rate is less than 90% if the congestion level reaches 2.

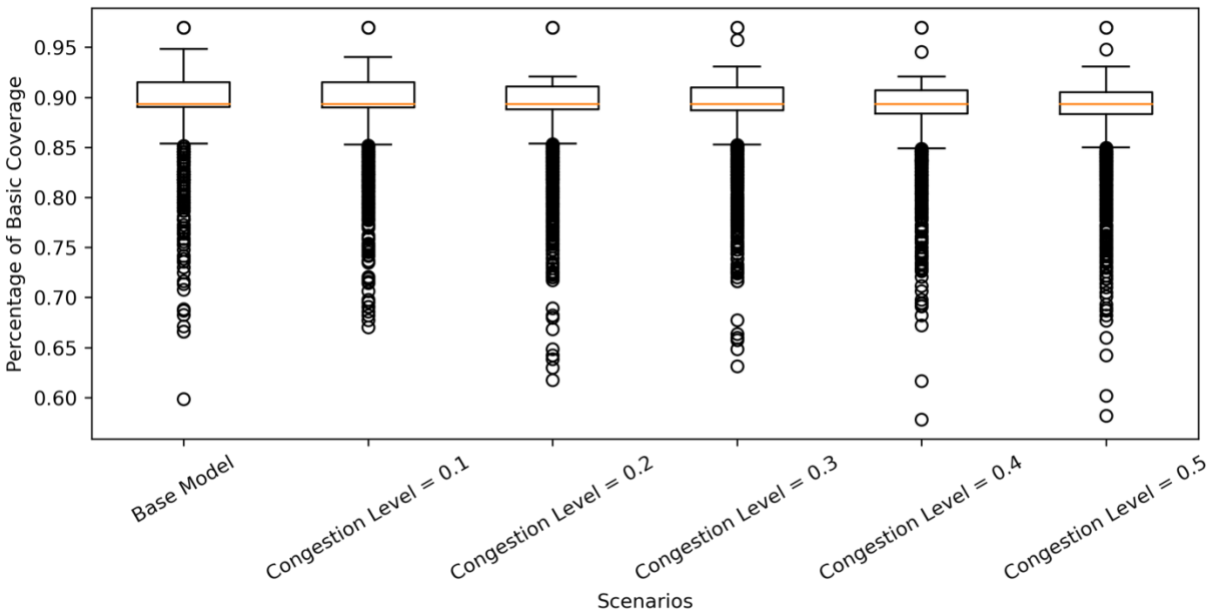


Figure 7.18 Comparison of the percentage of basic coverage for the base model and models with different congestion levels

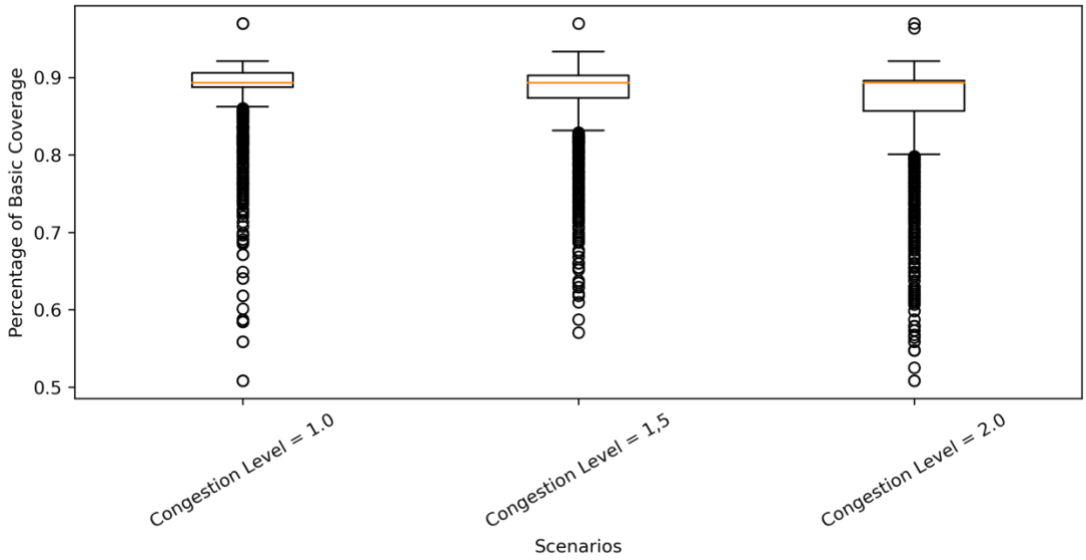


Figure 7.19 Comparison of the percentage of basic coverage for the base model and models with different congestion levels (continued)

Figure 7.20 and Figure 7.21 illustrate the comparison of the percentage of double coverage between the base model and models with different congestion levels. They show that for the base model, the system can maintain around 66 percent of the double coverage rate, regardless of the

impact of some special cases. With the increase in congestion levels, the median value of the double coverage rate will decrease accordingly. Moreover, in extreme conditions, the minimum value of the double coverage rate is also reduced, which is around 50 percent.

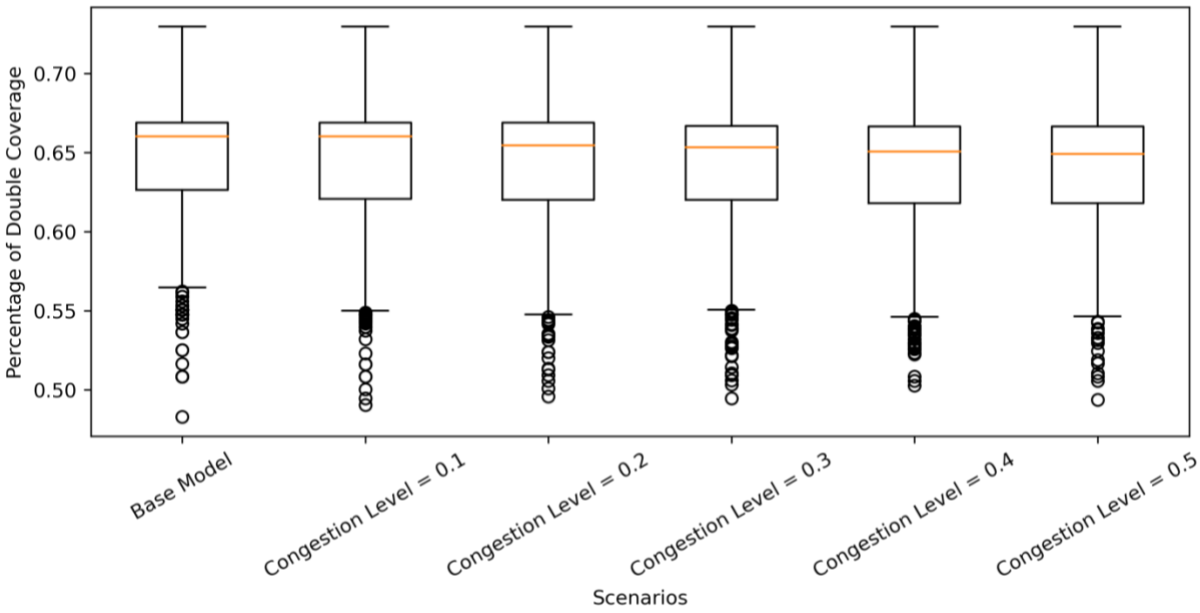


Figure 7.20 Comparison of the percentage of double coverage for the base model and models with different congestion levels

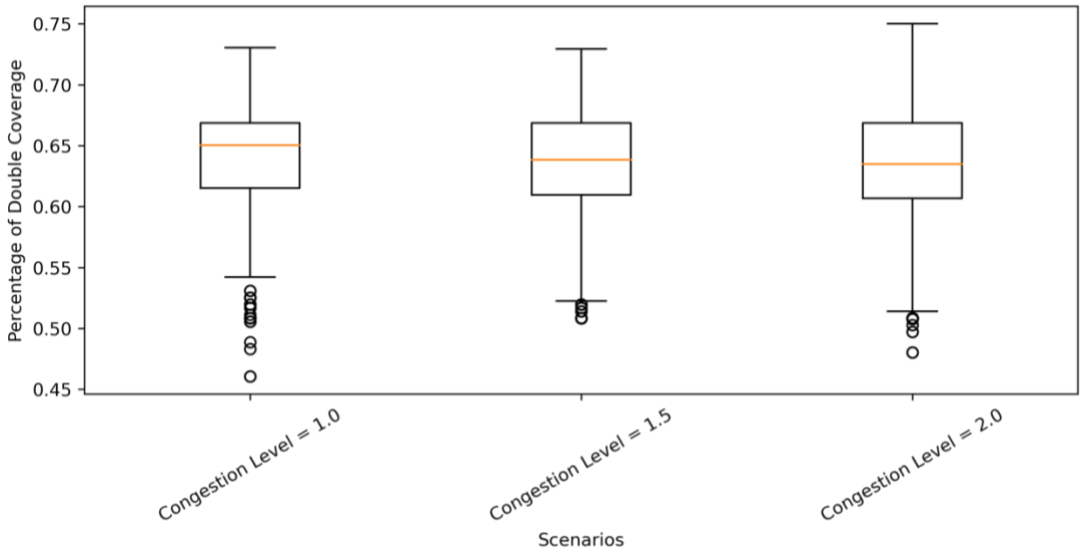


Figure 7.21 Comparison of the percentage of double coverage for the base model and models with different congestion levels (continued)

Figure 7.22 and Figure 7.23 compare the percentage of fully functional double coverage between the base model and models with different congestion levels. They show that for the base model, the system can maintain around 57.5 percent of the fully functional double coverage, regardless of the impact of some special cases. With the increase in congestion levels, the median value of the fully functional double coverage rate will decrease accordingly.

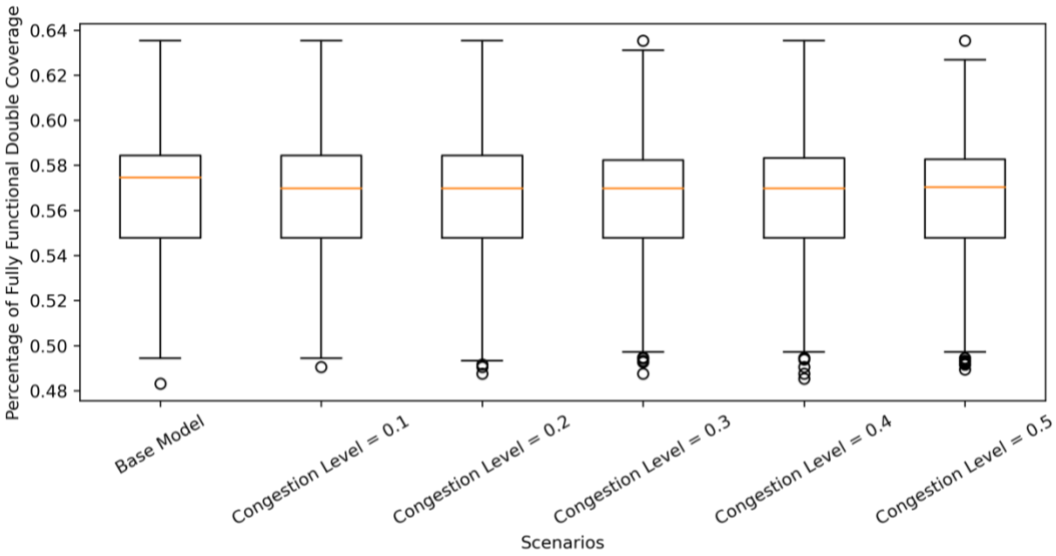


Figure 7.22 Comparison of the percentage of fully functional double coverage for the base model and models with different congestion levels

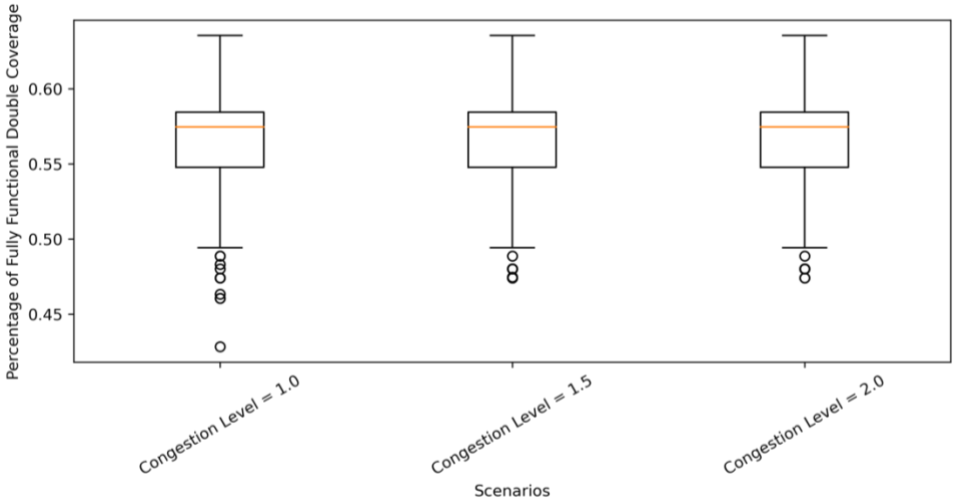


Figure 7.23 Comparison of percentage of fully functional double coverage for the base model and models with different congestion levels (continued)

Figure 7.24 and Figure 7.25 compare the workload distribution for advanced life support (ALS) ambulances between real case data, the base model, and different congestion levels. Because advanced level vehicles undertake more rescue tasks, the average number of dispatching per vehicle increases, but the workloads are more evenly distributed compared to the real-case data.

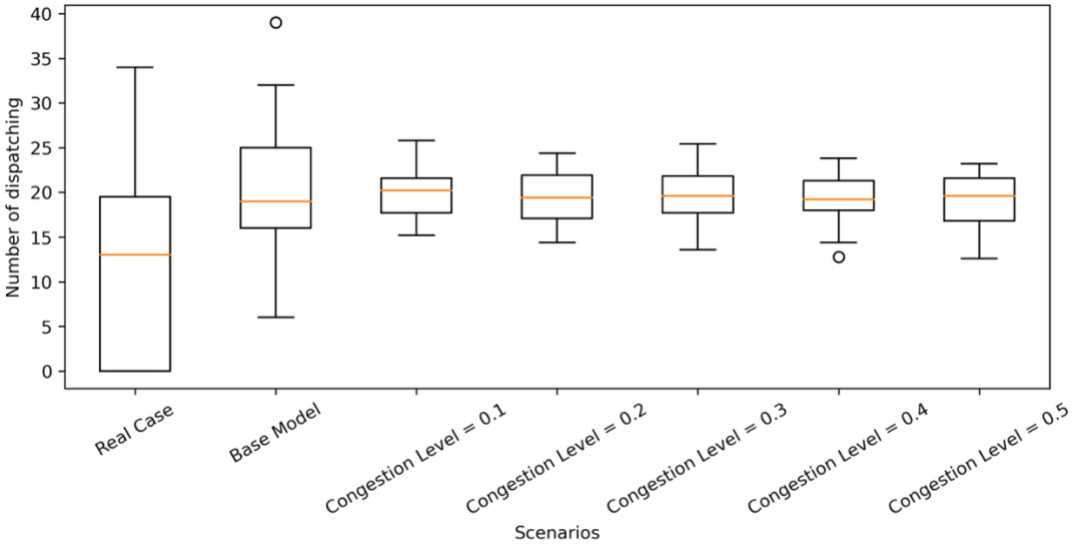


Figure 7.24 Comparison of the workload distribution of the ALS vehicles for the base model and models with different congestion levels

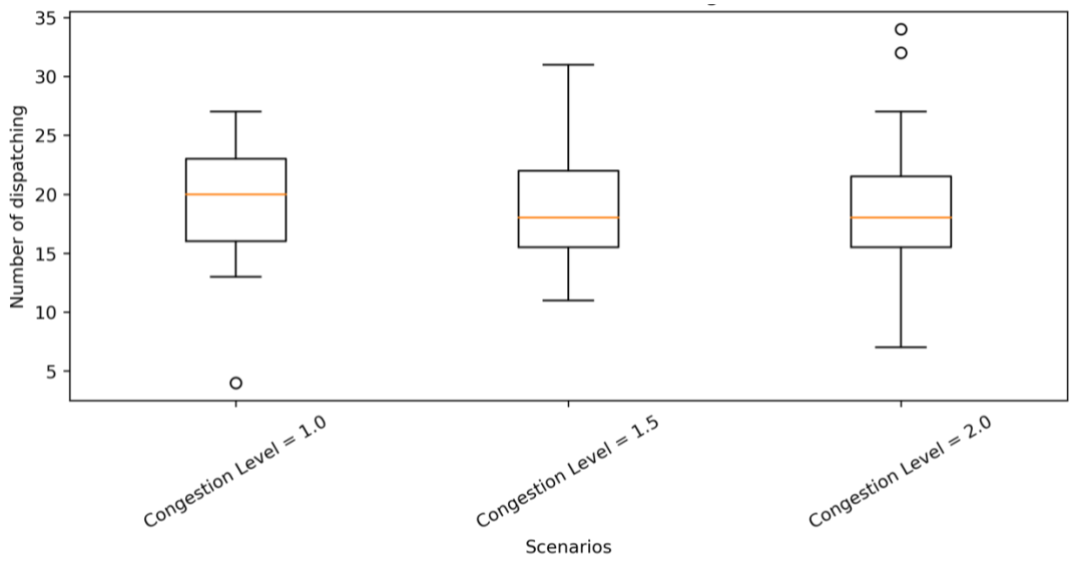


Figure 7.25 Comparison of the workload distribution of the ALS vehicles for the base model and models with different congestion levels (continued)

Figure 7.26 and Figure 7.27 compare the workload distribution for basic life support (BLS) ambulances between real case data, the base model, and different congestion levels. They show the variation of workload distribution in real-case data is extremely large, and the workload distribution is uneven. While for the proposed model, with and without considering congestion conditions, the results can achieve balanced scheduling. The huge improvement comes not only from the algorithm, which considers the workload factor, but also because the utilization of BLS is reduced. The model considers the functional division of the vehicles, which can increase the use of ALS vehicles to substitute the BLS vehicles and thus reduce the rescue time. The figures show that the median values can be maintained at a stable level of around 12 for the proposed models.

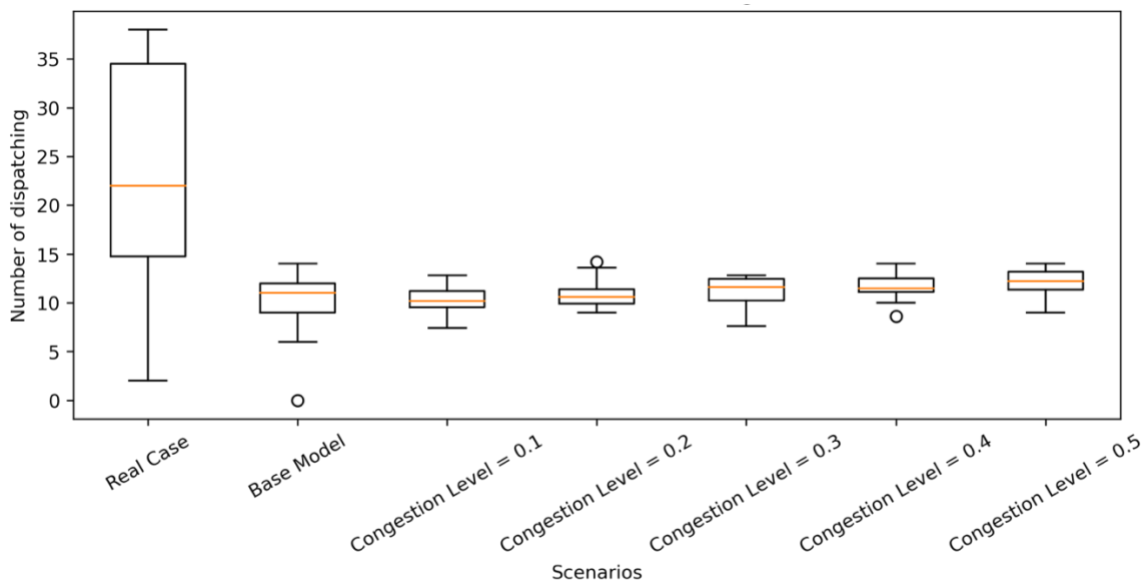


Figure 7.26 Comparison of the workload distribution of the BLS vehicles for the base model and models with different congestion levels

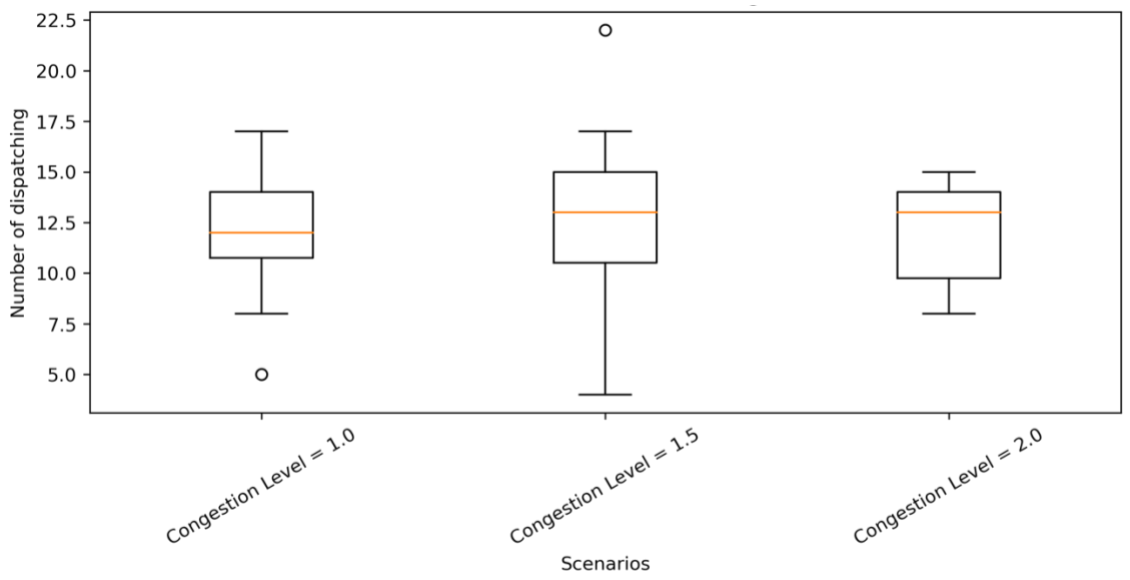


Figure 7.27 Comparison of the workload distribution of the BLS vehicles for the base model and models with different congestion levels (continued)

Figure 7.28 to Figure 7.35 compare the workload distribution between ALS and BLS ambulances under base condition and conditions with different congestion levels. These figures show that a more balanced distribution of rescue tasks among different vehicles is achieved.

Compared with the base model, different congestion levels do not create an unbalanced assignment of rescue tasks. On the contrary, the tasks assigned between different ambulances are more reasonable in some conditions. For example, in the base model, BLS ambulance A25 is not assigned any task.

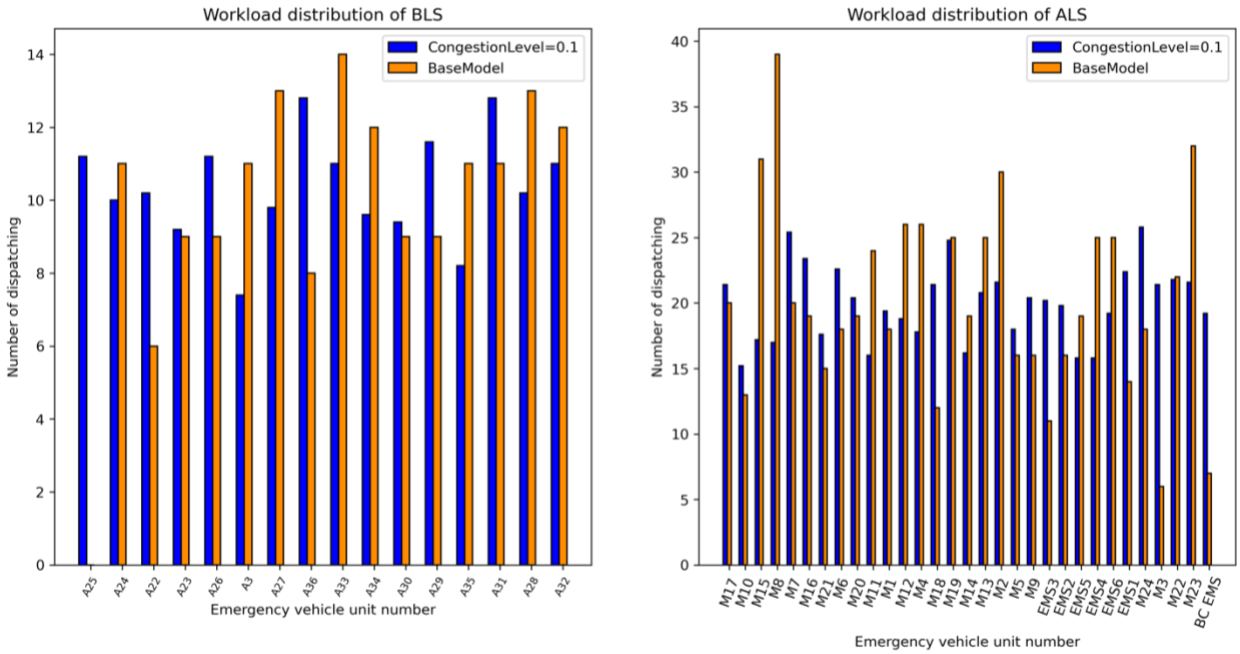


Figure 7.28 Comparison of the workload distribution of emergency vehicles for the base model and models with congestion level =0.1

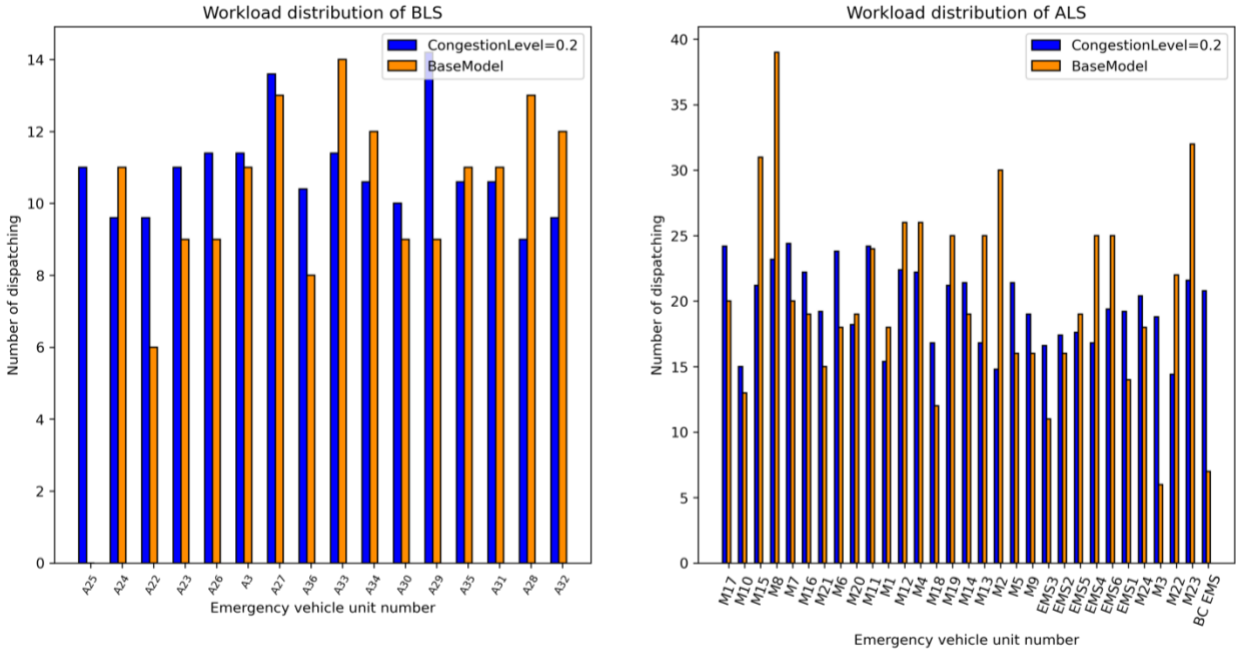


Figure 7.29 Comparison of the workload distribution of emergency vehicles for the base model and models with congestion level =0.2

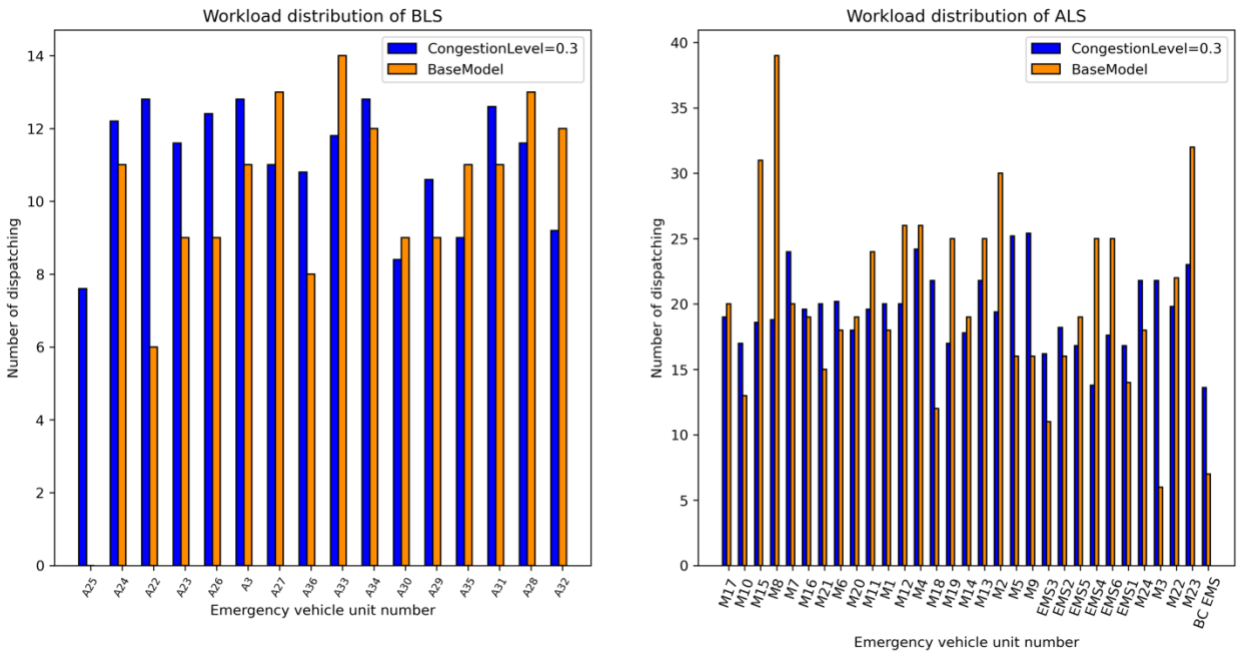


Figure 7.30 Comparison of the workload distribution of emergency vehicles for the base model and models with congestion level =0.3

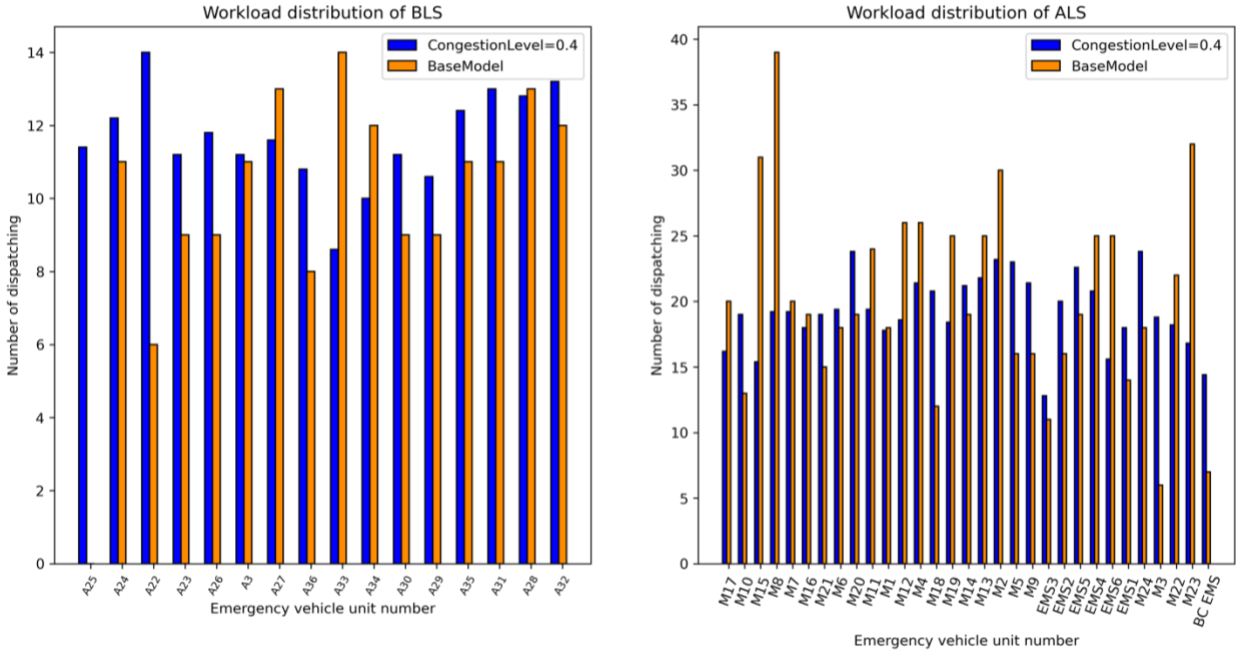


Figure 7.31 Comparison of the workload distribution of emergency vehicles for the base model and models with congestion level =0.4

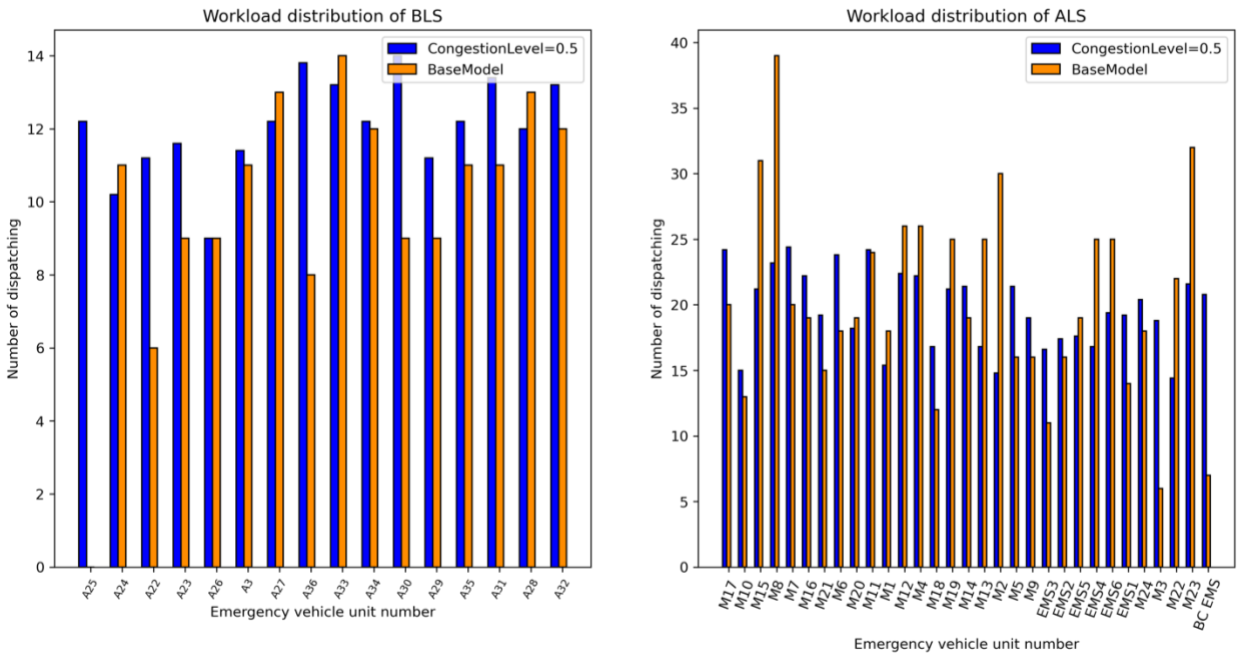


Figure 7.32 Comparison of the workload distribution of emergency vehicles for the base model and models with congestion level = 0.5

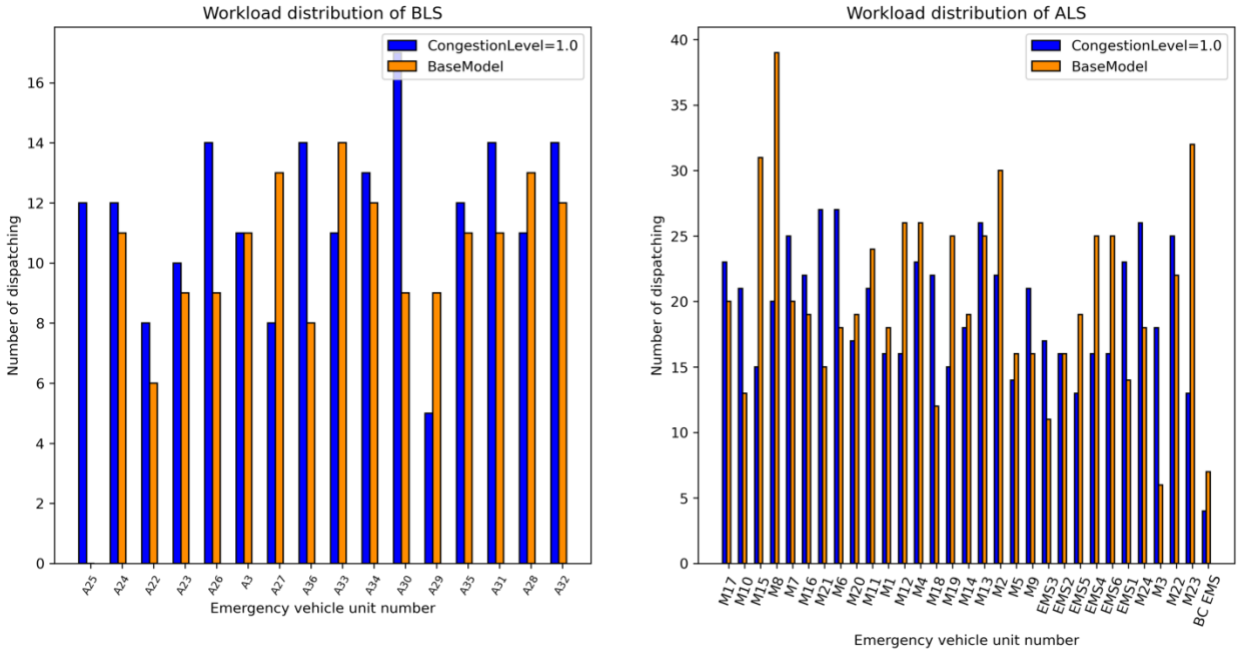


Figure 7.33 Comparison of the workload distribution of emergency vehicles for the base model and models with congestion level = 1.0

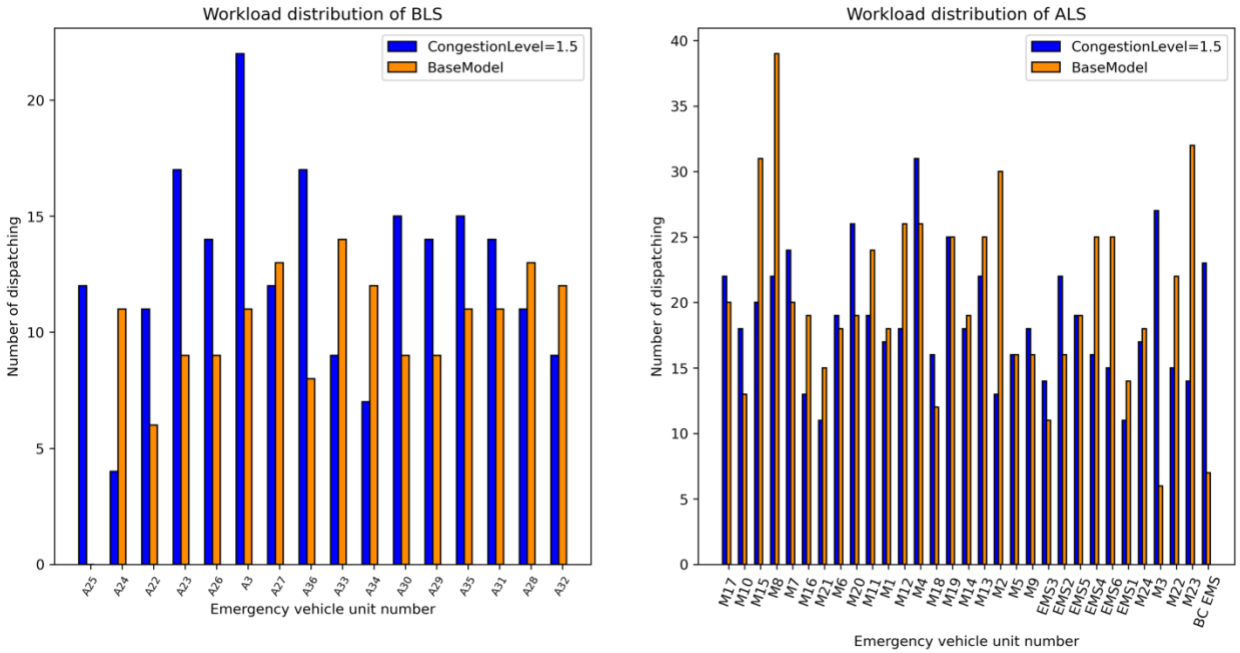


Figure 7.34 Comparison of the workload distribution of emergency vehicles for the base model and models with congestion level = 1.5

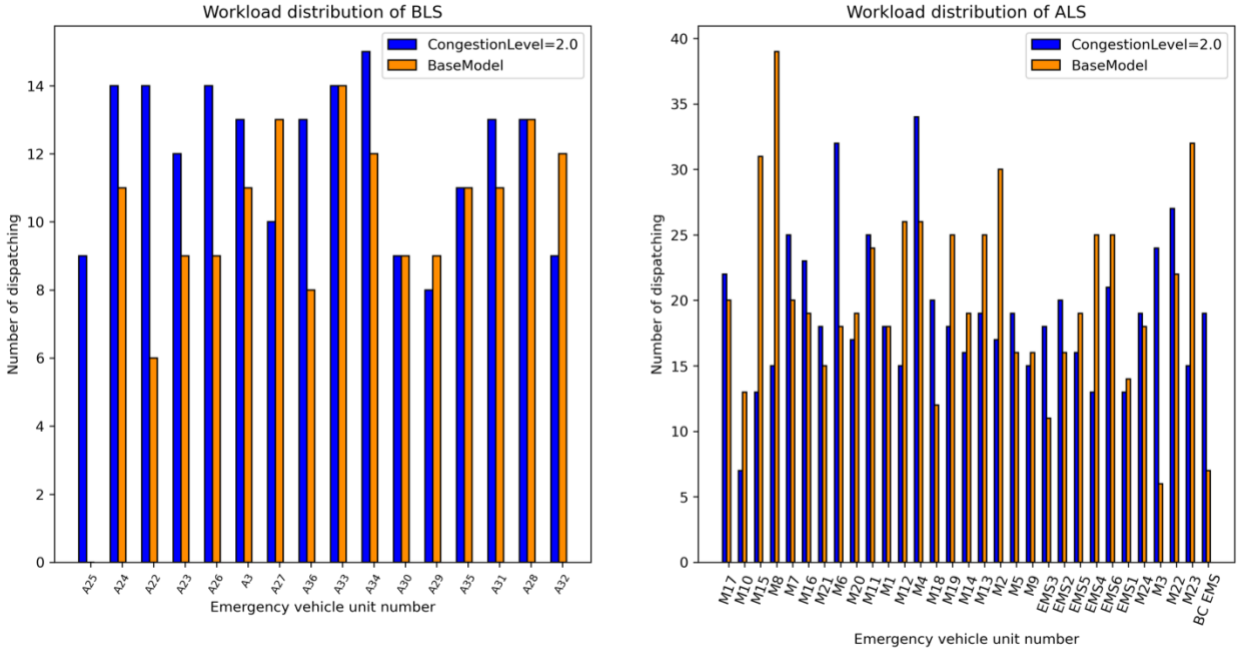


Figure 7.35 Comparison of the workload distribution of emergency vehicles for the base model and models with congestion level =2.0

7.4.2 Emergency vehicle busy factor

Another parameter that is worth analyzing is fleet size. It is evident that by increasing the fleet size, the results will improve a lot. The adequate number and type of emergency vehicles will decrease the rescue time, improve rescue efficiency, and enlarge the coverage for the whole area, which is equivalent to the effect of redeployment. It is interesting to see how the decisions will be influenced by changing the fleet size.

The vehicle sources have been analyzed in the previous chapter. In reality, the availability of the vehicle fleets at some specific timestamp is unknown. Moreover, the busy factor for each vehicle in the emergency fleet is also unknown. After serving an incident, how long the EMS fleet must recover is also inaccessible.

For this part of the analysis, a vehicle busy factor generator is developed and imported each time the decisions need to be made. The random busy factor generates a Boolean-valued outcome: available or busy according to a binomial distribution with given parameters. This module can define the status of all emergency vehicles. The limitation is that the same busy level is used for all stations to generate the outcomes. Like traffic congestion level information, the method used to generate the busy factors does not distinguish the stations located in busy areas and the non-busy areas.

But a general understanding will be summarized for the process and simulation. More accurate results can be obtained in the future with the support of comprehensive data. In this research, four different busy levels will be studied, which are: 0.05, 0.10, 0.15, 0.20. The parameter starts from 0.05 and increases gradually in increments of 0.05 to reflect different availability probabilities. For each busy level, five simulations are run to minimize the impact brought by the randomness. The results and analysis are shown below.

In the analysis, the base model will be used as a baseline to measure the effectiveness of the busy factor. When the busy factor is equal to 0.05, it means that for each vehicle, there is a probability of 0.05 that the vehicle status is unavailable.

Figure 7.36 and Figure 7.37 show the percentage of incidents that can be reached by emergency medical vehicles within 5 minutes and 8 minutes under different conditions. Different levels of busy factors are considered to simulate vehicle availability in different scenarios. Emergency medical vehicles located at stations may be unavailable in real situations due to repair or other conditions. In an ideal condition, all vehicles are available to be dispatched or redeployed when the events or incidents come into the system, which is the base model in the figures. About 65% of incidents can be reached within 5 minutes, and 83.5% within 8 minutes. If the busy factor increases to 0.20, 61% of incidents can be reached within 5 minutes, and 81% within 8 minutes. Five sets of experiments were conducted for each level of busy factor to exclude the interference of individual cases on the final results. Then the average percentage of incidents that can be reached within 5 minutes and 8 minutes were calculated and plotted.

As mentioned before, the percentage of incidents reached within 5 minutes drops to 56% if the congestion level reaches 0.5. The increase in congestion level will have a more significant impact on rescue time, so it will also affect the percentage of incidents that can be reached within the pre-defined time limit.

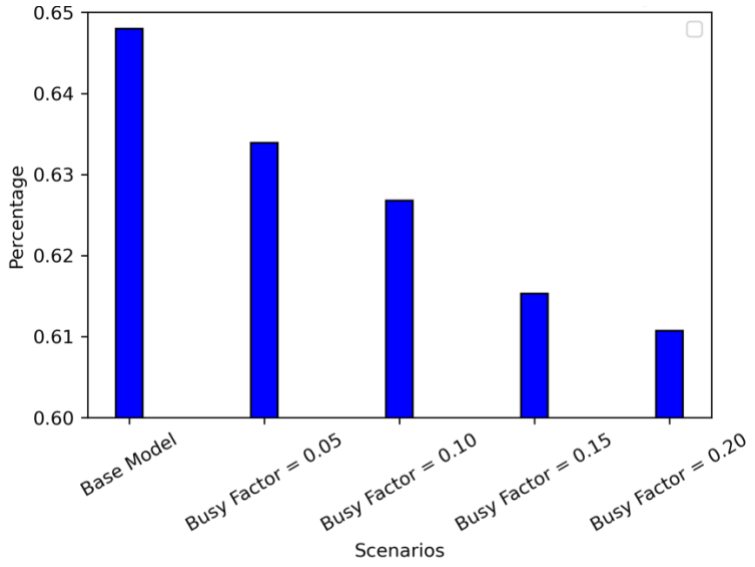


Figure 7.36 Comparison of the percentage of incidents reached within 5 minutes for the base model and models with different levels of busy factor

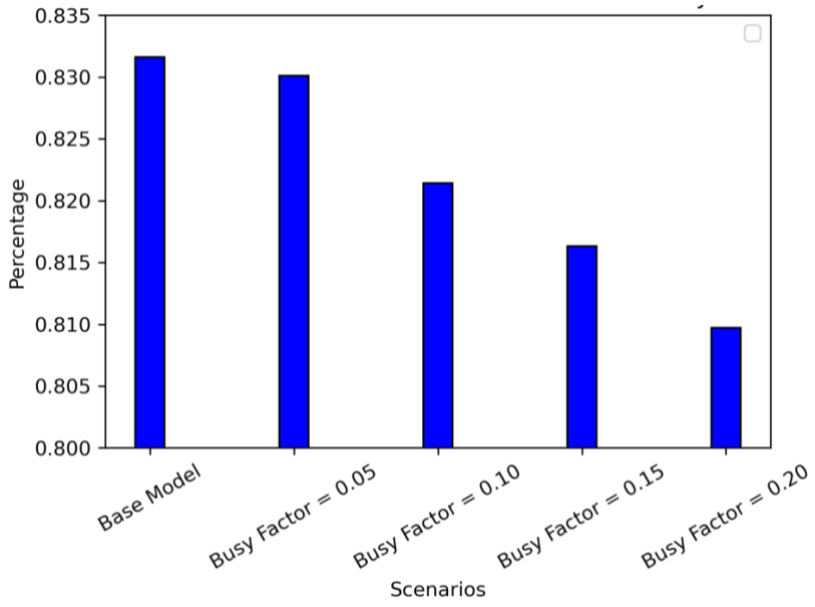


Figure 7.37 Comparison of the percentage of incidents reached within 8 minutes for the base model and models with different levels of busy factor

Figure 7.38 compares the average En Route time between the base condition and models with different levels of busy factors. Under the base condition, the average En Route time for emergency

medical vehicles is about 4.8 minutes. With the increase of the busy factor, the average En Route time increases and reaches 5 minutes.

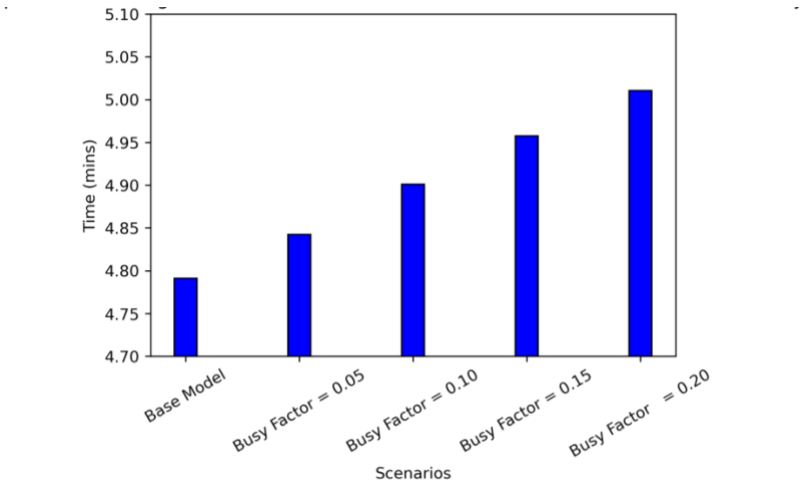


Figure 7.38 Comparison of the average En Route time for the base model and models with different levels of busy factor

Increasing busy factors or congestion levels will decrease rescue efficiency and increase rescue time. The increase in congestion level will have a greater impact on rescue time than the increase in a busy factor. This situation is because a station may be equipped with multiple emergency medical vehicles on standby, and if one vehicle becomes unavailable, other vehicles can be dispatched or redeployed to achieve the same effect. However, the increase in congestion level causes the rising of actual travel time on all routes, making the increase in rescue time inevitable. So as a reflection of the results, increasing the congestion level will lead to a longer rescue time.

Figure 7.39 compares the En Route time between the base condition, calculated from real case data, and different levels of busy factors.

Five sets of experiments are conducted for each condition of the busy factor, and the average value of the En Route time is calculated and represented in the box plot. The box plots can properly compare multiple data sets and visualize outlier values. There are some abnormal data for the real

case data, which means the En Route time is greater than 30 minutes. This is unacceptable in reality, and it greatly increases the rescue time. This issue can be effectively avoided by applying the proposed model. The maximal En Route time is about 20 minutes due to the dispatching distance or deficiency of emergency vehicles. Moreover, the average En Route time is decreased.

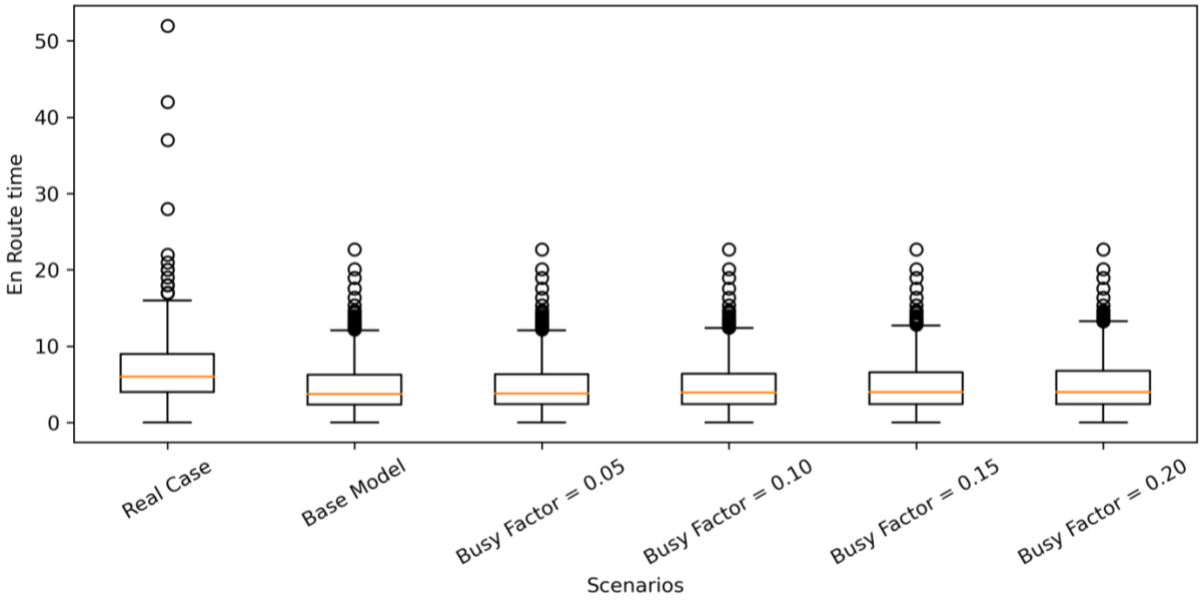


Figure 7.39 Comparison of the En Route time between real-case data, the base model, and models with different levels of busy factor

Table 7.8 summarizes the average En Route time between the base model without considering the unavailability of emergency medical vehicles and models with different levels of busy factors.

Table 7.9 compares the percentage of basic coverage between the base model without considering the unavailability of emergency medical vehicles and models with considering the busy factor. The results and conclusions are roughly the same as in the congestion level part. The base model can maintain higher coverage levels on other values except for the minimal value. For the minimal basic coverage level value, except for the condition that the busy factor equals 0.10, all other models can maintain a higher minimal value than the base condition. This phenomenon

is because the vehicle in the nearest station might be unavailable to be dispatched. Another difference is that the maximal value will be the same for all congestion levels, while the increase in the busy factor will cause a decrease in maximal basic coverage.

Table 7.8 Average En Route time between the base model and models with different levels of busy factor

	Base Model	Busy factor_0.05	Busy factor_0.10	Busy factor_0.15	Busy factor_0.20
Replica 1	4.79	4.830	4.918	4.927	5.038
Replica 2		4.852	4.910	4.989	5.001
Replica 3		4.842	4.911	4.977	5.034
Replica 4		4.842	4.879	4.927	4.970
Replica 5		4.844	4.888	4.966	5.011

Table 7.9 Comparison of percentage of basic coverage between base model and models with different levels of busy factor

	Integer Model	Busy factor_0.05	Busy factor_0.10	Busy factor_0.15	Busy factor_0.20
Minimal value	0.598	0.640	0.574	0.643	0.629
25th percentile	0.890	0.857	0.842	0.833	0.827
50th percentile	0.893	0.879	0.863	0.856	0.851
75th percentile	0.915	0.890	0.880	0.874	0.871
Maximal value	0.970	0.964	0.948	0.943	0.931

The busy factor and congestion level affect coverage level and rescue time completely differently. The results show that with the increase of the busy factor, the rescue time will increase slowly because there might be more than one vehicle located in a station. If the busy factor is not large enough, the probability of all rescue vehicles being unavailable simultaneously is very low. Stations can have multiple vehicles on standby, and the system can dispatch the vehicle from the nearest station. In comparison, congestion will have a more significant and immediate impact on rescue time.

The phenomenon is the opposite at the coverage level. The increase in congestion levels will cause a drop in coverage level slightly, while the increase in the busy factor will cause a significant reduction in coverage level. If the busy factor reaches 0.2, the maximal value will be reduced by about 5%.

Table 7.10 and Table 7.11 compare the percentage of double coverage and fully functional double coverage for the base model and models with different levels of busy factor. In most conditions, the coverage level will decrease with the increase of the busy factor. When the busy factor equals 0.2, the maximal value of double coverage is greater than the base condition. The maximal value remains the same for different conditions for the fully functional double coverage level.

Table 7.10 Comparison of percentage of double coverage between the base model and models with different levels of busy factor

	Integer Model	Busy factor_0.05	Busy factor_0.10	Busy factor_0.15	Busy factor_0.20
Minimal value	0.483	0.506	0.482	0.490	0.487
25th percentile	0.626	0.598	0.590	0.589	0.593
50th percentile	0.660	0.615	0.608	0.608	0.613
75th percentile	0.669	0.630	0.624	0.627	0.632
Maximal value	0.729	0.723	0.717	0.720	0.739

Table 7.11 Comparison of percentage of fully functional double coverage between the base model and models with different levels of busy factor

	Integer Model	Busy factor_0.05	Busy factor_0.10	Busy factor_0.15	Busy factor_0.20
Minimal value	0.483	0.488	0.481	0.484	0.487
25th percentile	0.548	0.548	0.548	0.547	0.548
50th percentile	0.574	0.568	0.564	0.565	0.565
75th percentile	0.584	0.582	0.582	0.580	0.579
Maximal value	0.635	0.635	0.635	0.635	0.635

Figure 7.40 compares the percentage of basic coverage between the base model and models with different levels of the busy factor. It shows that for the base model, the system can maintain 85 to 90 percent of the basic coverage, regardless of the impact of some special cases.

In some extreme conditions, the requests in the medical system increase. When multiple vehicles are either on-site or at the hospital to deal with tasks, the basic cover of the system will decrease to 60 percent. There will be a reduction in the system's coverage for different levels of busy factors, but the system can still maintain a high level of base coverage.

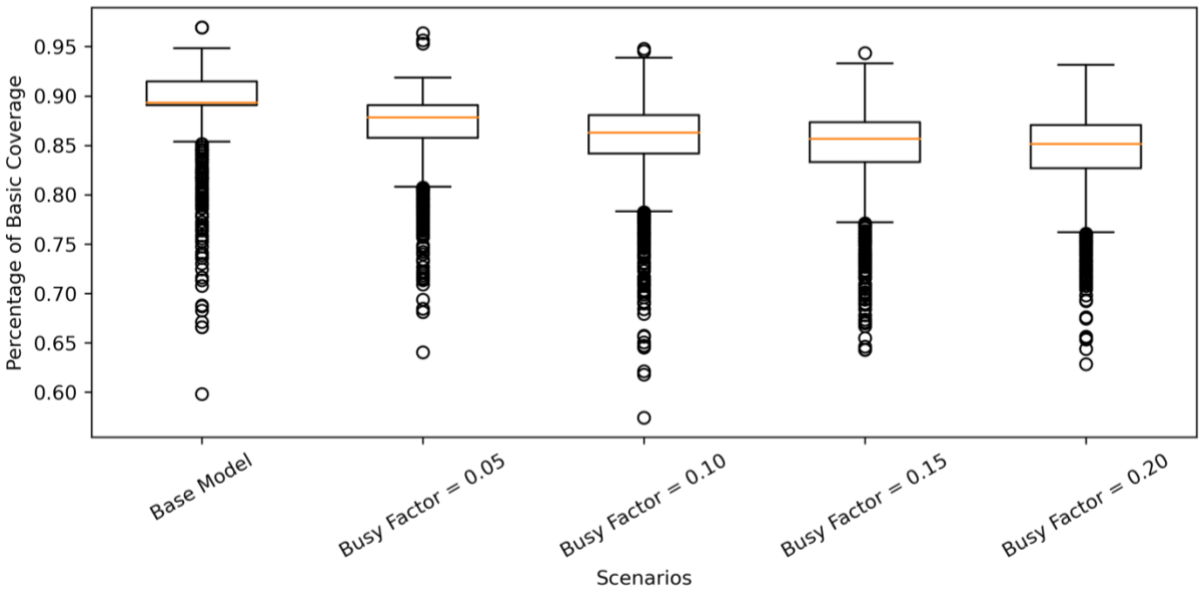


Figure 7.40 Comparison of the percentage of basic coverage for the base model and models with different levels of busy factors

Figure 7.41 compares the percentage of double coverage between the base model and models with different levels of busy factor. Compared with the base model, the median values of double coverage decrease by almost 7 percent for different levels of busy factors. Overall coverage levels decline due to the increase in busy factors. The minimal double coverage levels reach 50 percent.

The workload is also a critical metric for system evaluation. Figure 7.42 compares the percentage of fully functional double coverage levels for the base model and models with different levels of busy factor. Compared with the base model, the median value has a downward trend. Compared with the double coverage level, the decrease in fully functional double coverage between the base model and models with different levels of busy factor is smaller. Figure 7.43 compares the workload distribution for advanced life support (ALS) ambulances between the real case data, the base model, and models with different levels of busy factors. The box plots in the figure clearly illustrate the distribution in each experiment. In the real case data, the variation in workload is huge, which means the workload distribution is uneven.

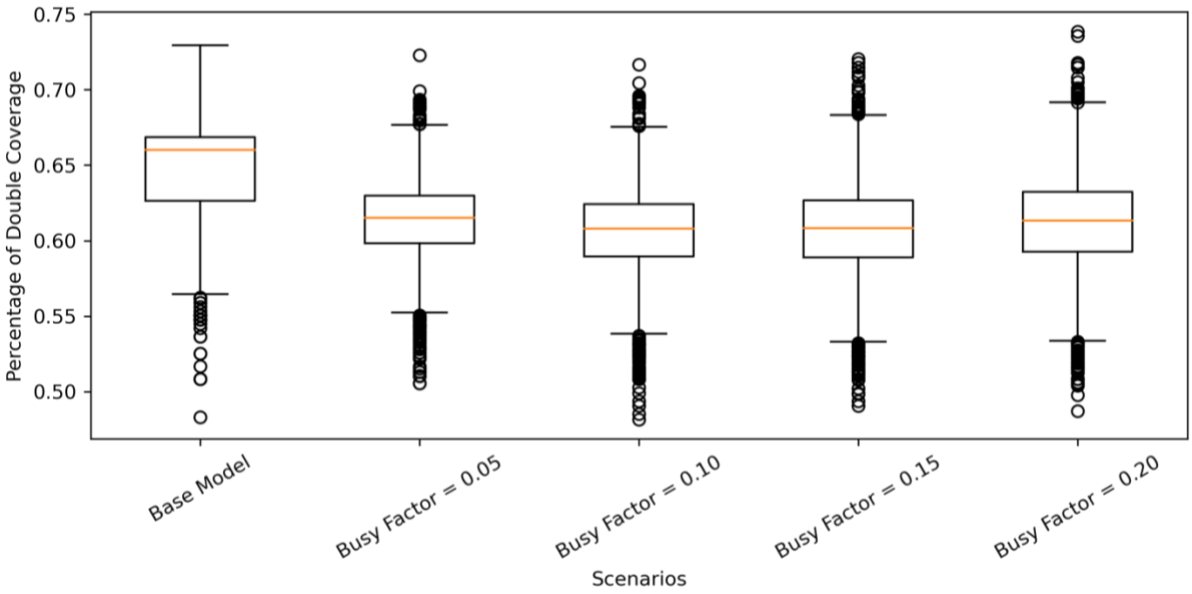


Figure 7.41 Comparison of the percentage of double coverage for the base model and models with different levels of busy factors

Applying the proposed model can largely decrease the variation in workload. Figure 7.43 illustrates that the median values for the base model and models with different levels of busy factors are approximately the same, which is about 20. Compared with the base model, adding different levels of busy factors to the system can narrow the variation significantly.

The results show the same trend as the congestion level. The workload distribution is more even than the real case data. The reasons for this phenomenon may be the original locations of the vehicles. After adding some perturbations to the system, the dispatch of vehicles achieves a more balanced scheduling.

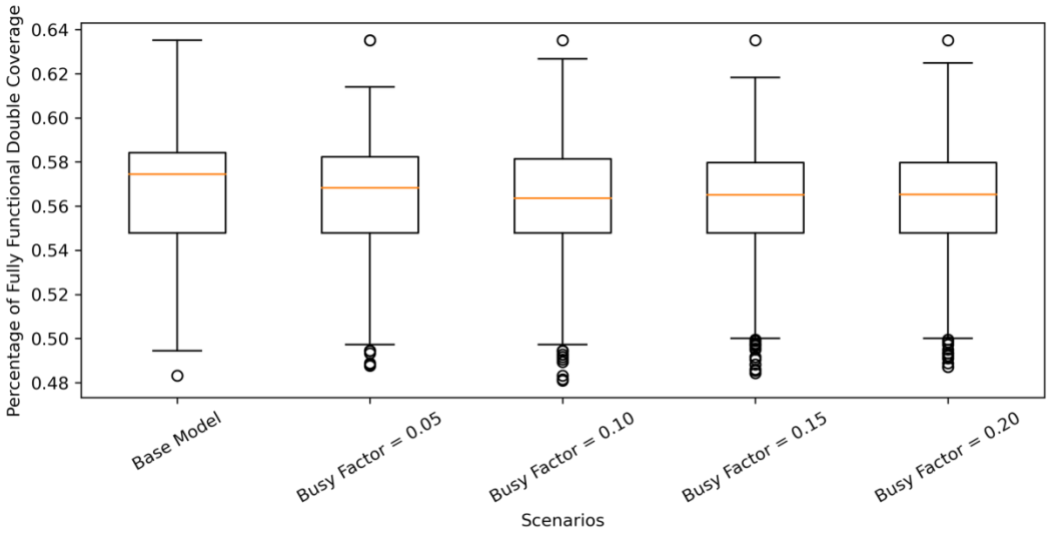


Figure 7.42 Comparison of the percentage of fully functional double coverage for the base model and models with different levels of busy factors

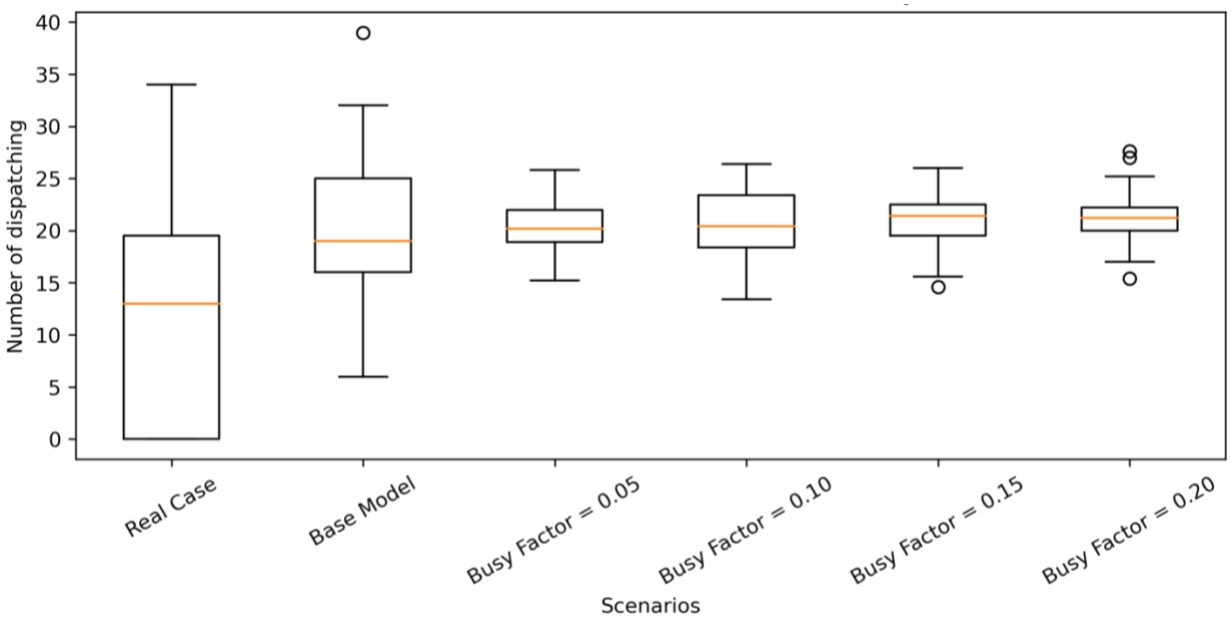


Figure 7.43 Comparison of the workload distribution of the ALS vehicles for the base model and models with different levels of busy factors

Figure 7.44 compares workload distribution for basic life support (BLS) ambulances between real case data, base model, and models with different levels of busy factors. The base model can largely decrease the variation of workload distribution. And the median values can be maintained at a stable level, which is about 12 for proposed models with and without considering the busy factors.

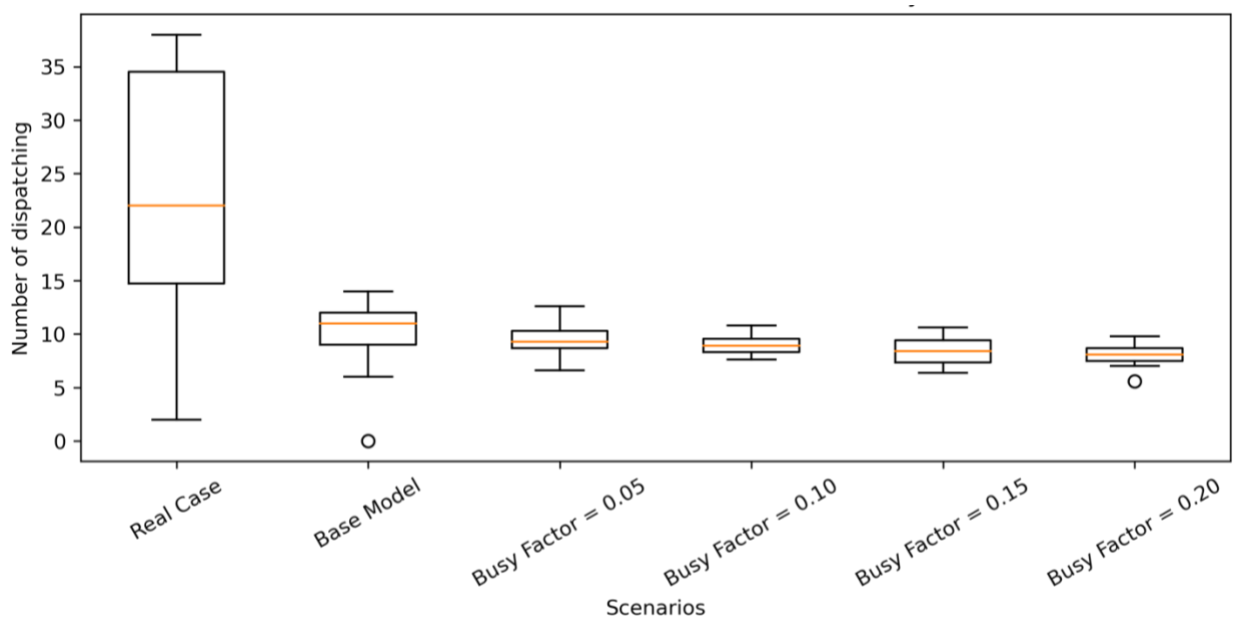


Figure 7.44 Comparison of workload distribution of the BLS vehicles for the base model and models with different levels of busy factors

Figure 7.45 to Figure 7.48 display the comparison of workload distribution between ALS and BLS ambulances under base condition and conditions with different levels of busy factor. These figures show that a more balanced distribution of rescue tasks among the different vehicles is achieved. Compared with the base model, different levels of busy factors do not cause the assignment of rescue tasks to be unbalanced. On the contrary, the tasks assigned between different ambulances are more reasonable in some conditions.

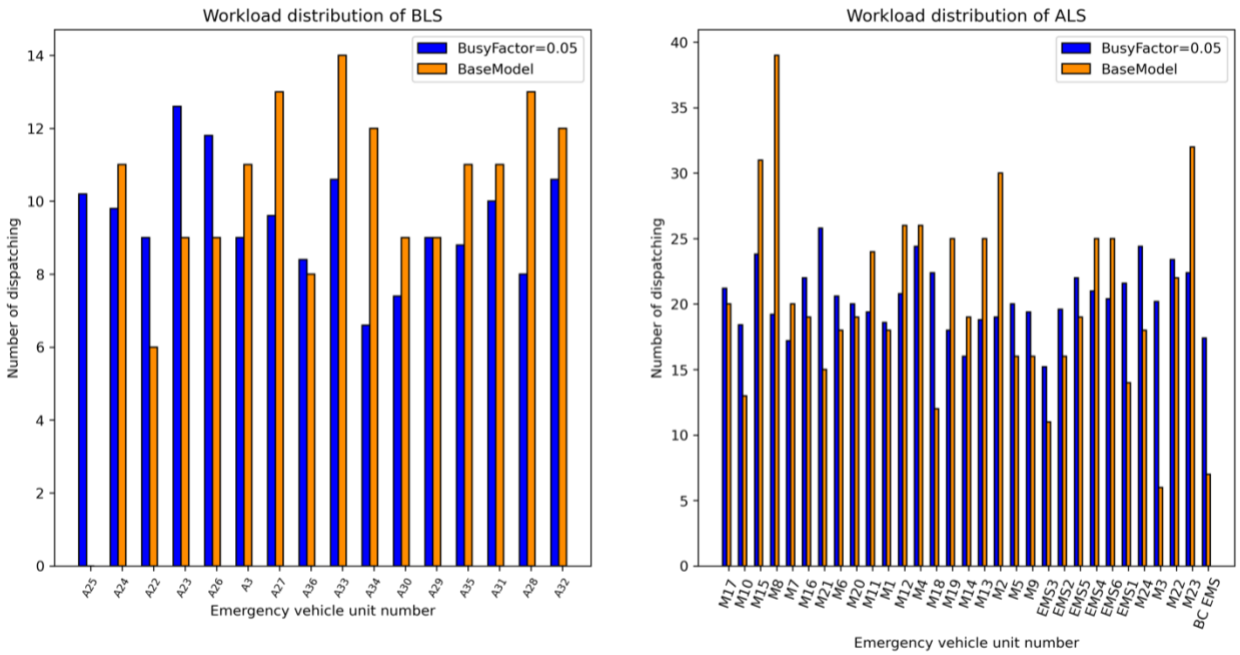


Figure 7.45 Comparison of the workload distribution of the emergency vehicles for the base model and model with busy factor = 0.05

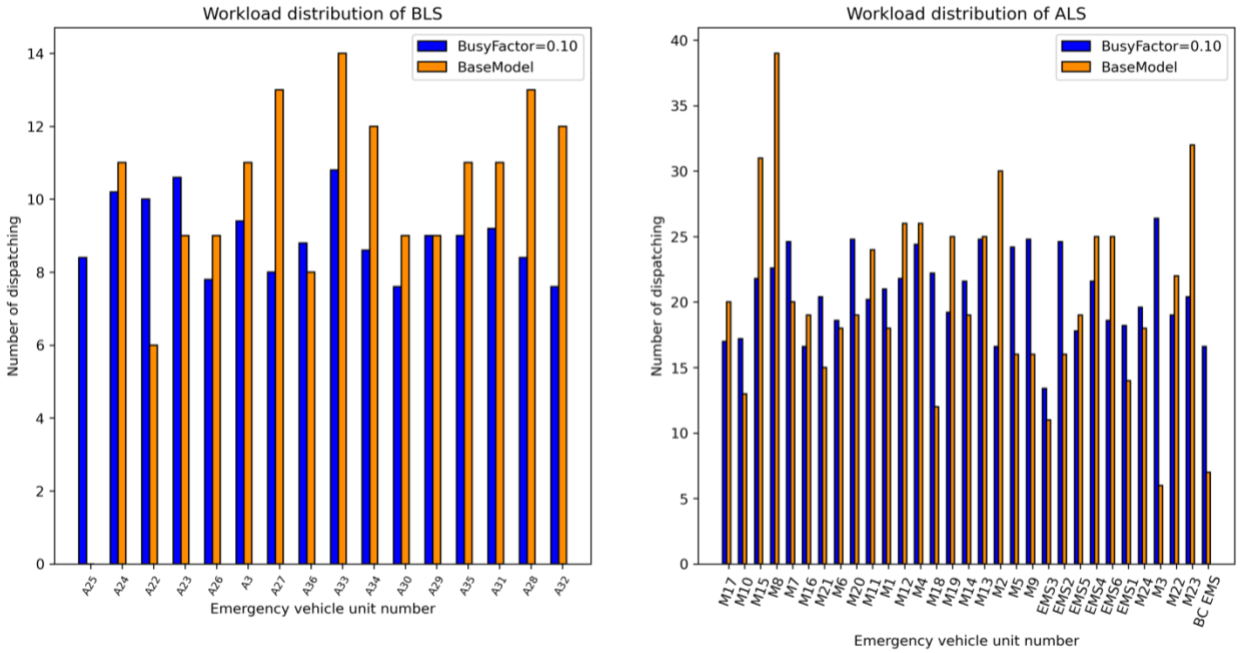


Figure 7.46 Comparison of the workload distribution of the emergency vehicles for the base model and model with busy factor = 0.10

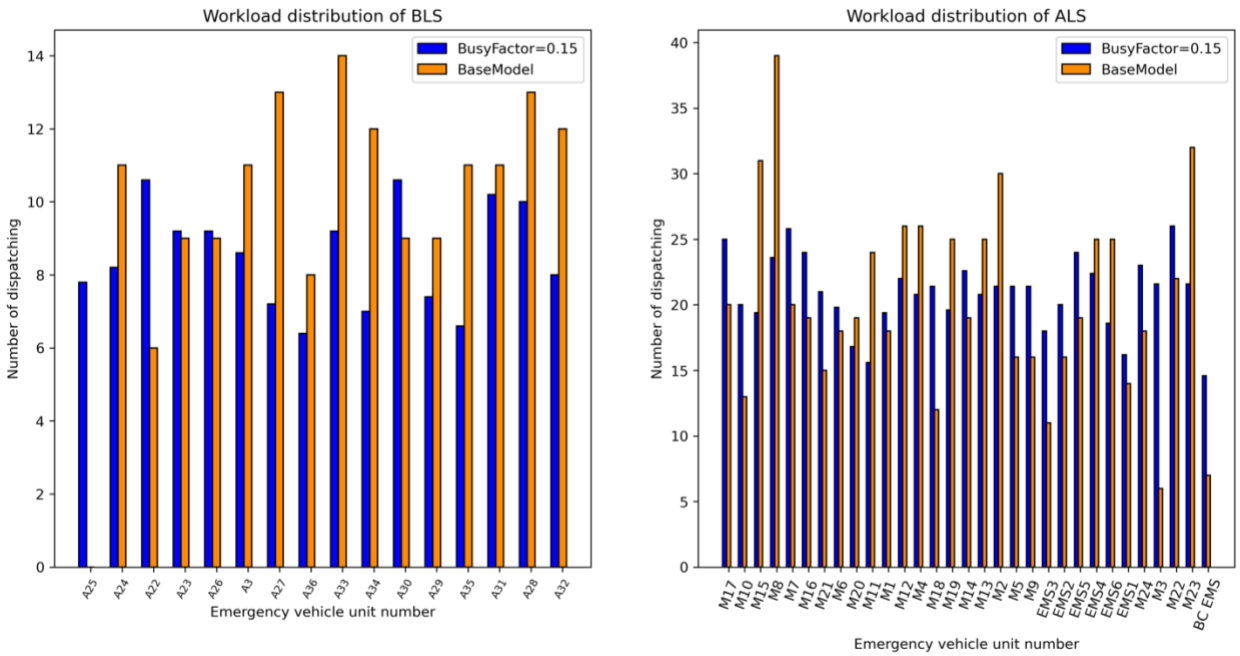


Figure 7.47 Comparison of the workload distribution of the emergency vehicles for the base model and model with busy factor = 0.15

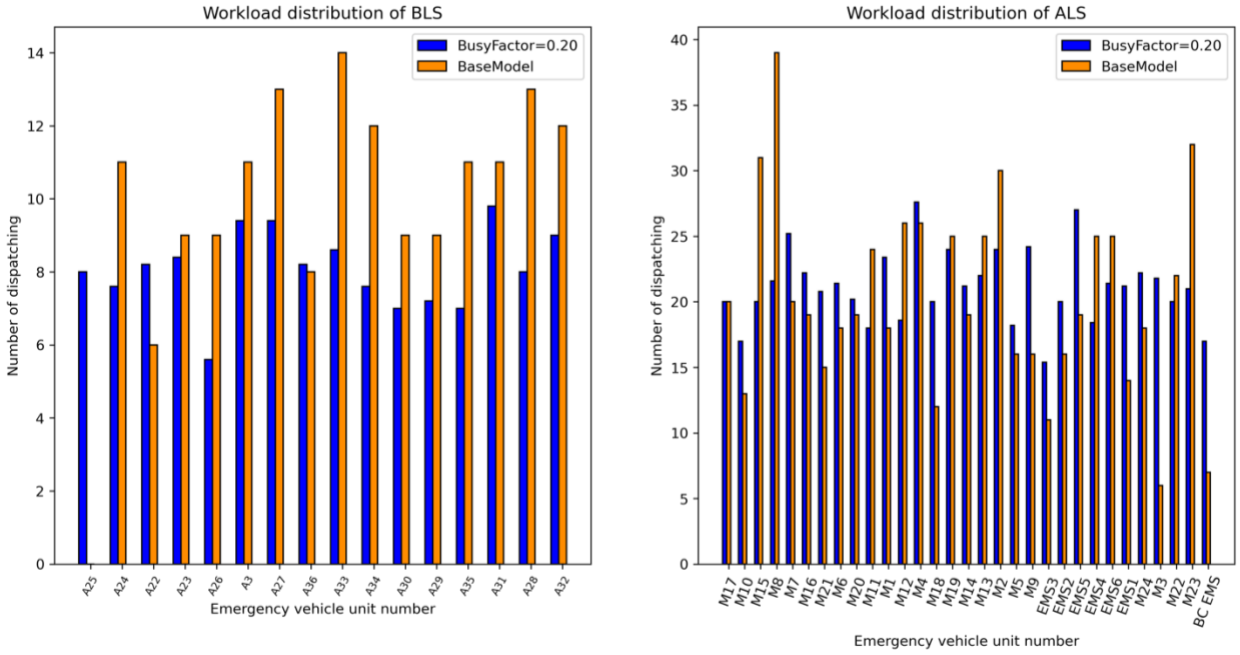


Figure 7.48 Comparison of the workload distribution of the emergency vehicles for the base model and model with busy factor = 0.20

These results illustrate that the proposed model can increase the use of advanced levels vehicles in rescue tasks. The increased ALS vehicle utilization does not result in uneven workload distribution. The daily workload of vehicles can be guaranteed to be smaller than the predetermined level.

The increased utilization of advanced level vehicles can provide more professional pre-hospital treatments to the patients on incident sites.

In the sensitivity analysis, the different levels of perturbations added to the system do not lead to an uneven workload. On the contrary, the workload balance between different vehicles is still balanced. These results also indicate that the model can not only take into account the workload balance but also it has the ability to handle a variety of different complex situations and perturbations.

7.5 Police vehicle analysis

The proposed model is designed to determine dispatching and redeployment decisions for a heterogeneous emergency vehicles fleet, which includes police vehicles, emergency medical vehicles, and fire vehicles. In the previous sections, the results of a comprehensive sensitivity analysis were reported for medical vehicles. Experiments need to be conducted on police and fire vehicles to test the model's performance on the whole heterogeneous emergency vehicles fleet.

In the real-case data, only emergency medical vehicle information is available. The police and fire vehicles dispatching information is not accessible to the research. Some reasonable assumptions and sensitivity analysis are needed to test the model's performance on the real-case data.

One of the most important parameters is the fleet size of each vehicle type. Detailed information about police vehicles, which includes the number of police vehicles, their home

locations, and routing information, is not accessible to the public. It is evident that by increasing the fleet size, the results will improve. However, due to funding and workforce issues, more police vehicles will increase the cost significantly.

In this study, different numbers of police vehicles are tested, and corresponding data is collected and analyzed during the computation. Several basic data metrics will be used to determine and analyze the model's performance. The metrics include average En Route time, percentage of incidents that can be reached within 5 minutes, percentage of incidents that can be reached within 8 minutes, number of redeployments for each incident, and the basic coverage level brought by the police vehicles.

This section considers four scenarios with different numbers of police vehicles, which are 10, 20, 30, and 40. The number of police vehicles is assumed to start from 10 and increase gradually in increments of 10 to reflect the different scenarios of real situations. The section aims to compare the impact of different numbers of police vehicles on the final results and the coverage level provided. The original locations for police vehicles are randomly selected. The original locations of the police vehicles will not affect the final result. The model can find the best configuration for the deployment of police vehicles. In this model, Baltimore City is divided into 356 zones, which means the problem size is small enough compared with using the nodes in the calculation. We can assume the nodes in each zone have the same properties. It is assumed that police vehicles will be deployed in the zones and patrol within each zone to reduce the size of the problem.

Figure 7.49 compares the En Route time between models with different numbers of police vehicles. The figure can give a basic illustration of the distribution of En Route time.

Figure 7.50 shows the average En Route time for models with different numbers of police vehicles. The figure shows that the average En Route time when the police vehicle is 10 is more

than 7.5 minutes. With the increase in the number of vehicles, the average En Route time decreases to about 5 minutes when there are 40 police vehicles.

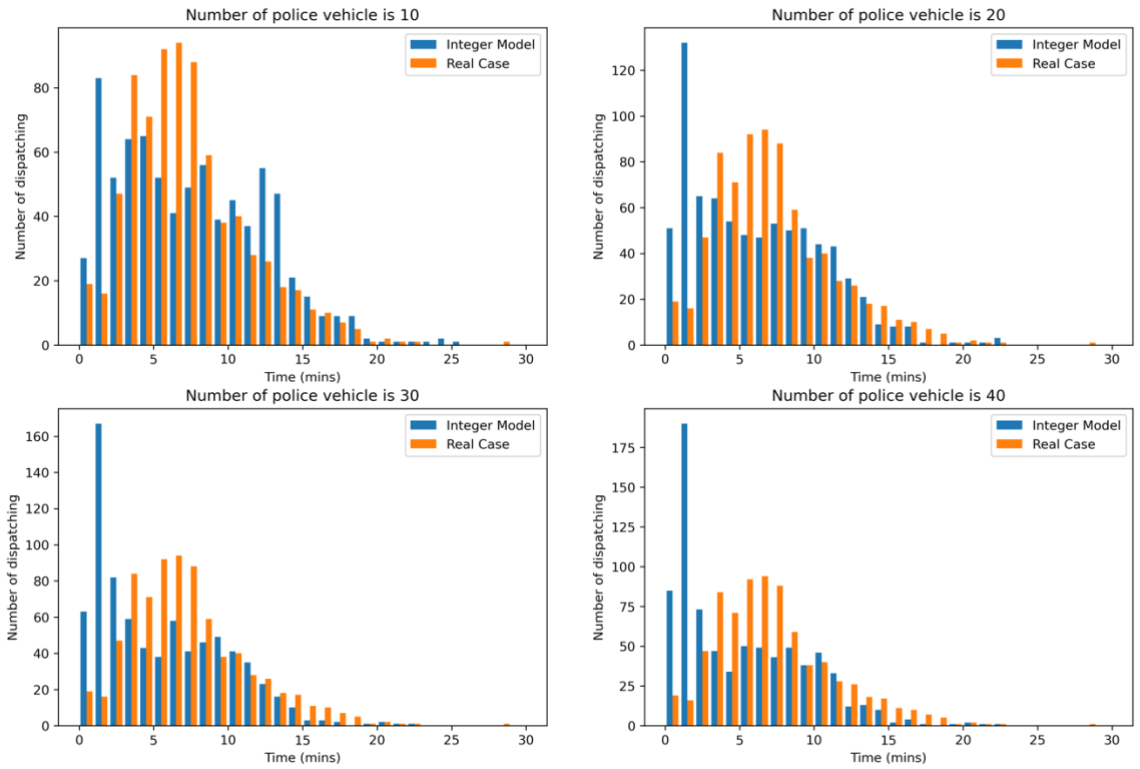


Figure 7.49 Comparison of the En Route time for models with different numbers of police vehicles

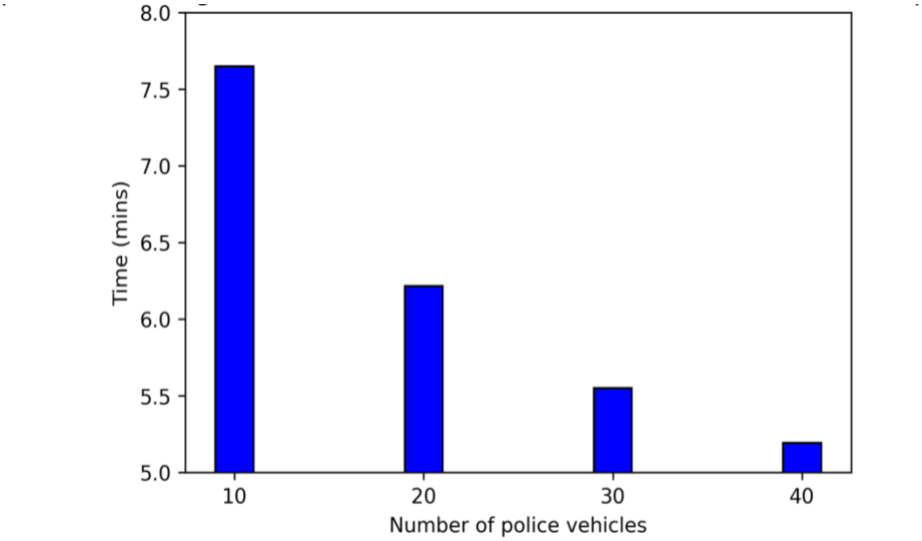


Figure 7.50 The average En Route time for models with different numbers of police vehicles

Figure 7.51 shows the percentage of incidents that can be reached within 5 minutes for different scenarios. Only 37.5 percent of incidents can be reached within 5 minutes if there are ten police vehicles. With the increase in the number of police vehicles, the coverage level will increase, and the percentage of incidents that can be reached within 5 minutes will increase. While the rate of increase is gradually decreasing. When increasing the number of police vehicles from 10 to 20, the percentage will increase by 27%, while if increasing the number of police vehicles from 30 to 40, the percentage will increase by 5%.

Figure 7.52 illustrate the percentage of incidents that can be reached within 8 minutes for the models. Only 55% of incidents can be reached within 8 minutes if there are ten police vehicles. The percentage will increase to 72.5 if there are 40 police vehicles. The increase of percentage brought by the vehicle number is decreasing.

Table 7.12 compares the average number of redeployments between situations with different numbers of police vehicles. When there are 10 or 20 police vehicles, the total coverage level of the system is relatively low. Police vehicles are required to patrol the incident-intensive areas. A relatively small number of deployments are necessary to achieve good coverage. When the number of police vehicles reaches 30 or 40, the system has more flexibility. Then the number of redeployments will increase.

Table 7.12 Comparison of the average number of redeployments between models with different numbers of police vehicles

Number of police vehicles	10	20	30	40
The average number of redeployments	1.2397	1.2257	1.7602	1.7602

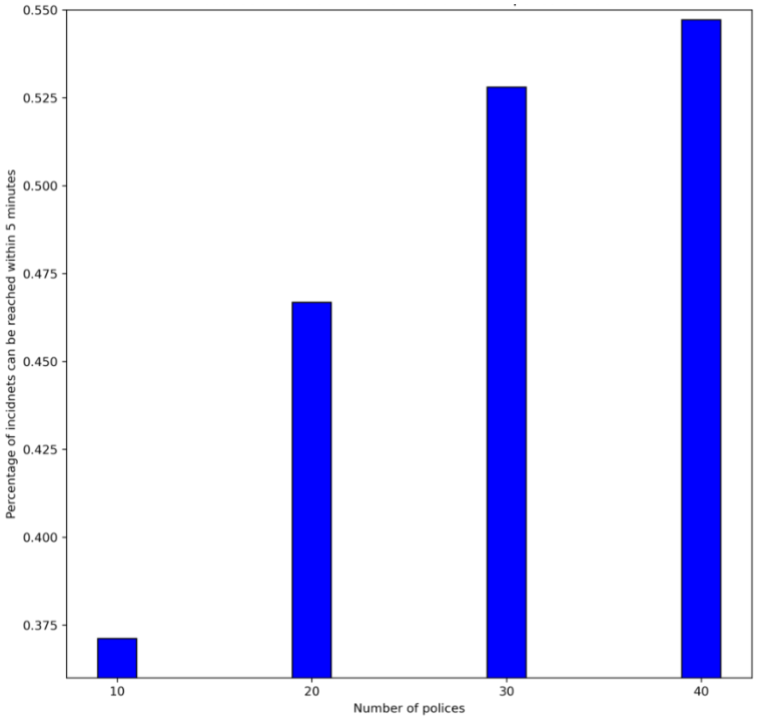


Figure 7.51 Comparison of the percentage of incidents that can be reached within 5 minutes for models with different numbers of police vehicles

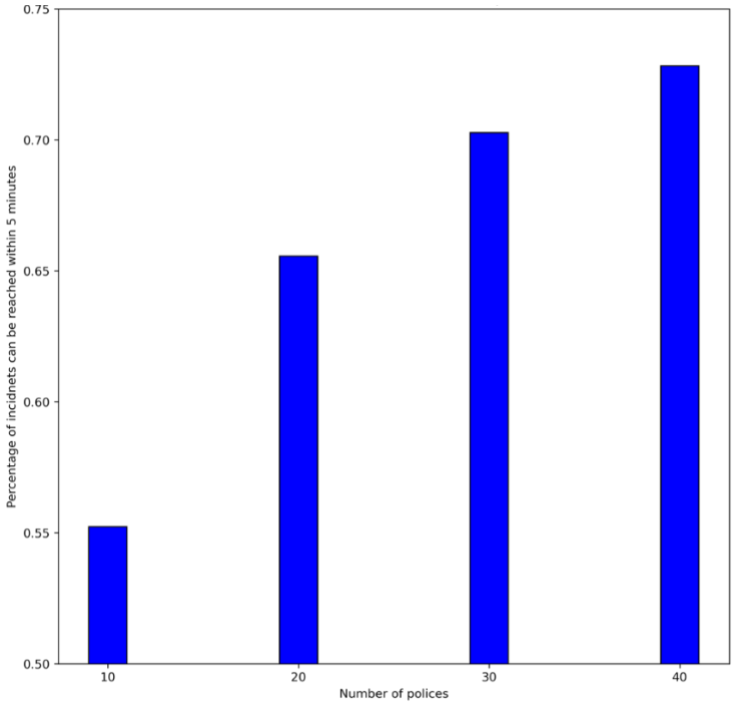


Figure 7.52 Comparison of the percentage of incidents that can be reached within 8 minutes for models with different numbers of police vehicles

7.6 Fire vehicle analysis

The previous section presented the results of a sensitivity analysis conducted on police vehicles. This section presents the results of a sensitivity analysis conducted on fire vehicles. The real requests and dispatching information for fire vehicles are also unavailable, similar to the police vehicle. Unlike police vehicles, the types and number of fire vehicles are known according to the Baltimore Fire Department website. The home station for each fire vehicle is already defined. It is assumed that one type of police vehicle is considered in the model. However, there are three types of fire vehicles according to their functions.

According to the official website, 32 fire engines, 18 fire trucks, and 4 fire quints are strategically located throughout the region.

Differences in the distribution of vehicle types and numbers and variations in actual demand requests can lead to different results with huge differences. In this research, the actual demand requests are for different types of ambulance vehicles. Reasonable assumptions are made to generate demand requests for different types of fire vehicles. The strategy for generating the demand requests is that if the priority of the incident is 1, which means a critically ill or injured person requiring immediate attention, the required fire vehicle type is a fire quint. Otherwise, the required fire vehicle type is randomly selected. Then multiple experiments are repeated to cancel the effect of extreme value and randomness.

Table 7.13 records the detailed results of the experiments. Various metrics are collected. Table 7.13 illustrates that the average En Route time is about 3.25 minutes. Figure 7.53 shows the distribution of En Route time for fire vehicles. The figure demonstrates that most incidents can get access to immediate rescue by fire vehicles within 5 minutes. The median value is about 3 minutes.

Fire vehicles have a wider distribution of stations, and there are more of them than ambulance vehicles, so they have a better performance than ambulance vehicles.

Table 7.13 Summary of performance measurements of fire vehicles

	Fire vehicle
Average En Route time	3.2462
Basic coverage Rate	0.9552
Double coverage Rate	0.9216
Fully functional double coverage Rate	0.7268
Reached within 5 minutes	0.8877
Reached within 8 minutes	0.9681

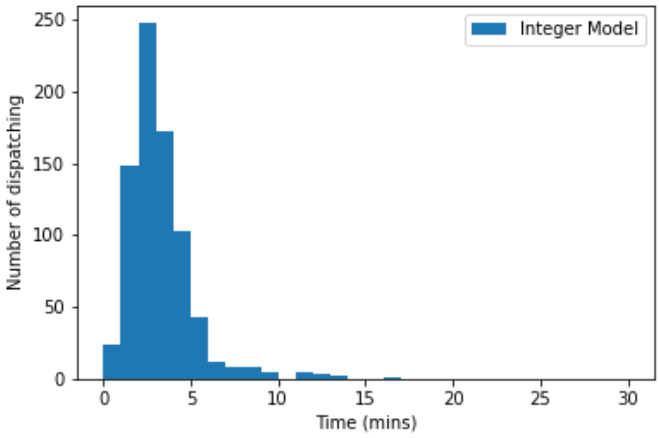


Figure 7.53 En Route time for fire vehicles

The average basic coverage rate for the system can be maintained above 95 percent. Almost all the zones can be covered by fire vehicles. The double coverage rate can also be maintained at a high level. If a zone can be covered by multiple fire vehicles, it is considered double coverage. The average double coverage for the system is 92 percent.

Figure 7.54,

Figure 7.55, and

Figure 7.56 show the basic coverage level, double coverage level, and fully functional double coverage level for the system correspondingly. These three figures clearly demonstrate that the basic coverage rate brought by fire vehicles is around 0.96. In most conditions, the basic coverage rate can be maintained at a high level. Even under certain extreme conditions, the system can still maintain a basic coverage rate of around 90%. Double coverage brought by the fire vehicles is around 92%. The double coverage rate fluctuates more than the basic coverage rate. In some conditions, the double coverage rate drops to below 90%. The Fully functional double coverage brought by the fire vehicles is around 73%. The fully functional double coverage rate is more sensitive to the incidents. The fluctuation is more frequent and dramatic compared with the basic and double coverage rates.

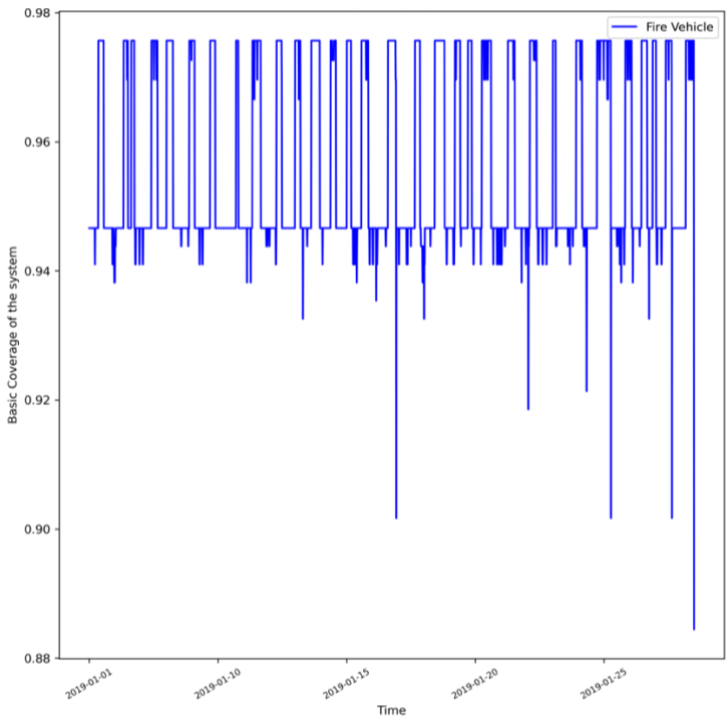


Figure 7.54 Basic coverage brought by fire vehicles

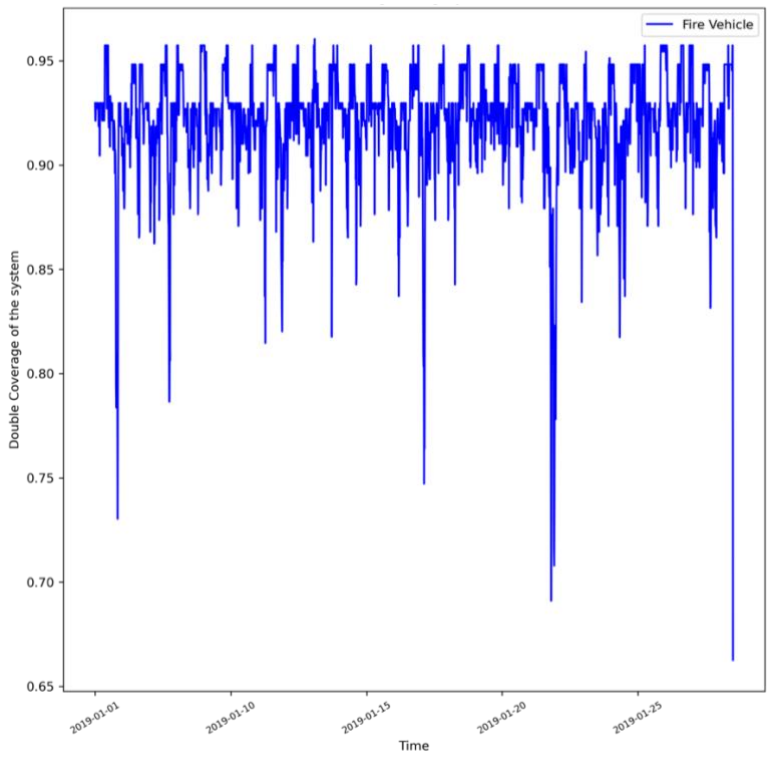


Figure 7.55 Double coverage brought by fire vehicles

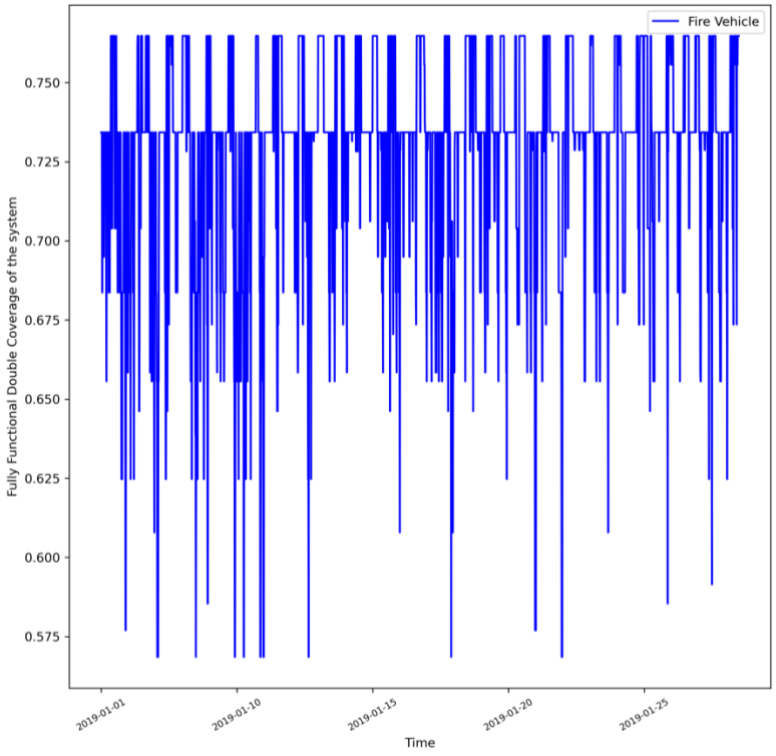


Figure 7.56 Fully functional double coverage brought by fire vehicles

In this research, even when the number of police vehicles is assumed to be 40, compared with ambulance vehicles and fire vehicles, the average En Route time is still larger. The various coverage levels for police vehicles are also lower than the other two types of vehicles. This is because the number of police vehicles is less than the other two types of vehicles. Another reason is the other two types of vehicles have various stations strategically distributed in the city, and the corresponding vehicles can only be located at the stations. This extends the basic coverage rate, reduces the En Route time, and avoids excessive vehicle concentration in certain areas. Other circumstances may lead to a performance improvement. For example, the whole area can be divided into subareas which contain various zones, and different numbers of police vehicles can be assigned to different subareas. This might avoid excessive concentration of police vehicles. In future studies, it would be beneficial if more detailed information is available and can be used.

In this study, there are 24 stations for ambulance vehicles and 38 stations for fire vehicles. The number of fire vehicles is also greater than ambulance vehicles. As a result, fire vehicles can achieve a better performance than ambulances in some metrics. Fire vehicles can achieve shorter rescue times and maintain higher levels of coverage. This is partly due to a more extensive distribution of stations.

In future research, if ambulance stations could be extended to a broader area, this could enhance the basic coverage to a greater extent. For example, if ambulances can use these stations and hospitals for temporary stations, the overall performance might be improved. Of course, corresponding performance tests need to be conducted. This is a trade-off analysis. More extensive stations mean better infrastructure development and more budget.

7.7 Summary

This chapter compared the results of applying the proposed model to the real-case dispatching requests with the real-case operational results. The results illustrated that the proposed model performed very well and often showed much improvement in all the metrics. Also, a comprehensive sensitivity analysis was performed on essential parameters in the model. The analysis intended to investigate how the results will change by modifying those parameters.

In the experiments, three different categories of performance measures were collected to test and quantify the performance of the models: metrics related to the dispatching efficiency, metrics related to the system operating cost, and those related to the workload.

When different levels of congestion or vehicle busy probability are considered, the dispatching time will increase gradually. Meanwhile, rescue efficiency will be reduced, and the number of incidents that can be reached within a predefined time will decrease. These parameters will also cause a drop in the coverage level of the system.

The proposed model is built for a heterogeneous emergency vehicles fleet with well-defined assumptions and constraints. The parameters might need to be tuned to apply the model in other real situations. Once the detailed information about vehicles and incidents is known, the proposed model can be implemented to provide dispatching and redeployment decisions. The model mainly focuses on applying and coordinating three common types of emergency vehicles to mimic field operations. It can be extended to incorporate other rescue methods, such as helicopters. Land-based modes of transportation rely heavily on the road network. Scenarios for using helicopters exist in some extreme cases, where the patients need access to more specialized hospital rescue equipment and professional medicians immediately. The use of helicopters is usually preceded by ambulances on-site to provide basic pre-hospital treatments to the patients. When the required

input information, such as the request demands and travel time matrix, is well defined, the model can provide dispatching guidance for helicopters.

The proposed model aims to provide proper numbers and types of emergency vehicles for medical request demands. The model can be implemented iteratively to find the best fleet size and mix of emergency vehicles indirectly. This would involve using the model with several different fleet size and mix on the same input data to determine the best fleet size and mix.

Chapter 8: Summary, Conclusion, and Future Research

8.1 Summary

In this research, the Emergency Vehicle Management System is studied and analyzed. A performance evaluation system consisting of various categories of statistics and metrics is developed. One of the key practical measurements is response time. Response time not only depends on the vehicle dispatching and routing choice but also relies on the configuration of the deployment of vehicle fleets. So, a comprehensive dispatching and redeployment decision-making model for emergency vehicle operations management is developed in this study. The proposed integrated decision-making model can simultaneously give dispatching and redeployment decisions within a short time. The proposed model can not only improve rescue efficiency but also provide better coverage for the whole area to prepare for the incidents which may happen in the future. This model can develop the best dispatching and redeployment strategy based on real-time information about the system's state.

The contributions of this research can be summarized as follows:

- Heterogeneous emergency vehicle fleets are considered in the system, which are comprised of one type of police vehicle, two types of ambulances: Advanced Life Support (ALS) ambulance and Basic Life Support (BLS) ambulance, and three different types of fire vehicles: Fire Truck, Fire Engine, and Fire Quint. No dispatching and redeployment model that makes decisions for these three types of vehicles simultaneously exists in the literature.
- The model aims to provide different predefined time thresholds of dispatching requirements and redeployment limitations for each type of vehicle. According to the regulation, two different levels of coverage have been considered for demand zones.

- The model attempts to provide basic coverage by using the basic level vehicles for the whole region to shorten the first response time and increase the coverage capacity. Moreover, the model also attempts to provide more coverage by using advanced level vehicles for the critical zones to provide more professional pre-hospital medical treatment for the patients.
- In the model, the whole region is divided into different demand zones. Various strategies are examined to generate the importance of each demand zone by using historical medical data. A simulation system is designed and implemented using these strategies, and the final results are analyzed to compare the performance.
- The model also attempts to balance the workload between different vehicle crews. A new strategy is proposed to restrict dispatching and redeployment actions to avoid excessive work.
- A new mathematical model is formulated and tested in the study that incorporates all the above factors and simultaneously makes dispatching and redeployment decisions.
- A new Discrete Event Simulation (DES) model is developed to test the performance of the proposed model with different parameters over a long time horizon.
- The simulation model is applied in a real-case study to examine the performance of the proposed model. Then a comprehensive sensitivity analysis is conducted to test the ability of the model to handle various scenarios and its response to extreme parameter changes.

8.2 Conclusion

This study conducted comprehensive data analysis on real-case historical medical data. The characteristics and features of medical data were extracted. The number of incidents in the zones

was used to represent the importance of the zone. The proposed model was applied to the real-case data. Three different categories of performance measures were collected and compared to test the accuracy of the results.

- Performance related to the dispatching efficiency: The proposed model makes dispatching and redeployment decisions for a heterogeneous emergency vehicle fleet and provides route guidance for the crew member at the same time. The following conclusions can be drawn from the final results of applying the proposed model to the real-case data.

  The system's overall performance is greatly improved when using the model for dispatching and redeployment decisions. Based on the computational results, the average En Route time of ambulances is reduced from 6.9 minutes in the real-case data to 4.8 minutes. En Route time is reduced by approximately 40 percent. The system coverage is significantly improved while increasing the efficiency of the rescue. The ratio of incidents that can be reached within 5 minutes increased from 0.45 to 0.65 when the proposed model was used, and the ratio of incidents that can be reached within 8 minutes increased from 0.75 to 0.83. The average total coverage rate of the system was increased from 9.3 to 11. These results illustrate that the proposed model can help decision-makers find the optimal configuration for dispatching and redeploying emergency vehicle fleets.

- Performance related to the workload: The model considers the workload balance for crew members and vehicles. Applying the proposed model can largely increase the utilization of advanced level vehicles and provide more professional pre-hospital treatments to patients. The advanced level ambulance utilization will increase from 50 percent to 80 percent. In the meantime, the system also ensures a balanced workload distribution between vehicles. The standard deviation for BLS ambulances in real-case data is 10.9137, while in the case

study is 3.2572. The standard deviation for ALS ambulances in real-case data is 10.0012, while in the case study is 7.1861.

Moreover, an extensive sensitivity analysis was conducted on some important parameters to test the model's performance. The parameters considered for sensitivity analysis in the case study and the results obtained from the analysis are as follows:

- The importance of zones: Four strategies for generating zone importance were developed and tested in the study. The importance of zones not only relies on the frequency of incidents in a period but also depends on the severity of the incidents. The results indicate that the strategy which uses the frequency of incidents to represent the importance of zones can achieve better performance (Strategy 1). This strategy can highlight the characteristics of the geographical distribution of incidents while guaranteeing basic coverage of low-frequency incident areas as much as possible. Except for total coverage improvement, four strategies can achieve similar performance in other metrics.
- Fleet size: To determine the coverage level and rescue efficiency, models with different numbers of police vehicles were tested. The results indicated that with 40 police vehicles, 75 percent of incidents can be reached within 8 minutes, and 55 percent can be reached within 5 minutes. Moreover, the average En Route time is about 5 minutes.
- Congestion level: Different levels of traffic conditions were randomly generated and added to the route segments. With the increase in congestion level, the rescue time would increase as expected. When the congestion level reached 0.5, the average En Route time increased by 21 percent. In some extreme conditions, when the congestion level reached 2, the average En Route time increased by 62.5 percent. It is clear that congestion can cause a significant increase in rescue time.

- Busy factor: This parameter was used to simulate the availability of vehicles in real life. Different levels of busy factors were randomly generated and assigned to each vehicle. With the increased busy factor, the number of incidents reached within 5 and 8 minutes decreased, and the average En Route time increased.

8.3 Future research

In this research, various factors and parameters related to real case situations were considered and added to the model to reflect the characteristics of the problem. Moreover, some operational settings were also taken into account in the model to simulate realistic situations. Even so, future studies still need to consider or improve some issues. In this section, some recommendations for future studies are discussed.

8.3.1 Mathematical formulation

The real-time dispatching and redeployment of heterogeneous emergency vehicle fleets model is formulated as a deterministic integer programming model. Various parameters are considered to simulate realistic scenarios in real life. Some simplified assumptions are made to deal with unknown conditions and improve computational efficiency. In the model, diversions are not considered. When the vehicles are dispatched to the incident sites, the destinations of vehicles are determined and cannot be changed. Diversions are reasonable measures to reduce rescue time in some extreme cases, but they heavily rely on computational efficiency and good communication. When an urgent incident happens, the management center needs to respond quickly to make a decision. If the model takes too long to compute, the optimal time for diversion may be missed. One avenue for future research is to develop efficient models and solution algorithms that allow for real-time diversions of vehicles if deemed necessary.

Another approach that can be investigated for future studies is to develop a two-stage model. The proposed model in this research can simultaneously make dispatching and redeployment decisions. There are three measures to control the number of redeployments in making the decisions. The first one is adding a constraint to limit redeployment distance. If the redeployment distance is longer than a pre-defined value, then the redeployment decision is unsuitable. Second, if a station has insufficient space, the vehicles cannot be dispatched to that station. Third, if the workload for a vehicle exceeds the limit, then this vehicle is not allowed to be redeployed. Otherwise, redeployment decisions will be made to improve the total coverage for the system, even if the improvement is minimal. Sometimes the benefits of system coverage enhancement may be less than the costs associated with redeployments. An approach to tackle this problem could be using a two-stage problem. In the first stage, the dispatching and redeployment decisions are made when the events come into the system. In the second stage, the coverage improvement can be calculated to evaluate whether redeployments are beneficial.

The third approach that can be investigated for future studies is to define more precise parameters in the model. This study assumes all stations have the same capacities for a specific type of vehicle. An approach to handle this issue would be to incorporate more precise parameters.

8.3.2 Simulation model

In this study, a discrete event simulation (DES) system was developed to mimic the entire operation of an emergency response system over a long time horizon. However, some parts are missing in the simulation model due to the absence of data and assumptions.

First, due to the limitation of real-case data, only emergency medical vehicle information and incident requests are accessible. Real operational information about police vehicles and fire

vehicles is not available. For the police vehicles, the basic dispatch and patrol logic is not clear. This study assumes that police vehicles can reach any zones and patrol around that area. This assumption can lead to the concertation of police vehicles in certain areas, such as downtown. In the future, approaches to handle this issue could be using more comprehensive real-case data or establishing a more realistic dispatching and redeployment logic for police vehicles.

Second, some assumptions about the medical vehicles need to be made to continue the study. In this study, it is assumed that ambulances can only be redeployed to other stations. In future studies, it can be assumed that ambulances can be located at any station and hospital if capacity allows. This measure can add more flexibility to the system and enlarge the coverage of the system.

Third, because historical traffic information is not available, the travel time matrix used in the computation cannot reflect the real situation. Two methods were used in this study to tackle this issue. First, the shortest travel times without considering traffic conditions were calculated by implementing Dijkstra's algorithm using a Python package. Dijkstra's method is a simple but powerful algorithm to calculate the shortest path. Other shortest path algorithms can be investigated in future research, and the best algorithm can be selected to generate the travel time matrix. Second, different congestion levels were generated and added to the model to test the model's performance and check the effect of the traffic condition. However, during the random congestion information generation, generating the corresponding traffic condition information for a specific road segment at a particular time is impossible due to the insufficiency of historical data. More sufficient data can be collected and used to handle this issue in future research. Accurate traffic data or the congestion distribution at different zones or road segments throughout the period can be obtained at the moment of the incident.

8.3.3 Sensitivity analysis

A sensitivity analysis is conducted on some critical parameters in this research. Some parameters that may affect the model's performance still need to be tuned, such as the penalties for deficiency of all kinds of vehicles for dispatching and redeployment and penalties for late arrival. The values of these parameters affect decision-making. For example, if the penalty for coverage deficiency is too large, keeping the vehicles on standby at the stations would be preferred. If the penalty for late arrival is larger than the penalty for deficiency of dispatching, then keeping the vehicles at the stations would be preferred if the En Route time is larger than the predefined values. Therefore, these parameters must be appropriately set to ensure the smooth operation of the entire system. A more extensive sensitivity analysis on important parameters and their combinations can be conducted in future studies.

8.3.4 Crew scheduling

The crew scheduling problem is a part of the vehicle assignment problem. This research considers some basic assumptions and preferences related to the crew and workload balance, but combining vehicle assignment and crew scheduling problems is complex and challenging. This issue can be taken into account in the model in future research.

8.3.5 Economic analysis

The core of the proposed model is finding the best configuration for deploying all vehicles to obtain maximum system coverage. This practice will increase the working time of the vehicles and crews. In the workload balance analysis, only the redeployment numbers and cumulated work hours have been considered in the model. The corresponding economic analysis of the tradeoffs between the benefits of coverage improvement and operational cost needs to be conducted in future studies.

Reference

A Brief History of Emergency Medical Services in the United States EMRA. (n.d.). Retrieved May 31, 2021, from https://www.emra.org/about-emra/history/ems-history/

Alanis, R., Ingolfsson, A., & Kolfal, B. (2013). A markov chain model for an EMS system with repositioning. *Production and Operations Management*, *22*(1), 216–231. https://doi.org/10.1111/j.1937-5956.2012.01362.x

Ali Nasrollahzadeh, A., Khademi, A., & Mayorga, M. E. (2018). Real-time ambulance dispatching and relocation. *Manufacturing and Service Operations Management*, *20*(3), 467–480. https://doi.org/10.1287/msom.2017.0649

Andersson, T., & Värbrand, P. (2007). Decision support tools for ambulance dispatch and relocation. *Journal of the Operational Research Society*, *58*(2), 195–201. https://doi.org/10.1057/palgrave.jors.2602174

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 and Operations Research*, Vol. 78. https://doi.org/10.1016/j.cor.2016.09.016

Ball, M. O., & Lin, F. L. (1993). Reliability model applied to emergency service vehicle location. *Operations Research*, *41*(1), 18–36. https://doi.org/10.1287/opre.41.1.18

Başar, A., Çatay, B., & Ünlüyurt, T. (2011). A multi-period double coverage approach for locating the emergency medical service stations in Istanbul. *Journal of the Operational Research Society*, *62*(4), 627–637. https://doi.org/10.1057/jors.2010.5

Batta, R., Dolan, J. M., & Krishnamurthy, N. N. (1989). Maximal expected covering location problem. Revisited. *Transportation Science*, *23*(4), 277–287. https://doi.org/10.1287/trsc.23.4.277

Bélanger, V., Kergosien, Y., Ruiz, A., & Soriano, P. (2016). An empirical comparison of relocation strategies in real-time ambulance fleet management. *Computers and Industrial Engineering*, *94*, 216–229. https://doi.org/10.1016/j.cie.2016.01.023

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. https://doi.org/10.1016/j.ejor.2018.02.055

Bélanger, Valérie, Lanzarone, E., Ruiz, A., & Soriano, P. (2015a). *The Ambulance Relocation and Dispatching Problem*. (November).

Bélanger, Valérie, Lanzarone, E., Ruiz, A., & Soriano, P. (2015b). *The Ambulance Relocation and Dispatching Problem*. Retrieved from https://sci-hub.do/https://www.cirrelt.ca/documentstravail/cirrelt-2015-59.pdf

Beraldi, P., & Bruni, M. E. (2009). A probabilistic model applied to emergency service vehicle location. *European Journal of Operational Research*, *196*(1), 323–331. https://doi.org/10.1016/j.ejor.2008.02.027

Carvalho, A. S., Captivo, M. E., & Marques, I. (2020). Integrating the ambulance dispatching and relocation problems to maximize system's preparedness. *European Journal of Operational Research*, *283*(3), 1064–1080. https://doi.org/10.1016/j.ejor.2019.11.056

Church, R. L., & Davis, R. R. (1992). The fixed charge maximal covering location problem. *Papers in Regional Science*, *71*(3), 199–215. https://doi.org/10.1007/BF01434264

Daskin, M. S. (1983). Maximum expected covering location model: formulation, properties and heuristic solution. *Transportation Science*, *17*(1), 48–70. https://doi.org/10.1287/trsc.17.1.48

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. https://doi.org/10.1007/s10729-014-9271-5

Eaton, D. J., Ml Sánchez, H. U., & Morgan, J. (1986). Determining Ambulance Deployment in Santo Domingo, Dominican Republic. In *Source: The Journal of the Operational Research Society* (Vol. 37).

EMS.gov | What is EMS? (n.d.). Retrieved May 31, 2021, from https://www.ems.gov/whatisems.html

EMS Response Time | fems. (n.d.). Retrieved May 31, 2021, from https://fems.dc.gov/page/ems-response-time

Enayati, S., Mayorga, M. E., Rajagopalan, H. K., & Saydam, C. (2018). Real-time ambulance redeployment approach to improve service coverage with fair and restricted workload for EMS providers. *Omega (United Kingdom)*. https://doi.org/10.1016/j.omega.2017.08.001

Enayati, S., Özaltın, O. Y., Mayorga, M. E., & Saydam, C. (2018). Ambulance redeployment and dispatching under uncertainty with personnel workload limitations. *IISE Transactions*, *50*(9), 777–788. https://doi.org/10.1080/24725854.2018.1446105

Erdös, P., & Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae*, *6*, 290–297.

Gendreau, M., Laporte, G., & Semet, F. (2006). The maximal expected coverage relocation problem for emergency vehicles. *Journal of the Operational Research Society*, *57*(1), 22–28. https://doi.org/10.1057/palgrave.jors.2601991

Gendreau, Michel. (1997). Solving an ambulance location model by tabu search. In *Locutmtr Sckwe* (Vol. 5).

Gendreau, Michel, Laporte, G., & Semet, F. (2001). A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Computing*. https://doi.org/10.1016/S0167-8191(01)00103-X

Geroliminis, N., Karlaftis, M. G., & Skabardonis, A. (2009). A spatial queuing model for the emergency vehicle districting and location problem. *Transportation Research Part B: Methodological*, *43*(7), 798–811. https://doi.org/10.1016/j.trb.2009.01.006

Goldberg, J., Dietrich, R., Chen, J. M., Mitwasi, M. G., Valenzuela, T., & Criss, E. (1990). Validating and applying a model for locating emergency medical vehicles in Tucson, AZ. In *European Journal of Operational Research* (Vol. 49).

Hogan, K., & ReVelle, C. (1986). Concepts and Applications of Backup Coverage. *Management Science*, *32*(11), 1434–1444. https://doi.org/10.1287/mnsc.32.11.1434

Iannoni, A. P., & Morabito, R. (2007). A multiple dispatch and partial backup hypercube queuing model to analyze emergency medical systems on highways. *Transportation Research Part E: Logistics and Transportation Review*, *43*(6), 755–771. https://doi.org/10.1016/j.tre.2006.05.005

Iannoni, A. P., Morabito, R., & Saydam, C. (2009). An optimization approach for ambulance location and the districting of the response segments on highways. *European Journal of Operational Research*, *195*(2), 528–542. https://doi.org/10.1016/j.ejor.2008.02.003

Ingolfsson, A. (2013). EMS planning and management. *International Series in Operations Research and Management Science*, *190*, 105–128. https://doi.org/10.1007/978-1-4614-6507-2_6

Jagtenberg, C. J., Bhulai, S., & van der Mei, R. D. (2015). An efficient heuristic for real-time ambulance redeployment. *Operations Research for Health Care*. https://doi.org/10.1016/j.orhc.2015.01.001

Jagtenberg, C. J., Bhulai, S., & van der Mei, R. D. (2017). Dynamic ambulance dispatching: is the closest-idle policy always optimal? *Health Care Management Science*, *20*(4), 517–531. https://doi.org/10.1007/s10729-016-9368-0

Jagtenberg, C. J., van den Berg, P. L., & van der Mei, R. D. (2017). Benchmarking online dispatch algorithms for Emergency Medical Services. *European Journal of Operational Research*, *258*(2). https://doi.org/10.1016/j.ejor.2016.08.061

Lam, S. S. W., Zhang, J., Zhang, Z. C., Oh, H. C., Overton, J., Ng, Y. Y., & Ong, M. E. H. (2015). Dynamic ambulance reallocation for the reduction of ambulance response times using system status management. *American Journal of Emergency Medicine*. https://doi.org/10.1016/j.ajem.2014.10.044

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. https://doi.org/10.1016/j.trc.2016.05.023

Mandell, M. B. (1998). Covering models for two-tiered emergency medical services systems. *Location Science*, *6*(1–4), 355–368. https://doi.org/10.1016/S0966-8349(98)00058-8

Marianov, V., & Revelle, C. (1994). The queuing probabilistic location set covering problem and some extensions. *Socio-Economic Planning Sciences*, *28*(3), 167–178. https://doi.org/10.1016/0038-0121(94)90003-5

Mark S. Daskin, E. H. S. (1981). A hierarchical objective set covering model for emergency medical service vehicle deployment. *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699. https://doi.org/10.1017/CBO9781107415324.004

Maxwell, M. S., Restrepo, M., Henderson, S. G., & Topaloglu, H. (2010). Approximate dynamic programming for ambulance redeployment. *INFORMS Journal on Computing*, *22*(2), 266–281. https://doi.org/10.1287/ijoc.1090.0345

Morabito, R., Chiyoshi, F., & Galvão, R. D. (2008). Non-homogeneous servers in emergency medical systems: Practical applications using the hypercube queueing model. *Socio-Economic Planning Sciences*, *42*(4), 255–270. https://doi.org/10.1016/j.seps.2007.04.002

Naoum-Sawaya, J., & Elhedhli, S. (2013). A stochastic optimization model for real-time ambulance redeployment. *Computers and Operations Research*. https://doi.org/10.1016/j.cor.2013.02.006

Rajagopalan, H. K., Saydam, C., & Xiao, J. (2008). A multiperiod set covering location model for dynamic redeployment of ambulances. *Computers and Operations Research*. https://doi.org/10.1016/j.cor.2006.04.003

Repede, J. F., & Bernardo, J. J. (1994). Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. In *European Journal of Operational Research* (Vol. 75).

Revelle, C., & Hogan, K. (1989). The maximum reliability location problem and a-reliable p-center problem: derivatives of the probabilistic location set covering problem. In *Annals of Operations Research* (Vol. 18).

Roemer, R., Kramer, C., & Frink, J. E. (1977). Planning Urban Health Services. *Medical Care*, *15*(5), 450–451. https://doi.org/10.1097/00005650-197705000-00014

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

Schilling, D., Elzinga, D. J., Cohon, J., Church, R., & Schilling, D. (1979). *The TeamFleet*

Models for Simultaneous Facility and Equipment Siting-annotated.pdf. *13*(2), 163–175.
Schmid, V. (2012). Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. *European Journal of Operational Research*, *219*(3), 611–621. https://doi.org/10.1016/j.ejor.2011.10.043
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. https://doi.org/10.1016/j.ejor.2010.06.033
Sharifi, E. (2014). *Active relocation and dispatching of heterogeneous emergency vehicles*.
Sofianopoulos, S., Williams, B., Archer, F., & Thompson, B. (2011). The exploration of physical fatigue, sleep and Depression in paramedics: A pilot study. *Journal of Emergency Primary Health Care*, *9*(1), 2011–990435. https://doi.org/10.33151/ajp.9.1.37
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. https://doi.org/10.1016/j.seps.2009.04.002
Stout, J. (1989). System status management. *J Emerg Med Serv*, *14*(April), 65–67.
Sudtachat, K., Mayorga, M. E., Chanta, S., & Albert, L. A. (2020). Joint relocation and districting using a nested compliance model for EMS systems. *Computers and Industrial Engineering*, *142*(January), 106327. https://doi.org/10.1016/j.cie.2020.106327
Sudtachat, K., Mayorga, M. E., & Mclay, L. A. (2016). A nested-compliance table policy for emergency medical service systems under relocation. *Omega (United Kingdom)*. https://doi.org/10.1016/j.omega.2015.06.001
Tassone, J., & Choudhury, S. (2020). *A Comprehensive Survey on the Ambulance Routing and Location Problems*. 1–30. Retrieved from http://arxiv.org/abs/2001.05288
Toro-Díaz, H., Mayorga, M. E., Chanta, S., & McLay, L. A. (2013). Joint location and dispatching decisions for Emergency Medical Services. *Computers and Industrial Engineering*, *64*(4), 917–928. https://doi.org/10.1016/j.cie.2013.01.002
Ullrich, O., & Lückerath, D. (2017). An Introduction to Discrete-Event Modeling and Simulation. *SNE Simulation Notes Europe*, *27*(1), 9–16. https://doi.org/10.11128/sne.27.on.10362
Understanding and Measuring Fire Department Response Times. (n.d.). Retrieved May 31, 2021, from https://www.lexipol.com/resources/blog/understanding-and-measuring-fire-department-response-times/
van Barneveld, T. C., Bhulai, S., & van der Mei, R. D. (2017). A dynamic ambulance management model for rural areas: Computing redeployment actions for relevant performance measures. *Health Care Management Science*, *20*(2), 165–186. https://doi.org/10.1007/s10729-015-9341-3
van Barneveld, T. C., van der Mei, R. D., & Bhulai, S. (2017). Compliance tables for an EMS system with two types of medical response units. *Computers and Operations Research*, *80*. https://doi.org/10.1016/j.cor.2016.11.013
Van Den Berg, P. L., & Aardal, K. (2015). Time-dependent MEXCLP with start-up and relocation cost. *European Journal of Operational Research*, *242*(2). https://doi.org/10.1016/j.ejor.2014.10.013
Who We Are. (n.d.). Retrieved May 31, 2021, from https://www.miemss.org/home/who-we-are
Yang, S. (2006). *Integrated management of emergency vehicle fleet*.
Yang, S., Hamedi, M., & Haghani, A. (2004). Integrated approach for emergency medical service location and assignment problem. *Transportation Research Record*, (1882), 184–

192. https://doi.org/10.3141/1882-22
Yang, S., Hamedi, M., & Haghani, A. (2005). Online dispatching and routing model for emergency vehicles with area coverage constraints. *Transportation Research Record*, (1923), 1–8. https://doi.org/10.3141/1923-01