5-2014

STRATEGIES TO IMPROVE THE EFFICIENCY OF EMERGENCY MEDICAL SERVICE (EMS) SYSTEMS UNDER MORE REALISTIC CONDITIONS

Kanchala Sudtachat
Clemson University, kanchas@g.clemson.edu

STRATEGIES TO IMPROVE THE EFFICIENCY OF EMERGENCY MEDICAL SERVICE (EMS) SYSTEMS UNDER MORE REALISTIC CONDITIONS

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Industrial Engineering

---

by
Kanchala Sudtachat
May 2014

Accepted by:
Dr. Scott J. Mason, Committee Chair
Dr. William G. Ferrell
Dr. Kevin M. Taaffe
Dr. Maria E. Mayorga

ABSTRACT

Emergency medical service (EMS) systems provide medical care to pre-hospital patients who need rapid response and transportation. This dissertation proposes a new realistic approach for EMS systems in two major focuses: multiple unit dispatching and relocation strategies.

This work makes recommendations for multiple-unit dispatch to multiple call priorities based on simulation optimization and heuristics. The objective is to maximize the expected survival rate. Simulation models are proposed to determine the optimization. A heuristic algorithm is developed for large-scale problems. Numerical results show that dispatching while considering call priorities, rather than always dispatching the closest medical units, could improve the effectiveness of EMS systems. Additionally, we extend the model of multiple-unit dispatch to examine fairness between call priorities. We consider the potentially-life-threatening calls which could be upgraded to life-threatening. We formulate the fairness problem as an integer programming model solved using simulation optimization. Taking into account fairness between priorities improves the performance of EMS systems while still operating at high efficiency.

As another focus, we consider dynamic relocation strategy using a nested-compliance table policy. For each state of the EMS systems, a decision must be made regarding exactly which ambulances will be allocated to which stations. We determine the optimal nested-compliance table in order to maximize the expected coverage, in the binary sense, as will be later discussed. We formulate the nested-compliance table model

as an integer program, for which we approximate the steady-state probabilities of EMS system to use as parameters to our model. Simulation is used to investigate the performance of the model and to compare the results to a static policy based on the adjusted maximum expected covering location problem (AMEXCLP). Additionally, we extend the nested-compliance table model to consider an upper bound on relocation time. We analyze the decision regarding how to partition the service area into smaller sub-areas (districts) in which each sub-area operates independently under separate relocation strategies. We embed the nested-compliance table model into a tabu search heuristic algorithm. Iteration is used to search for a near-optimal solution. The performance of the tabu search heuristic and AMEXCLP are compared in terms of the realized expected coverage of EMS systems.

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Maria E. Mayorga who guided and showed me how to analyze and develop the mathematical formulations. Her proof-writing helps me to fruitfully complete this research. Her patience encourages me to work in the right way to complete this dissertation. I extend my thanks to Dr. Scott J. Mason, Dr. William G. Ferrell and Dr. Kevin M. Taaffe for sharing idea as my committees and Dr. Laura A Mclay for her real-world data. My gratitude goes to Kendall McKenzie for her proof-writing, Nittaya Muangnak for advice on coding java programming language, and all laboratory mates at Clemson University and North Carolina State University for proof-writing and suggestions. Finally this paper manuscript cannot be successful without my family for support and stand beside me.

TABLE OF CONTENTS

Page

Table of Contents (Continued)

Page

LIST OF TABLES

List of Tables (Continued)

Table Page

LIST OF FIGURES

List of Figures (Continued)

Figure Page

CHAPTER ONE

INTRODUCTION

Emergency medical service (EMS) systems are health care systems that provide medical care and transportation of patients to hospitals when needed, thus potentially saving lives. Gibson [1] discussed the development of EMS systems. During the late 1960s, early studies of EMS systems focused on planning for fire and police departments and on assigning appropriate ambulances to assist on-scene at an accident. Since then, literature related to planning of EMS systems has experienced substantial growth and development. EMS access to patients is crucial to developing new strategies to improve the current EMS systems. EMS planning is challenging because of the varying severity of emergency calls on the scene of accidents, uncertainty in response time of the ambulances, uncertainty in demand, etc. The goal of EMS systems is to improve performance by increasing survival probability or reducing response time. Response time refers to the interval between the arrival of a call and the time at which the ambulance reaches the scene. Therefore, rapid response to a call can have a dramatic effect on the outcome of the patient and the performance of EMS systems.

Few research studies consider taking realistic features of EMS problems into account because many complexities will inherently be introduced into their formulations. However, these features cannot be ignored if the goal is to provide solutions that can be implemented in practice to real-world problem. *This dissertation proposes two major models to improve efficiency of EMS systems by taking into account some realities.*

Specifically, our primary focus is on improving the performance of EMS systems using multiple unit dispatching strategies and real-time relocation strategies. We investigate the benefits of our models via computational results based on datasets from real-world problems.

The contribution of this dissertation is to improve the efficiency of EMS systems in two ways. One area provides the structure of optimal and near optimal multiple dispatching policies to multiple call priorities under realistic on-scene conditions when the goal is to maximize the expected survival probability. We also propose the fairness strategy between call priorities. The near optimal multiple dispatching policy is implemented for each call priority and each call zone. Another area of study is to propose the optimal nested-compliance table policy under real-time relocation conditions that maximizes the expected binary coverage. Later, when implementing the optimal nested-compliance table policy whole service system is partitioned into small sub-systems, we determine the relocation time boundaries for the EMS system that maximizes the realized expected coverage. Both the multiple dispatching policy and the nested-compliance table policy are pre-specified policies, which are implemented in practice in EMS systems. In short, we are able to improve EMS system performance under more realistic conditions that have thus far been ignored in the literature.

When focusing on ambulance dispatching strategies, the dispatch center plays a key role in EMS systems. A dispatcher determines the severity of calls in order to dispatch the appropriate ambulance unit(s). Dispatchers typically have an ordered

preference list for each demand zone. The list assigns a ranking position to available ambulances so that the dispatcher would send the first unit on the ranking if available; or the second ambulance on the ranking if the first is busy, and so on. Dispatch decisions can largely impact system outcomes such as patient survival rate and response time.

To improve the response time of the dispatch center, EMS systems may implement priority dispatch. Priority dispatch relies on the idea of properly matching servers with severity of calls when ambulances are limited. Priority dispatch helps increase the number of available ambulances, and it improves the utilization of resources, which in turn impacts patient outcomes. For example, Kuisma et al. [2] studied the impact of medical priority dispatch on pre-hospital mortality. The results showed that pre-hospital mortality of lower urgency calls did not depend on the order in which ambulances were dispatched. These results suggested that it may be possible to tailor responses to patients without hurting their chances of survival. Nicholl et al. [3] discussed the call priority classification of the advanced medical priority dispatch (AMPD) system. AMPD classified emergency calls into four types based on severity and type of conditions, among other things. The suggested classifications were DELTA, CHARLIE, BRAVO and ALPHA levels. Nicholl et al. [3] compared AMPD classification with that of a review panel that categorized severity of calls. They discussed the aggregation of the AMPD classification into three levels. DELTA and CHARLIE were identifiers for life-threatening calls in which paramedic units and rapid transports were needed. BRAVO identified potentially-life-threatening calls that needed non-paramedic units and rapid transport. And ALPHA identified non-life-threatening

calls that needed non-paramedic units. For our research purposes, we consider two types of ambulances based on skills of staff and equipment: advanced life support (ALS), or paramedic, units and basic life support (BLS), or non-paramedic, units. Clawson et al. [4] suggested an example of response configuration of EMS systems where both ALS and BLS units would be dispatched in a critical situation (hot situation) in which rapid transport was needed. They referred to DELTA and CHARLIE calls. They suggested that the closest BLS unit was dispatched to respond to BRAVO calls whereas for ALPHA calls (non-critical situation - cold situation) a different decision would be made.

Our goal is to determine an optimal policy for multiple unit dispatch and call priorities to increase the overall patient survival probability. We investigate how to improve the performance of EMS systems in two major areas; multi-unit dispatch and relocation strategies. The first two studies focus on multi-unit dispatch. In the first study, proposed dispatching models are examined under stochastic behavior of emergency calls with three levels of priorities which need different medical care. In addition, we present some extensions to the model by considering real on-scene conditions, such as the fact that dispatch decisions can be changed. The simulation models for multiple unit dispatch with multiple call priorities are used to investigate the performance of all possible policies for dispatching ambulances by using an enumeration method. We study the optimal dispatch policies through several small examples. A heuristic algorithm is developed to dispatch ambulances for larger-scale problems. These details are described in Chapter 2. In the second study, we extend the models in Chapter 2 by considering alternative policies for priority2 calls, which is located in Chapter 3. In this chapter,

simulation models with fairness constraints are formulated as integer programming models in order to obtain the optimal dispatching policy for priority2 calls. The objective is to maximize the overall survival probability.

In another study, we consider the compliance table strategies for relocation of ambulances. We determine the best compliance table to use for improving the performance of the EMS systems. The compliance table is a table that shows the assigned ambulance locations for a given number of busy ambulances. Most of the early EMS models allocated ambulances to only one fixed location. The main assumption of a static location model is a fixed home station base. The dispatched ambulance would return back to a home station base after providing service to patients. This differs from ambulance relocation models, which are developed to more closely mimic the operations of actual EMS systems. Ambulances are repositioned in real-time after their finished service on the scenes of accidents or at hospitals. The compliance table is a commonly implemented relocation strategy in practice. Alanis et al. [5] analyzed the performance of EMS systems that repositioned ambulances under a compliance table policy. Their results showed that there were impacts of performance by changing the compliance table by analyzing a Markov chain with relocation model. In this paper, we study how to design the best compliance table to obtain the optimal expected coverage of EMS systems. There are a few studies in this area of research. In Chapter 4, we present an integer programming approach to the compliance table problem. The formulation applies a Markov chain model with relocation. The application of the Markov chain model with relocation is based on an approximation. We investigate the performances of the EMS

systems by using real-world data. The objective is to maximize expected coverage. Repositioning ambulances is a powerful way to improve expected coverage. Analysis of real world problems suggests that the design of compliance tables should place an upper bound on the moving time of repositioning ambulances. Because of this, in Chapter 5 we study the repositioning of ambulances while restricting the areas in which they can move. We determine the districting areas for repositioning ambulances. We embedded the nested-compliance table model into a heuristic algorithm to examine districting and relocation strategies for EMS systems.The rest of this doctoral dissertation is organized as follows. In the following section, we review the related work on models for improving the performance of EMS systems. In Chapter 2, we describe the enhancement of a model of EMS systems by considering the real on-scene conditions. We present the recommendations for dispatching emergency vehicles under multi-tiered response via simulation. In Chapter 3, we describe an extension of the model of EMS systems that can aid in determining the reaction to real on-scene conditions of priority2 calls. A simulation model for fairly dispatching emergency vehicles under multi-tiered response is proposed in Chapter 3. In Chapter 4, we present a nested compliance table policy for EMS systems under relocation. We propose an integer programming model for determining the best compliance table policy. In Chapter 5, we present a nested compliance table model embedded into a heuristic algorithm for districting and relocation strategies in EMS systems. In Chapter 6, we present conclusions and future work.

1.1 Literature Review

Since the late 1960's the rapid US population growth has generated an increasing demand for ambulance services. In 1967, the study of EMS systems began to determine the distribution and workload of the existing systems. King and Sox [6] were the first to conduct a study to evaluate the workload of EMS systems in order to improve performance. In 1972, EMS systems were analyzed in a study of a location model in order to minimize average response time, as seen in Carter et al. [7]. This study considered two ambulance units that were dispatched to respond to calls, given the different locations of the units. The study then determined the district boundary for each unit to respond to calls. The EMS planners then studied the number and type of ambulances to deploy to certain locations, as seen in Eaton et al. [8]. This study researched how to design the EMS systems to reduce cost. Two strategies were considered. In the first strategy, the EMS system operated with two types of ambulances: basic life support (BLS) and advance life support (ALS). The BLS unit would respond to non-life threatening calls, while the ALS unit would respond to the life-threatening calls. The second strategy considered the number of station bases and how to allocate ambulances to their stations based on cost trade-off. Later, in the early 1990s, the EMS systems began to use computer-aided dispatch (CAD) systems. CAD systems have been used to collect information from emergency calls. They monitor locations and availability of ambulances. The study of early implementation of the CAD systems, shown in Hougham [9], discussed the implementation of CAD to real-world problem. In 1996, collected data showed that 30 – 50 % of calls were non-life threatening calls. To improve

the performance of the system, EMS planners considered modeling the dispatch priority decision. Palumbo et al. [10] designed a comparison study of priority and non-priority dispatch systems. The results of this study indicated that the dispatch priority provided better performance in terms of lives saved on serious calls and increased the utilization of ALS units. Most previous work of EMS systems was focused on reducing response time, which has a crucial effect on the efficiency of EMS systems.

As it is most related to the work presented in the remaining chapters, and since a more in-depth review will be conducted within each chapter, here we briefly outline the development of the most relevant literature. This is presented in terms of the literature on dispatching strategies and the literature on relocation models.

The previous literature studied EMS strategies based on dispatching the closest ambulances. When implementing dispatching strategies, the ordered preference lists for each demand zone are different in order to maximize outcomes of the EMS systems. The significant strategies to improve response include policies on how to send an appropriate ambulance according to severity of the call. There have only been a few studies conducted that considered the dispatch strategies of EMS systems. Considering the dispatch of EMS vehicles, one strategy is to make better use of available ambulances by having close-by ambulances respond to serious calls and sending farther ambulances to non-serious calls. The idea to study dispatch policies was proposed by Lim et al. [11]. They studied the impact of dispatch policies on the performance of EMS systems. The effect of dispatch strategies on the performance of EMS systems was based on the

urgency of calls. The Maine EMS Service and Emergency Medical Dispatch (EMD) center, 2011 guidelines provided the priority dispatch implementation. The guidelines discussed details for determining the severity level of calls. A successful priority dispatch depended on organization and communication between staff and dispatching planners. They discussed the results of the implementation of priority dispatch that provided better use of ambulances types to matching the requirement of patients. Most dispatching ALS units would respond to life-threatening calls whereas non-life-threatening calls could require basic ambulances.

Only a few early studies of EMS systems dealt with ambulance relocation models. In reality, sometimes the closest ambulances are unavailable to respond to a call. The unavailable ambulances are a critical factor affecting the performance of EMS systems when a serious call arrives. Dynamic models were developed to reallocate idle ambulances to compensate for stations in which most of the ambulances are busy and unable to respond to arrivals of serious calls. The first relocation model was introduced by Kolesar and Walker [12]. This model examined the relocation of fire resources. The recent work in relocation models deals with three classes of models: integer programming, simulation, and Markov decision process. Unfortunately, there were known weaknesses of implementing the relocation models. The optimal solution of the relocation model frequently changed the destinations of ambulances, which in return required more powerful technology to keep track of the ambulances' current locations. Recently in the practice of EMS systems, computer-aided dispatch (CAD) systems are applied, making it possible to indicate the current locations of all ambulances. Therefore,

the increasing probability of available ambulances ready at their station bases provides a better performance measure of the EMS systems.

While some recommendations exist, the national academy of emergency medical dispatch (EMD) guidelines provides no details for operating prioritization recommendations. Thus, the priority dispatch problem motivates this research. In Chapter 2, to maintain fidelity to the real problem we propose to study simulation models with multiple unit dispatch and call priorities. We consider the optimal multiple unit dispatch strategies through several small examples in order to guide us in developing heuristic policies for use in real-world problems. The dispatch strategies of EMS systems are examined under stochastic behavior of the emergency calls with three levels of priority each needing different medical care. We consider two types of medical units, ALS and BLS units. Priority1 calls require multiple units. On the other hand, a single dispatch is used to respond to priority2 and 3 calls. The main focus of this paper is to develop a model showing how to dispatch two types of ambulances depending on the priorities. In addition, we present some extensions of the model by considering real on-scene conditions. For example, the fact dispatch decisions can be changed on the spot. A heuristic algorithm is developed to dispatch ambulances on larger-scale problems. This heuristic procedure is based on the hypercube model. The proposed heuristic uses the principle of balance of call volumes on servers. A comparison between the heuristic and the closest policy are demonstrated using real world data.

In Chapter 3, we also use a simulation model to determine the performance measure of EMS systems. We extend the model in Chapter 2 by considering real on-scene conditions of priority2 calls. The dispatching decision can be changed to BLS upgrade when the BLS unit arrives on-scene to service priority2 calls. We evaluate two alternative policies of dispatching BLS unit; the dispatching BLS unit of priority2 calls should be treated like the dispatching BLS policy of priority1 or 3 calls. This is based on always sending the closest ALS unit for priority2 calls. The goals considered are the same outcomes from Chapter 2, where the goal was to maximize the patient survival probability. In addition, the equity between patient priorities and time is crucial for dispatching decisions. The dispatching ambulances to respond to priority 1 calls forces busy ambulances to respond to priority2 calls. The priority2 calls on-scene situation may be changed to life-threatening and require the care of an ALS unit. To improve inequities of EMS systems, finding the balance between dispatching ambulances for each priority is a challenge in this research. As an extension of this Chapter, we consider the notions of fairness between priority1 and 2 calls by limiting the wait time of first response for priority1 and 2 calls. The average waiting time of first response should be equalized between priority1 and 2 calls. As the main focus of the fairness problem, we determine how to dispatch ALS units to respond priority2 calls while the waiting time between priority1 and 2 calls is restricted.

In Chapter 4, we focus on the compliance table strategies under a relocation policy. The compliance table is useful for relocation problems. Relocation problems refer to the dispatched ambulances that could travel back to different home station bases. There

are significant influences on relocation strategies when changing a compliance table. The decision of how we design the best compliance table, given number of available ambulances, is complicated. Therefore, in Chapter 4 we focus on developing compliance table strategies. Our goal is to maximize expected coverage. We formulate this problem as integer programming.

In Chapter 5, we present the districting and the compliance table strategies under a relocation policy. The studied compliance table model in Chapter 4 showed adverse effects of the repositioning time of moving ambulance between stations to the performance of EMS systems. We consider an upper limit on repositioning time of moving ambulance as a districting problem. The decisions of how we determine the moving areas for repositioning ambulances and allocate the ambulances to each moving area are proposed. The compliance table model is embedded into a heuristic algorithm to obtain the better configuration of compliance table for each district. The EMS systems operate under districting and relocation strategies.

CHAPTER TWO

RECOMMENDATIONS FOR DISPATCHING EMERGENCY VEHICLES UNDER MULTI-TIERED RESPONSE VIA SIMULATION

2.1 Introduction

Emergency medical service (EMS) systems are operated with the underlying goal of maximizing survival probability of patients. However, most EMS systems use measures of efficiency to evaluate their performance, such as average response time and expected coverage, which could in turn affect patient survivability. Coverage refers to the proportion of patients who can be attended by ambulances within a predetermined time or distance. Response time refers to the time from when an ambulance dispatches to when an ambulance arrives on scene (see Figure 2.1 below). Most operations decisions involved in EMS systems affect response time, such as ambulance location, ambulance relocation and ambulance dispatching. As the main focus of this paper, we consider the ambulance dispatching decision; that is, our goal is to determine the appropriate ambulance/s to assign to respond to a call. EMS systems operate in dynamic conditions. When a call arrives to the system, the dispatcher asks a series of questions to determine the severity level of the call. The dispatcher must then assign an ambulance (or set of ambulances) based on the severity and location of the call and on the availability and location of resources in the system.

Our goal is to design dispatching strategies for multi-tiered responses that maximize patient survival probability. In this paper, we consider two types of ambulances

based on skills of staff and equipment: advanced life support (ALS), or paramedic, units and basic life support (BLS), or non-paramedic, units (usually these are staffed by EMTs). Several recommendations as well as current best-practices exist regarding which types of units should be sent to different types of calls. The dispatcher determines the severity level of the call using the answers to key questions from the initial phone call and additional information (e.g. weather). Clawson et al. [4], suggest that both a paramedic unit and basic life support unit be dispatched to the most serious calls (classified as DELTA and CHARLIE). They suggest that the closest basic life support unit be assigned to respond to BRAVO, or not believed to be life-threatening calls. The ALPHA calls are considered non-critical. While standard practice is to send the closest BLS unit to these calls, dispatching a nearby ambulance to the lower priority patient may make ambulances unavailable for future nearby life-threatening calls. Thus, EMS systems managers need to make a decision about how to dispatch a basic life support unit for the ALPHA calls. Our work is motivated by determining the optimal policy for multiple-unit dispatch and call priorities in order to increase the overall survival probability of patients. As will be explained in Section 2.3, we focus on the decisions regarding how to dispatch a BLS unit for priority1 and priority3 calls, while assuming a fixed dispatching policy for priority2 calls. The dispatching strategies of EMS systems are examined under stochastic behavior of the emergency calls with three levels of priorities, each of which need different medical care. In addition, we allow for realistic on-scene conditions that are often ignored in the literature, such as the fact that dispatching decisions can be changed

as more information is revealed on-scene (ALS downgrade and/or the BLS upgrade). The details are described in Section 2.3.

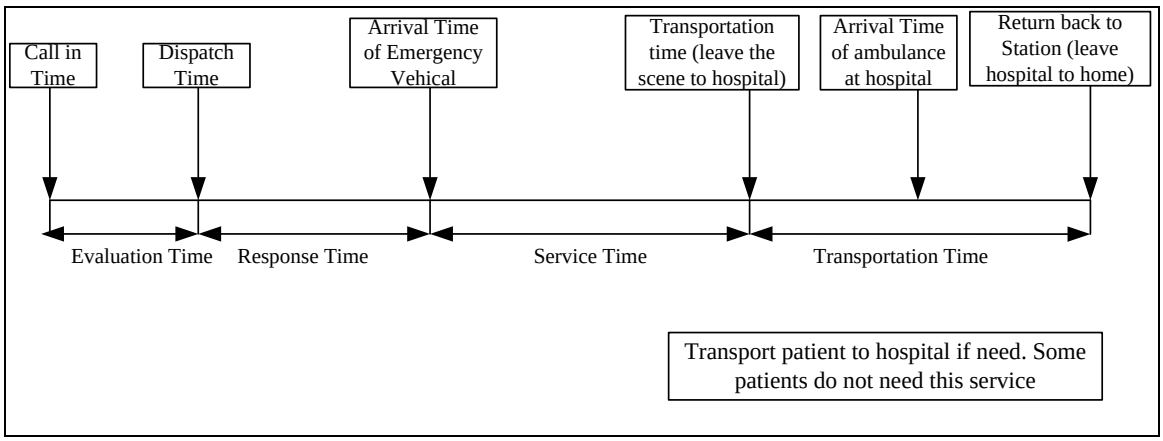


Figure 2.1: The time line for calls to an EMS system

In this study we:

- Develop and analyze simulation models that consider how to dispatch two types of ambulances to three call priorities.
- Use the simulation model to study the performance of all possible policies for small problems and study the structure of the best policy.
- Propose a heuristic algorithm for designing dispatching strategies for large-scale problems. The hypercube model (HQM) is used to develop the iteration method for the heuristic algorithm.
- Show, through the numerical results, how the heuristic compares with the closest dispatch policy in real world problems, and specify conditions under which it yields the highest improvements.

The rest of this paper is organized as follows. In Section 2.2 the relevant literature is discussed; in Section 2.3 we present the model for EMS systems with multi-tiered response; in Section 2.4 we show illustrative examples; in Section 2.5 we present the heuristic approach; in Section 2.6 we show the case study and computational results; and in Section 2.7 we present conclusions and outline future work.

2.2 Literature Review

The literature related to EMS systems is extensive. Some models are descriptive, where the goal is to more accurately measure the performance of a system given known operating conditions, while in other models, decisions are made to allocate resources. These decisions are often ambulance location or ambulance dispatching. First we present some descriptive models, and then models are categorized by the decision (e.g. location, dispatching) being made, though some look at multiple decisions.

The hypercube model is a model that describes the queuing dynamics of EMS systems with multiple servers and allows for the evaluation of busy probabilities among non-identical servers. Larson [13] introduced the hypercube model with a single dispatching problem to determine the busy probabilities of units of EMS systems. Larson [14] later developed an approximation of the hypercube model by using an iterative procedure. His approximation procedure relied on the *M/M/N* queue and the *M/M/N* with 0 queue to obtain approximate workload of servers as well as other performance measures. Larson [14] assumed one type of customer that was generated according to a

Poisson process. Jarvis [15] considered the Erlang loss system and generated an approximate workload of servers by using the hypercube model. He assumed multiple types of customers that arrive to the system according to a Poisson process. The approximation algorithm used an iterative procedure to obtain servers' workloads. More recently, Budge et al. [16] proposed approximating the dispatching probabilities by using an *M/MN/N* system. Their study is different from previous studies since multiple servers can be assigned to each station. Chiyoshi et al. [17] provided a survey of the hypercube model applied to EMS systems.

Early prescriptive works in EMS systems focused on ambulance location problems, many of which use the hypercube model approximations mentioned above to account for the busy probability of servers. In these problems there are a limited number of ambulances to be located at stations to maximize the coverage, or proportion, of calls responded to within a given time/distance standard. Church and ReVelle [18] introduced the maximal covering location model. Daskin [19] developed the traditional maximum covering location model, incorporating the congestion phenomena by considering the busy probability of the servers. He formulated the maximum expected covering location problems (MEXCLP). Gendreau et al. [20] suggested a dynamic ambulance relocation model, referred to as the maximal expected covering relocation problem (MECRP) that allowed the number of ambulances in stations to be changed. ReVelle and Hogan [21] formulated an integer programming model for the maximal availability location problem, which sought to maximize the probability of an ambulance being available within a coverage standard. Because of the congestion of the system, calls can arrive while all

servers are busy. Marianov and ReVelle [22] developed the maximal availability location problem, in which the busy probability of servers is taken into account and the goal is to maximize the expected availability of servers. There are several extensions to location problems. Erkut et al. [23] formulated ambulance location problems incorporating a survival function, referred to as the maximal survival location problem (MSLP). They used the MECRP formulation of Daskin [19], extending the model to the maximal expected survival location problem (MEXSLP). Both the MSLP and the MEXSLP were developed by incorporating probabilistic response times.

Other works combine the location decision with a relocation or districting decision. Mendonça and Morabito [24] analyzed ambulance deployment on highways by using the hypercube model. They considered a list of two preferred ambulances for each demand zone. If both of ambulances were busy, then the call was said to be lost. The objective was to balance workload among bases and to minimize mean travel time. Atkinson et al. [25] analyzed EMS deployments on highways introducing two heuristic methods. Both heuristics had embedded the *M/M/N* queuing model with loss. Iannoni et al. [26] presented two decision problems related to EMS operation on highways. They developed an exact partial backup model, extending the work by Mendonça and Morabito [24]. The model was embedded into greedy algorithms to solve two small problems combining location and districting decisions.

Other works consider the decision of which vehicle to dispatch to a call, given fixed vehicle locations. For the dispatching problem, recent papers focus on priority

dispatch in which ambulances are assigned based on the severity classification of each emergency call. Considering priority dispatching problems, McLay and Mayorga [27] examined how to optimally dispatch ambulances by using a Markov decision process (MDP) approach. They investigated the impact of response time thresholds (RTTs) on outcomes, as well as how the best dispatching policy changed according to a given RTT. Bandara et al. [28] formulated an MDP model for dispatching problems with call priorities to determine the optimal dispatching strategies. Their approach maximizes the overall expected survival probability. McLay and Mayorga [29] extended the basic MDP model. Their formulation focused on balancing the equity and efficiency of servers and other fairness constraints related to customers. They solved the MDP formulation by using equivalent linear programming models. McLay and Mayorga [30] also considered an MDP approach for a system in which call priority assignments are subject to classification errors. The objective was to maximize the overall coverage rate. They noted that dispatching the closest vehicle is not always optimal. In all of these previous works, it was assumed that there was one type of server and that only one vehicle was dispatched per call.

Extensions of the priority dispatching model considering multiple-unit dispatch with call priorities have also been developed. Chelst and Barlach [31] introduced the hypercube model for multiple unit dispatch. They presented an exact and an approximate version of the hypercube model. The exact model is based on an *M/M/N* system given single and bulk arrivals. In addition, the approximate model was formulated by using the *M/M/N/0* queuing model. They assumed independent servers. Gau and Larson [32]

developed the model for multiple unit dispatch for *N*-patrol-unit systems. They considered calls with two priorities. Type I calls needed only a single unit, and type II calls required two units. Each demand zone had a preference list for ranking the ambulances. They formulated the hypercube model for the exact solution. In addition, they developed approximate models that assumed a fixed-preference dispatching list. Iannoni and Morabito [33] analyzed the operation of EMS systems on highways with a zero queue. They presented two types of vehicles: medical units and rescue ambulances. Furthermore, they considered a single dispatch, either an ambulance or a medical unit, double dispatch and triple dispatch. The dispatching policies depended on requirements of calls and call locations. In addition, they considered in-site service when patients could be served at the ambulance stations. Iannoni et al. [34] considered multiple unit dispatch by using the hypercube model. Assumptions were the same as those made by Iannoni and Morabito [33]. They verified their results by using simulation models. A genetic algorithm (GA) was presented to search for near optimal solutions. The GA approach produced configurations that provided input data to the hypercube model. These previous works assumed that the real on-scene conditions could not be changed; in this paper, we consider that information available on-scene may be used to update or change the call priority.

In several studies of priority dispatching models in realistic EMS systems, authors analyzed models by using heuristic and simulation approaches. Andersson and Värbrand [35] proposed a new way to dispatch ambulances in which the outcome was preparedness. Call priorities were considered in their models. They formulated an integer

programming model to solve the ambulance relocation problem. A tree-search heuristic was used to solve these problems in reasonable computational running times. Simulation models were also used to evaluate dispatching strategies. Lee [36] considered a dispatching problem without including the priority of calls. The objective was to investigate the preparedness of the system given that there was a non-zero queue. Bandara et al. [37] studied the dispatch of a single type of ambulance considering call priorities and using simulation models. These models were developed to allow for a non-zero queue. In addition, they used a heuristic approach to determine dispatching strategies for large problems. The results showed that considering call priorities provided higher efficiency for EMS systems than the use of the closest rule, regardless of the assumption of a zero-queue. Most of the previous work in the simulation approaches to the dispatching problem considered a single dispatch. In contrast to previous work, this paper considers the multiple unit dispatch.

In this work we extend the priority dispatching strategies proposed by Bandara et al. [37]. The modification considers multiple-unit dispatch with multiple call priorities. Recent studies considering multiple-unit dispatch and partial backup were presented by Iannoni and Morabito [33] and Iannoni et al. [34]. Their performance measures included busy probability, loss probability and fraction of dispatch (of a specific server), among others. However, our work differs in that our objective is to maximize expected survival probability. We include the multiple-unit dispatch in the model where the service time of units is not independent (one unit may need to wait for back-up). Our study requires the first ambulance to have to wait for arrival of a second or more appropriate unit before

returning to service. Furthermore, we allow for upgrades and downgrades based on on-scene conditions. The second unit can be canceled based on information from the first unit.

2.3 EMS Systems with Multi-tiered Response

Given the characteristics of real EMS systems, we consider EMS systems with multiple-unit dispatch and multiple call priorities. We assume that there are three call priorities, indexed by (m), and two types of ambulances, indexed by (j), ALS units (j=1,..,J) and BLS units (j=J+1,…,J+K). The ALS units are paramedic units that can provide patient transport to hospitals. On the other hand, the BLS units cannot provide patient transport to hospitals. Furthermore, we allow system status updates at the scene of the accident; in other words, there is a chance that the severity of the situation can be changed. We consider ambulance dispatching policies that depend on call priorities and the availability of ambulances, while taking into account that on-scene updates are possible. We model this as a zero-queue system. That is, if a call arrives to the system when all ambulances are busy, the dispatcher transfers it to other systems (such as sending a fire truck or asking for assistance from a neighboring county). The sequence of events is described as follows:

- Call arrivals: When a call from zone i arrives, we know the location and the priority of the call. The call priority (priority1, priority2 or priority3) is assigned during the initial phone call to emergency service by the dispatch operator.

- Vehicle dispatch decisions: We consider ambulance dispatching policies that depend on call priorities and available ambulances. We assume that each ambulance is located at a fixed station. At least one ambulance is assigned to each call, if at least one ambulance is available. When a call with the highest priority (priority1) arrives, two types of medical units are dispatched if both of them are available. On the other hand, when a call with lower priority (priority2 or priority3) arrives, only a single unit is dispatched.
- Response time: The time observed between the moments that ambulance/s dispatches until the first ambulance arrives on the scene of the accident is referred to as the *response time (γ).*
- On-Scene: in the case of double-dispatch, when the first ambulance arrives on the scene, there is a chance that the severity of the situation can be changed. The model allows for BLS upgrade or ALS downgrade. There are four possible situations during a double-dispatch mode (in which both ALS and BLS units are sent). The EMS system process for priority1 calls is described in Figure 2.2 :

  1) Both ALS and BLS are available and ALS arrives first: The ALS provides care, if the patient truly needs ALS, the BLS unit is called off and the ALS transports to the hospital if needed. If the ALS determines that BLS is sufficient then the ALS waits for the BLS, the ALS unit helps the BLS unit once it arrives on the scene to finish working on the patient together. Both units return to their original station when service is complete.

2) Both ALS and BLS are available and BLS arrives first: BLS provides initial care and waits for the ALS unit. When the ALS unit arrives it determines if the patient needs ALS & transport. If BLS is sufficient, the ALS units stays with the BLS unit to provide training but both are able to return to their home station once service is complete. If ALS transport is needed, the BLS unit takes over and the BLS unit returns to its home station while the ALS unit serves and then transports the patient.
3) Only an ALS unit is available: The ALS unit is sent, and it serves and transports the patient if necessary. No BLS unit is used in this situation.
4) Only a BLS unit is available: The BLS will provide service and wait for the ALS unit to determine if patients need transportation to hospitals. Then we proceed as in case (2) above.

The time that the first unit spends waiting for backup unit is referred to as the *waiting time* (ω). The total time spent providing service to patients is referred to as the *service time* (μ).

- Transportation: After providing service to patients, units will provide patient transport to hospitals if needed. Upon completing service, the ambulance will return to its original ("home") station. The time that the ambulance takes to return back to its original station, including the transportation time of the patients to hospitals if needed, is referred to as the *transportation time* (τ).

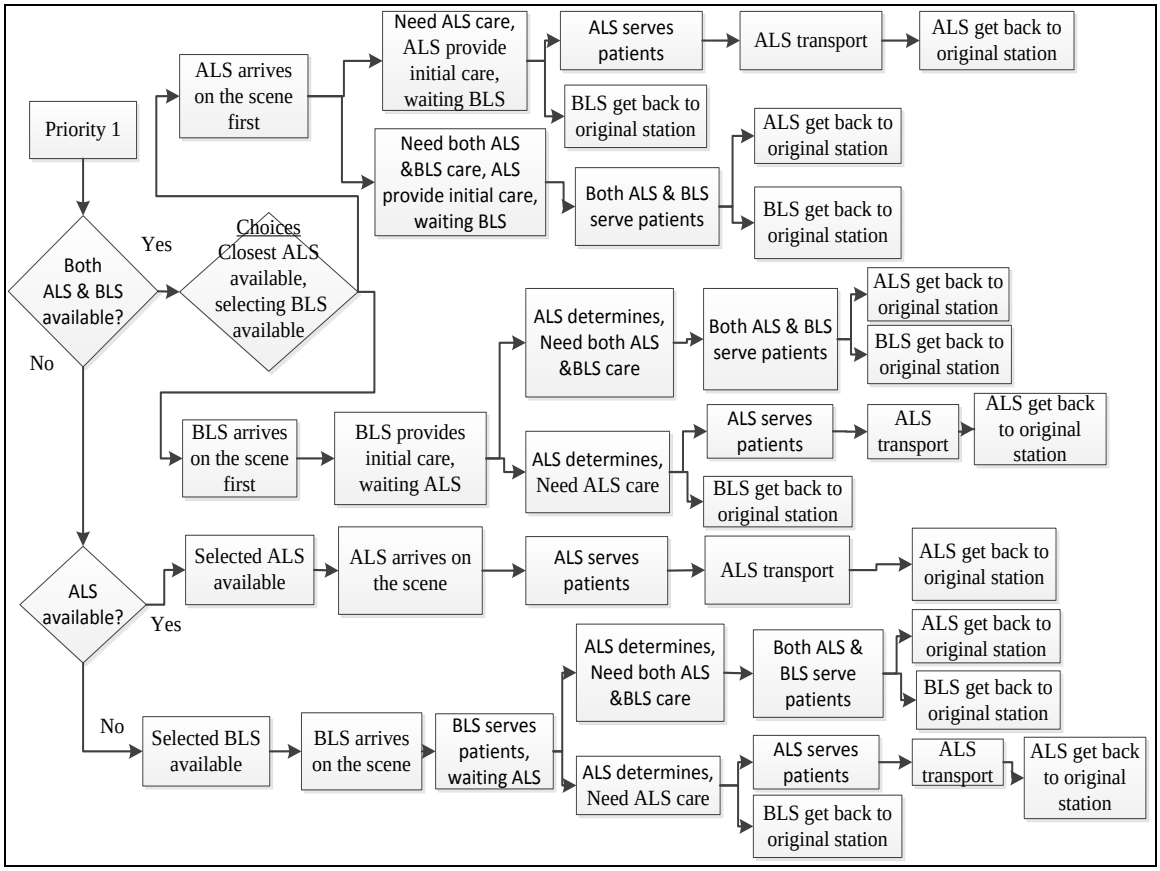


Figure 2.2: The EMS system process for priority1 calls

2.3.1 Model Description

Emergency calls in each zone *i* arrive to the system according to a Poisson process with rate λ_i, where λ represents the total call arrival rate and λ_i represent the call arrival rate from zone *i*. Calls are classified by priorities and mapped to the type of ambulance/s (ALS unit and BLS unit) required. The dispatching policies differ by call priority since the priority1 calls need two distinct ambulances (the ALS unit and the BLS unit), and priority2 and 3 calls need one ambulance (the BLS unit). As we describe below, for some calls the closest unit is sent, for other calls ambulances will be dispatched according to an ordered preference list. An ordered preference list details the

preferred order in which to send units based on availability. In other words, the unit that is first on the list is sent if it is available, otherwise the second on the list is sent, and so forth. If all backup units are busy, the call is transferred to another system. The objective is to maximize the overall expected survival rate of patients, which is related to the response time of the first ambulance to attend priority1 calls. The details of the multiple-dispatch decision tree are described in Figure A.1, Appendix A.1 and a summary of the EMS process and dispatching policies for priority1 calls is described in Figure 2.2. The model notation and description are shown in Table 2.1. The multiple unit dispatching policies are given as follows (note that some policies are fixed, while others will be optimized; the policies we can control are shown in italics):

(i) Priority1 calls require double dispatch.

   a. If both ALS and BLS are available, we dispatch both of them. However, if only one type of ambulance is available, we dispatch the available ambulance. We dispatch the closest ALS unit, while **we select BLS units according to a rank ordered preference list** to maximize the expected survival rate.

   b. If only one type of ambulance is available, we send the one which is available. The situation at the dispatching center will follow the "on-scene" scenarios listed above.

(ii) Priority2 and 3 calls require a single dispatch (BLS unit). We assume that patients do not need transportation to hospitals in these types of calls. For priority2 calls, we dispatch the closest available BLS unit. To maximize the expected survival rate, **we make the decision to dispatch a proper (not necessarily the closest) BLS unit**

when priority3 calls arrive to systems. If all BLS are busy, the calls are transferred to other systems.

Times between events in the EMS system are explained in Figure 2.1 The time between the dispatch of ambulance/s and the arrival of the first ambulance on-scene is referred to as response time, and it follows a Lognormal distribution with mean response time γ_{ij}, which depends on the call zone *i* and the responding unit *j*. In addition, we allow a service time distribution of the EMS system that can depend on the call zones, the available ambulance types and the priority of calls. The time during which ambulances provide medical care to patients is referred to as service time, and it follows an Exponential distribution with mean service time μ_{mij} which depends on the call priority *m*, the call zone *i*, and the unit providing service *j*. The mean transportation time required for an ambulance to provide patient transport to a hospital if needed is assumed to follow a Lognormal distribution with mean transportation time τ_{ij} which depends on the call zone *i* and the responding unit *j*.

Table 2.1: The parameters of multiple types of ambulances with multiple call priorities

Notation	Description
λ	call arrival rate
n	total number of demand zones
i	indicator of demand zone
m	indicator of call priority as $m = 1,2,3$.
J	number of ALS medical units
K	number of BLS medical units
j	indicator of medical unit with known location, ALS units are numbered $j = 1,\ldots,J$; *BLS* as $j = J+1,\ldots,J+K$.
μ_{mij}	mean service time of priority m calls for ambulance j for demand zone i
γ_{ij}	mean response time for ambulance j for demand zone i
τ_{ij}	mean transportation time for ambulance j for demand zone i
λ_i	call arrival rate from demand zone $\sum_{i=1}^{n} \lambda_i = \lambda$
p_i^m	proportion of priority m calls from demand zone i: such that $\sum_{l=1}^{3} p_i^m = 1$
q	probability that priority1 calls require the ALS medical unit on the scene.
Additional parameters for the heuristic approach	
ρ	traffic intensity or fraction of time server is busy
r	the expected offered load
τ	the expected total time for independent server
p_u	probability u server are busy
a_{iml}	the l^{th} preferred server for priority m of call zone i
$k = a_{iml}$	server k is assigned to priority m of call zone i for which it is l^{th} preferred is given
$\lambda_{im,aiml}$	arrival rate of priority m of call zone i is served by server a_{iml} for which it is l^{th} preferred
$t_{im,aiml}$	the expected total time for server a_{iml} that dispatch to serve priority m of call zone i for which it is l^{th} preferred
$f_{im,aiml}$	total rate from priority m of call zone i that server a_{iml} is assigned to a call of priority m of call zone i. The server a_{iml} is the l^{th} preferred server.
g_v	the probability of the first v servers are busy
$\upsilon_{BLS:k}$	total call volume that are served by server k (BLS)
B	mean absolute deviation of call volume are severed by BLS
$r_u \in R$	the rank of BLS matrix (1 x K) that call volumes are sorted from *Max* to *Min*
$k = r_u$	preferred as BLS: k is sorted as u^{th} in the matrix of rank of call volumes $\upsilon_{BLS:\,r1} \geq \upsilon_{BLS:\,r2} \geq \upsilon_{BLS:\,r3} \geq \ldots \geq \upsilon_{BLS:\,rK}$
r_K	preferred as BLS: K is sorted as a last server in the matrix of rank of call volumes
z	position of the lowest call volume of server (r_κ)
w	position of the chosen server to swap (k)
vdev	deviation of call volume between the lowest call volume of among servers and the chosen server to swap
vol	total increased call volume
$Q(K, \rho, v)$	the correction factor that indicate the probability of obtain v busy servers given by *M/M/K* systems
$Q^*(K, \rho, v)$	the correction factor that indicate the probability of obtain v busy servers given by *M/M/K/K* systems

In this paper, the objective is to maximize the patient survival rate, where survival rate is defined based on the work of Larsen et al. [38]. They formulated multiple linear regressions from real data. The patient survival is a function of the response time t_R. The patient survivability was represented by a probability which varied significantly between zero and one with respect to response times between zero and 9 minutes. Let $s(t_R)$ denote the probability of patient survival under response time t_R. McLay and Mayorga [39] used the response time interval as a proxy for patient survival probability. They formulated the maximum expected coverage location model with multiple classes of patients and two types of servers. Knight et al. [40] considered the maximal expected coverage location problem with multiple call priorities. They proposed an iterative approach by using queuing theory to determine the utilization of ambulances. Bandara et al. [37] used the patient survival function based on the work of Larsen et al. [38] and McLay and Mayorga [39] as the objective function. Based on previously mentioned works, a fast response for priority1 calls affects the overall probability of survival of life-threatening patients. We consider the survival probability of patients as a function of the response time for priority1 calls using the equation shown below, a given in McLay and Mayorga (2010).

$$s(t_R) = \max[(0.594 - 0.055 * t_R), 0] \tag{2.1}$$

2.3.2 Simulation Model

Following the description provided in Section 2.3.1, we developed a discrete event simulation model of an EMS system. The simulation model was implemented using Arena Version14. The simulation model was then used to investigate the performance of

a given dispatching policy. The status of the EMS system is described by the state-space of multi-server queuing systems in order to represent each ambulance individually. The model consists of seven subsystems: call generating, dispatching, response, calculating the patient survival, waiting for another unit, waiting for next available ALS unit, service and transportation. The simulation flow chart is described in Figure A.2, Appendix A.1 for priority1 calls and Figure 2.3 for priority2 and 3 calls.

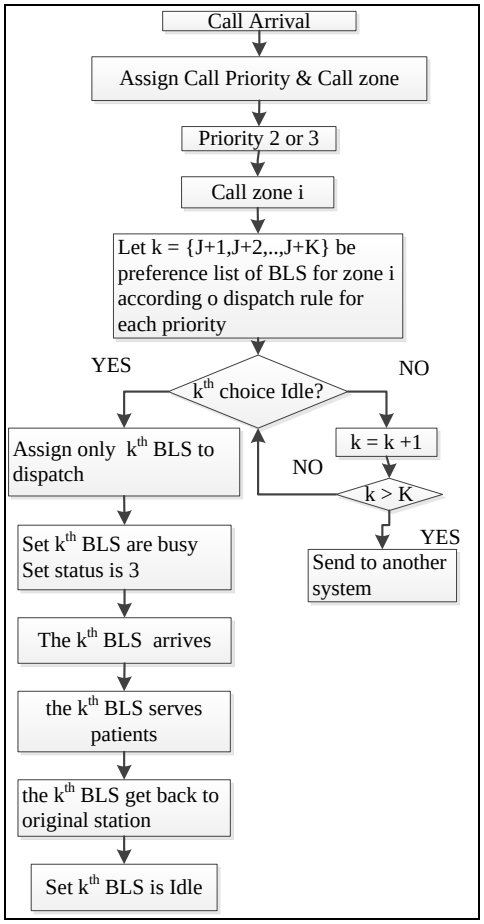


Figure 2.3: Simulation flow chart of EMS systems for priority2 and 3 calls

(i) Call generating system: The inter-arrival time of calls is generated randomly based on a known distribution and total arrival rate. The calls are be assigned a specific priority, location and initial vehicle assignment based on attributes of the call.

(ii) Dispatching system: The call is classified by its call zone and priority and assigned ambulance (which could be the closest available or based on an ordered preference list). This system considers the states of ambulances to make this decision. The potential states are: "idle" at station base, "busy" serving a priority1 call with double dispatch (status 1), "busy" serving a priority1 call with only available dispatch of ALS unit (status 2), "busy" serving a priority1 call with only available dispatch of BLS unit (status 4), "busy" serving priority2 call or priority3 call (status 3).

(iii) Response system: The ambulance is dispatched from its home station to the scene, given the response time distribution is based on call zone and home station of the ambulance.

(iv) Calculating the patient survival rate: When the first ambulance arrives on the scene, we calculate the survival rate by using the response time of this ambulance for priority1 calls. The survival probability is calculated by using equation (2.1).

(v) Waiting for another unit: For double dispatch, when the first ambulance arrives on the scene, the state of this ambulance is "waiting" for another unit. When the second unit arrives on the scene, the status of both of ambulances is changed to "busy" (offering service to patients).

(vi) Waiting for next available ALS unit: If only BLS units are available to respond to a priority1 call, we dispatch a BLS unit. When the BLS unit arrives on the scene, the state of this ambulance is "waiting" for an available ALS unit. When an ALS unit becomes available, it is dispatched to the scene. Then the states of waiting BLS and ALS units are changed to "busy" again.

(vii) Service and transportation systems: After the required ambulances arrive on the scene, the service and transportation times are generated randomly based on a known distribution. The state of ambulance is "busy" when they are in service and providing patient transport to a hospital. After an ambulance provides service to patients, patients are transported to a hospital in necessary, and then the ambulance return to its original station. The state of ambulance is "idle" again.

2.3.2.1 The state space of EMS systems

There are $J+K$ ambulances in the EMS system. The vector $A = (a_1,.....,a_J,a_{J+1},....,a_{J+K})$ is the state of the system, where a_j: $j \in [1,\ldots, J]$ contains information about the status ALS units and a_j: $j \in [J+1,\ldots, J+K]$ contains information about the status of BLS units. We use a two – dimensional state to represent the status of each ambulance. The state of ambulance j is given by $a_j = (\sigma_j, \beta_j)$ where σ_j is the status of the ambulance j and β_j is the associated ambulance unit which is dispatched with the ambulance j (in case of double dispatch) as described in Table 2.2. For example a_1 = (1, 3) represents the double dispatch of ALS unit1 and BLS unit3 to respond to a priority1 call, a_1 = (2, 0) represents that an available ALS unit is dispatched alone to respond to a

priority1 call (which would happen if all BLS units are busy), a_1 = (5, 4) represents that ALS unit1 is dispatched to respond to a priority1 call after the system had only available BLS units, we dispatched BLS unit4, and the BLS unit4 is waiting for ALS unit1 at the scene of the accident, and a_4 = (3, 0) represents the dispatch of BLS unit4 to respond to a priority2 or priority3 calls. The state space of EMS systems is described be Table 2.2.

Table 2.2: The possible status of ambulances in EMS systems

Indicator	σ_j	Status of ambulance
$j \in [1,\ldots, J]$: ALS	0	Idle at base
	1	Double dispatch of ALS for priority1 calls
	2	Only ALS unit dispatch to respond to priority1 calls
	5	ALS unit dispatched to priority1 call following a BLS unit which was sent when no ALS units were available
$j \in [J+1,\ldots, J+K]$:BLS	0	Idle at base
	1	Double dispatch of BLS for priority1 calls
	3	BLS unit dispatch to respond to priority2 or 3 calls
	4	Only BLS unit dispatch to respond to priority1 calls

2.4 Illustrative Example of an Optimal Policy

In this section, we develop a simulation model and analyze its solution. We illustrate our model with an example of an EMS system with multiple-tiered responses of size 2x2x2. That is, we assume the EMS system has two demand zones, two ALS units and two BLS units. In addition, we assume the EMS system has two ambulance stations with one ALS unit and one BLS unit located at each station. We investigate the optimal dispatching policy for priority1 calls (how to dispatch BLS units given that we fix the closest dispatch for ALS units); and the optimal dispatching policy for prioirty3 calls (how to dispatch BLS units); whereas we fix the closest dispatch of BLS units for

prority2 calls. We enumerate all possible policies. There are $(K!)^n(K!)^n$ possible dispatching orders for EMS systems dealing with three priority types, n call zones and K BLS units. In this example, there are 16 possible dispatching policies. Suppose call arrivals follow a Poisson process with rate $\lambda = 1$ call per hour. We assign the proportion of calls for each priority type for each zone i to be $p_i^1 = 0.5$ for priority1 calls, $p_i^2 = 0.25$ for priority2 calls and $p_i^3 = 0.25$ priority3 calls. Let q, the probability that priority1 calls require the ALS medical care on the scene, be 0.5 of call arrivals. The EMS system operates 24 hours per day. Arena Version14 is utilized to develop a model that ran 336 simulated hours for each replication to obtain a steady-state result. For each policy, we ran 1500 replications to obtain the expected survival rate with half-width less than 0.0001. The input parameters are shown in Table 2.3.

Table 2.3: Input parameters (Ambulances 1 and 2 are ALS units, and ambulances 3 and 4 are BLS units)

Zone i	Response Times		Service Times		Transportation Times	
	Ambulance 1 and 3	Ambulance 2 and 4	Ambulance 1 and 3	Ambulance 2 and 4	Ambulance 1 and 3	Ambulance 2 and 4
Zone 1	logn(9.07, 4.19)	logn(14.03, 6.48)	Expo(41)	Expo(40)	Expo(9)	Expo(14)
Zone 2	logn(14.03, 6.48)	logn(9.02, 6.48)	Expo(46)	Expo(41)	Expo(14)	Expo(9)

The results in Table 2.4 show the optimal dispatching policies for BLS units to respond to priority1 and priority3 calls compared with the closest dispatching policies given that we fix the closest dispatching policies for BLS units to priority2 calls. In addition, we also fix the closest dispatching policies for ALS units to priority1 calls. In Table 2.4, a dispatching policy for BLS units for priority1 calls is shown on the left. For example, the numbers in the first row and column (3, 4) indicate that for the closest policy, the first choice assigned to call zone1 (when the percent of calls from zone1 is

10%) is the BLS3 (unit3), and the second choice assigned to call zone1 is the BLS4 (unit4). In other words, send unit3, and if it's not available send unit4. We investigate the optimal dispatching policy by simulating all policies. When the optimal policy is the same as the closest policy for priority1 calls, they are underlined, and when the optimal and closest policies are the same for priority1 and 3 calls, they are underlined and bolded. The 95% confidence interval provides an estimate of the accuracy of the expected survival rate point estimates resulting from the optimal and closest policies. We use the statistical 95% confidence interval in testing the difference between optimal and closest policies; that is, if the intervals do not overlap we say the differences are significant. The results show that the closest policy is optimal in all instances for prority1 calls. In contrast, the closest policy is not always optimal for prority3 calls; it depends upon the proportion of calls to each zone. For priority3 calls, the optimal dispatching policies are the same for instances when the proportion of demand from zone 1 is between 10% and 40%, in which case the optimal dispatching policy is always send BLS3 (unit3) first for calls from zone1 and zone2. When the proportion of demand from zone1 increases from 40% to 50%, the optimal policy changes to the same as the closest dispatching policy. When the call volume from zone1 increases from 60% to 70%, the optimal dispatching policy changes to always send BLS4 (unit4) first for both demand zones. The boxed rows show that the optimal policy is the same as the closest dispatching policy for both priority1 and 3 calls when the call volume is balanced between demand zones (50%-60% of calls from zone1). Comparing these policies, the evidence suggests that the closest dispatching policy for BLS units for priority1 calls is optimal, whereas the optimal

dispatching policy for priority3 calls depends on the proportion of calls between demand zones.

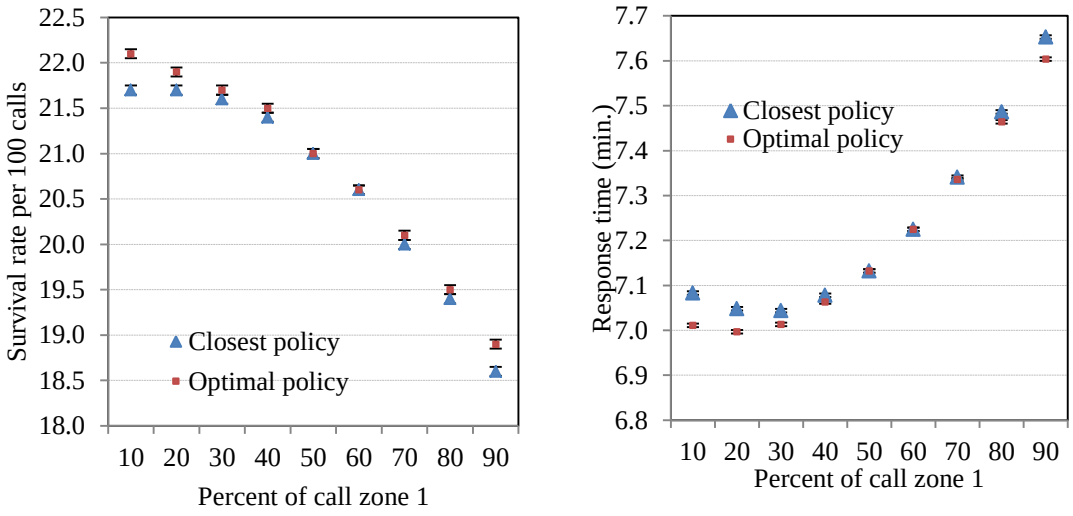
Table 2.4: Comparison of dispatching policies for BLS units between the closest policy and the optimal policy for priority1 and priority3 calls

Dispatch policy		Base policy of BLS medical units				BLS medical units for priority1 calls				BLS medical units for priority3 calls			
		Closest policy				Optimal policy				Optimal policy			
		Call zone 1		Call zone 2		Call zone 1		Call zone 2		Call zone 1		Call zone 2	
Choice of dispatch		1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd
Percent of call zone1 (%)	10	3	4	4	3	3	4	4	3	3	4	3	4
	20	3	4	4	3	3	4	4	3	3	4	3	4
	30	3	4	4	3	3	4	4	3	3	4	3	4
	40	3	4	4	3	3	4	4	3	3	4	3	4
	50	**3**	**4**	**4**	**3**	**3**	**4**	**4**	**3**	**3**	**4**	**4**	**3**
	60	**3**	**4**	**4**	**3**	**3**	**4**	**4**	**3**	**3**	**4**	**4**	**3**
	70	3	4	4	3	3	4	4	3	4	3	4	3
	80	3	4	4	3	3	4	4	3	4	3	4	3
	90	3	4	4	3	3	4	4	3	4	3	4	3

Recommendations: underlined rows indicate optimal and closest policies are the same for priority1, bold rows indicate instances for which the optimal and closest policies are the same for priority3 calls, and boxed rows indicate where optimal and closest are the same for both priorities.

In Figure 2.4 we show a comparison of the closest dispatching policies and the optimal dispatching policies for the 2x2x2 case in terms of the resulting expected survival rate and the expected response time for priority1 calls. The results show that the optimal policy performs better (statistically significant difference) when the percentage of calls from zone1 range from 0% to 20% and from 90% to 100%. In addition, when we consider minimizing response time for priority1 calls as the objective, the resulting optimal policies are the same as those when we consider

maximizing the expected survival rate.



(a) The expected survival rate

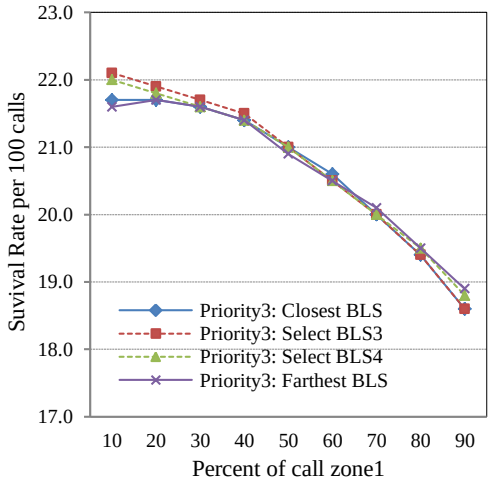
(b) The expected response time for prority1 calls

Figure 2.4: Comparison of the expected survival rate and the expected response time for priority1 calls under the closest dispatching policy versus the optimal dispatching policy

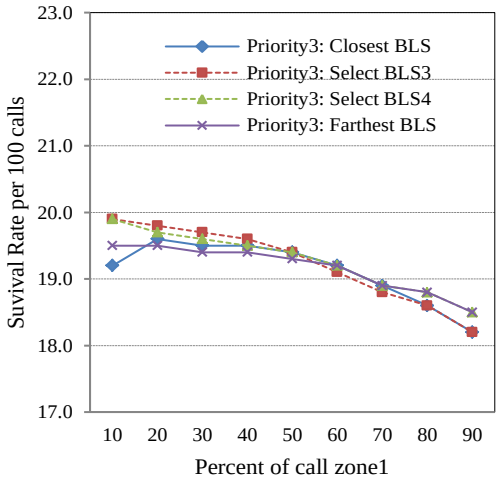
In Figure 2.5, for each panel we fix the policy for priority1 calls (there are 4 possibilities) and look at the performance of all possible policies for dispatching BLS units for priority3 calls. There are 4 possibilities of how we dispatch BLS units for priority3 calls in each case. Therefore, there are 16 possible dispatching policies for dispatching BLS units for priority1 and 3 calls. The results in Figure 2.5(a) show that, regardless of the policy applied to priority3 calls, the system performance is best when the closest unit is sent to priority1 calls. In other words, the lines in 2.5(a) for each case are higher than the corresponding lines in 2.5(b)-2.5(d) for all possible call volume distributions. Results also indicate that the closest dispatching policy is not always

optimal for priority3 calls (the blue line is not always highest). However, when the call volume is balanced between zones (50% to 60% in zone1), the optimal dispatching policy for priority3 calls is the closest dispatching policy.

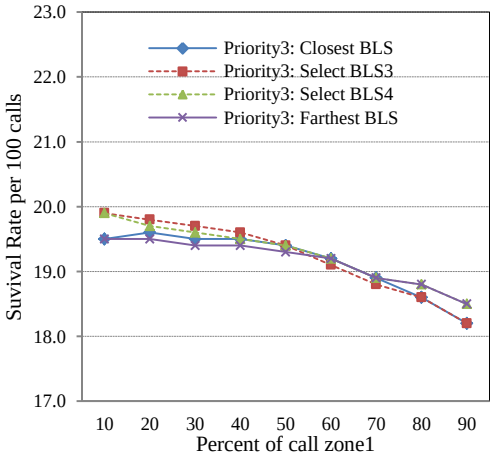
(a) Priority1: Closest ALS and always send closest BLS first to priority1



(b) Priority1: Closest ALS and always send BLS4 first to priority1



(c) Priority1: Closest ALS and always send BLS3 first to priority1



(d) Priority1: Closest ALS and always send farthest BLS first to priority1

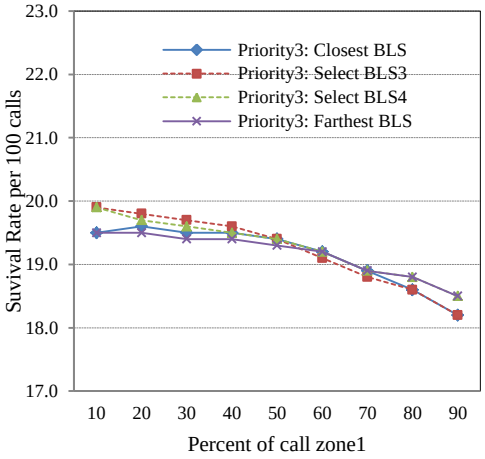


Figure 2.5: Comparison of the expected survival rate of dispatching policies for prirority3 calls given a fixed dispatching policy for priority1 calls

2.5 Heuristic Approach

In previous sections, we used simulation to study the nature of the optimal policy for multiple unit dispatch with call priorities. The model considered the situations of upgrade or downgrade, based on the on-scene conditions. There was a chance that the severity of the situation could be changed. Although the simulation model was flexible and allowed us to realistically represent the complexity of this problem, it was better suited for analyzing system performance than for optimization due to running time. Therefore, a heuristic approach was developed for realistic EMS system implementation. For example, the running time to find the optimal policy for the problem of size 2x2x2 (16 possible dispatching policies) was 208 minutes, while the heuristic approach, which we will present, was 1 second for a problem of the same size. By studying the small problem, we learned that there are service time dependencies between the ALS and BLS medical units. After comparing the closest dispatching policy to the optimal policy, the results show that the closest dispatching policy of both ALS and BLS medical units for priority1 calls is optimal. However, when demand zones are not balanced, the closest dispatching policy is not optimal for priority3 calls. When congestion is considered (e.g., having busy servers), EMS systems approach the optimal survival rate of patients by trying to send nearby ambulances to respond to priority1 calls, and sending ambulances that are far away to respond to priority3 calls, when demand between zones is not balanced. In other words, the system tries to balance the workloads among ambulances when the arrival rate of each demand zone is imbalanced. This evidence shows that we approach the optimal policy by balancing workloads or call volumes among ambulances.

Therefore, a heuristic algorithm is developed to provide an ordered preference list for BLS units for priority3 calls, given that we dispatch the closest ALS and BLS units to respond to priority1 calls and dispatch the closest BLS unit to respond to priority2 calls. The main idea of the heuristic algorithm is to arrange the ordered preference list of priority3 calls to balance call volumes among servers.

Our heuristic consists of several embedded iterative procedures. The highest level is denoted as the "main" procedure. In the main procedure, we update the candidate solution, which provides an ordered preference list of priority3 calls for each zone for any iteration. The main procedure is run until the stopping criterion is met or a fixed number of iterations are completed. Our main procedure has two steps as sub-routines: in *Step I*, we calculate the expected service time of servers (t), the approximated busy probability for servers (ρ_k), the approximated call volume for servers ($v_{BLS:k}$) and the mean absolute deviation of call volumes for servers (B). Within *Step I*, we use an iterative procedure to estimate the busy probability for servers. We update ρ_k for any iteration and for each server k until ρ_k approaches convergence for each server k. In *Step II*, we improve the current solution by using a swapping procedure. The new arrangement of the ordered preference list of proirty3 calls minimizes the mean absolute deviation of call volumes for servers. The algorithm allows for acceptance of unimproved solutions when a better result is not found. Figure 2.6 shows the logic of the heuristic algorithm in a flowchart. The main procedure consists of the following steps:

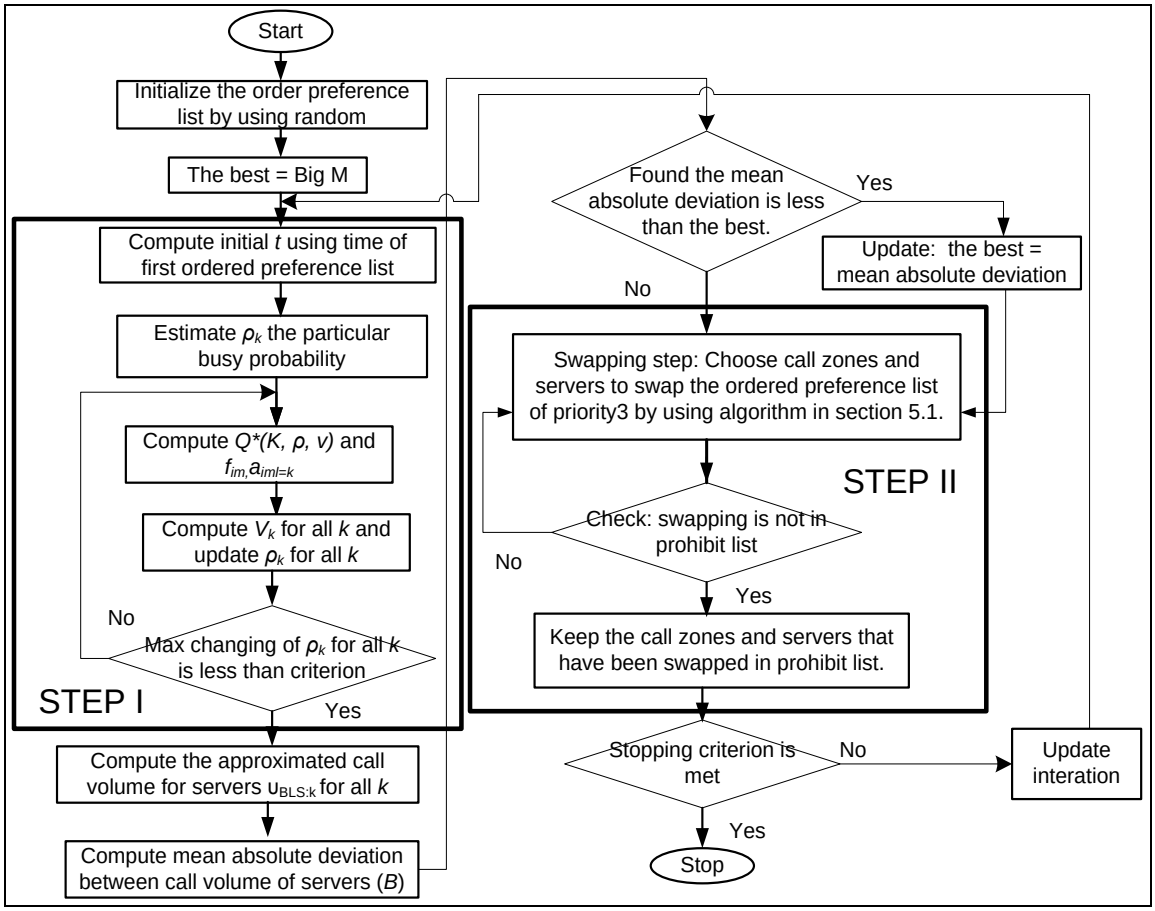


Figure 2.6: Flow chart describing the heuristic algorithm

"Main" Procedure Steps

Step1: Initialization step, we start with initial solution by using a randomized ordered preference list for priority3 calls.

Step2: Compute ρ_k and $f_{im,}a_{iml=k}$, using *Step I(1) – I(8)* below.

Step3: Calculate the initial call volume ($\upsilon_{BLS:k}$) for each server using (2.15).

Step4: Compute the mean absolute deviation (*B*) by using (2.16) and (2.17).

Step5: Use *Step II*. Swapping the arrangement of ordered preference list of prority3 calls by using the swapping procedure in Section 2.5.1.

Step6: Stop if the stopping criterion is met. Otherwise, we return to *Steps 2 – 4*. Compute ρ_k, $f_{im,}a_{iml}$ and $\upsilon_{BLS:k}$ by using *Step I(1) – I(8)* below, and equation (2.15). Compute *B* by using equation (2.16) and (2.17). For our algorithm, the stopping criterion is met or the fixed number of iterations is completed.

In *Step I*, the hypercube model is applied to approximate the busy probability among servers based on the approximation for the hypercube model given by Larson [14]. We approximate the busy probabilities by considering a general *M/M/K/K* system (multi-server with zero queue) for which the EMS system operates in steady-state. A summary of the heuristic parameters is given in Table 2.1. We estimate ρ, or average offered load per server, by $\rho = r/K$, where r is average server utilization (λ/τ), and K is the number of BLS units. In addition, we estimate the average server workload ρ_k for all k by ρ (1- p_K), where p_K is the busy probability for all servers based on an *M/M/K/K* system. The service rate τ is determined based on servers, priorities and call zones where t is the average service time. The $t_{im,}a_{iml}$ is the expected service time at which priority m of call zone i is assigned to server a_{iml} (i.e. server $k = a_{iml}$). The server a_{iml} is the l^{th} preferred server. The server a_{iml} is available for priority m of call zone i, and the first (l-1) preferred servers are busy. The initial approximated t is calculated by using the expected service time of the first ordered preference list given by

$$t = \sum_{i=1}^{n}\sum_{m=1}^{3}(\lambda_{im,a_{im1}} / \lambda)\cdot t_{im,a_{im1}} \qquad for \quad i = 1, 2,..., n;\ m = 1, 2, 3;\ l = 1, 2,...,K \qquad (2.2)$$

$$\tau = 1/t \qquad (2.3)$$

The busy probability that u servers are busy is then calculated based on the *M/M/K/K* system. Let p_0 denote the probability that all servers are available and p_u denote the probability that u servers are busy. The steady-state of an *M/M/K/K* system is given by

$$p_0 = [\frac{r^K}{K!}(\frac{1-\rho^{K-K+1}}{1-\rho}) + \sum_{u=0}^{K-1}\frac{r^u}{u!}]^{-1} \tag{2.4}$$

$$p_u = \frac{(\lambda/\tau)^u}{u!} / \left(\sum_{w=0}^{K}\frac{(\lambda/\tau)^w}{w!}\right) \qquad for \quad u = 1, 2,..., K \tag{2.5}$$

We start the procedure by calculating the initial values of particular busy probabilities for each server. We estimate ρ_k as the busy probability of server k, assuming one server per station. We assume that the system operates with t average service time for the *M/M/K/K* system. We use equation (2.6) for approximating the busy probability for each server.

$$\rho_k = \lambda \cdot t \cdot (1 - p_K) / K \qquad for \quad k = 1, 2,..., K \tag{2.6}$$

Our model begins with the server independence assumption. The $f_{im,}a_{iml}$ is the total rate at which priority m of call zone i are assigned to server a_{iml} (i.e. server $k = a_{iml}$). The server a_{iml} is the l^{th} preferred server. The server a_{iml} is available for priority m of call zone i, and the first (l-1) preferred servers are busy. Larson [14] developed the approximation method for $f_{im,aiml}$ as given by

$$f_{im,}a_{iml} \sim g_{l-1} \cdot \lambda_{im,}a_{iml} \qquad for\ l = 1, 2, \ldots, K\text{-}1 \tag{2.7}$$

Estimating the probability g_v that the server $v+1$ is available and the first (v) servers are busy as given by

$$g_v \sim \sum_{v=0}^{K-1} Q^*(K, \rho, v) \cdot (1 - \rho_{a_{im,v+1}}) \cdot \prod_{l=1}^{v} \rho_{a_{iml}} \qquad for\ v = 0, 1, 2, \ldots, K\text{-}1 \tag{2.8}$$

For example, to estimate g_{l-1} from equation (2.7) where l^{th} is 3^{rd}, we use equation (2.8), where g_v is g_{l-1} ($g_v = g_{3-1} = g_2$). The correction factor $Q^*(K, \rho, v)$ indicates that the probability of obtaining v servers are busy. The ρ is the average offered load per server. Let $Q^*(K, \rho, 0)$ be 1 given by the *M/M/K/K* system. Larson [14] developed a correction factor $Q^*(K, \rho, v)$ for *M/M/K/K* systems based on the expression of the correction factor $Q(K, \rho, v)$ for the *M/M/K* system. Larson [14] defined $Q^*(K, \rho, v)$ as

$$Q^*(K,\rho,v) = Q(K,\rho,v)\left(\frac{1}{1-p_K}\right)^v\left(\frac{1}{1+\frac{\rho p_K}{1-\rho}}\right) \tag{2.9}$$

Where $Q(K,\rho,v) = \dfrac{\sum_{k=v}^{K-1}\left((K-v-1)!(K-k)/(k-v)!\right)(K^k/K!)\rho^{k-v}}{(1-\rho)\left(\sum_{w=0}^{K-1}(K^w/w!)\rho^w\right)+(K^K\rho^K/K!)}$ *for v =0, 1, 2,…, K-1* (2.10)

Next, we use an iterative procedure as developed by Jarvis [15]. For any iteration, we update ρ_k, the particular busy probability for each server until the maximum change in ρ_k is less than the convergence criterion. We update $\rho_k(new)$ by using equations (2.11) and (2.12).

$$\rho_k(new) = V_k/(1+V_k) \qquad \text{for } k = 1, 2,..,K \tag{2.11}$$

Where V_k is given by

$$V_k = \sum_{i=1}^{n}\sum_{m=1}^{3} f_{im,a_{iml}=k}\cdot t_{im,a_{iml}=k}/(1-\rho_{a_{iml}=k}) \qquad \text{for } k = 1, 2,..,K \tag{2.12}$$

In addition, similar to Jarvis [15], we approximate t at the end of any iteration by

$$t(new) = \sum_{i=1}^{n}\sum_{m=1}^{3}\left(\begin{array}{l}(f_{im,a_{im1}}/\lambda)\cdot t_{im,a_{im1}} + (f_{im,a_{im2}}/\lambda)\cdot t_{im,a_{im2}} + \ldots \\ +(f_{im,a_{imK}}/\lambda)\cdot t_{im,a_{imK}}\end{array}\right) \tag{2.13}$$

$$\rho(new) = \lambda \cdot t(new) / K \tag{2.14}$$

Each iteration of *Step I* consists of the iterative algorithm for approximating ρ_k, the particular busy probability for each server. The procedure consists of the following steps:

Step I:

Step 1: Initialize t and ρ using (2.2) and (2.3).

Step 2: Compute $p_{0,}\, p_u$ for all u and ρ_k for all k, using (2.4) – (2.6).

Step 3: Compute $Q^*(K, \rho, v)$ for v =0, 1, 2,…, K-1, using (2.9) and (2.10).

Step 4: Compute $f_{im,}a_{iml=k}$ by using (2.7) and (2.8).

Step 5: Compute V_k for all k, using (2.12) where $f_{im,}a_{iml=k}$ is obtained from Step I(2.4).

Step 6: Update ρ_k for all k, using (2.11).

Step 7: Stop if maximum change in ρ_k for all k is less than criterion. Otherwise, update t(new) and ρ(new), using (2.13) and (2.14).

Step 8: Return to Step I(2.2).

The result of *Step I* is the approximated busy probability for servers (ρ_k) and the final value of $f_{im,}a_{iml=k}$. The evaluation of the heuristic algorithm uses the mean absolute deviation of call volumes for servers (B). The approximated call volumes are calculated in each iteration by using the equation below.

$$\upsilon_{BLS:k} = \sum_{i=1}^{n}\sum_{m=1}^{3}\begin{pmatrix} f_{im,a_{im1}=k} + f_{im,a_{im2}=k} + f_{im,a_{im3}=k} + .. \\ + f_{im,a_{imK}=k} \end{pmatrix} \quad for\ \upsilon_{BLS:1}\ ,\ \upsilon_{BLS:2}\ ,\ldots,\ \upsilon_{BLS:\,K} \tag{2.15}$$

The mean absolute deviation of call volumes for servers is calculated by

$$B = \sum_{k=1}^{K}\left|\upsilon_{BLS:k} - \overline{\upsilon_{BLS}}\right| \tag{2.16}$$

Where $\overline{\upsilon_{BLS}} = \left(\sum_{k=1}^{K} \upsilon_{BLS:k} \right) / K$ (2.17)

2.5.1 Step II: Swapping Procedure

We improve a current solution by swapping the arrangement of the ordered preference list for priority3 calls**.** The heuristic parameters are given in Table 2.1. The swapping procedure reorders the preference list by evaluating the increase in the call volumes (*vol*) for each server resulting from moving the server with the current lowest call volume (r_K) to a different position on the preference list. In Figure 2.7, we show an illustrative example of the swapping procedure executed for a problem of size 2x2x3 (2 demand zones, two ALS units and 3 BLS units). For this instance, suppose we have 3 BLS units (numbered 3, 4 and 5) in the ordered preference list with proportion for priority3 calls for each zone to be $p_1^3 = 0.10$ and $p_2^3 = 0.15$. Information about the current solution is given in panel A, including the current preference list, call volumes per unit, deviation in call volumes between all units and the unit with the current lowest call volume, the fitness value, and the probability of the first *v-1* servers being busy (g_v). The fitness is the mean absolute deviation of call volumes for servers (*B*), calculated using equation (2.15). In the current solution, the server with the lowest call volume is server 5. Since there are three BLS units, for each zone, we have 2 possible swaps in the ordered preference list ((3, 5) and (4, 5)). Next we evaluate the effects of these swaps, as shown in Panels B and C. If we swap the pair (3, 5), the increase in value of call volumes to unit 5 from zone1 (*vol*) is 0.049 and the increase in call volume to unit 5 from zone2 is

0.016. The total increase in volume to unit 5 from both zones is 0.065, which is less than the current deviation in call volumes between units 3 and 5 (*vdev* (3, 5) = 0.113). Therefore, swapping the order between units (3, 5) for both zones becomes a candidate solution. Next, we consider the pair of (4, 5) in the same fashion. The equation for calculating the increased call volume is given in (18). In general, we evaluate swapping unit r_K with all other units that are higher up in the preference list for each zone. We only allow swapping the order in the preference list for zones in which increasing the call volumes (*vol*) is less than the value of the deviation of call volumes between swapping pairs (*vdev)*. Then we choose the best solution of all possible solutions, based on the fitness value. The algorithm allows for acceptance of unimproved solutions, in terms of better fitness (compared to the current solution), though the new solution will be the best among all possible options. A more detailed description of the swapping algorithm is shown in Appendix A.2.

$$\textit{Increased call volume (vol)} = \sum_{i=1}^{vol \le vdev} \lambda_{im} \cdot (g_{w-1} - g_{z-2}) \qquad \textit{for} \quad i=1, 2, 3, \ldots, n \qquad (2.18)$$

where g_{w-1} the probability of the first *w-1* servers are busy

g_{z-2} the probability of the first *z-2* servers are busy

w the w^{th} preferred server in which server r_K is assigned to respond to priority m of call zone i (the server r_K is the position w^{th} in the rank of ordered preference list for priority3 calls)

z the z^{th} preferred server in which server k is assigned to respond to priority m of call zone i (the server k is the position z^{th} in the rank of ordered preference list for priority3 calls)

Current Solution					
Preference List	Zone 1 { 3, 4, 5} Zone 2 { 4, 3, 5}				
Call volume per unit	{0.353, 0.427, **0.240**}, Unit 5 has lowest call volume				
Fitness value:	0.200				
$\boldsymbol{g}_v$	$g_1 = 0.571$, $g_2 = 0.186$, $g_3 = 0.079$				
Deviation in call volumes	vdev (3, 5): (0.353 - 0.240) = 0.113; vdev (4, 5): (0.427 - 0.240) = 0.187				
Evaluate Swapping Location on Preference list of Unit 3 with Unit 5					
				vol	vdev
	zone1	{ 5, 4, 3 }	0.10 *(0.571 - 0.079) =	0.0492	< 0.113
	zone2	{ 4, 5, 3 }	0.15 *(0.186 - 0.079) =	0.01605	< 0.113
			Sum vol =	0.06525	< 0.113
Candidate solution: swap (3, 5) for both zone1 and 2					
Evaluate Swapping Location on Preference list of Unit 4 with Unit 5					
				vol	vdev
	zone1	{ 3, 5, 4 }	0.10 *(0.186 - 0.079) =	0.0107	< 0.187
		r_K k		vol	vdev
	zone2	{ 5, 3, 4 }	0.15 *(0.571 - 0.079) =	0.0738	< 0.187
			Sum vol =	0.0845	< 0.187
Candidate solution: swap (4, 5) for both zone1 and 2					
Next, we will calculate the fitness value and choose the better of these two candidate solutions					

Figure 2.7: An illustrative example of the swapping procedure for a problem of size 2x2x3

In the swapping procedure, we use a prohibit list to avoid selecting a previously-considered solution. We record the old solutions in a prohibit list. This list consists of zones and pairs of swapped servers. In each iteration, we create a new solution that is not

in a prohibit list. The size of the list is managed to provide a better solution while not resulting in too much additional computational time (list size is between 4 and 15).

Illustrative Example of the Heuristic Policy Implementation

Here, we briefly compare the results of the heuristic policy with the optimal policy for several small problems of size 2x2x2. We use the same input parameters as the examples in Section 2.4. Our heuristic algorithm is programmed in the Java programming language. The NetBeans IDE 7.3.1 is used to implement on an Intel® Core(TM)2 Duo CPU. The results in Table 2.5 show the heuristic dispatching policies compared with the optimal dispatching policies for BLS units to respond to priority3 calls where we fix the closest dispatching policies for BLS units to respond to priority1 and 2 calls. In Table 2.5, the results show that the heuristic dispatching policies are the same as the optimal dispatching policies. It does not depend upon the proportion of calls to each zone. There are two instances in which the results of the heuristic dispatching policies are not the optimal dispatching policies. They are underlined. However, these results show that there are small deviations between results of the heuristic dispatching policies and the optimal dispatching policies with average percent error 0.10%. Comparing the running times, the heuristic approach takes < 1 second to find a solutions in all cases, the simulation models finds the optimal policy by enumerating all possible policies, which takes 208 minutes on the problem of size 2x2x2. Therefore, it suggests that the heuristic approach could improve more the survival rate for EMS systems when the enumeration of all possible policies is not practical, as is the case in most real world problems.

Table 2.5: Comparison of dispatching policies for BLS units between the optimal policy and the heuristic policy for priority3 calls (underlined rows indicate instances for which the optimal and heuristic policies are different policies)

Dispatch policy		Optimal Policy									Heuristic Policy									
		Closest policy				Optimal policy				E[Sur. Rate] /100 calls	Closest policy				Optimal policy				E[Sur. Rate] / 100 calls	% Error
		Call zone 1		Call zone 2		Call zone 1		Call zone 2			Call zone 1		Call zone 2		Call zone 1		Call zone 2			
Choice of dispatch		1st	2nd	1st	2nd	1st	2nd	1st	2nd		1st	2nd	1st	2nd	1st	2nd	1st	2nd		
Percent of call zone1 (%)	10	3	4	4	3	3	4	4	3	22.1	3	4	4	3	3	4	3	4	22.1	0.00
	20	3	4	4	3	3	4	4	3	21.9	3	4	4	3	3	4	3	4	21.9	0.00
	30	3	4	4	3	3	4	4	3	21.7	3	4	4	3	3	4	3	4	21.7	0.00
	40	3	4	4	3	3	4	4	3	21.5	3	4	4	3	3	4	3	4	21.5	0.00
	50	3	4	4	3	3	4	4	3	21.0	3	4	4	3	3	4	4	3	21.0	0.00
	60	3	4	4	3	**3**	**4**	4	3	20.6	3	4	4	3	**4**	**3**	4	3	20.5	0.49
	70	3	4	4	3	3	4	4	3	20.1	3	4	4	3	4	3	4	3	20.1	0.00
	80	3	4	4	3	3	4	4	3	19.5	3	4	4	3	4	3	4	3	19.5	0.00
	90	3	4	4	3	3	4	**4**	**3**	18.9	3	4	4	3	4	3	**3**	**4**	18.8	0.53

The average % error = 0.10%

2.6 Case Study and Computational Results

Real-world data was collected from Hanover Fire and EMS department, which is located in Hanover County, Virginia. The data was collected in 2007 and consists of approximately 10,000 calls. The system operates 24 hours per day to respond to 911 calls in an area of about 474 square miles with a population of approximately 100,000. We partitioned this area into twelve demand zones and considered four rescue stations. The four rescue stations are shown in Figure 2.8. We studied the performance of the system as we varied the number and locations of the ALS units. The ALS and BLS units were randomly allocated to four stations. We varied the number of ALS units between 1 and 3 and fixed the number of BLS units at 3. We considered three call priorities (priority1, 2 and 3 calls) and allowed the proportion of calls from each priority to vary by call zone. We assume that the probability that a priority1 call requires the ALS medical unit on the scene is 0.5 for priority1 calls. All input parameters are shown in Appendix A.3, where

response times and transportation time used a distribution by based on call zones. We noted that the mean service time for priority1 calls is higher than the mean service time for priority2 and 3 calls.

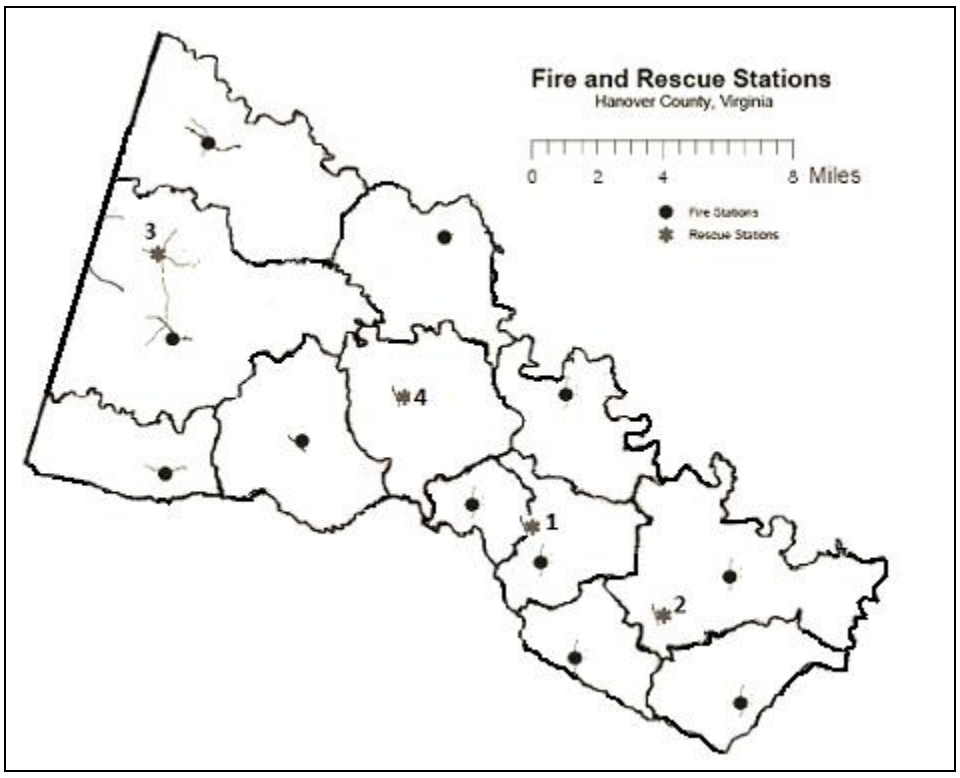


Figure 2.8: Map of fire and rescue stations in Hanover County, Virginia

Table 2.6 shows results for different number of ALS units and a fixed number of BLS units. We allocate the BLS units to four fixed stations and allocate the ALS units to different possible stations. We compare the results of using the heuristic policy to always sending closest dispatching policy, with a varied number of calls per hour. We observe that dispatching units according to the heuristic policy provides better outcomes for the mean absolute deviation of the busy probability of BLS units, as shown in column11 of Table 2.6. These outcomes imply that the heuristic policy achieves better balance in

utilization of BLS units as compare to the closest dispatching policy. We observe that dispatching multiple units according to the heuristic policy can increase the survival probability in comparison to always sending closest policy. Table 2.6 shows the improvement in terms of number of lives saved per 10,000 calls (the approximate annual call volume in Hanover) when using the heuristic approach in comparison to always sending closest dispatching policy (last column). The results show that the heuristic policy can increase the number of lives saved. We also observe that the number of ALS units provides a large effect to the number of lives saved according to the heuristic policy in comparison to always sending closest dispatching policy. Fewer available ALS units results in a larger difference in outcomes between the heuristic and closest policies. That is, given that there is one ALS unit, the results in Figure 2.9 show a larger difference in outcomes when using the heuristic policy in comparison to always sending closest dispatching policy. When there are three ALS units in the system, the results show only a slight difference between the heuristic and closest policies. These results suggest that the efficiency of the heuristic algorithm depends on the number of ALS units in the system. This implies that the heuristic policy provides more value in resource-constrained systems (limited number of ambulances). In addition, Figure 2.9 shows a decreasing function relationship between the number of lives saved per 10,000 calls and call arrival rate, which is due to the increasing busy probability of servers as the call arrival rate increases. We also observe that as call volume increases, the number of lives saved will decrease in a non-monotone fashion depending on the number of ALS units in the system. When there are two ALS units (in 10(a)), in comparison to one ALS unit (in

10(b)), an increase in call arrivals per hour decreases the benefit of the heuristic policy. This observation suggests that when a system is close to capacity, increasing the number of ALS units results in only slight improvements.

Table 2.6: Comparison of performance of heuristic policy to closest policy as we vary the number of ALS units and call arrival rate

ID	Demand (calls/hr)	Policy	Utilization						Mean utilization of BLS	Mean absolute deviation of BLS	# of lives saved /10,000 calls	% Imp.	# of the imp. of lives saved /10,000 calls
3 ALS 3 BLS 12 Zones													
			ALS1 :St4	ALS2 :St1	ALS3: St 3	BLS4 : St 4	BLS5: St1	BLS6 : St1					
		Closest	14.43	3.10	6.244	23.87	12.601	3.123	13.196	21.3385	1629		
1	0.25	Heuristic	14.42	3.18	6.171	16.16	13.762	11.37	13.764	4.7899	**1656**	1.657	27
		Closest	33.29	14.78	22.837	43.81	35.008	21.38	33.398	24.0423	1468		
2	0.50	Heuristic	33.07	14.92	22.088	33.29	31.402	36.99	33.890	6.1885	**1507**	2.657	39
		Closest	57.16	40.37	49.713	64.76	62.512	51.87	59.713	15.6815	1278		
3	0.75	Heuristic	55.96	39.03	48.152	57.72	60.973	58.61	59.100	3.7473	**1316**	2.973	38
		Closest	75.57	64.12	72.76	81.18	81.396	76.43	79.671	6.4729	1141		
4	1.00	Heuristic	75.22	64.09	72.406	77.43	81.101	80.59	79.707	4.5529	**1163**	1.928	22
		Closest	84.89	76.00	83.859	88.94	90.008	87.32	88.757	2.8823	1050		
5	1.25	Heuristic	84.58	76.12	83.661	87.23	90.070	89.08	88.792	3.1342	**1069**	1.810	19
2 ALS 3 BLS 12 Zones													
			ALS1 : St3	ALS2 : St1	ALS3	BLS4 : St 4	BLS5: St1	BLS6 : St1					
		Closest	27.92	15.41	N/A	30.77	22.017	12.35	21.712	18.7269	1303		
1	0.25	Heuristic	27.80	15.21	N/A	23.80	23.044	19.34	22.059	5.4439	**1340**	2.840	37
		Closest	67.35	60.69	N/A	69.87	66.125	57.87	64.623	13.5071	1114		
2	0.50	Heuristic	66.74	59.24	N/A	63.48	63.774	64.39	63.880	1.0173	**1163**	4.399	49
		Closest	85.30	82.72	N/A	86.91	85.847	81.59	84.782	6.3778	1014		
3	0.75	Heuristic	84.96	82.43	N/A	84.62	85.796	83.66	84.692	2.2065	**1047**	3.254	33
		Closest	92.06	91.14	N/A	93.26	92.810	90.76	92.279	3.0342	948		
4	1.00	Heuristic	91.97	90.86	N/A	91.81	92.732	91.92	92.153	1.1574	**982**	3.586	34
		Closest	94.64	94.46	N/A	95.70	95.656	94.28	95.213	1.8590	903		
5	1.25	Heuristic	94.69	94.59	N/A	95.23	95.762	94.99	95.328	0.8683	**924**	2.326	21
1 ALS 3 BLS 12 Zones													
			ALS1 :St1	ALS2	ALS3	BLS4 : St 4	BLS5: St1	BLS6 :St1					
		Closest	61.25	N/A	N/A	62.07	56.415	50.27	56.253	11.9599	1098		
1	0.25	Heuristic	60.81	N/A	N/A	57.84	56.421	53.98	56.078	4.2067	**1134**	3.279	36
		Closest	89.43	N/A	N/A	89.18	87.246	83.81	86.744	5.8728	957		
2	0.50	Heuristic	89.18	N/A	N/A	87.08	86.205	86.59	86.623	0.9145	**1011**	5.643	54
		Closest	95.40	N/A	N/A	95.05	94.148	92.19	93.796	3.2167	877		
3	0.75	Heuristic	95.01	N/A	N/A	93.69	93.588	92.83	93.367	1.0841	**922**	5.131	45
		Closest	97.44	N/A	N/A	97.09	96.598	95.37	96.353	1.9659	818		
4	1.00	Heuristic	97.28	N/A	N/A	96.25	96.267	95.97	96.161	0.3840	**887**	8.435	69
		Closest	98.31	N/A	N/A	97.95	97.625	96.77	97.448	1.3489	794		
5	1.25	Heuristic	98.21	N/A	N/A	97.47	97.485	97.05	97.336	0.5693	**833**	4.912	39

Next, we also investigated how the locations of ALS units impacted our results, as shown in Figure 2.10. Given that there are two ALS units in the system, the graph shows that the heuristic policy yields an improvement (over the closest policy) when the ALS units are sited in station 3 and station 1. Even though locating the ALS units to stations 3 and 1 is the worst location options in terms of survivability, allocating units properly can improve system survival. In addition, the heuristic policy still provides improvement in the number of lives saved when we allocate one ALS unit to station 1. However, locating the unit at station 1 provides the least reward in comparison to other locations. These results imply that the heuristic policy provides gains in the number of lives saved regardless of where the ALS units are located, and it provides the highest gains in the number of lives saved when the ALS units are poorly located. These results are shown in Appendix A.4. However, the results also show that the locations of ALS units have a stronger impact on the number of lives saved. Therefore, in future work we will incorporate joint location into our multiple dispatching model.

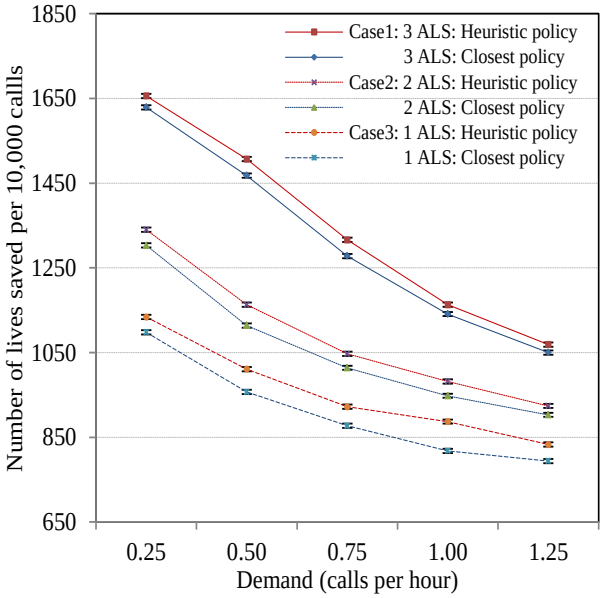


Figure 2.9: Comparison of the efficiency of the heuristic policy with closest policy for different numbers of ALS units

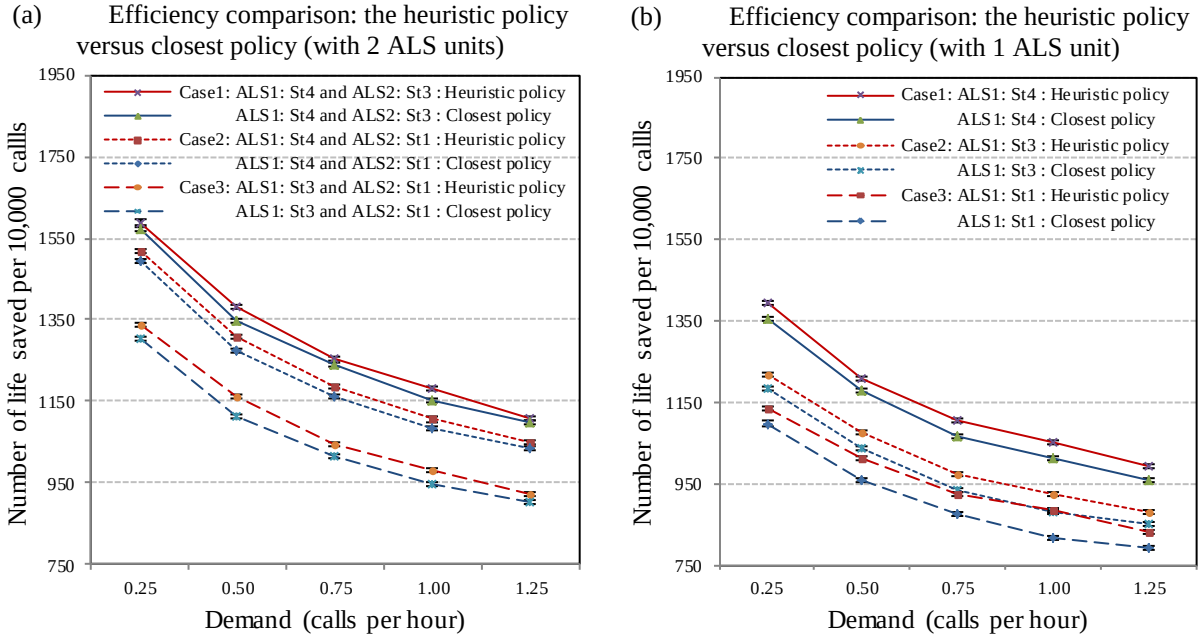


Figure 2.10: Comparison of the efficiency of the heuristic policy with closest dispatching policy for different locations of the ALS unit

2.7 Conclusions and Future Research

In this paper, we developed a simulation model for multiple unit dispatch. Emergency calls are classified into three types based on their perceived severity, which require different response modes. We consider that these classifications can be changed based on information on-scene of accidents. The simulation is formulated given particular dispatching policies such as single and multiple dispatching of ALS and BLS ambulance types. First, we study the optimal policy for small problems by using the full enumeration approach, taking into account the survival probability of priority1 calls. We present the comparison of the closest dispatching policy with the optimal policy. Numerical results on these instances show that the closest dispatching policy of both ALS and BLS medical units for priority1 calls is the optimal policy. However, the closest dispatching policy is not optimal for priority3 calls. The optimal policy tends to send father ambulances for priority3 calls to balance workloads among servers. A heuristic algorithm has been proposed for multiple unit dispatch in large-scale problems. In this heuristic, we dispatch the closest available ALS unit and the closest available BLS unit for priority1 calls given that we dispatch the closest available BLS for priority2 calls. The hypercube model is adapted to determine the busy probability for servers. We developed the heuristic algorithm by following the balanced call volume for servers. The proposed rule provides an ordered preference list for priority3 calls to minimize the deviation in the busy probability of servers by sending a less busy BLS unit.

We have demonstrated the heuristic algorithm by using data supplied from Hanover Fire and EMS department, in Hanover County, Virginia. The case study shows that fewer ALS units provide more efficiency in the heuristic policy than a larger of the number of ALS units. The number of ALS units in the system has a strong effect on the efficiency of the heuristic policy. This study highlights the importance of acquiring the proper number of ALS units that are reasonable within a limited budget. We present other sensitivity analyses, which show that the location of ALS units has a large effect on the benefit of the heuristic policy.

In future research, we will develop an exact Markov Chain Model for multiple tiered responses with consideration for different on-scene conditions of accidents. However, the model will likely be complex. In addition, we will further expand upon the multiple dispatching model by using a weight for each patient priority, or by using multiple coverage criteria. In addition, the results indicate that investigating a redeployment problem would improve the efficiency of EMS systems. We will develop a dynamic ambulance location (deployment) problem. Work is in progress to address these issues.

CHAPTER THREE

A SIMULATION MODEL FOR FAIRLY DISPATCHING EMERGENCY VEHICLES UNDER MULTI-TIERED RESPONSE

3.1 Introduction

Medical priority dispatching is used to improve efficiency of EMS systems. The strategy of medical priority dispatching is to consider a faster response time to life-threatening patients. The study of pre-hospital mortality in EMS systems by Kuisma et al. [2] showed that dispatching a far ambulance to low priority patients does not negatively impact pre-hospital mortality rates. Therefore, the decisions regarding how to dispatch ambulances do not adversely affect low priority patients in terms of survival rates, since these patients are non-critical. Medical priority dispatching may make the closest ambulances unavailable to non-serious patients. Emergency calls are classified into three priority levels upon dispatch, and their classification may be updated once a responder reaches the scene and makes further assessment. For example, BRAVO calls (prioirty2) are potentially life-threatening calls that could be upgraded to life-threatening (priority1). In this case, priority2 calls need a paramedic unit and rapid transport.

In this chapter, balancing equity between priorities while still providing a high efficiency is investigated. We consider fairness in patient outcomes; that is, the waiting time until the first response should be balanced between life-threatening calls (priority1) and potentially life-threatening calls (priority2). Considering the decision regarding how to dispatch ambulances based on assumptions of Chapter1, the optimal policy tended to

dispatch nearby ambulances to the priority1 calls and a father ambulances to the priority2 calls. However, in this chapter we allow for priority2 calls that could be upgraded to life-threatening. Dispatching the father ambulance to the priority2 calls may affect the death rate of priority2 calls that are later upgraded. The fairness in patient waiting times will enforce the first response time of priority1 and priority2 calls that result in increasing a probability of nearby ambulances available to respond to priority2 calls and decreasing the first response time of priority2 calls, potentially saving lives of priority2 patients.

Fairness or equity may be impacted by the dispatching decision, as seen in Marsh [41]. Many fairness measures were suggested in several works investigating the facility location problem. The study regarding how to allocate resources to facilities in order to improve efficiency of systems leads to inequities for customers. The works on equity considered distributional equity, which may still improve of the overall outcomes, as seen in Savas [42], Bodily [43] and Henderson and Schilling [44]. Other studies analyzed fairness between customers. Keeney and Winkler [45] presented different ways to evaluate equity as both ex-ante customer equity and ex-post customer equity. How ambulances were allocated to stations resulted in unfairness between patients of different priorities in EMS systems. A review of equity measurements in facility location was presented in Marsh [41]. Several measures were proposed on the issue of fairness. Marsh analyzed how to choose the equity measurement alternatives based on several criteria. The decision regarding which ambulances to dispatch while incorporating equity was presented in McLay and Mayorga [29]. They formulated a model of dispatching ambulances given a set of equity constraints. Four different equity measures were

considered. The first equity measure considered the fraction of calls for which the closest ambulances could be dispatched to patients. The second equity measure considered the patient survival rates. The third equity measure evaluated server busy probabilities between a lower bound and an upper bound. The last equity measure considered the proportion of instances in which servers respond to life-threatening calls. In this chapter, we consider fairness in patient outcomes when the waiting time until the first response should be balanced between life-threatening calls and potentially life-threatening calls.

In this chapter, we extend the model of multiple unit dispatch from Chapter 2. We consider EMS systems with multiple unit dispatch, multiple call priorities, and a zero-queue. We assume that once the ambulances complete their service, they return to their original ("home") station. The proposed simulation model determines how to dispatch a basic life support unit (BLS unit) for prority2 calls by considering two alternative policies. In the case of using a single dispatch in response to priority2 calls, we take into account the changing situations on-scene based on information of the first arriving ambulance. We examine the condition of BLS upgrade in which the patients need ALS care (an advance life support unit). In this case, the BLS unit provides initial care and waits for the arrival of the next available ALS unit. In addition, we examine an ambulance dispatching policy for the ALS unit for the priority2 calls that affect the death rate for priority2 calls. We extend the model by considering how to dispatch the ALS unit for priority2 calls in order to improve the outcome of EMS systems. We determine the better dispatching policy for dispatching the ALS unit for priority2 calls while balancing the waiting time until the first response between patient priorities. The simulation model

of multi-tiered responses is formulated by incorporating fairness constraints into the model.

3.2 Literature Review

Fairness is of critical importance to the management in EMS systems. The study of distributional equity introduced by Mandell [46] showed trade – offs between efficiency and equity. They formulated a bi-criteria mathematical programming model for how to allocate resources. The objectives were overall output and equity. Ogryczak [47] considered multiple criteria models for the location problem. These performance measures were evaluated under the view of the clients. They introduced the model for the location problem that incorporated an inequality measurement. The objective was to minimize total distance while minimizing the three inequality measures in the model. The inequality measures shown were; the maximum deviation, the mean deviation and the mean difference. Felder and Brinkmann [48] considered the equity – efficiency trade off in EMS systems. The equity measure was the difference in cost across the regions. Several research studies of the equity issues were then presented to multiple criteria optimization problem by incorporating equitable criteria into assumptions. Kostreva et al. [49] studied the equity problem through a capital budgeting problem. They used the Pareto – optimality to present equity into their model. Heshmati [50] reviewed inequalities measured in economics. They suggested that the inequalities of two different areas: income and non – income. The income inequity was the output or efficiency of systems that were easily evaluated while the non-income inequality was skills, education,

opportunities, etc. The result from the review showed how to account for the relationship between inequality in income and non-income measures.

The study of equity measures in EMS systems has been mostly done in terms of the location problem. The primary objective was to maximize equality. The median function was introduced in Hakimi [51] evaluated equity of the location problem. Other works, such as Halpern [52] presented the median and center function. Ogryczak [53] considered the location model by analyzing the model of two criteria: mean overall efficiency and mean equity measure. He generated multiple criteria minimization. The outcome was to minimize distance as the equity measure was incorporated into the model. He discussed the direct use of general inequity measures that might contradict minimization of efficiency. The results of multiple criteria optimization that considered both the Pareto – efficiency and inequity measures provided a good solution for the location problem. Furthermore, he focused on several inequity measures and showed how the optimal solution of several inequity measures can be incorporated in the location model. Lorenz [54] introduced the Lorenz curve to measure equitable distribution. This graph presented the cumulative proportion of demand based on the cumulative proportion of income. The Lorenz curve was a convex function which the area between this curve and the straight equity line was the inequity measure of resource to demand. The Gini coefficient, introduced by Gini [55], was the ratio of area between the Lorenz curve and straight equity to the whole area below the equity line. The adaptation of the Gini coefficient into location models was presented by Drezner et al. [56]. They analyzed the model of the location problem that incorporated equity measures. The Gini coefficient

was used to measure inequity distribution. The objective was to minimize the sum of the Gini coefficient which represented the deviation between the income of each demand point and the income of other demand points. The income was the distance between demand point and the closest facility. The median problem (DOMP) was introduced by Marín et al. [57]. The objective was to minimize the median cost. The median function was a function of weight on cost in which resources at each location satisfied the total demand from each zone. In addition, Drezner et al. [58] considered the facility location problem by minimizing the deviation of equality among demand groups. In multiple facilities location, they presented a tabu search that was based on descent solutions to obtain an improved solution. Marín [59] proposed a facility location problem that considered the equity measure into the model. He considered the equity by calculating the difference between the maximum and minimum number of customers allocated to any plant. The two integer programming formulations were then developed. The first formulation considered the p-median problem which included variables representing the maximum and minimum number of customers allocated to any plant. The second formulation was used to obtain a different way the maximum and minimum numbers of customers allocated to any plant. He considered the constraints that ensured each customer was allocated to the closest plant. Bertsimas et al. [60] proposed resource allocation problems with fairness considerations. They enforced a bound on fairness for both maximum and minimum proportional fairness. Their results showed the relative outcome loss under fairness measures. Chanta et al. [61] considered the minimum p-envy for equity in EMS systems. The decision was to allocate ambulances to stations. The

"envy" of each demand zone was a level of customer's dissatisfaction which compared to other demand zones. Toro-Díaz et al. [62] considered the combination of the location and dispatching problem which included the fairness performance indicator in their conclusion. They formulated the combined an integer programming model representing location and dispatching problems. They developed the optimization based on genetic algorithms.

In other related works of fairness, Bertsimas et al. [63] proposed the allocation of donor kidneys to patients on a waiting queue. They considered the incorporation of fairness and efficiency in the model. The ordered rank of patients was considered the priority criteria. They formulated the maximum medical efficiency that included the fairness constraints, a lower bound on the percentage of kidneys that were allocated to each patient priority. Noyan [64] considered a stochastic model of a location problem where demand was uncertain. They developed two stochastic optimizations that included risk measures into their models. The risk constraint was the magnitude of violation in coverage. The two stage problem was then formulated with the constraint being change of risk.

In this work we extend the multiple-unit dispatch with multiple call priorities proposed in Chapter 2. The modification considers the fairness between priorities into the model. Recent studies considered the fairness among demand zones was presented by Chanta et al. [61]. Several previous works relevant to fairness above analyzed model by not taking account the real conditions at on-scene of accidents into the model. However, our work differs in which we consider the realistic on-scene conditions that the

potentially-life-threatening calls might need the paramedic unit. We also considered the fairness outcomes between call priorities which rapidly responding to the potentially-life-threatening calls requires in the model.

3.3 Model Description

In this section, we discuss the EMS systems which are extended from the original model in Chapter 2. This chapter proposes the multiple unit dispatch of EMS systems while considering on-scene conditions. The systems have three call priorities and two types of ambulances (the ALS unit and BLS units). The response area is partitioned into demand zones, each with a distinct dispatch preference list. When a call arrives at the dispatch center, the dispatch planners make the decision about which ambulances to assign in response to the call according to the preference lists. In the case when all ambulances in the preference list are busy, the call will transfer to another dispatch center. The classification of call priorities are also considered in this chapter. The dispatching of different types of ambulances depends on call priorities. The characteristics of the EMS systems, described in Chapter 2, showed that priority1 calls require a double dispatch of the ALS unit and the BLS unit. Single dispatching of the BLS unit is when the BLS unit is assigned to respond the priority2 or 3 calls. In this chapter, we consider the dispatching policy for priority2 calls. The configuration of the EMS system process with BLS-upgrade of priority2 calls is described in Figure 3.1. The main assumptions of priority1 and 3 calls are the same as the original studied in Chapter

2, except for the situation of on-scene upgrades/downgrades for priority2 calls. The adapted models of possible situations at on-scene priority2 calls are:

(i) Vehicle dispatch decisions: Priority2 calls require a single dispatch (BLS unit). We dispatch the available BLS unit for priority2 calls according to two possible policies; priority1 (closest policy) or 3 (heuristic policy) calls in which the inputs for the dispatching policies of priority1 or 3 calls are based on results from Chapter 2. To obtain high efficiency of EMS systems, we compare the two alternative policies. We make a decision to dispatch the BLS unit by choosing the policy that provides the better overall expected survival probability of life-threatening patients. If the first ambulance in the rank of ordered preference list is busy, the next one will be dispatched. In case of all BLS unit are busy, a call would be transferred to another dispatch center.

(ii) On-Scene: If patients require BLS care at on-scene priority2 calls, BLS serves and then returns back to the home station base. However, if patients require ALS care, judged by the BLS personnel, the BLS unit will provide the initial care, wait for the ALS unit to determine if patients need transportation to hospitals, and then head back to their original station. The dispatch of the ALS unit is assigned according to the available ALS units in rank of the ordered preference list for prirority2 calls. We refer to this as BLS-upgrade.

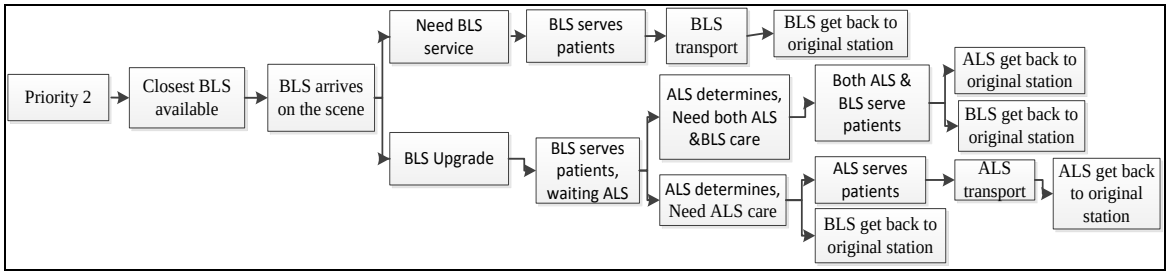


Figure 3.1: The EMS system process with BLS upgrade for priority2 calls

To dispatch different types of ambulances depends on the severity levels of calls. We introduced a specific dispatching preference list to each priority type and each type of required different type of ambulances. Table 3.1 shows the policies for dispatching ambulances for each call priority. Different performance measures are considered when we make a decision on how to dispatch the BLS unit for priority2 calls that should be treated like priority1 or 3 calls. The objective is to maximize the expected survival probability by comparing two alternative policies. We consider the expected survival probability of patients as a function of priority1 calls response time, as discussed in Chapter 2.

Table 3.1: Types of calls, types of ambulances and their corresponding dispatching policies

Types of call	Types of ambulances	
	ALS unit	BLS unit
priority1	closest policy	closest policy
priority2	closest policy	policy of priority1(closest) or priority3 (heuristic)
priority3	not needed	heuristic policy

3.3.1 Simulation Model

The simulation models are implemented using Arena Version14. The designed simulation models are then used to investigate the performance of a given dispatching

policy. The status of EMS systems is described by the state of each ambulance. The states could be: “idle” (at station base), “busy” (on the way to respond a call), or “busy” (serving and providing transportation a call). Table 3.2 shows the state space of EMS systems. We consider integer numbers to represent status of the ambulances. The definition of state space in this table is described in Chapter 2. We generate different modules to dispatch ambulances according to attributes of the calls (priority and location). When a call arrives on scene, we assign a call priority and a demand zone. In addition, the dispatch centers will decide which units to dispatch depending on call priorities. We then assign a status to dispatch ambulances according to the state of ambulances shown in Table 3.2. Double dispatch is when assigning a pair of dispatched ambulances. When the first ambulance arrives on the scene, we calculate the survival rate by using the response time of the first ambulance to priority1 calls. The survival probability is then calculated using equation (2.1) in Chapter 2. Considering a single dispatch for prority2 calls with status3, as the BLS arrives on-scene of accident, the state would be “waiting” for another unit with status4, if patients need ALS care. When the ALS unit arrives on scene, the status of both ambulances would be changed to “busy” (offering service to patients). After the ambulances provide service to the patients and go head to their original stations the state would become “idle” again. We investigate the better dispatch policy when the EMS systems reach the steady-state.

Table 3.2: The status of ambulances in EMS systems

Indicator	σ_j	Status of ambulance
$j \in [1,\ldots, J]$: ALS	0	Idle at base
	1	Double dispatch of ALS for priority1 calls
	2	Only ALS unit dispatch to respond to priority1 calls
	5	ALS unit dispatched to priority1 or 2 calls following a BLS unit which was sent when no ALS units were available
$j \in [J+1,\ldots, J+K]$:BLS	0	Idle at base
	1	Double dispatch of BLS for priority1 calls
	3	BLS unit dispatch to respond to priority2 or 3 calls
	4	Only BLS unit dispatch and waiting for ALS unit to respond to priority1 calls Waiting for ALS unit to respond to priority2 calls

The simulation models analyze different dispatching policies and evaluate patient survival probability. The simulation flow chart is described in Appendix A.2. When a priority2 call arrives to systems, we dispatch the BLS unit according to the dispatching policy like prirority1 or 3 calls. When the BLS unit arrives on the scene of an accident, dispatchers make a decision to upgrade or not. In case of no upgrade, BLS unit provides care to patient and then return back to home station. If BLS upgrade occurs, we will dispatch the ALS unit according to the policy where the closest ALS is always sent. In the case where all ALS units are busy, the BLS unit provides initial care and waits for the next available ALS unit. The simulation models assume they operate 24 hours per day. In this study, we investigate the better policy of dispatching BLS unit for priority2 calls, where we treat the policy of dispatching the BLS unit for priority2 calls like the policy for priority1 or 3 calls. The Process Analyzer in Arena Version14 is used to obtain the better policy. These simulators run 1800 replications per one simulation with half width of 0.0001 the 95% confidence interval around the survival probability. Each replication

runs 10 weeks to reach steady-state results. The performance of two alternative policies is compared to obtain the better policy.

3.4 Computational Results with Dispatching Policy of the BLS Unit for Priority2 Calls

In this section, we investigated the alternative policies by using collected real-world data at Hanover Fire and EMS department. The system operates 24 hour per day. The data set contains response time and transportation time from 4 stations to 12 demand zones seen in Table 3.3. The service times are shown in Table 3.4. We study the performance of systems in which the number of ALS and BLS units are fixed at three. They are located at different stations: ALS1 is located at Station4, ALS2 is located at Station1, and ALS3 is located at Station3. In addition, BLS4 is located at Station4, BLS5 is located at Station1, and BLS6 is located at Station1. There are three priorities where a proportion of call priorities depends on the demand zone.

Table 3.3: Response times (Lognormal distribution), transportation times and proportion of calls from each zones

Demand Zone	Call proportion	Station 1	Station 2	Station 3	Station 4
zone1	0.226034	(13.42,12.47)	(12.344,11.47)	(10.704,9.95)	(6.424,5.97)
zone2	0.019513	(25.71,23.89)	(25.712,23.89)	(15.896,14.77)	(25.712,23.89)
zone3	0.060281	(18.98,17.64)	(7.936,7.38)	(10.736,9.97)	(15.072,14.01)
zone4	0.043914	(21.01,19.52)	(25.712,23.89)	(20.856,19.38)	(12.312,11.44)
zone5	0.02657	(13.51,12.56)	(19.648,18.26)	(13.728,12.76)	(22.752,21.14)
zone6	0.09327	(8.06,7.48)	(13.056,12.13)	(25.712,23.89)	(12.472,11.59)
zone7	0.326744	(20.02,18.61)	(7.88,7.32)	(11.344,10.54)	(12.032,11.18)
zone8	0.065128	(15.06,13.99)	(25.712,23.89)	(10.992,10.21)	(20.632,19.17)
zone9	0.007525	(25.71,23.89)	(25.712,23.89)	(21.872,20.32)	(16.72,15.53)
zone10	0.077626	(10.08,9.36)	(15.696,14.59)	(11.704,10.87)	(10.16,9.44)
zone11	0.029886	(18.38,17.08)	(14.624,13.59)	(15.816,14.70)	(15.752,14.63)
zone12	0.023509	(25.71,23.89)	(14.904,13.85)	(25.712,23.89)	(15.776,14.66)

Table 3.4: Service times (Exponential distribution) and proportion of priority1, 2 and 3 calls

Demand Zone	Proportion of Priority1 calls	Proportion of Priority2 calls	Proportion of Priority3 calls	Service times	
				Priority1	Priority2,3
zone1	0.394	0.098	0.508	67.07	60.24
zone2	0.452	0.113	0.435	100.32	90.29
zone3	0.394	0.098	0.508	62.44	55.86
zone4	0.425	0.106	0.469	66.90	59.42
zone5	0.409	0.102	0.489	65.25	57.76
zone6	0.404	0.101	0.495	56.32	49.78
zone7	0.443	0.111	0.446	54.18	48.36
zone8	0.438	0.109	0.453	84.42	75.5
zone9	0.417	0.104	0.479	104.31	92.93
zone10	0.442	0.111	0.447	58.27	51.82
zone11	0.434	0.109	0.457	81.38	72.32
zone12	0.446	0.112	0.442	59.60	52.49

Regarding the improvements from Chapter 2, we fixed the closest policy for priority1 calls and the heuristic policy for priority3 calls. Note that we obtained the heuristic policy from the results in Chapter 2. We study the policy of priority2 calls that could be treated like priority1 or 3 calls by varying the percent of BLS upgrade for priority2 calls. Similar to a previous study, the objective is to maximize the patient survival probability. Table 3.4 showed the comparison of two alternative policies with the closest policy. In addition, Table 3.5 showed the "busy" probability of each ambulance given the different policies for priority2 calls. The underlines indicate the expected survival probability according to the closest policy, and bolded indicate the expected survival probability according to the better dispatching policy for each case. When the proportion of priority1 and 3 calls were close to balanced, the better policy for priority2 calls was to treat them like priority3 calls. However, when systems provided service for higher demand rate such as 1.25 calls per hour, the better policy for priority2 calls was treating them like priority1 calls.

Table 3.5: Comparison of two alternative policies and closest policy for priority2 calls with 30% upgrades.

ID	Arrival rate Calls / hr.	Policy Treat like	Resp. time P1 :mins	Resp. time P2 :mins	Resp. time P3 :mins	Percent of covered P1 (< 9 mins)	Percent of covered P2 (< 15 mins)	Percent of covered P3 (< 22 mins)	Percent of total coverage	Sur. Prob.	% Imp.	# of the imp. of lives saved /10,000 calls
		Closest	7.28	13.31	13.25	0.7373	0.7161	0.8449	0.7799	0.2545		
		Priority1	7.19	12.85	17.52	0.7425	0.7299	0.7517	0.7451	0.2576		
1	0.25	Priority3	7.17	17.55	17.49	0.7442	0.587	0.7524	0.7322	**0.2584**	1.532	39
		Closest	8.22	14.27	14.19	0.7046	0.6871	0.8273	0.7457	0.2354		
		Priority1	8.08	13.63	17.66	0.7142	0.7075	0.7488	0.7212	0.2398		
2	0.50	Priority3	8.03	17.67	17.61	0.7167	0.5807	0.7505	0.7119	**0.2411**	2.421	57
		Closest	9.72	14.79	14.72	0.6647	0.6724	0.8162	0.7018	0.2112		
		Priority1	9.56	14.28	16.88	0.6742	0.692	0.764	0.6905	0.2149		
3	0.75	Priority3	9.59	16.80	16.82	0.6743	0.6022	0.7648	0.6827	**0.2149**	1.752	37
		Closest	11.06	15.08	15.01	0.6273	0.6718	0.8118	0.6621	0.1919		
		Priority1	10.89	14.62	16.88	0.6379	0.6865	0.7616	0.6556	0.1958		
4	1.00	Priority3	10.82	16.98	16.87	0.6409	0.5966	0.763	0.6507	**0.1967**	2.501	48
		Closest	11.83	15.25	15.16	0.6001	0.6601	0.808	0.6369	0.1811		
		Priority1	11.70	14.90	16.56	0.6087	0.6749	0.7698	0.6327	**0.1842**		
5	1.25	Priority3	11.73	16.48	16.51	0.6077	0.6084	0.7717	0.6271	0.1837	1.712	31

In Figure 3.2 we investigated the two alternative policies and the closest policy. There were slight differences in performance of priority2 calls when we treated them like priority1 or 3 calls. In Table 3.4, the proportion of priority2 calls was very low when compared with priority1 and 3 calls. The results indicated that there was a slight impact on number of lies saved. When the percent of BLS upgrade was changed, the trends in the graphs showed no difference between upgrades at 20 and 30 percent.

Table 3.6 showed the comparison of the “busy” probabilities for two alternative policies and the closest policy. We observed that multiple unit dispatch for priority2 calls according to the heuristic policy could increase the patient survival probability as compared to the closest policy. We observed the “busy” probability of each ambulance in Table 3.6. The better policy for priority2 calls from changed from being treated like priority3 calls to being treated like priority1 calls when the busy probability of each

server was over 78 percent. When systems were full there was no difference between the closest policy and the heuristic policy.

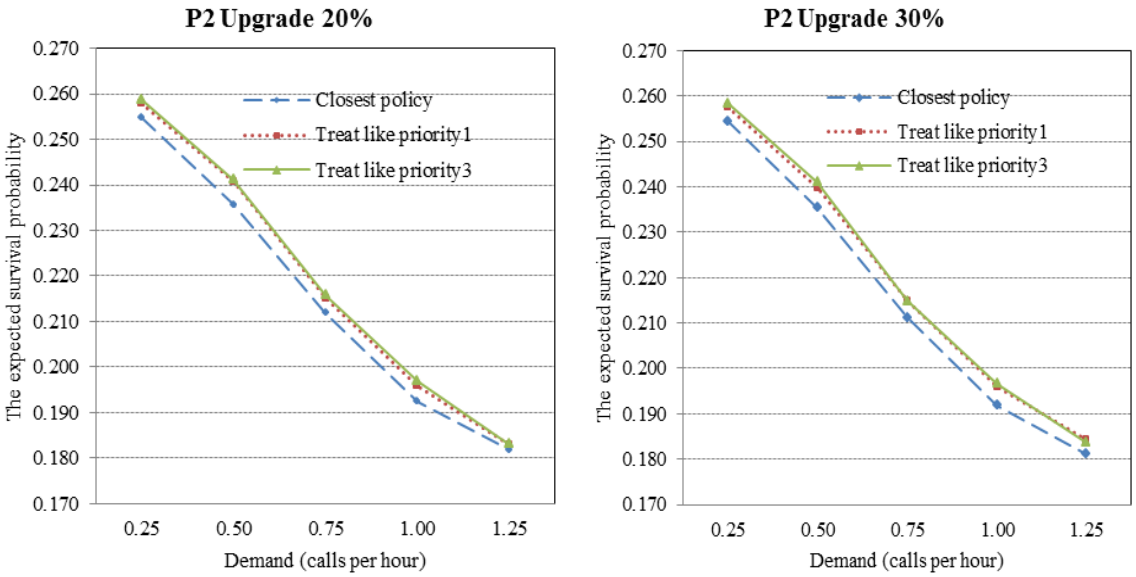


Figure 3.2: Comparison of the expected survival probability for two alternative policies and the closest policy

Table 3.6: Utilization of each ambulance under two alternative policies and the closest policy for priority2 calls with 30% upgrades

ID	Demand (calls /hour)	Policy Treat like	Utilization						Survival Prob.	% Imp.	# of the improvement of lives saved /10,000 calls
			ALS1	ALS2	ALS3	BLS4	BLS5	BLS6			
		Closest	14.57	3.44	6.36	22.94	12.08	2.778	0.2545		
		Priority1	14.47	3.49	6.24	15.25	13.06	10.74	0.2576		
1	0.25	Priority3	14.51	3.46	6.24	13.69	13.67	12.12	**0.2584**	1.532	39
		Closest	33.22	15.83	23.13	42.61	33.88	20.82	0.2354		
		Priority1	33.23	15.47	22.89	32.35	29.42	37.00	0.2398		
2	0.50	Priority3	32.82	15.15	21.99	29.66	28.90	39.13	**0.2411**	2.421	57
		Closest	58.29	42.13	51.21	65.33	62.39	52.69	0.2112		
		Priority1	57.46	42.55	50.87	59.34	61.81	59.24	0.2149		
3	0.75	Priority3	57.75	40.96	50.86	57.74	62.23	60.33	**0.2149**	1.752	37
		Closest	76.07	65.19	74.60	81.99	81.92	77.09	0.1919		
		Priority1	76.11	66.12	74.89	78.84	81.63	81.47	0.1958		
4	1.00	Priority3	75.76	65.16	73.59	77.50	81.15	81.23	**0.1967**	2.501	48
		Closest	84.93	78.12	85.43	89.82	90.65	87.88	0.1811		
		Priority1	84.62	78.91	85.49	88.59	90.63	89.70	**0.1842**		
5	1.25	Priority3	84.06	78.06	85.12	88.12	90.54	89.50	0.1837	1.712	31

3.5 Fairness and Efficiency of EMS Systems

Fairness is a crucial factor in deciding on how to dispatch ambulances. An important consideration of fairness arises when a serious call arrives to EMS systems, then the closest available ambulances are dispatched to respond. This makes the closest units unavailable to other patients. When a lower priority could be upgraded to a highest priority, the level of faster response needed between priorities is a critical issue in decision making. Dispatching far away ambulances to respond to the upgraded patients might increase the number of pre-hospital deaths for the lower patient priority. In this section, we consider fairness in patients waiting time for first response between priority1 and 2 calls. We focus on the mathematical formulation of constraints to balance fairness. The better dispatching policy on how to dispatch the ALS unit for priority2 calls is considered to maximize the expected patient survival probability. We modified the simulation model in the previous section by adding some constraints which indicate the ranked ordered preference lists of ALS unit for priority2 calls. The fairness measures, analyzed in the simulation models, added an equity constraint which is not a linear constraint.

The objective of multiple unit dispatch is to maximize the expected patient survival probability, as shown in Equation (3.1). This outcome is a response value that is obtained from running the simulation model. Equation (3.2) is the function of survival probability based on the work of Larsen et al. [48]. The response variable of fairness e represents the equity of waiting times for the first response between priroirty1 and 2 calls.

Equation (3.3) is a fairness constraint in which the difference of waiting time for the first response between priroirty1 and 2 calls is less than *e*. Equation (3.4) – (3.6) work together to indicate the rank ordered preference list of dispatching ALS unit for priority 2 calls. Equation (3.4) ensures that each ALS unit is assigned by exactly one rank order in the preference list for each priority and each demand zone. Equation (3.5) ensures that each rank of l^{th} order in the list is exactly one ALS unit for each priority and demand zone. The control variable x_{milj} is a binary variable that indicates whether an ALS unit *j* is the l^{th} rank order in the preference list in order to assign a priority *m* call and demand zone *i*. Equation (3.6) assigns the rank order preference list to simulation models where the control variable $ALSpolicy_{mil}$ would be assigned to a call. The ALS unit is dispatched according to attribute of a call.

The maximum expected patient survival probability with fairness model:

Maximize $f(x) = g(t_R)$ (3.1)

Subject to:

$$g(t_R) = \max(0.594 - 0.055t_R, 0) \quad (3.2)$$

$$wt_2 - wt_1 \le e \quad (3.3)$$

$$\sum_{l=1}^{J} x_{milj} = 1 \qquad \forall i = 1, 2,..., n \ \ \forall j = 1, 2,..., J \ \ \forall m = 1, 2 \quad (3.4)$$

$$\sum_{j=1}^{J} x_{milj} = 1 \qquad \forall i = 1, 2,..., n \ \ \forall j = 1, 2,..., J \ \ \forall m = 1, 2 \quad (3.5)$$

$$ALSpolicy_{mil} = \sum_{j=1}^{J} jx_{milj} \qquad \forall i = 1, 2,..., n \ \ \forall j = 1, 2,..., J \ \ \forall m = 1, 2 \quad (3.6)$$

Where:

x_{milj} = 1 if ALS unit *j* is assigned to priority *m* zone *i* in the l^{th} preferred server.

0 otherwise

$ALSpolicy_{mil} = j$ if ALS unit *j* is assigned to priority *m* zone *i* in the l^{th} preferred server.

wt_m average waiting time for priority *m* that waits for service from ALS unit.

$g(t_R)$ the survival probability of patients as function of response time for priority1 calls.

t_R the response time of first ambulance arriving on the scene for priority1 calls.

e the upper bound of deviation between waiting time for ALS unit for priority1 and 2 calls

n total number of demand zones

m indicator of priority as *m* = 1, 2, 3

l indicator of ranked in preference list as *l* = 1, 2, 3,.., *K*

i indicator of demand zone

j indicator of a ALS medical unit with known locations as *j = 1,2,…,J.*

k indicator of a BLS medical unit with known locations as *k = J+1,J+2,…,J+K.*

J number of ALS medical units

K number of BLS medical units

3.5.1 OptQuest for Simulation Model

OptQuest is a powerful tool in Arena Version14 which searches for the better dispatching solution for the simulation model. We find a solution which satisfies the constraints using OptQuest. OptQuest uses the outputs from running the simulation model to be response values which are inputted into an optimization model. The OptQuest adapts a Meta–heuristic method to find the better dispatching solutions by using the stopping criteria. The simulation models will then be stopped when stopping criteria is met. We use a stopping criteria based on improvement, specifically, we stop when the solutions do not improve within 100 simulations. The efficiency of finding the best solution depends on many factors. We use a closest dispatching policy for initial values of control variables. The initial values are located in the suggested values of control variables in OptQuest. OptQuest starts to search for the best solution by evaluating the initial values first.

3.6 Computational Results with Fairness Constraints

In this section, we implemented the simulation model with fairness constraints to real-world data. The data set from Hanover Fire and EMS department, Hanover County, Virginia was the same as the one implemented data set in Chapter 2. The city area was partitioned into twelve demand zones with four rescue stations. The illustration of the Hanover Fire and EMS department was shown in Figure 2.8. The 3 ALS units and 3 BLS units were randomly allocated to different four stations, as presented in Section 3.4.

Whereas we implemented the resulting policies from Chapter 2 for priority1 and 3 calls, we used the heuristic policy for priority2 calls for dispatching policy of the BLS unit at a call arrival rate of 1 call per hour. We considered the better dispatching policy of the ALS units for priority2 calls.

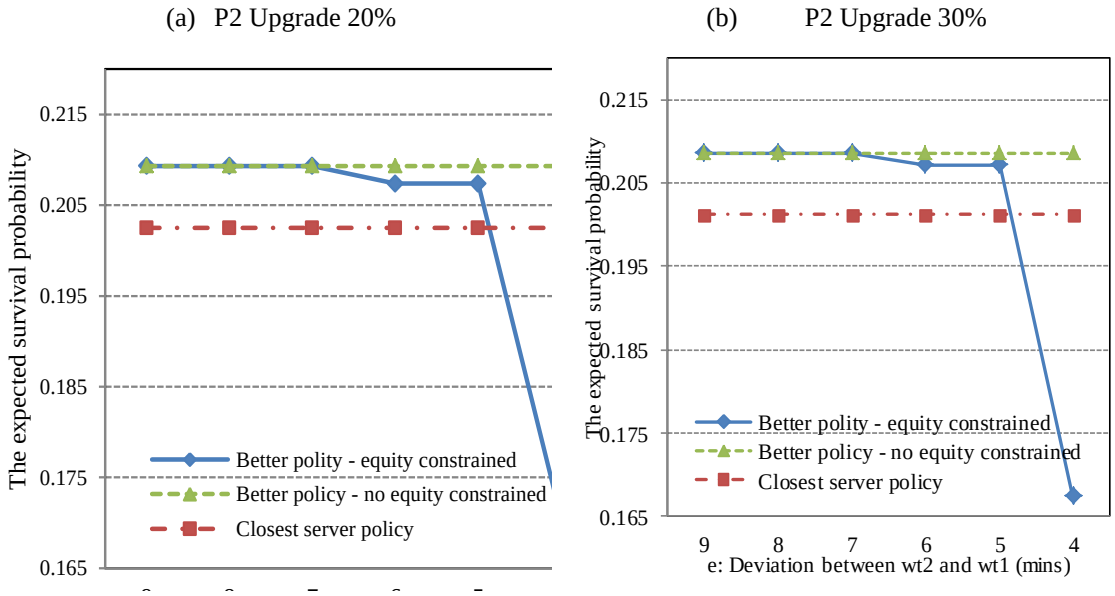


Figure 3.3: Comparison of the expected survival probability under the better dispatch policy with equity constraints and without equity constraints and the closest dispatching policy for priority2 calls

Figure 3.3 showed the results for 3 ALS units and 3 BLS units with upgrade priority2 calls to life-threatening at 20 and 30 percent. That is, we assumed that at on-scene of accidents the priority2 calls required ALS care 20 or 30 percent of the time. We studied the performance of EMS systems as we varied the allowable difference in waiting time of first response time between priority1 and 2 calls. We compared the results of using the better dispatch policy for priority2 calls with equity constraint and no equity constraint, and dispatch policy of always sending the closest ALS unit. We observed that dispatching ALS units according to the better dispatch policy with equity constraints

provided better outcome over the closest dispatch policy as the deviation of the first response time between priority1 and 2 calls changed within 5 to 6 minutes. When we forced the fairness constraints to 4 minutes, the results showed outcome lower than the closest dispatch policy. In addition, the results showed no difference between upgrade 20 and 30 percent for priority2 calls.

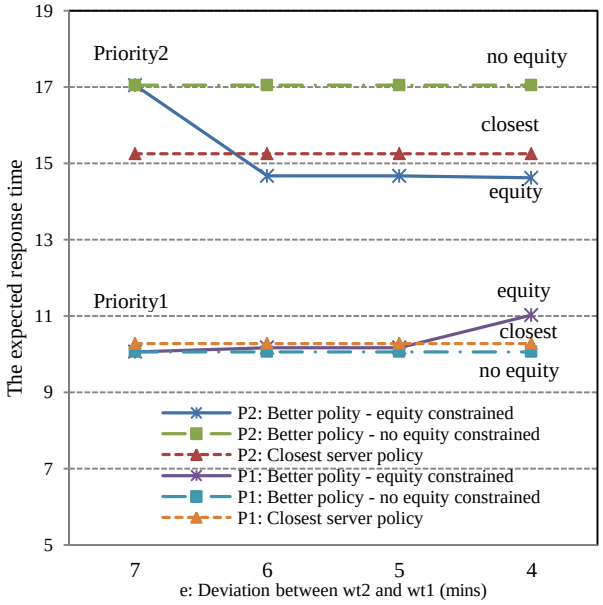


Figure 3.4: Comparison of the expected response time under the better dispatch policy with equity constraints and without equity constraints and the closest dispatching policy for priority2 calls with upgrade 20%

Figure 3.4 showed results of the expected response time for priority1 and 2 calls for 3 ALS units 3 BLS units with upgrade priority2 calls at 20 percent. We compared the results of using the better policy with equity constraints, the better policy without equity constraints and the closest dispatching policy. The results showed that the equity constraints provided better outcomes than no equity constraints and the closest dispatching policy in terms of the expected response time for priority2 calls. These

observations indicated that we could reach on-scene of accident for priority2 calls faster resulting in increasing the expected survival probability of priority2 calls. However, imposing equity constraints increased the expected response time resulting in decreasing the expected survival probability for priority1 calls but still offers improvement over the closest dispatching policy when we enforced the deviation of response time between priority1 and 2 calls at 5 and 6 minutes.

3.7 Conclusions and Future Research

In this chapter we analyzed a simulation model for multiple unit dispatch in EMS systems. We consider classifications of call priorities and two types of ambulances. The simulation model is formulated as a model given a particular dispatching policy. We consider the dispatching policy of BLS units based on possible situations that can be changed at on-scene of accidents for priority2 calls. We compare two alternative dispatching policies of BLS units for priority2 calls. Numerical results showed that the dispatching policy of the BLS unit for priority2 calls treated like priority3 calls (heuristic policy) provided improvement over the closest dispatching policy. When average busy probability of servers was over 80 percent, there was no difference between the heuristic and the closest dispatching policy for priority2 call.

We consider the simulation model with fairness by using the OptQuest. We implement the model using a real world example. The simulation model with a linear programming model is formulated for EMS system. The observations showed that the equity constraints decreased the expected survival probability but still offered

improvements over the closest dispatch policy for priority1 calls. The results also showed that the equity constraints decreased the expected response time, resulting in increasing the expected survival probability, for priority2 calls. The results suggested that imposing equity constraints often leads to an infeasible solution and that we should be careful in trying to enforce the deviation of first response time between priority1 and 2 calls while improving survivability of EMS systems. In future work, we will expand the dispatching model to consider the location of ambulances that lead to increasing the expected survival probability of EMS systems.

CHAPTER FOUR

A NESTED-COMPLIANCE TABLE POLICY FOR EMERGENCY MEDICAL SERVICE SYSTEMS UNDER RELOCATION

4.1 Introduction

The goal of emergency medical service (EMS) systems is to save the lives of emergency patients. The potential for improving performance of EMS systems is directly related to reducing response time, which is in turn related to increasing coverage. The decisions regarding ambulance location strategies can improve expected coverage. The ambulance location problem refers to the assignment of a limited number of ambulances to maximize coverage, given that the system has a fixed number of potential locations, and a demand zone is considered to be covered when an ambulance is located within a predetermined time standard. However, in reality the demand changes over time. Dynamic ambulance relocation can improve the performance of systems in situations with fluctuating demand. The growth and development of EMS systems literature shows a drastic increase in the percentage of dynamic strategies used. This is discussed in Alanis et al. [5]. The current analysis of relocation strategies deals with a compliance table. A compliance table refers to a particular table that shows the number of available ambulances in relation to the choices of open stations. That is, a compliance table shows where ambulances should be located when there are a certain number of ambulances available. Considering the example in Table 4.1 below; in this case, when only one ambulance is available, it is located at station A. In scenario 1, once a second ambulance becomes free, it will go to station C, and the first ambulance will stay at station A. In

scenario 2, once the second unit becomes available, ambulance 1 will need to relocate so that stations B and C can be open. Scenario 1 maintains what we refer to as a *nested* structure, which we will discuss later in further detail. One way to operationalize a relocation policy, which is not too computationally intensive, is via a compliance table policy. That is, vehicles will be relocated as calls come in and vehicles become available, but the policy is pre-determined and thus is easy to implement in real time. The challenge for EMS planners is that it may be difficult to identify the best compliance table.

Table 4.1: Sample compliance table

#of available units	Open stations Scenario 1	Open Stations Scenario 2
1	A	A
2	A, C	B, C
3	A, B, C	A, B, C

In this paper, we determine the best nested-compliance table for dynamic strategies in EMS systems. The compliance table policies in dynamic ambulance relocation include consideration for the real-time movement of idle ambulances to new locations. The decision to assign new locations to the available ambulances depends on the compliance table. To assess the best compliance tables, we consider only the possible compliance tables in a set of nested-compliance tables. Suppose we have K number of ambulances and v number of busy ambulances. A nested-compliance table refers to a compliance table in which the set of open stations when there are K-v-1 available ambulances is a subset of the open stations when there are K-v available ambulances, as shown in scenario 1 in Table 4.1. The original available ambulances from the previous state will be sent to their original home station, while a newly available ambulance will be given choice of stations. The benefit of nested policies is that only one ambulance,

which is already on the move, is relocated, avoiding unnecessary moves that can result in more accidents. We consider a single type of ambulance (paramedic units) and a single type of call priority when determining the best compliance table policy. We formulate an integer programming model to maximize the expected coverage with respect to the best compliance table. Real world data is used to validate the models.

In this study we:

- Modify a Markov chain model based on Alanis et al. [5] that considered the steady-state probabilities of EMS system in order to approximate coverage.
- Propose the nested-compliance table formulation as an integer programming model to determine the maximum expected coverage using a binary notion of coverage.
- Show, through the numerical results, how the solutions from our nested-compliance table formulation compare with a static (non-relocation) policy based on the adjusted maximum expected covering location problem (AMEXCLP) of Batta et al. [65] in real world problems.

This chapter is organized as follows. In Section 4.2 we review the related work on the nested-compliance table problem in EMS systems. Section 4.3 presents a description of EMS systems with relocation strategies and explains how to implement the nested-compliance table in EMS systems. Section 4.4 presents the application of a Markov chain model with relocation. Section 4.5 presents an integer programming approach to obtaining the optimal nested-compliance table with relocation. Section 4.6 presents the

efficiency of the nested-compliance table solutions. Finally, Section 4.7 presents conclusions and a discussion of future works.

4.2 Literature Review

The literature related to EMS vehicles is extensive. We limit our discussion to works related to ambulance location problems. We categorize models in terms of the decisions (e.g. location, relocation) being made, the objective (e.g. minimum number of servers, maximum coverage) function, and the methods (e.g. integer programming model, heuristic model, and Markov chain model) used.

In essence, the decision of a compliance table model deals with the ambulance location. One of the early works related to location decisions is the set covering problem, introduced by Hakimi [51]. The objective was to minimize the sum of distances between locations and nodes. The first mathematical formulation was developed by Toregas et al. [66] and Toregas and ReVelle [67]. The location set covering problem (LSCP) was to minimize number of vehicles which required to covering all demand nodes. The decision was where the resources were to be located in order to cover all demand nodes. The objective was to minimize the total number of resources. Aly and White [68] further developed the LSCP; they considered a random variable of location of call arrivals into model. In the LSCP, the goal is to minimize the number of resources; on the other hand, several facility location problems seek to minimize some cost with a fixed number of resources. Ingolfsson et al. [69] considered the location problem with random delay and

travel times. The objective was to minimize the number of ambulance so as to maintain a specified service level.

Other works considered the extensions of LSCP. The *p*-center and *p*-median problems are two common objectives in facility location problems, for reviews see (Tansel et al., [70] and [71], Krarup and Pruzan, [72]). The objective of the *p*-center problem is to minimize the maximum distance between nodes and their closest locations, while the objective of *p*-median is to minimize the total distance between nodes and their closest locations. Brandeau et al. [73] provided an overview of location problems. They focused on optimization problems such as p-center, p-median and other location problems. Other works considered a probability of set covering problem as probabilistic location set covering problem (PLSCP). ReVelle and Hogan [21] considered conditions when servers were busy during arrival of calls. They denoted α as reliable service, while the objective was to mini-max time between nodes and locations. Beraldi and Ruszczynski [74] developed PLSCP. The objective was to maximize the minimum reliable services. They considered the random binary ξ (0, 1). If ambulances were available at station bases, the random binary ξ was equally one. The results show that some discrete distributions provide an extremely large number of *p*, reliability level of available ambulances. Saxena et al. [75] considered the mixed integer programming for PLSCP. They use the PLSCP formulation of Beraldi and Ruszczynski [74], extending the model to improve the PLSCP. The random variable ξ could be decomposed into *L* blocks say $\{\xi^1,\ldots, \xi^L\}$ as ξ^t still being a 0-1 random for $t \in \{1\ldots L\}$.

In many other facility location problems, the objective is to maximize some notion of coverage, with a fixed number of resources. Church and ReVelle [18] proposed the maximal covering location problem (MCLP). This model assumed no "busy" ambulances in the systems. Daskin [19] introduced an extension of the MCLP, known as the Maximum Expected Coverage Location Problem (MEXCLP), by considering the "busy" probability for resources. This model assumed that all servers in the systems had the same "busy" probability. They used the binomial distribution to estimate the busy probability. ReVelle and Hogan [21] considered the Maximal Availability Location Problem (MALP). They assumed that server availability was independent of the number of servers. Batta et al. [65] relaxed some assumptions of the MEXCLP; such as, servers operating independently, and same busy probabilities for servers. This model referred to the Adjusted Maximum Expected Coverage Location Problem (AMEXCLP). The "busy" probability for servers was estimated using the hypercube queuing model. They presented the heuristic procedure for this adjusted model. The hypercube model was a model with multiple servers that described the queuing dynamics of systems. Larson [13] introduced the hypercube model to determine the busy probability according to a particular preference list for dispatch servers for each demand zone. Jarvis [15] generated the approximate workload of servers by using the hypercube model. Pirkul and Schilling [76] expanded the MCLP by incorporating capacities on servers and prioritized s for emergency calls. Marianov and ReVelle [22] developed the queuing maximal availability location problem (Q-MALP) based on ReVelle and Hogan [21] This model considered that the demand nodes were classified as covered when the probability of an ambulance

being available within time standard was at least α. The parameter α is referred to reliability level of available ambulances. Gendreau et al. [77] proposed a location problem using a double standard model (DSM). In the DSM all demand must be covered by ambulances located within r_1 time, and a proportion of demand α must be covered by ambulances located within r_2 time. They formulated the linear programming model, in which the objective was to maximize the total coverage of demand. A tabu search heuristic was developed for real world problems. Roberto et al. ([78] and [79]) considered the similarities and dissimilarities between the MEXCLP and the MALP, and presented an extension of the MALP, coined EMALP. Erkut et al. [80] also provided comparisons of existing maximum covering location models and developed an extension to the MEXCLP by considering the probability that demand node i was covered by the ambulance sited in j^{th} preferred station. Erkut et al. [81] also extended the Q-MALP by allowing for multiple servers at some stations and combined dispatch probability. The works discussed above did not consider relocation models. The relocation of emergency medical service (EMS) systems is the one possible strategy to increase coverage and improve patient outcomes. The use of relocation strategies could improve performance measures of EMS systems. Relocation strategies may be used to determine if available ambulances should relocate to better cover densely populated locations that have been left vulnerable by buy ambulances. In the relocation models dispatched ambulances might return back to a new station different from their originating station.

Early work of the relocation problem began with the formulation of an exact model in 1972. Kolesar and Walker [82] studied the dynamic relocation of fire resources.

Later, Gendreau et al. [83] proposed the dynamic double standard model (DDSM) which solves the repositioning problem based on the objective of the DSM. They considered a parallel tabu heuristic search to solve the DDSM in a reasonable computer running time. They developed a model of real time decisions on two levels; allocation problem and redeployment problem. The allocation problem determines which ambulance responds to a call using the closets dispatch. The redeployment problem relocates available ambulances to new locations to be better prepared to respond to future calls. They used a simulation model to evaluate the efficiency of the heuristic. Anderson and Varbrand [35] proposed the development decision tools for dispatching ambulances under dynamic ambulance relocation. The aim was to increase the preparedness for arrival of emergency calls. A tree search algorithm was used to find a solution for the dynamic relocation problem. In addition, Rajagopalan et al. [84] considered the covering location model for dynamic redeployment problem. The objective was to minimize the number of ambulances. They formulated the dynamic available coverage location (DACL) model under fluctuating demand and considered the "busy" probability of servers using Jarvis' [15] algorithm. The decision variable was the number of ambulances at each location at a certain time period. The large scale problems were then solved by using a search algorithm. A simulation model was used to validate the mathematical model. Recent work on the relocation problem focused on how to formulate the model for large scale problems. Majzoubi et al. [85] proposed the dispatching ambulance problem with relocating ambulances for EMS systems. Their model allowed for serving more than one patient per dispatch. The objective was to minimize the total costs to the EMS systems.

They formulated a non-linear program and developed a linear programming model approximation to obtain the solutions. Most previous works focused on integer programming and heuristic models. To maintain fidelity to the real problem of ambulance relocation problem, others use Markov Decision Process (MDP) models to analyze the EMS system

Berman [86] considered the repositioning of emergency units using an MDP formulation. The state of systems was represented by the status of each ambulance. The decision was to design where and when the EMS planners would move the servers to other locations from any possible state. Maxwell et al. [87] proposed an approximate dynamic programming model for redeployment of EMS systems. The objective was to maximize the number of covered calls. They formulated a state space that represented the status of each ambulance with two components; the first component was information on ambulances and second component was the number of waiting calls. The multiple dimensional states represented the status of ambulances; “idle” or “busy”, original locations, destination locations, and starting time of movement for each ambulance. Alanis et al. [5] analyzed Markov chain models of EMS systems to analyze their performance under repositioning. They presented a two-dimensional state space to represent the status of the EMS systems. The first component was the number of “busy” ambulances and the second component represented whether the system was in “compliance” or not. They validated the mathematical model using simulation models.

In this work we extend the relocation strategies proposed by Alanis et al. [5]. The modification determines the best of nested-compliance table with a single type of ambulance and a single type of call priority. Recent studies of Alanis et al. [5] determined the steady-state probabilities and estimated all service rates according to where exactly ambulances were located for any state of EMS system. However, our work differs in that we determine the steady-state probabilities and estimate all service rates independent of where exactly ambulances are located for any state of EMS system. We apply the output of steady-state probabilities based on Alanis et al. [5] as input parameters to our integer programming model. Our approximated formulation finds the nested-compliance table that maximizes the coverage of EMS systems.

4.3 EMS Systems with a Nested-Compliance Table Policy

In this section, we discuss EMS systems which are operated under relocation policies based on a nested-compliance table. We consider EMS systems with a single unit type and a single type of call. We assume that there is one ambulance located at each station. The EMS systems are a zero-queue system. When a call arrives at the dispatch center, the dispatch planners assign the closest ambulance in response to the call. In the case when all ambulances in the system are busy, the call will transfer to another dispatch center. As stated in Section 4.1, a compliance table is a nested-compliance table. Table 4.2 provides an example of a nested-compliance table. There are two events that could result in a repositioning move: call arrivals and call completions.

- Call arrivals: When a call from zone *i* arrives, the closest ambulance responds to call requiring service. If the closest is busy, the second closest responds to the call and so on.
- Relocation via call arrivals: When the number of busy ambulances changes, the dispatchers consider which ambulance to move (if any) to be better pre-positioned, since the dispatched ambulance may have left some critical areas uncovered. For example, based on Table 4.2, suppose a call arrives to a system with zero busy ambulances, and the ambulance in station 2 responds to the call. The system state changes from zero busy ambulances to one busy ambulance. The located ambulance in station 12 moves to replace the ambulance at station 2 in the new system state.
- Service time: We define the service time as the time between the EMS staff arriving on-scene and completing service, including providing transportation to a hospital if needed.
- Relocation via call completions: After the EMS staff has completed service to patients the ambulance may return to any open station, not necessarily its previous station. The dispatchers consider which station the now available ambulance should be located to. The system state changes to decrease number of busy ambulances. During this time the ambulance is free, but cannot be assigned to a new call. If a call arrives, it will transfer to the next closest ambulance. For example, in Table 4.2, while the system state is two busy ambulances, the dispatched ambulance travels back to station11. The system state changes from two to one busy ambulance.

Other important times in the EMS system include Response time and travel time between stations.

- Response time: The travel time between the stations of a dispatched ambulance to the scene of the incident.
- Travel time between stations: The travel time between the original stations to the new stations when system states are changed.

Table 4.2: The nested-compliance table

# of busy servers	Stations															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0		1	1	1	1	1		1	1		1	1				1
1		1	1	1	1	1		1	1		1					1
2		1	1	1	1	1		1	1							1
3		1	1		1	1		1	1							1
4		1	1			1		1	1							1
5		1	1			1			1							1
6		1				1			1							1
7		1				1			1							
8		1							1							
9									1							

4.4 The Application of a Markov Chain Model with Relocation for EMS Systems

To assess the performance of a specific compliance table in terms of coverage, we need to approximate the steady-state probabilities of an EMS system. To approximate the steady-state probabilities, we build upon a Markov chain model with relocation developed by Alanis et al. [5]. They formulated the model as a finite, continuous time Markov chain according to a compliance table policy. The state variable $V(t)$ denoted the number of busy ambulances at time t, and the state variable $C(t)$ denoted the status of the EMS system, whether the system was in compliance or out of compliance at time t. The

state space of $V(t)$ was given by the set $V(t) = (0, 1, 2, \ldots, K)$ and $C(t)$ was given by the set $C(t) = (0, 1)$. The $C(t) = 0$ indicated when the system was out of compliance and $C(t) = 1$ indicated when the system was in compliance. In compliance referred to all available ambulances being at their assigned stations. On the other hand, out of compliance referred to the status that not all available ambulances are at their assigned stations; that is, an ambulance being en-route to its home station. They assumed that the arrival process of calls to EMS systems was Poisson and all service times were exponentially distributed. They assumed a zero-queue system. When calls arrived to system when all ambulances were busy, they would transfer to another system. They assumed that the system reached out of compliance when a call arrived to systems or an ambulance completed service at on-scene. In work of Alanis et al. [5], they considered the relocation model where the state transitions occurred due to one of three event types, call arrival, call completion and a moving ambulance reaching compliance. Suppose we were in state (v, 1), in compliance with v busy ambulances. The transition to reach out of compliance state (v+1, 0) occurred via a call arrival, where λ was call arrival rate. The transition to reach out of compliance state (v-1, 0) occurred via a call completion with rate $v\mu_1$, where μ_1 was the completion rate given that the system was in compliance state. Similarly, suppose we were in state (v, 0), out of compliance with v busy ambulances. The transition to reach out of compliance state (v+1, 0) occurred via call arrival. The transition to reach out of compliance state (v-1, 0) occurred via a call completion with rate $v\mu_{v,0}$ where $\mu_{v,0}$ was the call completion rate given that system was out of compliance state. When the system was out of compliance in state (v, 0), the transition rate γ resulted in a transition to

state (v, 1), in compliance state. While the ambulance moved to new home station, the ambulance could not be dispatched to respond to a new call. When a call arrived to system during this time, it would transfer to the next closest ambulance. The details of these transitions were explained in Alanis et al. [5]. Table 4.3 shows the notation of the parameters of the nested-compliance table model under relocation.

In this paper, we formulate a nested-compliance table model under the same assumptions as those studied in Alanis et al. [5]. However, *our work differs in that we approximate the transition rates not according to the exact nested-compliance table policy, but rather based only on the number of busy ambulances.* If the approximation of transition rates is known and not according to an exact nested-compliance table policy, the nested-compliance table model can be solved as an integer programming model. Otherwise, we have to consider a meta-heuristic or enumerate all solutions which require long computational running time. When estimating the transition rates, we relax the assumption of approximating parameters given by Alanis et al. [5] such that our approximation of transition rate γ, μ_1 and $\mu_{v,0}$ are independent of the exact nested-compliance table. The approximations of transition rates are calculated based on the total covered arrival intensity for each station, as discussed in Section 4.4.1. Figure 4.1 illustrates our modification of the transition diagram of Markov chain model with relocation based on Alanis et al. [5] for K= 5, number of ambulances.

We describe the process related to situations of the EMS system with the state-transition network in Figure 4.1. The system starts with all idle ambulances at assigned

stations in state (0, 1). As a call arrives, the state reaches out of compliance immediately resulting in increasing the number of busy ambulances to 1, a transition to state (1, 0) and potentially relocating an ambulance to replace at station of the dispatched ambulance. Suppose no new call arrives in the meantime and we relocate the ambulance completely before the dispatched ambulance finishes at on-scene of accident, the system reaches compliance at rate γ, resulting in a transition to state (1, 1). However, in case we could not relocate the ambulance to replace at station of the dispatched ambulance completely resulting in system state still out of compliance state while the ambulance completes to reach at on-scene of accident and completely provides service to patients with call completion rate $\mu_{1,0}$ resulting in decreasing number of busy ambulances and a transition to state (0, 0). Similarly when system is in state (1, 1) and the ambulance completely reaches on-scene of accident and completely provides service to patients resulting in transition to state (0, 0). After the call completion, the ambulance travels back a home station (possible new home station) with relocation rate γ resulting in a transition to state (0, 1). As another possible situation, suppose a new call arrives when in state (1, 0) or (1, 1) both result in out of compliance immediately and a transition to state (2, 0). Therefore, we need to relocate an ambulance to new home station in order to achieve compliance state.

Table 4.3: The parameters of the nested-compliance table model under relocation

Notation	Description
i	indicator of demand zone
m	indicator of station
v	indicator of the state of EMS system-- number of busy servers
c	indicator of the status of EMS system = 0 system is out of compliance = 1 system is in compliance
λ	call arrival rate
λ_i	call arrival rate from demand zone i
K	total number of ambulances in the EMS system
M	total number of stations
γ	the rates at which compliance is reached
$\mu_{v,0}$	the service rate or call completion rate at which each individual busy ambulance completes its call, given that system is out of compliance
μ_1	the service rate or call completion rate at which each individual busy ambulance complete its call, given that system is in compliance
$\pi_{v,0}$	the steady-state probability that the system is out of compliance and in state v
$\pi_{v,1}$	the steady-state probability that the system is in compliance and in state v

The notations for the approximation of parameters

fd_m	the total covered arrival intensity for station m.
$Pr(A_m)$	the probability of covered arrival intensity for station m.
r_{im}	the response time from station m to demand zone i.
t_{jm}	the travel time from station j to station m.
d_{im}	the travel time between station m and demand zone i
a_{im}	indicator of ambulance at station m can respond to demand zone i within specified response time RTT the specified response time thresholds (RTTs) a_{im} = 0 if d_{im} > RTT a server at station m does not cover demand zone i = 1 if $d_{im} \leq$ RTT a server at station m covers demand zone i
T_m	the mean travel time between any station to station m
α_v	the rates of call arrival into state (v, 0).
β_v	the rates of call completion into state (v, 0).
$\tau_{0,\,arrival}$	the mean service time to enter the state (v, 0) via a call arrival
$\tau_{0,\,completion}$	the mean service time to enter the state (v, 0) via a call completion
$\tau_{0,arrival,i}$	the composition of the expected travel time entering state (v,0) via a call arrival from any station to demand zone i and the expected service time at on-scene of accident
$\tau_{0,completion,i}$	the composition of the expected travel time entering state (v,0) via a call completion from demand zone i to any station and the expected service time at on-scene of accident
$E[S_{i,\,on\text{-}scene}]$	estimated from empirical data which is the composition of the service time on-scene and the time to transport patients to hospital if needed.
$E[R_i]$	the mean response time of any station to demand zone i.
$E[S_{0,Travel,i}\|(v, 0)$ entered via a call arrival$]$	the expected travel time entering state (v,0) via a call arrival from any station to demand zone i.
$E[S_{0,Travel,i}\|(v, 0)$ entered via a call completion$]$	the expected travel time entering state (v,0) via a call completion from demand zone i to any station

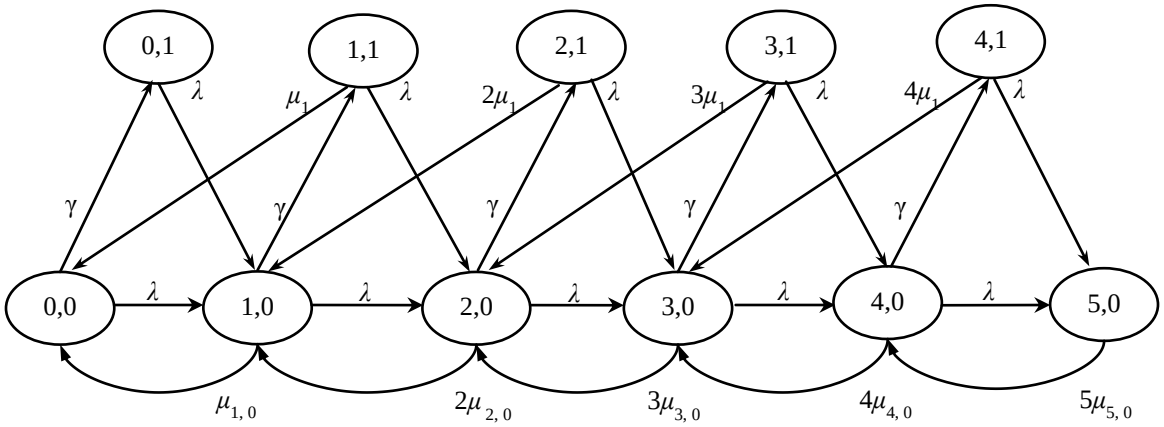


Figure 4.1: The modified state transition of EMS systems with relocation, based on Alanis et al. [5]

The Markov chain model with relocation was applied to approximate the steady-state probabilities $\pi_{v,c}$ based on the model given by Alanis et al. [5]. They formulated the flow balance equations for state (v, 1) and (v, 0) that were given by

$$\pi_{v,1} = \frac{\gamma}{\lambda + v\mu_1}\pi_{v,0} \qquad \textit{for} \quad v = 0, 1, 2, \ldots, K\text{-}1 \qquad (4.1)$$

$$\pi_{v,0} = \frac{(v+1)(\lambda + v\mu_1)(\mu_{v+1,0}\pi_{v+1,0} + \mu_1\pi_{v+1,1})}{\lambda(\lambda + \gamma + v\mu_1)} \qquad \textit{for} \quad v = 0, 1, 2, \ldots, K\text{-}1 \qquad (4.2)$$

In order to compute the steady-state probability, Alanis et al. [5] used the recursive method by starting with state (K, 0) and the normalization method in the last step so that the sum of the steady-state probabilities equaled to one. The steady-state probabilities could be obtained by using the following recursive algorithm which presented in Alanis et al. [5].

In order to determine the steady-state probability, we need to approximate the completion rate $\mu_1$and the rate to reach compliance, γ, which will be shown in Section

4.4.1. In this section, we consider how to calculate the completion rate $\mu_{v,0}$. Alanis et al. [5] discussed the rate $\mu_{v,0}$ which depended on a particular compliance table. Our work differs in that we use the total arrival intensity to weigh service time. However, Alanis et al. [5] considered the two possible situations to reach out of compliance state (v,0) via a call arrival and a call completion. They defined two parameters α_v and β_v as the rates of call arrival and call completion into state (v, 0). The rate $\mu_{v,0}$ were obtained by weighing these two parameters. The rate α_v and β_v are shown in equation (4.3) and (4.4). The $\tau_{0,\,arrival}$ is the service time to enter the state (v, 0) via a call arrival. The $\tau_{0,\,completion}$ is the service time to enter the state (v, 0) via a call completion. Alanis et al. [5] estimated the $\tau_{0,\,arrival}$ and $\tau_{0,\,completion}$ depending on system state v. However, we estimate $\tau_{0,arrival}$ and $\tau_{0,\,completion}$ do independent of the system state v. The estimation of $\tau_{0,\,arrival}$ and $\tau_{0,\,completion}$ are discussed in Section 4.4.1.2. We modify the algorithm *A* based on the work in Alanis et al. [5] to determine the service rate $\mu_{v,0}$ in that not according to the exact nested-compliance table policy. The service rate $\mu_{v,0}$ can be obtained by using the iterative algorithm *A* as following.

$$\alpha_v = \lambda(\pi_{v-1,0} + \pi_{v-1,1}) \tag{4.3}$$

$$\beta_v = (v+1)(\mu_{v+1,0}\pi_{v+1,0} + \mu_1\pi_{v+1,1}) \tag{4.4}$$

$$\mu_{v,0} = \frac{\alpha_v + \beta_v}{\alpha_v\tau_{0,arrival} + \beta_v\tau_{0,completion}} \tag{4.5}$$

where $\tau_{0,\,arrival}$: E[S_0| any state (v, 0) entered via a call arrival]

$\tau_{0,\,completion}$: E[S_0| any state (v, 0) entered via a call completion]

Algorithm A (Modification based on Alanis et al. [5])

Set for all $\pi_{v,c} = 1/(2K+1)$

Step1: Set $\mu_{K,0} = 1/\tau_{0,\text{arrival}}$

Step2: Compute α_K by using equation (4.3)

Step3: Decrease v from K to K-1, using equation (4.4) to compute β_{K-1}

Step4: Use equation (4.5) to obtain the value of the rate $\mu_{K-1,0}$

Step5: Compute α_{K-1} by using equation (4.3)

Step6: Decrement v from K-1 to 0 in step of 1, using equation (4.4) to obtain the value of β_v, using equation (4.5) to obtain the value of the rate $\mu_{v,0}$ and using equation (4.3) to obtain the value of α_v

4.4.1 Parameters Approximating for the Markov Chain Model

The purpose of this section is to describe of how we calculate the service time and the travel time between stations (relocation time) of EMS systems. The Markov chain model requires estimating average service times and average travel time between stations as input parameters. In Alanis et al. [5] defined the average travel time between stations as the rates at which in compliance states were reached, γ. The γ depended on exactly where the available ambulances were located and depended on the system state v. However, our model is different in that we estimate the average travel time between stations by using the covered arrival intensity for each station in order to weigh located ambulances to stations. In addition, the Markov chain model with relocation of Alanis et al. [5] assumed the transition states occurred at rate $v\mu_{v,0}$ when in state (v, 0) and rate $v\mu_1$

when in state (v, 1). In work of Alanis et al. [5], the service rate $\mu_{v,0}$ depended on exactly where the available ambulances were located. However, our model considers the covered arrival intensity for each station to estimate the probabilities where stations will be assigned the available ambulances to. The transition state of entering state (v, 0) occurs via a call arrival and a call completion. Therefore, we have to compute the expected service time entered state (v, 0) via a call arrival and the expected service time entered state (v, 0) via a call completion. However, the service rate μ_1 is the average rate of call completion, from arrival of a call to service completion.

4.4.1.1 Approximating Relocation Time between Stations

For the average travel time between stations (relocation time), we consider the rate γ which does not depend on where the available ambulances are located exactly. We estimate the probability of which available ambulances are located to station m by using total covered arrival intensity for station m. Suppose that $\lambda_1, \lambda_2, \ldots, \lambda_n$ are proportion of call arrivals from demand zones1 through n. The fd_m refers to the total covered arrival intensity for station m. The $Pr(A_m)$ refers to the probability of covered arrival intensity for station m.

$$\Pr(A_m) = \frac{fd_m}{\sum_{m=1}^{M} fd_m} \qquad for \quad m = 1, 2, \ldots, M \qquad (4.6)$$

$$fd_m = \sum_{i=1}^{n} a_{im} \cdot \lambda_i \qquad for \quad m = 1, 2, \ldots, M \qquad (4.7)$$

a_{im} = 0 if d_{im} > RTT a server at station m does not cover demand zone i

1 if $d_{im} \leq$ RTT a server at station m covers demand zone i

d_{im} the travel time between station m and demand zone i

Estimating the γ, the travel rate per hour is simply calculated by the mean travel time between stations of M stations. The T_m refers to the mean travel time between any station to station m. The t_{jm} refers to the travel time from station j to station m. The γ is obtained by using equation (4.8) and (4.9)

$$\gamma = 60 / \frac{\sum_{m=1}^{M} T_m}{M} \tag{4.8}$$

$$T_m = \sum_{j=1}^{M} \Pr(A_j) \cdot t_{jm} \tag{4.9}$$

4.4.1.2 Approximating Service Time

The Markov chain model with relocation requires to estimate the service times where the system enters out of compliance via a call arrival, $\tau_{0,arrival}$, and system enters out of compliance via a call completion, $\tau_{0,completion}$, and service rate per hour where the system enters in compliance, μ_1. The parameter estimation for Markov chain model is modified to approximate the average service times based on the approximation given by Alanis et al. [5]. They estimated the services time and service rate $\tau_{0,arrival}$, $\tau_{0,completion}$ and μ_1 depending on exact location configurations of compliance table being in any the $K - v$ available ambulance states. However, our approximations differ in that we estimate these service times and service rate does not depend on the location configurations of

compliance table. Our estimating $\tau_{0,arrival}$, $\tau_{0,completion}$ and μ_1 depend on the covered arrival intensity of each station. We consider the covered arrival intensity to each station in order to weigh average response time from any station to a demand zone. We assume that the service time at on-scene and the time of transportation to hospital is the same for all demand zones. The service rate at which the system enters in compliance, μ_1 is simply the arithmetic mean of the *n* demand zones for total service time of the expected response time and the expected service time at on-scene. It is straightforward to estimate the expected response time to demand zone *i*, $E[R_i]$ from the mean response time of the *M* stations to demand zone *i*. The expected service time $E[S_{i,\ on\text{-}scene}]$ is estimated from empirical data that is the composition of the service time on-scene and the time to transport patients to hospital if needed. We describe how we estimate μ_1 in Equation (4.10) and (4.11).

$$\mu_1 = 60 / \frac{\sum_{i=1}^{n}\left(E[R_i] + E[S_{i,on-scene}]\right)}{n} \tag{4.10}$$

$$E[R_i] = \frac{\sum_{m=1}^{M} r_{im}}{M} \tag{4.11}$$

Where r_{im}: response time from station *m* to demand zone *i*.

The service times, where the system enters out of compliance via a call arrival, $\tau_{0,arrival}$ is simply the arithmetic mean service time of the *n* demand zones. We estimate the service time entered out of compliance via call arrival corresponding to demand zone *i*, $\tau_{0,arrival,i}$ from the composition of the expected service time entered out of compliance

via call arrival, $E[S_{0,Travel,i}|(v, 0)$ entered via a call arrival] and the expected service time at on-scene of accident, $E[S_{i,\ on\text{-}scene}]$ corresponding to demand zone *i*. For the purpose of estimating the $E[S_{0,Travel,i}|(v, 0)$ entered via a call arrival], Alanis et al. [5] assumed a known the configurations of ambulance locations specified in the compliance table. However, our model differs in that we estimate the $E[S_{0,Travel,i}|(v, 0)$ entered via a call arrival] by using the probability of covered arrival intensity for station *m*. The $Pr(A_m)$ is used in order to weigh average travel time from any station *m* to demand zone *i*. The r_{im} refers to response time from station m to demand zone *i*. In using equation (4.12) - (4.14), we obtains the $\tau_{0,arrival}$.

$$\tau_{0,arrival} = \frac{\sum_{i=1}^{n} \tau_{0,arrival,i}}{n} \tag{4.12}$$

$$\tau_{0,arrival,i} = E\left[S_{0,Travel,i} \,|\, (v,0) \text{ entered via a call arrival} \right] + E[S_{i,on-scene}] \tag{4.13}$$

$$E\left[S_{0,Travel,i} \,|\, (v,0) \text{ entered via a call arrival} \right] = \sum_{m=1}^{M} \Pr(A_m) \cdot r_{im} \tag{4.14}$$

Similarly, to estimate the $\tau_{0,completion}$, we use the probability of covered arrival intensity for station *m*, $Pr(A_m)$ in order to weight average travel time from a demand zone *i* to station. The service time where system entered out of compliance via a call completion, $\tau_{0,completion}$ is also the mean travel time of the *n* demand zone. We estimate the service time entered out of compliance via call completion corresponding to demand zone *i*, $\tau_{0,completion,i}$ from the composition of the expected service time entered out of compliance via call complete, $E[S_{0,Travel,i}|(v, 0)$ entered via a call completion] and the expected

service time at on-scene of accident, $E[S_{i,\ on\text{-}scene}]$ corresponding to demand zone i. The $E[S_{0,Travel,i}|(v, 0)$ entered via a call completion] is composed of the mean travel time from demand zone to ambulance station and the mean travel time between stations. The $T_{,m}$ refers to the mean travel time between any station to station m by using equation (4.9). The r_{im} refers to response time from station m to demand zone i. We use equation (4.15) – (4.17) for estimating $\tau_{0,completion}$.

$$\tau_{0,completion} = \frac{\sum_{i=1}^{n} \tau_{0,completion,i}}{n} \tag{4.15}$$

$$\tau_{0,completion,i} = E\left[S_{0,Completion,i} \,|\, (v,0) \text{ entered via a call completion} \right] + E[S_{i,on-scene}] \tag{4.16}$$

$$E\left[S_{0,Travel,i} \,|\, (v,0) \text{ entered via a call completion} \right] = \sum_{m=1}^{M} \Pr(A_m) \cdot r_{im} + \sum_{m=1}^{M} \Pr(A_m) \cdot T_m \tag{4.17}$$

4.5 The Formulation of the Nested-Compliance Table Model

The Markov chain model with relocation was a powerful tool that could be used to approximate the steady-state probability of systems based on Alanis et al. [5]. This model provided the approximation of performance measure of EMS systems as well, such as response time distribution for a given compliance table policy (knowing exactly where ambulances are located for each state v). Therefore, if we knew the distribution of response time, we could estimate the expected coverage. However, the exact expected coverage cannot practically be used in an optimization framework. The expression will be a non-linear formulation. In this paper, we develop the nested-compliance table model

given the output of the Markov chain model with relocation, such as steady-state probabilities. The steady-state probabilities can be approximated independent of the exact compliance table policy. Consequently, the steady-state probabilities will be input parameters to the nested-compliance table formulation. The objective is to determine the maximum expected coverage using a binary notion of coverage. The covered calls refer to the calls in which an ambulance from stations can respond to the call within a specified amount of time. We calculate the expected coverage not considering the variability of response time. Figure 4.2 shows the flow process of the compliance table model.

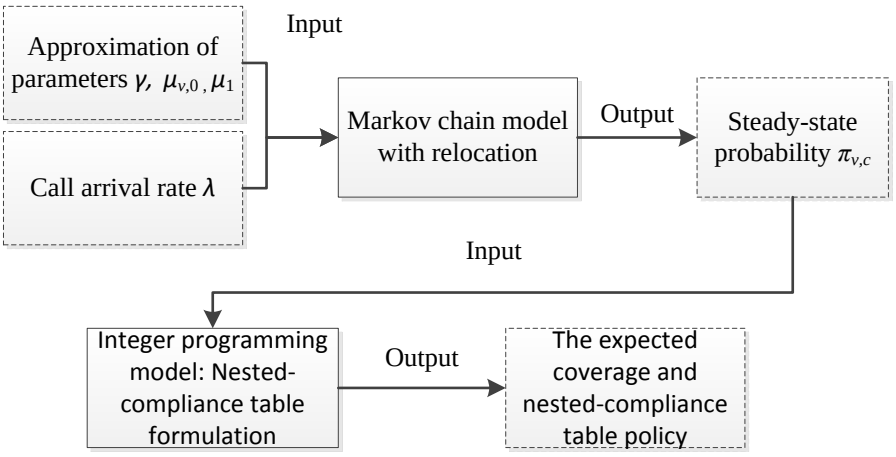


Figure 4.2: The process flow of the nested-compliance table model

In this section, we formulate the nested-compliance table model for the ambulance relocation problem. We consider a single type of patient calls and a single type of ambulances (paramedic units). The nested-compliance table model under relocation problem is introduced as an integer programming model. A Markov chain model is applied to approximate the steady-state probabilities when system state is in compliance $\pi_{v,1}$ and out of compliance $\pi_{v,0}$ for each state v (number of busy ambulances)

based on our modification of Alanis et al. [5].We consider the nested-compliance table model with n demand zones, K ambulance units and M ambulance stations. We assume that the EMS system operates with a relocation policy according to a nested-compliance table. We assume that one ambulance is located in each station for each state of system. When a call arrives to system, the closest ambulance responds to a call. If the first closest ambulance is busy, the second will respond to a call and so on. We assume that the EMS system operates as a zero-queue system. The model is formulated as an integer linear programming model with the approximate steady-state probabilities serving as an inputs to the model. We define the decision variable x_{mv} as a binary variable. If $x_{mv} = 1$, it indicates that an ambulance is located at station m when the system is in state v. The decision variable $y_{i,v}$ is binary variable also. If $y_{iv} = 1$, it indicates zone i is covered when the system is in state v, For each demand zone, we define M_i as the set of ambulance stations that can respond to calls from demand zone i within a specific time. We use the following notation:

Indices

$i = 1, 2, \ldots, n$	demand zone
$m = 1, 2, \ldots, M$	ambulance station
$v = 0, 1, 2, \ldots, K\text{-}1$	state of EMS system (number of busy vehicles)

Parameters

λ	call arrival rate
n	total number of demand zones
K	number of paramedic units

M total number of ambulance stations

M_i set of locations that can respond to calls at demand zone i within the specific time

λ_i call arrival rate from demand zone i, such that

$$\sum_{i=1}^{n} \lambda_i = \lambda$$

$\pi_{v,0}$ the steady-state probability that the system is out of compliance when in state v (number of available servers is K-v)

$\pi_{v,1}$ the steady-state probability that the system is in of compliance when in state v (number of available servers is K-v)

Decision Variables

x_{mv} = 1 if an ambulance is located to station m when system being in state v (number of available servers is K-v)

= 0 otherwise

y_{iv} = 1 if demand zone i is covered when system being in state v (number of available servers is K-v), if all vehicles are at their assigned locations

= 0 otherwise

Objective function:

$$Maximize \quad \sum_{i=1}^{n} \sum_{v=0}^{K-1} (\lambda_i / \lambda) \cdot \left(\pi_{v,1} + \left(\frac{K-v-1}{K-v} \right) \cdot \pi_{v,0} \right) \cdot y_{iv} \qquad (4.18)$$

Subject to

$$\sum_{m=1}^{M} x_{mv} = K - v \qquad for \quad v = 0, 1, 2, \ldots, K\text{-}1 \qquad (4.19)$$

$$y_{iv} \leq \sum_{m \in M_i} x_{mv} \qquad for \quad i = 1, 2, \ldots, n \ for \quad v = 0, 1, 2, \ldots, K\text{-}1 \qquad (4.20)$$

$$x_{m,v-1} \geq x_{mv} \qquad for \quad m = 1, 2, \ldots, M \ for \quad v = 1, 2, 3, \ldots, K\text{-}1 \qquad (4.21)$$

$x_{mv} \in \{0, 1\}$ $y_{iv} \in \{0, 1\}$

The maximum expected coverage of the nested-compliance table model under relocation is introduced as an integer programming model. The objective function is to maximize the demand that is covered as shown in equation (4.18). The equation consists of products of the decision variable $y_{i,v}$ and the probability of covering call zone i when the system is in state v, and the proportion of calls from demand zone i. The parameter $\pi_{v,1}$ indicates the probability that all available ambulances are at their assigned stations (the system is in compliance). Therefore, if a call from demand zone i arrives, all available ambulances K-v are at their assigned stations, thus we know which demand zones are covered directly from $y_{i,v}$. On the other hand, the system will be out of compliance in state v with probability $\pi_{v,0}$. We do not know of which ambulance is not at its located station. We assume that it is equally likely that one ambulance from K-v available ambulances is not available at its located station. The term (K-v-1)/(K-v) indicates the likelihood of K-v-1 available ambulances being in their stations and one ambulance being in en-route to its home station when the system is in state v. The constraint (4.19) ensures that we allocate the number of ambulances equal to the number of available ambulances K-v for each state of the EMS system. Constraint (4.20) indicates that the demand zone i is covered when at least one ambulance is located in a station in set M_i at each state v. M_i is the set of locations that can cover demand zone i. Constraint (4.21) ensures that the optimal solution is in the set of the nested-compliance table solutions. The integer programming model requires us to approximate the steady-state probabilities of the system being out of compliance $\pi_{v,0}$ and of the system being in

compliance $\pi_{v,1}$ for each state v. We discussed how to calculate these steady-state probabilities in Section 4.4.

4.6 The Efficiency of Nested-Compliance Table Model under Relocation

In this section, we present the results of our model applied to real-world data. The data was collected from Hanover Fire and EMS department, Hanover County, Virginia. The data consisted of approximately 12,000 calls per hour. The city covered is an area of about 474 square miles with 122 demand zones and has a population of about 100,000. We considered an EMS system with varied number of ambulances from 6 to 10 ambulances, 16 station bases. The EMS system operated 24 hours per day. The data set matched our assumption that call arrivals were Poisson during peak times, with a mean arrival rate of 1.5 calls per hour. The data set includes the distances between stations and demand zones. We assume that ambulances use vehicle speed 50 miles per hour to travel between stations to demand zones or stations to stations. The response time was an exponential distribution, with rates which depend on call zones and stations of dispatched ambulances. The service time was the total time in which ambulances were busy on-scene of the accident and provided transportation to hospital if needed. Service times were also assumed to be exponentially distributed. We assumed that the service time did not depend upon call zones and stations of dispatched ambulances. The returning time also followed an exponential distribution based on call zones and which station a traveling ambulance would return to in the new system state. The Markov chain model with relocation was programmed in the Java programming language. The NetBeans IDE

7.3.1 was used to implement the model. The outputs of the Markov chain model with relocation were inputs to integer programming model, which we programmed in IBM ILOG CPLEX Optimization Studio 11.2.

4.6.1 Nested-Compliance Table Model Validation

We developed a discrete event simulation to validate the integer programming model. The simulation model was implemented using Arena Version14, running on an Intel® Core(TM)2 Duo CPU. We used the previously described data set from Hanover Fire and EMS department. We formulated the simulation model where data sets being along with our assumptions; call arrival rate was Poisson distribution and all interval times were exponential distribution. *The objective is to maximize the expected coverage using a binary notion of coverage. The binary coverage refers to a call being covered if we dispatch an ambulance from stations within a pre-specified response time threshold (RTT) to respond to the call. We calculate the expected coverage not considering variability in response time.* We used the closest policy to respond to calls. We ran the simulation model with 1680 simulated hours for each replication, and 500 replications. The simulated time was 19 minutes for each policy, compared to the integer programming model taking 20 seconds to obtain the optimal policy. The input parameters are shown in Appendix A.3.

We compared the results of integer programming model to the results of the simulation model. Table 4.4 shows the absolute error and percent error of the coverage in comparison between the integer programming model and simulation model based on

same the nested-compliance policy. These results indicated the average percent error 2.2% with the mean service time 60 minutes and 3.2% with the mean service time 70 minutes for systems with an arrival rate of 1.5 calls per hour and a response time threshold of 9 minutes. Thus the approximation of our objective function used in the integer programming model was close to the coverage obtained from simulation model. The percent errors tended to be higher when the mean service time was increased. These results suggest that when ambulances spend more time providing service to patients, the resulting increases in the busy probabilities result in increasing percent error of the approximated coverage using our Markov model.

Table 4.4: Comparison of the results of the integer programming model to results of the simulation model at arrival rate 1.5 call per hour, and response time threshold (RTTs) of 9 minutes

Service Time (mins)	# of Servers	Relocation Model with Nested Cons.				Service Time (mins)	# of Servers	Relocation Model with Nested Cons.			
		Math	Simu.	Abs. error	% error			Math	Simu.	Abs. error	% error
60	6	0.88	0.91	0.03	3.18	70	6	0.86	0.90	0.05	5.19
	7	0.92	0.95	0.02	2.27		7	0.90	0.94	0.03	3.46
	8	0.95	0.96	0.02	1.79		8	0.94	0.96	0.03	2.74
	9	0.96	0.98	0.02	2.21		9	0.95	0.98	0.02	2.33
	10	0.97	0.98	0.02	1.70		10	0.96	0.99	0.02	2.51

Figure 4.3a and 4.3b showed that the expected coverage increases with increasing the number of ambulances. They also showed that the error of our approximated coverage was higher for a larger RTT (7 compared to 9 minutes) for both 60 and 70 minute service times. These observations suggested that the impact of the nested-compliance table model on the expected coverage of systems was how to set the response time thresholds (RTTs) in which a smaller RTT provided higher accuracy and a smaller service time provided higher accuracy of the nested-compliance model. The results also showed when a system

has larger number of ambulance, the results of the simulation model provided better than the results of the integer programming model. These observations resulted from the likelihood of out of compliance state. We assumed that ambulance being in en-route cannot respond to a call. In realistic condition, we might dispatch back up ambulance to respond to the call that simulation model allows for this assumption. However, when the system has a fewer number of ambulances, there is higher possibility that backup ambulance is not available or cannot respond to the call within pre-specified response time thresholds (RTTs). Therefore, the backup ambulance does affect to accuracy of our nested-compliance table model in case of the larger number of ambulances but in case of the fewer number of ambulances do not affect to accuracy of the model.

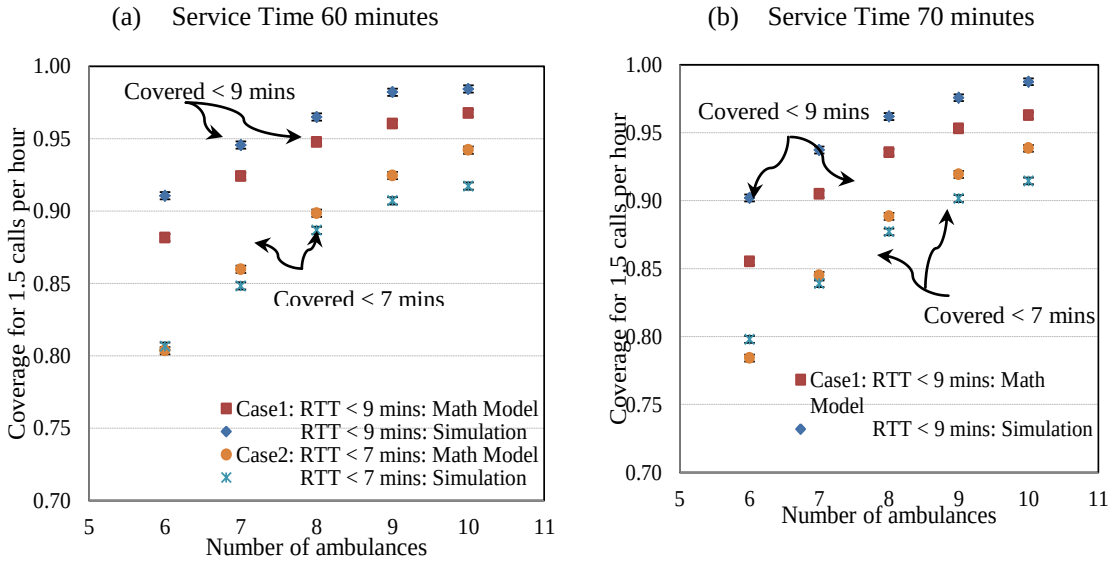


Figure 4.3: Comparison of the coverage of 1.5 calls per hour under the integer programming model versus the simulation model

4.6.2 Comparison with Non-Relocation Model based on the Adjusted Maximum Expected Covering Location Model

While our model seeks to determine the nested-compliance table through the Markov chain model with relocation embedded into an integer programming model, we have to verify the efficiency of the nested-compliance model for use in real-world EMS system. That is, we wish to know if there is any benefit of relocating vehicles. In this section, we compared the nested-compliance table model to a traditional adjusted maximal expected covering location problem (AMEXCLP) based on Batta et al. [65]. They modified the MEXCLP objective function developed by Daskin [19]. Their idea was to relax the independence assumption of server busy probabilities for the hypercube model. The correction factors, $Q(M, p, v)$ based on Larson [14] were included in the MEXCLP model. The correction factors indicated that the probability of having v busy servers. The adjusted maximal expected covering location problem (AMEXCLP) was formulated as a baseline for non-relocation model to compare the expected coverage. The objective function was to maximize the expected proportion of demand that could be covered. The formulation of AMEXCLP was showed below.

AMEXCLP Model

Objective function:

$$\textit{Maximize} \quad \sum_{i=1}^{n}\sum_{v=1}^{K}(\lambda_i / \lambda)Q(K, p, v-1)(1-p)\,p^{v-1}\cdot y_{vi} \tag{4.22}$$

Subject to

$$\sum_{v=1}^{K} y_{vi} - \sum_{v=1}^{K} a_{vi}x_v \le 0 \qquad \forall i \tag{4.23}$$

$$\sum_{v=1}^{K} x_v = K \tag{4.24}$$

$x_v \in \{0, 1\}$ $\forall v$

$y_{vi} \in \{0, 1\}$ $\forall v, i$

a_{vi} = 1 if $d_{vi} > D$ a server at station v does not cover demand zone at i

0 if $d_{vi} \le D$ a server at station v covers demand zone at i

y_{vi} = 1 if demand zone i is covered by at least v servers

0 otherwise

x_v = 1 if server locates at station v

0 otherwise

p server busy probability

K number of servers to be located

n number of demand zones

We used the real-world data from Hanover Fire and EMS department to compare the two models. We consider two instances of the data set, where the first data set is data from real world problem and the second is data set from random proportions of demand

zones based on the first data set. We compared the results of integer programming model of AMEXCLP to results of the simulation model. Figure 4.4 showed the expected coverage with varied number of ambulance from 6 to 10. These results indicated the average percent error 1.5% with the mean service time 70 minutes. These observations showed that the approximations of AMEXCLP are pretty close to the simulated AMEXCLP policies. Figure 4.5a showed results from real world problem and Figure 5b showed results from random proportions of demand zones. We varied the number of ambulances from 7 to 10 given that response time thresholds were 7 and 9 minutes. We compared the results of our nested-compliance table model to non-relocation model based on the AMEXCLP, with call arrival rate 1.5 call per hour and the mean service time 70 minutes.

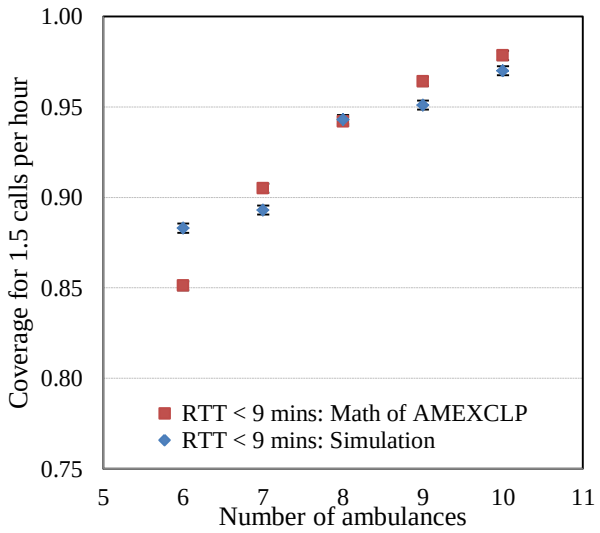


Figure 4.4: Comparison of the coverage at 1.5 calls per hour and RTT < 9 minutes under the AMEXCLP math model versus the simulated AMEXCLP policy

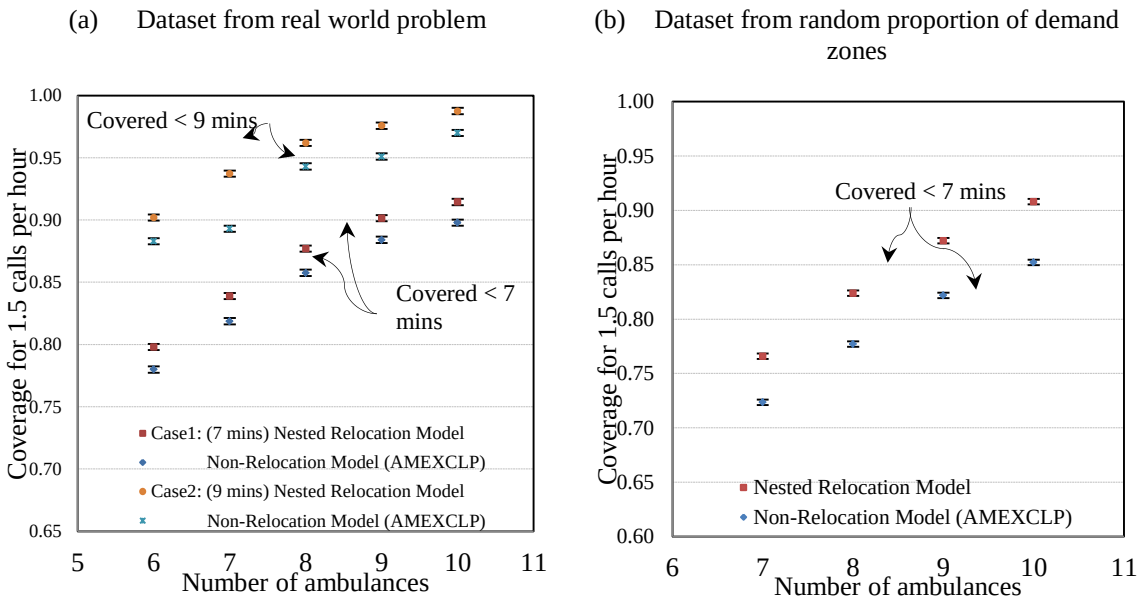


Figure 4.5: Comparison of the coverage of 1.5 calls per hour and service time 70 mins under the nested-compliance table model versus the non-relocation model (AMEXCLP)

These results showed improvement in outcomes when using the nested-compliance table model in comparison to non-relocation policy based on the AMEXCLP, coverage is calculated in the simulation model (since the AMEXCLP provides a different approximation of coverage which does not account for relocations). In Figure 4.5a when we used the criteria of response time threshold 9 minutes, the results showed improvement average 2.8% of the nested-relocation policies in comparison to non-relocation policies. When we reduced the criteria of response time threshold to 7 minutes, the results showed the slightly decreasing of the benefit of our nested-relocation policies to 2.1% improvement in comparison to non-relocation policies (AMEXCLP). These results showed higher efficiency of our nested-compliance model over the non-relocation based on AMEXCLP. In Figure 4.5b, we used the data set from random proportions of

demand zones based on the first data set. The results indicated that percent improvement of the nested-relocation policies in comparison to non-relocation policies were average 6.1% in which there were more percent improvement than dataset from real world problem. The observations of the data set from random proportions of demand zones showed lower number of stations that could cover for high proportion of demand zones than the data set from real world problem. These results suggested that the efficiency of the nested-compliance table model depends on number of stations that could cover for high proportion of demand zones. This implied that dataset of lower number of stations that could cover for high proportion of demand zones would provide more efficiency of the nested-relocation policies. The intuition behind this result can be explained in the following manner. In an EMS systems where there are fewer stations that can cover a high proportion of demand zones, if a call arrives to system, and the first closest ambulance is not available, without relocation, there is a higher probability that the second closest ambulance was not located in a station that could cover that demand zone. If we relocated an ambulance to stations that could cover this call, we would increase the expected coverage of EMS systems.

4.7 Conclusions and Future Research

In this paper, we formulated and validated a nested-compliance table model of an EMS system under relocation. We modeled the nested-compliance table as integer programming model. The model requires outcomes of steady state probabilities from a Markov chain model with relocation to be input parameters for our integer programming

model. The solutions were validated with data set from real-world problem. The mathematical model provided solutions the pretty close solutions to simulation solutions based on an objective of the maximum expected coverage using a binary coverage. The calls are covered if the assigned ambulance is located to station within a pre-specified response time. The validation showed that the nested-compliance table provided estimates of average error 2% - 3%. We have demonstrated the efficiency of the nested-compliance table model in comparisons to results from the non-relocation (AMEXCLP) model. The results showed that our model provided improvement of solutions over the results of the non-relocation (AMEXCLP) model of average 2.8% based on original data set from real-world problem and 6.1% based on data set from random proportion of demand zones. The performance of the nested-compliance table model depended on the pre-specified response time threshold (RTT) to calculate the expected coverage and data set of problems.

Implementing solutions to real-world problem suggests that the solutions of the nested-compliance table model provided improvement over the non-relocation (AMEXCLP) model depended on where the binary coverage is used. In the realistic problem, the distribution of response time might affect the realized expected coverage. The observation of results showed that the efficiency of the nested-compliance model depends on the average travel time between stations (relocation time). The application of the nested-compliance table model should be limiting the relocation time between stations. Thus, the possible way to impose an upper bound of relocation time is to partition service time in to small sub-areas (districts). The relocation rules which

allowing an ambulance moving within its district will further expend of the nested-compliance table model.

CHAPTER FIVE

A NESTED-COMPLIANCE TABLE MODEL EMBEDDED INTO A TABU SEARCH HEURISTIC FOR DISTRICTING AND RELOCATION IN EMS SYSTEMS

5.1 Introduction

The goal of emergency medical service (EMS) systems is to save the lives of out-of-hospital patients. The most common performance measure used to evaluate the efficiency of EMS systems is coverage, which is the proportion of calls that can be responded to within some pre-specified time standard. Coverage is related to the allocation of ambulances to stations to service areas in potential demand zones. Relocation, which involves moving ambulances to replace ambulances that have become busy in order to prevent some demand areas from being uncovered, is a well-known strategy to improve the performance of EMS systems. However, in Chapter 4 the objective of our work was to maximize the expected coverage using a binary notion of. A call will be covered if we dispatch an ambulance from a station within a pre-specified response time threshold (RTT) to respond to the call. In realistic EMS systems, the results of realized expected coverage (whether the call was actually reached within the time standard, not whether it should have been reached) might be different. The observation in Chapter 4 suggests that limiting relocation time is important for implementation of a relocation model in real-world systems based on the realized expected coverage measure. Long relocation time can result in the loss of calls that arrive during the move of an ambulance to a new station. Therefore, the decisions regarding relocation could be

improved by imposing some limitation on relocation time. One possible way to impose a relocation time restriction is to partition the whole service area into districts. In this work, we incorporate the districting problem into our relocation model. The service area is partitioned into small sub-areas (districts). Each sub-area operates under a particular relocation strategy based on a compliance table policy.

Our major contribution is to determine districting strategies that maximize the overall realized expected coverage among districts. The realized coverage refers to a call being covered if an ambulance responds to a call within a pre-specified response time threshold (RTT) that is coverage is calculated post, not pre, ambulance arrival. We calculate the expected coverage considering variability of response time. The decisions are two-fold. First, we determine how to partition service areas into districts and allocate ambulances to each district. Then, we determine the compliance table policy for each district; that is, we embed a relocation strategy into the districting model. The compliance table is a table that shows the choices of open stations depending upon the number of available ambulances. The details of the compliance table are presented in Chapter 4. The EMS systems operate under a dynamic strategy. Each district operates individually based on its own compliance table policy. We fix the dispatching policy to always send the closest ambulance to respond to a call. We consider an intra-district policy, which does not allow for ambulances to cross districts. The benefit of an intra-district policy is that relocating ambulances is forced to occur within a single district, which allows us to impose a limiting relocation time constraint. Thus, it is possible to increase the probability of availability of ambulances at potential locations. In this paper, we consider

a single call priority and a single type of ambulance (paramedic units). We combine two main decisions: districting and relocation. The algorithm is formulated by taking into account the compliance table model and embedding it into a tabu search heuristic. The objective is to maximize overall expected coverage.

In this study we:

- Develop a districting model for EMS systems that considers the number of districts and the allocation of ambulances to districts.
- Propose a tabu search heuristic to determine the maximum overall realized expected coverage using a searching method based on the optimization of the nested-compliance table formulation in Chapter 4.
- Show, through the numerical results of simulated realized coverage, how the solutions from our combination of districting and relocation strategies compare with the non-district and non-relocation strategies based on the adjusted maximum expected covering location problem (AMEXCLP) of Batta et al. (1989) in real world problems.

This article is organized as follows. In Section 5.2 we provide a brief review of the districting problem in service systems applications. Section 5.3 presents a description of EMS systems with districting and relocation strategies, as well as a description of how we developed the model. Section 5.4 presents the tabu search and nested-compliance table algorithms. Section 5.5 presents a more detailed discussion of the tabu search approach. Section 5.6 presents the efficiency of the districting and relocation solutions. Finally, Section 5.7 presents the conclusion and a discussion of future work.

5.2 Literature Review

Initial work described in service systems, Hess et al. [88] considered the service area of police patrol system into small sub-areas. It was referred to as a "districting problem" in which the region was partitioned into districts in order to improve outcomes of service systems. An integer programming model was formulated to minimize the sum of the squared distance given a particular number of districts. The decision was to assign the population to districts. Similar work of Gass [89] used a heuristic for police patrol problem presented by Hess et al. [88]. Bertolazzi et al. [90] formulated the districting problem as an integer programming problem. The objective was to minimize the overall travel time while providing workload balance. The decision was the allocation of calls to the stations. an application of the districting problem to transportation problem, Marlin [91] considered the districting problem to minimize total travel cost. The decisions were to assign locations and workload to districts. They considered upper and lower bounds of total workload for each district. They formulated the model as a linear programming model. Fleischmann and Paraschis [92] considered the application of districting problem in design of sales territories. They formulated an integer programming model. The objective was to minimize the total scores of products with distance between center coordinate and center locations of sales territories. Schoepfle and Church [93] considered the districting problem which applied to school systems. They introduced a network flow problem which was equivalent to a districting problem. They formulated their model as a linear programming model. They referred to it as the Generic Districting Problem (GDiP). Hojati [94] considered the optimal political districting problem given tolerance

of districts. The decision was to assign populations to districts by applying the transportation problem. They resolved the problem by splitting problem until convergence occurs. Geroliminis et al. [95] extended the districting problem to consider spatial and temporal demand. Their model accounted for the probability that a server is not available. The model considered server rate which is dependent on districting and dispatching policies. They formulated the model as an integer programming model using an embeded spatial queuing model. Several works considered restricting the districting problem by only allowing redistricting to occur between adjacent districts, which are more realistic, to improve the efficiency of systems.

Several works related to the districting problem allow activities to cross district boundaries. Larson and Stevenson [96] introduced the response redistricting problem in EMS systems. They considered the redistricting problem associated with facility location allowing for response across district boundaries. They first introduce a system in which servers did not cross district boundaries. Then, they assumed that a server might respond to a call from an adjacent district. Larson [97] considered a hypercube queuing model for location and redistricting problem. The model included inter-district (boundary crossing allowed) and intra-district (boundary crossing not allowed) response given dispatching policies. Traditional districting models considered the minimum sum of distance as objective functions; Plane [98] considered an alternative objective function in the redistricting problem. The alternative objective is to maximize interaction/minimize separation. The maximum interaction was referred to as the maximum intra-district spatial interaction or minimum intra-district interpersonal separation. The maximum

interaction was the maximization of total flow connecting all possible pairs of nodes with other districts. The minimum separation was the minimization of total flow connecting all possible pairs of nodes within districts. Justin and Williams [99] reviewed the redistricting problem. They mentioned the contribution of studied redistricting problem in several subtopics: possible criteria, methods (e.g. optimization, heuristic algorithm), and the extension of future works.

Other studies considered an optimization based on heuristic approach or a heuristic approach to the districting problem. Mehrotra et al. [100] considered the district boundaries within the state of South Carolina, US. They proposed an optimization based heuristic algorithm to solve the districting problem while providing population equality. They developed a mathematical model to obtain the district policies and used a branch-and-price method to determine the policy to yield equally size of populations for district. Muyldermans et al. [101] considered the districting problem in road networks. The road networks were partitioned to districts. The problem accounted for the different types of routing. The decisions were to choose the routing, balance in workload, configuration of sub-areas and center of the depot of each district. They considered the heuristic procedure for districting problem. Amico et al. [102] considered a redistricting problem to police command boundaries. They modeled the problem as a graph-partitioning problem subject to constraints of contiguity, compactness, convexity and size. The simulated annealing algorithm was proposed to search the partitions of districts. Bozkaya et al. [103] considered a political districting problem. The problem was modeled subject to several constraints such as contiguity, population equality, and compactness. The problem was

solved by using a tabu search algorithm. Ricca and Simeone [104] presented the districting problem for political elections. They applied a traditional political districting model to formulate their model. The mathematical model was very complicated. They used a local search heuristic to search a good solution. Iannoni et al. [105] consider the combination of location and districting problem on highway. The objective was to minimize average response time while considering balancing workload in systems. The problem approach used a spatial distributed queuing model embedded into a hybrid genetic algorithm.

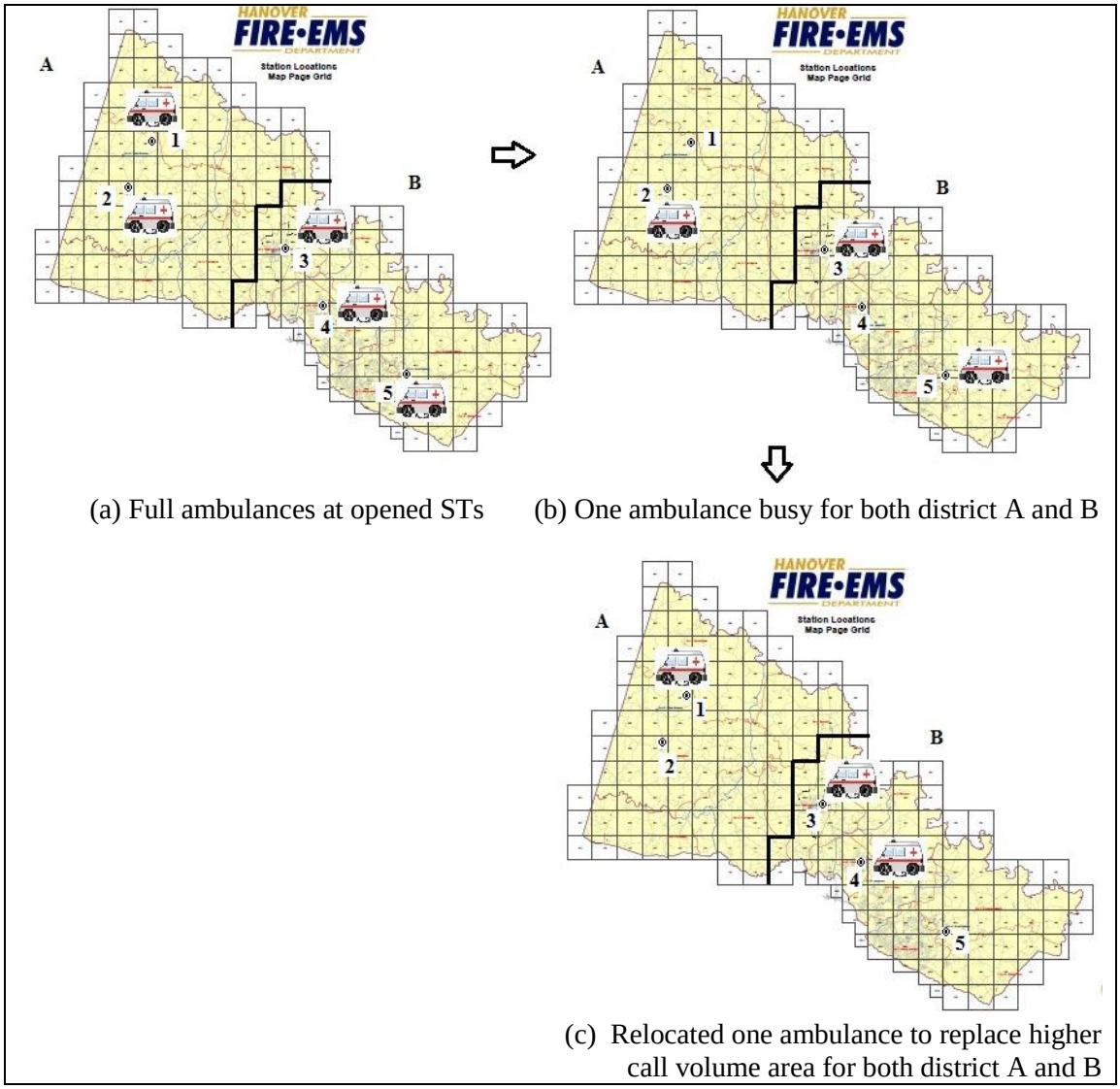
In this work we extend the relocation strategies proposed in Chapter 4. In particular, we developed a model which considers the partitioning of the service area into districts. We then applied the compliance table model into the sub-areas to maximize overall expected coverage. The decisions are number of districts, locations of ambulance for each district and the compliance table for each district. This work differs the previous work in that we consider the combination of districting and relocation in a single model.

5.3 EMS Systems with Districting and Relocation

EMS systems operate as a zero-queue system. We consider an EMS system with a single dispatch and single call priority. The EMS system operates under a relocation strategy with a nested-compliance table. That is, we consider a relocation policy in which we allow for relocation of at most one ambulance upon call arrivals and call completions. When the whole system operates as a single district, the nested-compliance table

approach may produce undesired relocation times. The upper bound on relocation time results in increasing the performance of implemented the nested-compliance model to EMS systems. In this chapter, we present EMS systems in which the service area is partitioned into sub-areas to limit the relocation time. This is referred to as the "districting problem". The districting model generates the service area to the small sub-areas. The number of ambulances is given to each sub-area. Each sub-area operates as a distinguishable sub-system, responding to calls and relocating within its sub-area. Given the partitioned service area, we operate each sub-area under relocation policies based on the nested-compliance table. Figure 5.1 provides an example of districting and the nested-compliance table model. We partition the whole service area into two districts given two ambulances for district A and three ambulances for district B shown in Figure 5.1(a). The district A and B operate independently. Suppose a call of demand zones in responsibility area of district B arrives to a system, and the ambulance in station4 respond to the call shown in Figure 5.1(b). The system state of district B changes to one busy ambulance. The located ambulance in station5 moves to replace the ambulance at station4 shown in Figure 5.1(c). No ambulances of district A in station1 and 2 are dispatched across district to respond the calls from demand zones in district B and no relocating ambulances of district A to replace the ambulance at stations in district B. The sequence of events of sub-system under relocation is discussed in Section 4.3, Chapter 4. The procedures and assumptions of how we approach the districting and the nested-compliance table problem are described as follows. Figure 5.2 shows the overall process, and is discussed in detail in Section 5.4

- Partitioning the service area: the service area is partitioned based on relocation time between stations. The model limits relocation time given the specified upper bound of relocation time.
- Assigning number of ambulances: we determined the number of ambulance for each sub-area based on call volume. Given some possible number of ambulances for each sub-area, comparison between policies is demonstrated to obtain the better policy.
- Determining the nested-compliance table policy: the optimization model of the nested-compliance table based on previous work in Chapter 4 is used to determine the optimal nested-compliance table for each sub-area. The objective is to maximize the realized expected coverage for each sub-area.
- Evaluating the solution: we consider the realized expected coverage based on the objective of the nested-compliance table under relocation model which is presented in Chapter 4 to evaluate the solution.
- Developing solution: we consider the tabu search heuristic to develop the solutions.
- Demonstrating model: the districting and the nested-compliance table models are demonstrated using real world data. A comparison between our model and adjust maximum expected coverage location problem (AMEXCLP) [24] is provided to show the efficiency of our tabu search heuristic.



(a) Full ambulances at opened STs

(b) One ambulance busy for both district A and B

(c) Relocated one ambulance to replace higher call volume area for both district A and B

Figure 5.1: The combination of the districting and relocation strategies

5.4 Tabu Search Heuristic and Nested-Compliance Table Policy for EMS Systems

In this section, we describe the algorithm of the nested-compliance table embedded into a tabu search heuristic for the districting and relocation problem. The tabu search is an iterative method to search for near optimal solutions where the objective function is to maximize the expected coverage throughout each solution. We classify

solutions into two categories: districting and nested-compliance table solutions. The efficiency of the algorithm to search for better nested-compliance table solutions depends on the districting solutions. In districting algorithm, the algorithm consists of two loops for improving two solutions: station solution and demand zone solution in which the loop of demand solution is contained inside the loop of station solution. The tabu search heuristic is a powerful tool for this nested problem structure while genetic algorithm requires generating large number of chromosomes for station solution and generating large number of chromosomes for demand zone solution that are contained inside of each chromosome for station solution resulting in longer computational running time. Figure 5.2 shows the flow process of our algorithm to search the solutions of districting and relocation problem. The procedure approach is described as follows:

- Generating the initial districting solutions consists of two solutions: stations for each district and demand zones for each district given number of districts and number of ambulances for each district.
- Generating the optimal nested-compliance table for each district, we use an optimization model of the nested-compliance table formulation to locate ambulances to stations for each state for each district given number of ambulances for each district.
- Developing solution of the demand zone solutions for each district, we consider the tabu search heuristic whereas the algorithm incorporates the method to determine the optimal nested-compliance table solution inside.

- Developing solution of the station solutions for each district, we also consider the tabu search heuristic whereas the searching demand zone solutions and optimal nested-compliance table solutions are embedded into this algorithm.

5.4.1 The Application of a Nested-Compliance Table Model

The nested-compliance table is a particular table in which shows the number of busy ambulances associated with where exact ambulances are located to. The dynamic relocation deals with the real-time movement of one ambulance to new location. We describe the EMS systems as the two-dimensional state spaces. The first state variable $V(t)$ denoted the status of number of ambulances busy at time t. The second state variable $C(t)$ indicated the status of systems in compliance, $C(t) = 1$ or out of compliance, $C(t) = 0$ at time t. The in compliance states mean all available ambulances are ready at their home stations to respond to call arrival, whereas out of compliance states mean one ambulances is not ready at its home station to respond to call arrival. It is during traveling to new home station. Suppose we have K_j ambulances in district j, there are particular combinations $2K_j$-1 possible states to system where the (K_j, 1) does not existing because of no any ambulance available at station. The assumptions of the nested-compliance table model for EMS systems are:

- The service area is partitioned into districts. Each district consists of demand zones which each district operates independent. Each demand zone i calls arrive according to a Poisson process. The λ is total call arrival rate and λ_i is call arrival rate of demand zone i. The calls require the dispatch of the closest ambulance within their district.

- There are K ambulances that are divided into K_j ambulances for each district. In general, the server has distinct mean service time depends on its home station, demand zone required service and the decision of a new home station which dispatched ambulance traveling back to.
- Relocated an ambulance for each district occurs when the number of busy ambulances changes; call arrival and call completed service. The one ambulance has to move to the new home station in the new system state. We described the event of relocation in Section 4.3.
- We determine the approximation steady-state probability $\pi_{v,c(j)}$ for district j based on our previous work in Sections 4.4. The assess of steady-state probability, we need to approximate some parameters; the average rate $\mu_{v,0(j)}$ of call arrival for district j when in state (v, 0), the average rate of call completion $\mu_{1(j)}$ for district j, and the average travel time between stations $\gamma_{(j)}$ for district j. The approximations of parameters of transition rates are described in Section 4.4.1

In application of nested-compliance table formulation we have two decision variables. We formulated the nested-compliance table model as integer programming model based on our previous work in Section 4.5. The maximum expected coverage is determine individually for each district. The details of the objective function and constraints are described in Section 4.5.

$x_{mv(j)}$ = 1 if an ambulance is located to station *m* in district *j* when system being in state *v*

= 0 otherwise

$y_{iv(d_i=j)}$ = 1 if demand zone *i* assigned to district *j* is covered when the system in district *j* is in state *v* if all vehicles are at their assigned locations

= 0 otherwise

Objective function for each district j:

$$\textit{Maximize} \quad \sum_{i=1}^{n} \sum_{v=0}^{K_j-1} (\lambda_{i(d_i=j)} / \lambda) \cdot \left(\pi_{v,1(j)} + \left(\frac{K_j - v - 1}{K_j - v} \right) \cdot \pi_{v,0(j)} \right) \cdot y_{iv(d_i=j)} \qquad (5.1)$$

$$\textit{for} \quad j = 0, 1, 2, \ldots, J$$

Subject to

$$\sum_{m=1}^{M} x_{mv(j)} = K_j - v \qquad \textit{for} \quad v = 0, 1, 2, \ldots, K_j\text{-}1 \qquad (5.2)$$

$$y_{iv(j)} \le \sum_{m \in M_{i(d_i=j)}} x_{mv(j)} \qquad \textit{for} \quad i = 1, 2, \ldots, n \qquad (5.3)$$

$$\textit{for} \quad v = 0, 1, 2, \ldots, K_j\text{-}1$$

$$x_{m,v-1(j)} \ge x_{mv(j)} \qquad \textit{for} \quad m = 1, 2, \ldots, M_j \qquad (5.4)$$

$$\textit{for} \quad v = 1, 2, 3, \ldots, K_j\text{-}1$$

$x_{mv(j)} \in \{0, 1\}$ $y_{iv(j)} \in \{0, 1\}$

Table 5.1 The parameters of the nested-compliance table model and the tabu search heuristic

Notation	Description
n	number of demand zones
J	number of districts
M_j	number of ambulance stations in district j
$M_{i(d_i=j)}$	set of locations in district j that can respond to calls at demand zone i within the specific time where demand zone i assigned to district j
M	number of ambulance stations in the EMS system
λ_i	call arrival rate from demand zone i, such that
d_i	indicate the district of demand zone i
J	number of districts
K_j	number of paramedic units at district j
K	number of paramedic units
i	indicator of demand zone as $i = 1, 2,..., n$
j	indicator of district as $j = 1, 2,..., J$
λ	arrival rate
$\lambda_{i(d_i = j)}$	arrival rate of call zone i assigned to district j
$\pi_{v,0(j)}$	the steady-state probability that the system in district j is out of compliance when in state v (number of available servers is k_j-v)
$\pi_{v,1(j)}$	the steady-state probability that the system in district j is in compliance when in state v (number of available servers is K-v)

5.4.2 Tabu Search Approach for Districting and Relocation Problem

In this section, we describe the tabu search algorithm which is developed for the districting and relocation problem. The iterative procedure is used to search the maximum expected coverage throughout two searching and an optimization steps: determining the optimal nested-compliance table solution, searching the demand zone solution and searching the station solution. We keep results of the maximum expected coverage given by the optimal nested-compliance table solution to the loop of searching the demand zone solution. We also keep in memory the maximum expected coverage given by both the optimal nested-compliance table and demand zone solution to the loop of searching the station solution. We start with descriptions of the components of the tabu search heuristic for the station solution (main algorithm), the demand zone solution

(algorithm A) and the optimization model of the nested-compliance solution (algorithm B). Figure 5.2 shows the process flow of the nested-compliance table embedded into the tabu search heuristic.

5.4.2.1 The Objective Function by Using the Nested-Compliance Table with Relocation Model

An application of the nested-compliance table with relocation in Section 4.5, Chapter 4 is applied to obtain the objective function (fitness). The objective function is to maximize the expected coverage throughout all districts. The v is state of system which indicates number of busy servers, for each district. The parameter $\pi_{v,1(j)}$ is the steady-state probability of district j that the system is in compliance when in state v. The parameter $\pi_{v,0(j)}$ is the steady-state probability of district j that the system is out of compliance when in state v. The λ is total arrival rate. The $\lambda_{i(di=j)}$ is the call arrival rate from demand zone i assigned to district j. The variable $y_{iv(di=j)}$ is 1, if demand zone i is in district j and covered in state v and otherwise is zero. We use the same notation following notations in Section 4.5, Chapter 4. Table 5.1 shows notations of our algorithm. We consider the constraints following equation (4.19) – (4.20) to provide the feasible solutions. The objective function is calculated by

$$\sum_{j=1}^{J}\sum_{i=1}^{n}\sum_{v=0}^{K_j-1}(\lambda_{i(d_i=j)}/\lambda)\cdot\left(\pi_{v,1(j)}+\left(\frac{K_j-v-1}{K_j-v}\right)\cdot\pi_{v,0(j)}\right)\cdot y_{iv(d_i=j)} \qquad (5.5)$$

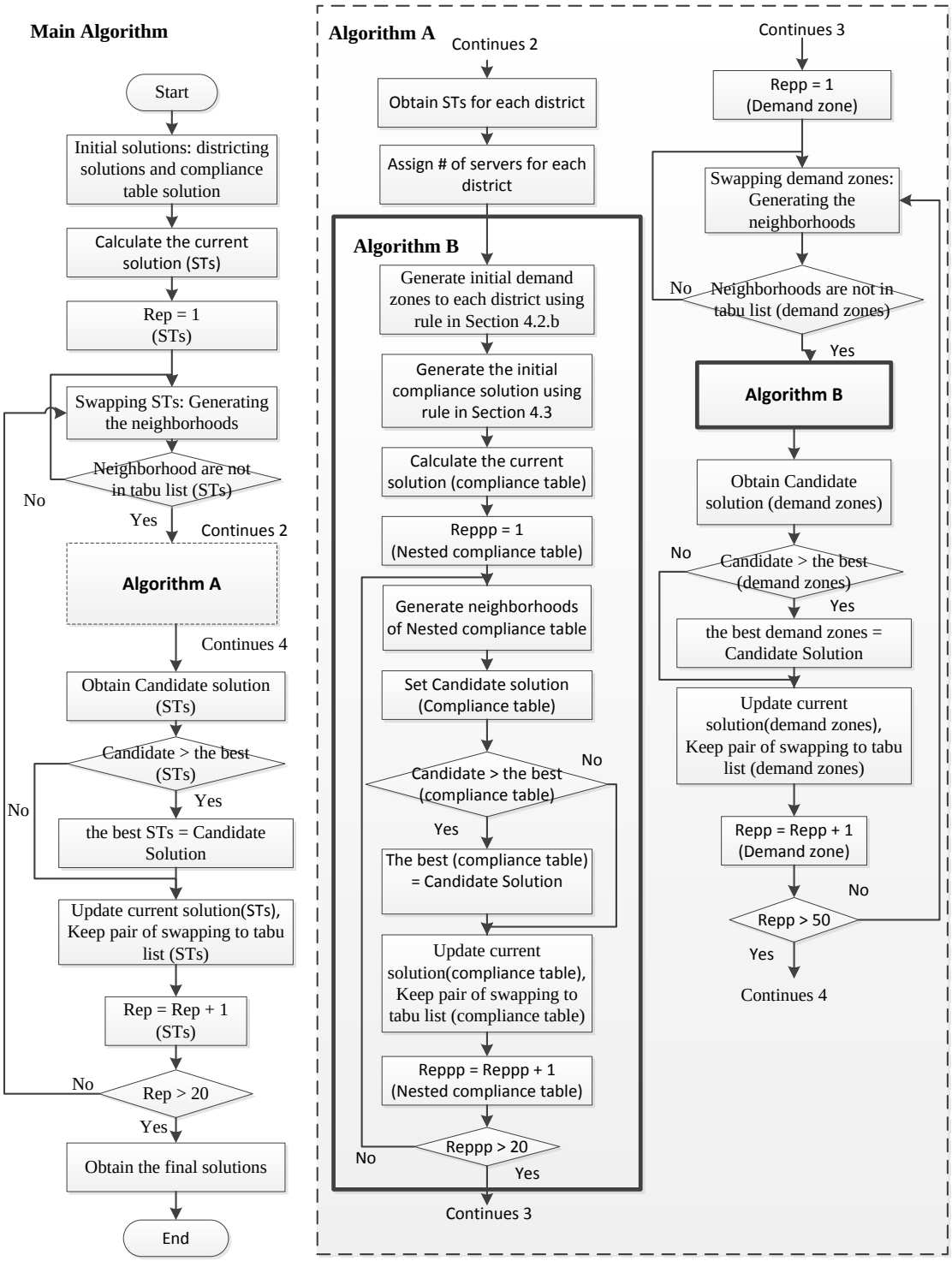


Figure 5.2: The process flow of the nested-compliance table embedded into the tabu search heuristic algorithm

5.4.2.2 Generating an Initial Districting Solution

The initial districting solution is generated by using a heuristic algorithm. The districting solutions found throughout the heuristic algorithm consist of three sequential steps.

(a) We first start with a number of districts equal to two. Starting with first district, we first select the station at a corner of map to assign to the set of the first district. We then gradually select a station by selecting its closest adjacent units. We consider an upper bound of relocation time between current assigned station and the next station assigned into the same district. The district is completed when no adjacent station- is near-by within the upper bound of the relocation time. If no any station is located in the current district for which travel time from the current assigned station to them is within the upper bound of relocation time, we then start to consider the next district and so on. We then update the number of districts when all districts are completed and remaining stations cannot assign to any district

(b) We consider demand zones to assign to each district. We determine the demand zones for each district using the maximum number of covered stations for each demand zone. Suppose we consider demand zone *i*. If more than one district has number of stations which can cover demand zone- *i* equally, we use the closest stations. The demand zone *i* will be allocated to the same district as its closest station.

(c) The number of ambulances is assigned to each district based on call volumes. We consider some possible solutions by increasing and decreasing number of ambulances.

Algorithm Initial ST solutions

begin
  initialize number of district equal to two
  for each district *j*
    select the station at the corner of the map to assign to the first district (district *j*)
    gradually select an adjacent station within relocation time which is the closest station
    if no adjacent station is in relocation time, update number of districts
  until all stations are assigned to district
  if all districts are completed and remaining stations cannot assign to any district, update the –
      number of district = number of district + 1
  repeat for loop again
end

Algorithm Initial demand zone solutions

begin
  for each demand zone *i*
    for each district, count the number of stations which cover the demand zone *i* (respond to demand zone *i* within a given pre-specified RTT)
    select the district which contains the maximum number of stations which can respond to demand zone *i* within the given RTT, Suppose it is district *j*
    assign demand zone *i* to district *j*
    if more than one district are selected, choose the closest station (Suppose the closest is in district *k*). We assign demand zone *i* to district *k*
  until all demand zones are assigned to district
end

5.4.2.3 Solution representation

The permutation representation is used to present our solutions. The representation shows three solutions in which there are relationships among three decisions: station solution, demand zone solution and nested-compliance table solution. Since the nested-compliance table solution depends on the station solution by considering the open stations of each district, which allows for assigned ambulances available to them for each state. While the demand zone solution provides the call arrival rates to calculate

the steady state probabilities of each district, we use this information to create the optimal nested-compliance table solution for each district. Figure 5.3 shows the instance of problem size n = 14, m = 9, and J = 3. The two servers are assigned to district A. The station solution is represented by station s2, s3 and s8 which are assigned to district A. The demand zone solution is represented by demand zone z1, z2, z4, z6 and z11 which are assigned to district A. The compliance table solution is represented by the ambulances available at station s2 and s8, when the system of district A is in state v = 0, and at station s8 where system of district A is in state v = 1.

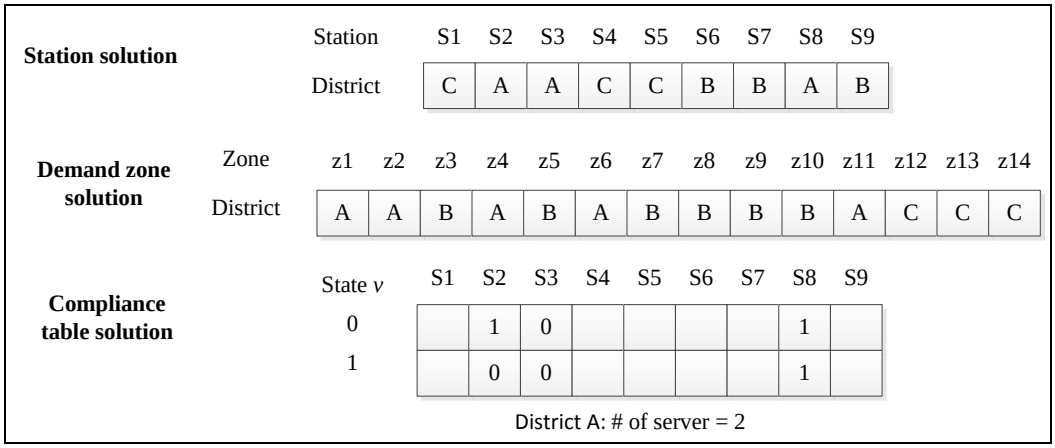


Figure 5.3: Permuted representation of the district and relocation problem

5.4.2.4 Improving process

We consider the improving procedure which the best neighborhood solution as a candidate solution to compare with the current solution. If the candidate solution is better, the candidate solution will be the best so far solution and the candidate solution will be the current solution also. If the candidate solution is lower than the current solution, the candidate solution will be the current solution but the best so far solution does not

change. In this section, we discuss the "neighborhood" solutions for each decision individually: the station solution, the demand zone solution and the optimal nested-compliance table solution.

5.4.2.4.1 Neighborhood of Main Algorithm

We consider the neighborhood solutions of a station solution. In Figure 5.4 we show the current solution and one of the neighborhood solutions. Each station (box) contains the letter that indicates its specified district. The neighborhood solution is to swap a pair of adjacent stations under a constraint. We consider the constraint of relocation time. The relocation time between a pair of swapping stations with other stations in the same district after swapped is less than the upper bound of relocation time. We consider all possible solutions in the neighborhood of the current solution. For example suppose we have four stations in Figure 5.4. The current solution {s3, s4} = {A, C} is swapped to {s3, s4} = {C, A} where relocation time between s1-s3 and s2-s4 are less than the upper bound of relocation time.

5.4.2.4.2 Neighborhood of Algorithm A (demand zone assignment)

We consider the neighborhood solutions of a demand zone solution. Figure 5.4 shows an example of a neighborhood solution of the demand zone solution. We start by randomly selecting a district for the swapping procedure. We examine a demand zone at a boundary area of the chosen district to swap with the demand zone of its adjacent district. Suppose we search the demand zone at the boundary area of the chosen district,

which we consider the demand zone: z2 (district A) which its adjacent demand zone: z3 (district B) indicates in different district. Thus, we consider {z2, z3} = {A, B} to swap to {z2, z3} = {B, A}. We examine all the solutions in the neighborhood of the current solution with restrictions above.

5.4.2.4.3 Algorithm B (compliance table assignment)

We consider the optimal solution of a nested-compliance table solution. The optimal solution is determined by using the formulation in Section 4.1. We determine the optimal nested-compliance table for each district. The overall expected coverage of current solution is composition of the expected coverage for each district.

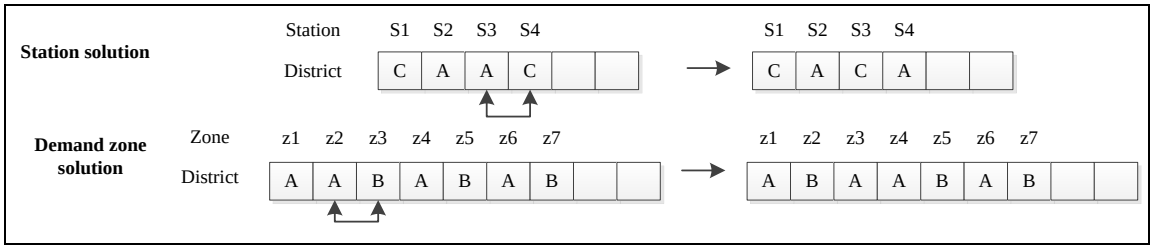


Figure 5.4: Permuted representation of swapping in the district and relocation problems

In improving solution, we use a tabu lists to record the old solutions in the lists. We considered separate tabu lists for station solution, demand zone solution and compliance table solution. Each tabu list consists of pairs of swapped solutions. Islam and Eksioglu (1997) recommended that the tabu list size was too small, the algorithm might be cycling and too large, good solutions might be skipped. They suggested that the appropriate size of a tabu list was five – ten recorded solutions to provid the better solution. We choose seven pairs of recorded solutions. For other parameters of traditional tabu search algorithms, previous work suggested that the stopping rules the solution. In

this work we consider a maximum CPU time to determine the number of iterations. We limit the running time of our algorithm for a size problem of five servers at an hour. The TS is run for the station solution at 10 iterations, and the demand zone solution at 300 iterations whereas the nested-compliance table solution we use the optimization model of formulation in Section 5.4.1.

5.5 Computational Results

In this section, we present computational results based on real-world data. A case study of data set from the Hanover Country Fire and EMS department, Hanover Country, Virginia is used to investigate our model. The service area of 474 square miles is partitioned into 122 demand zones and 16 station bases. The system handles approximately 1.2 calls per hour. The assumptions of our model were the same as assumptions in Section 4.6, Chapter 4. We assumed that call arrivals were Poisson during peak time, with mean arrival rate of 1.5 calls per hour. The service time follows an exponential distribution, with mean service time of 70 minutes. The response time and the returning time follow a lognormal distribution. The relocation time between stations follow an exponential distribution based on the current station and a new home station in the new system state. The algorithm was programed in the Java programming language. The NetBeans IDE 7.3.1 was used to implement the algorithm.We considered an EMS system with varied number of ambulances from 5 to 10. We terminated the program using the stopping criteria which was presented in Section 5.4.

We developed a discrete-event simulation to compare our nested-compliance table embedded into tabu search heuristic model to non-districting and non-relocation model based on the adjusted maximum expected covering location problem (AMEXCLP). Table 5.2 shows the comparison of the results of our nested-compliance table embedded into the tabu search heuristic to the AMEXCLP model. We varied the arrival rate from 1.0 to 2.5 calls per hour in a system with 7 ambulances, given a mean service time of 70 minutes, a pre-specified response time threshold of 9 minutes and a fixed maximum relocation time of 9 minutes. The column "Simu." shows the simulated expected coverage as a result of implementing the policies produced by either model, while the column "Resp T" shows the simulated average response time. These results show improvement in the realized expected coverage when using our algorithm over the AMEXCLP model. The observations of results indicate that the increasing of arrival rate results in increasing for the efficiency of our algorithm in which the average percent improvement was 3.26%. Figure 5.5 shows a comparison of our algorithm to the AMEXCLP model in terms of the resulting realized expected coverage and the expected response time. The results show that our algorithm provided the better statistically significant difference when the arrival rate was 1.5 to 2.5 calls per hour. The improvements are higher when the arrival rate is increasing or systems have higher busy probability of ambulances. Considering our model, as expected, when the arrival rate increased the expected coverage decreased while the expected response time increased (the slight decrease between arrival rate of 1.5 and 2.0 is not significant and is due to simulation uncertainty). This observation suggested that using the nested-compliance

table model based on a binary coverage the solutions might be not the optimal solution for realized expected coverage, but our solutions still provided the benefits over the AMEXCLP solutions. The results of applying the AMEXCLP solution to the simulated real system reveal unexpected results. In particular, the expected coverage and expected response time are not monotone in the arrival rate (these differences are significant). We believe this is driven by two assumptions of the AMEXCLP model which we relaxed in the simulation. First, the AMEXCLP is based on the binary coverage in which the solutions might be not the optimal solution for realized expected coverage. Another effect resulted from the realized response time distribution following a lognormal distribution while the AMEXCLP assumed the response time following an exponential distribution. However, the distributions of response time did not affect to the nested-compliance table model because of the steady-state probabilities of our model were insensitive to the shapes of response time. If the composition of response time and service time (total service time) were state-independent and relocation time approves infinity, then the model approaches an Erlang loss model.

Table 5.2: Comparison of the districting and relocation model (tabu search heuristic) to non-districting and non-relocation model (AMEXCLP) under varied arrival rate

Total # of Servers	# of Districts	Arrival Rate: calls per hour.	Districting and Relocation Based on Tabu Search Heuristic		AMEXCLP- Non Districting and Non Relocation		% Improved
			Simu.	RespT	Simu.	RespT	
7	3	1.0	0.93	3.89	0.92	4.42	0.87
		1.5	0.92	3.96	0.88	5.11	5.36
		2.0	0.92	3.93	0.89	4.55	3.55
		2.5	0.92	4.02	0.86	5.03	6.80
							3.26

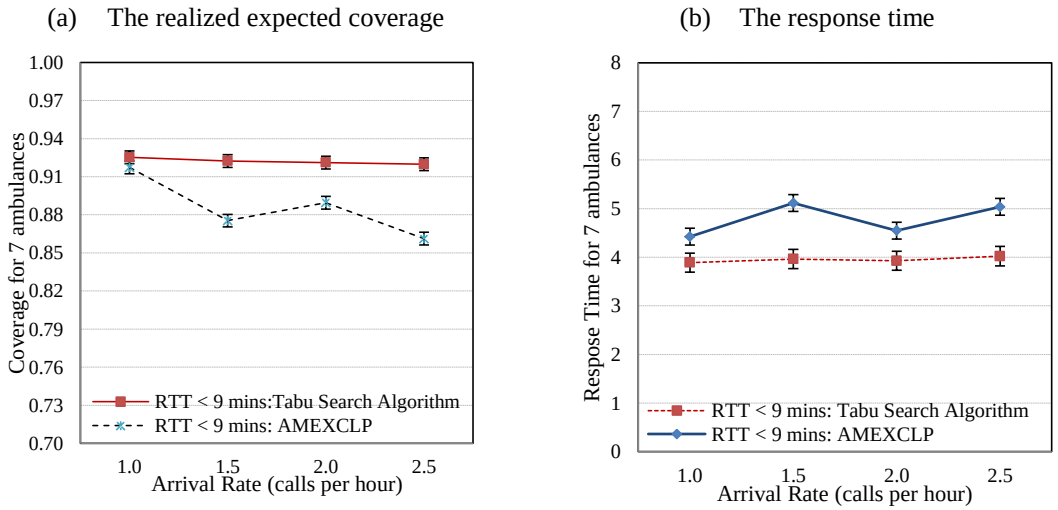


Figure 5.5: Comparison of the expected coverage and response time of the districting and relocation model versus the non-districting and non-relocation model (AMEXCLP)

Table 5.3 shows results for different number of ambulances. We compared the results of our algorithm to the AMEXCLP model with varied number of ambulances. We observed that the nested-compliance table embedded into the tabu search heuristic provided better outcomes for the realized expected coverage at average 2.29%. These outcomes showed that the smaller number of ambulances provided better efficiency of our algorithm than larger number of ambulances. These results suggest that when systems have higher busy probability of ambulances, the nested-compliance table embedded into the tabu search heuristic achieved better realized expected coverage. In cases of small number of ambulances (6 and 8 ambulances), the results of 6 ambulance showed a slight improvement and 8 ambulance showed negative improvement because of imbalance load between districts for assigned number of ambulances to districts. We can improve these solutions when implementing in practice by moving some demand zones to obtain balancing load between districts. The observations of our model showed that the results

of the realized expected coverage and the expected response time were not consistently a large improvement over the AMEXCLP solutions. We believe this is resulting from the districting solutions which provided imbalance load solutions and using the nested-compliance table model based on the binary coverage, that is, it is the nature of the heuristics algorithm which may not find the optimal solution. In Figure 5.6 presents in a different way the same results shown in Table 5.3. The graph shows an increasing function relationship between the number of ambulances and the realized expected coverage, which is related to decreasing busy probability of ambulances as the number of ambulance increases. These results showed decreasing the improvement of our algorithm over AMEXCLP when number of ambulances increased. The observations implied that decreasing busy probability of ambulances provided decreased benefit of our algorithm. When a call arrives to EMS systems with low busy probabilities, the dispatch center is likely to have available ambulances to respond to the call. Thus the relocation of ambulances will provide only a small benefit for these EMS systems. However, when EMS systems have small number of ambulances, the relocation of ambulances to potential high demand areas will provide higher benefit for EMS systems.

Table 5.3: Comparison of the districting and relocation model (tabu search heuristic) to non-districting and non-relocation model (AMEXCLP) under varied number of the ambulances

Total # of Servers	Districting and Relocation Based on Tabu Search Heuristic		AMEXCLP- Non Districting and Non Relocation		% Improved
	Simu.	RespT	Simu.	RespT	
5	0.92	4.10	0.85	5.47	7.84
6	0.88	4.88	0.87	5.04	1.16
7	0.92	3.96	0.89	4.77	3.76
8	0.91	4.30	0.93	3.99	-1.78
9	0.94	3.57	0.93	3.88	0.93
10	0.94	3.71	0.94	4.02	0.23
					2.02

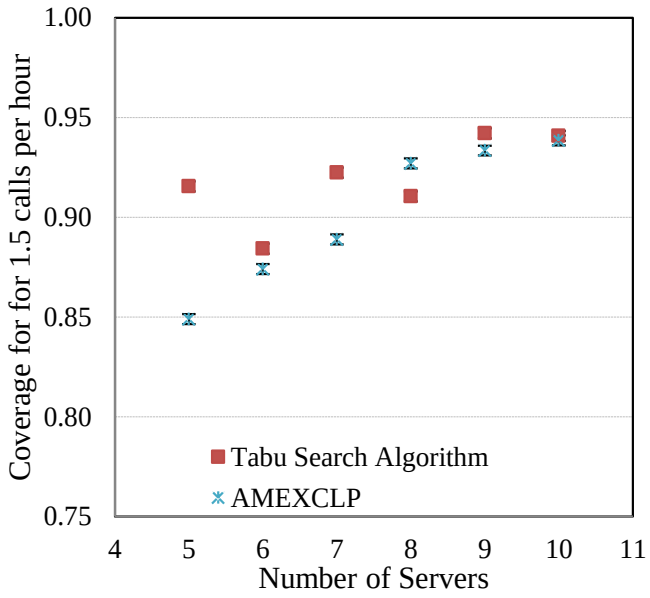


Figure 5.6: Comparison of the districting and relocation model versus the non-districting and non-relocation model (AMEXCLP)

5.6 Conclusions and Future Research

In this paper, we extend the nested-compliance table model to consider a combination of districting and relocation problem. We assumed that each district operates independently and in which each district operates under relocation. No ambulances are allowed across area boundaries. We developed the algorithm of the nested-compliance table model embedded into the tabu search heuristic for districting and relocation problem. The algorithm requires the optimization model of the nested-compliance table model throughout the searching method. The tabu search heuristic is used to search the solutions of districting problem; station solutions and demand zone solutions whereas each districting solution provides input parameters to determine the optimal nested-compliance table solution for each district. The results showed that optimization

embedded into the tabu search heuristic yields outcomes better than the AMEXCLP model policies for the realized expected coverage measure. The computational study showed that the realized expected coverage depends on number of ambulances and call arrival rate. When number of ambulances is smaller or call arrival rate is higher, our algorithm provides better outcome. We showed that there are higher benefits of combined districting and relocation problem when there are higher busy probabilities of ambulances in EMS systems. However, some results showed only a slight improvement of our algorithm compared to the AMEXCLP. We noted that implementing our algorithm in practice should consider the balancing load among districts which will provide more benefit.

In future research, we will develop the multiple-objectives for combination of districting and relocation problem. By partitioning the whole service area into small sub-areas, the algorithm results in unfairness among sub-areas. The fairness objective could be considered into the model. The contribution of this model will be helpful for realistic EMS systems.

CHAPTER SIX

CONCLUSION AND DISCUSSION

6.1 Conclusion

The goal of Emergency Medical service (EMS) systems is to provide quick pre-hospital care and transportation to patients, which in turn affects lives saved. The rapid response is important in reducing mortality rates of emergency patients. The purpose of this research is to improve the performance of EMS systems in terms of the expected survival probability and the expected coverage measures that are related to response time. We proposed two strategies to improve the efficiency of EMS systems: multiple unit dispatch and relocation strategies. Our primary focus is to consider models taking into account more realistic conditions; that is, we lift assumptions that are commonly made in the analysis of EMS systems. Multiple unit dispatching models are developed and analyzed to maximize outcomes based on dynamic conditions of real on-scene accidents. In another focus, we consider the relocation models that are implemented in real-world systems using the nested-compliance table policy. We used the real-world data collected from Hanover Fire and EMS department in Hanover County, Virginia, to evaluate the performance of our models.

First, we developed a discrete event simulation model for multiple unit dispatching and multiple call priorities. Emergency calls are classified into three types. We consider two types of medical units: ALS and BLS medical units. A decision must be made regarding how ambulances will be dispatched to respond to calls depending on call

priorities in order to maximize the expected survival probability. We consider the situation based on conditions at the scene of the accident. We used the closest dispatching policy for priority2 calls. Numerical results showed that the closest dispatching policy of double dispatching is optimal for prority1 calls, whereas the optimal dispatching policy for prirority3 calls is not the closest dispatching policy. A heuristic is developed to determine the near optimal policy for priority3 calls in large-scale problems. The proposed heuristic is to provide an ordered preference list for priority3 calls. We developed the heuristic by following the balanced call volume among servers. The results showed the efficiency of the heuristic was better than the closest dispatching policy.

We extend the model of multiple unit dispatching to consider fairness between call priorities. We consider the fairness in patient waiting time until the first response between pririty1 and 2 calls. We assumed that priority2 calls can be upgraded to priority1 calls based on information on-scene. We developed the optimization model based on simulation. The objective was to maximize the expected survival probability. The results showed that the optimal dispatching policy is better than the closest dispatching policy, where the imposed restriction on the deviation of waiting time until the first response between prority1 and 2 calls was set at 5 and 6 minutes.

Second, we formulated the nested-compliance table model under relocation as an integer programming model. The objective was to maximize the expected coverage based on binary coverage. We modified the Markov chain model with relocation based on Alanis et al. [5]. We approximated the transition rates by relaxing the assumption

proposed by Alanis et al. [5]. Our approximation of transition rates is independent of the exact nested-compliance table. We approximated the transition rates by using the covered arrival intensity to weigh potential stations. The benefit of our approximation is to calculate these parameters as input for the integer programming model. We validated our formulation by using a simulation. The results showed that the percent error of the expected coverage is 2% - 3%. We verified the efficiency of the nested-compliance table model by comparing our results with the AMEXCLP solutions. The results showed that the nested-compliance table solutions are better than AMEXCLP solutions by2% - 3% based on using real-world data.

Previous work with the nested-compliance table model considers binary coverage as the objective function, which may produce results that are different from the realized expected coverage. The results of the relocation model suggested that imposing an upper bound on relocation time can improve the performance of the system under relocation. Thus, we consider the whole service area that is partitioned into small sub-areas. We extend the nested-compliance table model to consider a districting problem. Each sub-area operates independently under its own nested-compliance table policy. We developed the nested-compliance table policy and embedded it into a tabu search heuristic. The objective was to maximize the realized expected coverage. We used an iterative method to search for near-optimal solutions to the districting problem, including station solutions and demand zone solutions. Each districting solution is used as input parameter for the nested-compliance table model. We determined the optimal nested-compliance table solution for each districting solution. We compared our solution to the AMEXCLP

solutions. The results showed that our tabu search heuristic yields outcomes better than AMEXCLP solutions in terms of the realized expected coverage.

6.2 Managerial Insights

The purpose of this research is to develop strategies to improve the performance of EMS systems by attempting to take into account the realistic conditions in EMS systems that are often ignored in the literature. The goal is to deliver medical units to patients in rapid response time. Several studies discussed the effect of delayed response time to survival probability of patients such as car crashes and cardiac arrest patients. The EMS administrators and providers continuously improve the performance of EMS system by focusing on decreasing the response time. The new strategies to dispatch rapid medical units to patients consider multiple unit dispatching and relocation strategies in practice. Suppose a cardiac arrest call arrives to EMS systems, we dispatch the closest two units to respond to the call. The direct effect is to increase survival probability of this patient.

The multiple unit dispatching policy we proposed can be used to implement in real-world EMS problems. Our assumption considers the realistic on-scene conditions of EMS systems, in which situation on-scene can be changed based on information. In term of improvement of performance measures, we found that implementing our multiple unit dispatching policy provides an additional 29 lives saved per 10,000 calls (3 ALS units 3 BLS units) and 49 lives saved per 10,000 calls (1 ALS unit 3 BLS units) in

comparison to a traditional policy (always send closest medical units). Our dispatching policy results in higher probability that the closest ambulance is available for life-threatening calls. In addition, our dispatching policy is easily implemented in EMS systems. The dispatch centers have only the ranked preference lists of dispatching medical units for each call priority. Suppose dispatch centers know where exact stations of available ambulances are located to, they can dispatch the particular medical units according to the ranked preference list of arrival call priorities.

Our combination of districting and relocation policy is possible to implement in practice by using the computer-aided dispatch (CAD) system and global positioning system (GPS). We proposed the specified nested-compliance policy for each district. Each district operates following its own nested-compliance table. The nested-compliance table is a particular table that indicates the exact stations for each state for each district of EMS systems. This table shows the assigned ambulance stations related to number of busy ambulances. In practice, the dispatchers have their own nested-compliance table lists and monitors that can track of status of all ambulance in systems and their current stations. When the number of ambulances of EMS systems changes, the dispatcher looks at the monitor and relocates ambulances to new stations in the new system state. No extra training course is required for using our nested-compliance table policy. In terms of outcomes, our districting and relocation policy provides better solutions than non-districting and non-relocation policy based on AMEXCLP model at 3.26% with 7 ambulances.

In terms of the costs associated with implementation of our policies, the EMS administrators do not need to invest in installing any system or training course for using our multiple-unit dispatching policy. Considering districting and relocation policy, the current practice of EMS systems already use the CAD systems. There is only investment for installing GPS to keep track of every ambulance; though since we only require the location of ambulances after service is complete, this can be achieved via radio. Thus we recommend that our multiple unit dispatching and combination of districting and relocation policies will provide high benefit to EMS systems.

APPENDICES

Appendix A

Additional Model and Results of Chapter 2

A.1 Flow process chart of Section 2.3

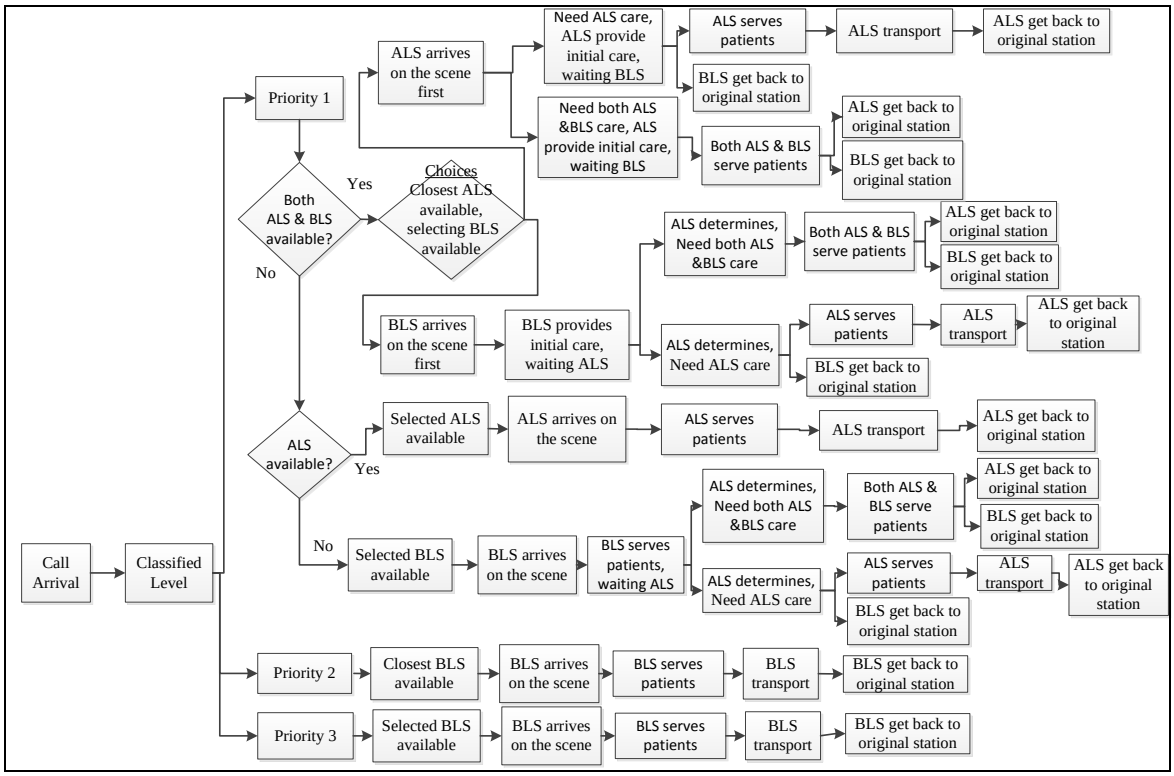


Figure A.1: The EMS system process

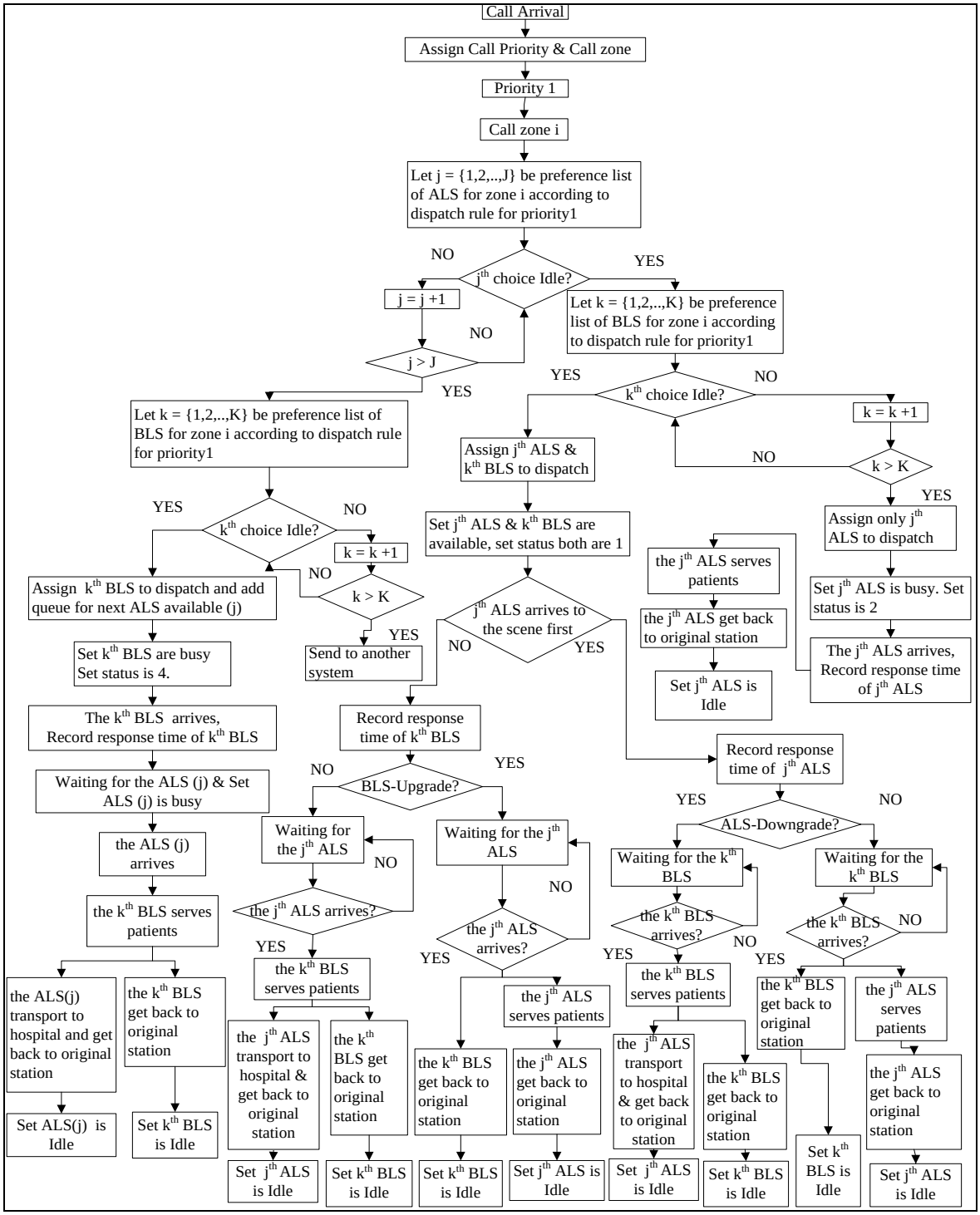


Figure A.2: Simulation flow chart of EMS systems

A.2 Swapping Procedure of Section 5

We improve our solution by swapping the arrangement of the ordered preference list for priority3 calls given that we have a fixed closest policy for priority1 and priority2 calls**.**

Let $a_{i31}, a_{i32}, \ldots, a_{i3K} \in A_i$ be a permutation of (1,2,…,*K*) that ranks in order preference list to respond priority 3 of call zone *i*

$c_{i31}, c_{i32}, \ldots, c_{i3K} \in C_i$ be a permutation of (1,2,…,*K*) that ranks in order preference list to respond priority 3 of call zone *i*

$d_{i31}, d_{i32}, \ldots, d_{i3K} \in D_i$ be a permutation of (1,2,…,*K*) that ranks in order preference list to respond priority 3 of call zone *i*

$r_u \in R$ be the rank of BLS matrix (*1* x *K*) that call volumes are sorted from Max to Min

$k = r_u$ preferred as BLS: *k* is sorted as u^{th} in the matrix of rank of call volumes.

```
Procedure Main {
    Prohibit List ∈ {ϕ}
    B = Big M;
    for  k ∈ K  Do {
        copy A_i to C_i
        if  (k != r_K) then {
            vdev = υ_BLS:k - υ_BLS: rK
            call Procedure swapping(k, r_K, C_i )
        }
        Loop to calculate busy probability {
            calculate busy probability using  Step I
        }
        calculate the call volume using equation 15
        calculate mean absolute deviation using equation 16 and 17
        if  ( B′ < B ) then {
            B = B′ ;
            copy C_i to A_i
        }
    }
}
Procedure swapping (k, r_k, C_i ) {
    B′ = BigM
    sum = 0;
    For i ∈ N Do {
        pos_k = position of server k in preference list of priority3 for call zone i
        pos_rK = position of server r_K in preference list of priority3 for call zone i
    if (pos_k < pos_rK) then {
            calculate the approximated  increasing of call volume for server  r_K (denoted as vol)
            if (pos_k = w && pos_rK = z)
                vol= λ_im · (g_w−1 − g_z−2)
        if  (i,k,r_K) ∉ Prohibit List
            sum = sum + vol
            if  (sum <  vdev)  then
                swap positions of ordered preference list between (k, r_K) of call zone i
```

ex. (...,k ,..., r_K ,...) swap to (...,r_K ,..., k ,...)

copy new swapping to Di
} }
calculate the busy probability using step2
calculate the call volume using step3
calculate the mean absolute deviation(B'') using step 4
if ($B'' < B'$) then {
$B' = B''$;
copy D_i to C_i
}
if (no call zone i can be swapped)
For i ϵ N Do {
pos_k = position of server k in preference list of priority 3 forcall zone i
pos_{rK} = position of server r_K in preference list of priority 3 forcall zone i
if ($pos_k < pos_{rK}$) then {
if $(i,k,r_K) \notin$ Prohibit List
swap positions of ordered preference list between (k, r_K) of call zone i
ex. (...,k ,..., r_K ,...) swap to (...,r_K ,..., k ,...)
copy new swapping to Di

calculate the busy probability using step2
calculate the call volume using step3
calculate the mean absolute deviation(B'') using step 4
}
if ($B'' < B'$) then {
$B' = B''$;
copy D_i to C_i
}
}
}

A.3 Input Data of Section 6

Response Times, Transportation Times and Proportion of call zones: Lognormal Distribution

Demand Zone	Call proportion	Station 1	Station 2	Station 3	Station 4
zone1	0.226034	(16.77,12.47)	(15.43,11.47)	(13.38,9.95)	(8.03,5.97)
zone2	0.019513	(32.14,23.89)	(32.14,23.89)	(19.87,14.77)	(32.14,23.89)
zone3	0.060281	(23.72,17.64)	(9.92,7.38)	(13.42,9.97)	(18.84,14.01)
zone4	0.043914	(26.26,19.52)	(32.14,23.89)	(26.07,19.38)	(15.39,11.44)
zone5	0.02657	(16.89,12.56)	(24.56,18.26)	(17.16,12.76)	(28.44,21.14)
zone6	0.09327	(10.07,7.48)	(16.32,12.13)	(32.14,23.89)	(15.59,11.59)
zone7	0.326744	(25.03,18.61)	(9.85,7.32)	(14.18,10.54)	(15.04,11.18)
zone8	0.065128	(18.82,13.99)	(32.14,23.89)	(13.74,10.21)	(25.79,19.17)
zone9	0.007525	(32.14,23.89)	(32.14,23.89	(27.34,20.32)	(20.9,15.53)
zone10	0.077626	(12.6,9.36)	(19.62,14.59)	(14.63,10.87)	(12.7,9.44)
zone11	0.029886	(22.98,17.08)	(18.28,13.59)	(19.77,14.70)	(19.69,14.63)
zone12	0.023509	(32.14,23.89)	(18.63,13.85)	(32.14,23.89)	(19.72,14.66)

Service times and Proportion of Priority1, 2 and 3 calls: Exponential Distribution

Demand Zone	Proportion of Priority1 calls	Proportion of Priority2 calls	Proportion of Priority3 calls	Service times	
				Priority1	Priority2,3
zone1	0.394	0.098	0.508	67.07	60.24
zone2	0.452	0.113	0.435	100.32	90.29
zone3	0.394	0.098	0.508	62.44	55.86
zone4	0.425	0.106	0.469	66.90	59.42
zone5	0.409	0.102	0.489	65.25	57.76
zone6	0.404	0.101	0.495	56.32	49.78
zone7	0.443	0.111	0.446	54.18	48.36
zone8	0.438	0.109	0.453	84.42	75.5
zone9	0.417	0.104	0.479	104.31	92.93
zone10	0.442	0.111	0.447	58.27	51.82
zone11	0.434	0.109	0.457	81.38	72.32
zone12	0.446	0.112	0.442	59.60	52.49

A.4 Results of Section 6

Comparison of performance of closest policy and heuristic policy

3 ALS 3 BLS 12 Zones

ALS1: Station 4, ALS2: Station 1 and ALS6: Station 4 BLS3: Station 4, BLS4: Station 1 and BLS5: Station 1

ID	Demand (calls/ hour)	Policy	Utilization						Mean BLS	Mean absolute deviation of BLS	Prob Survival	% Imp.	# of the imp. of lives saved /10,000 calls
			ALS1 :St4	ALS2 :St1	ALS3 :St3	BLS4 : St 4	BLS5 : St1	BLS6: St1					
		Closest	14.426	3.095	6.244	23.866	12.601	3.123	13.196	21.3385	0.1629		
1	0.25	Heuristic	14.418	3.178	6.171	16.159	13.762	11.372	13.764	4.7899	0.1656	1.657	27
		Closest	33.291	14.775	22.837	43.809	35.008	21.377	33.398	24.0423	0.1468		
2	0.50	Heuristic	33.070	14.919	22.088	33.285	31.402	36.985	33.890	6.1885	0.1507	2.657	39
		Closest	57.157	40.372	49.713	64.755	62.512	51.873	59.713	15.6815	0.1278		
3	0.75	Heuristic	55.962	39.028	48.152	57.720	60.973	58.605	59.100	3.7473	0.1316	2.973	38
		Closest	75.573	64.116	72.760	81.182	81.396	76.434	79.671	6.4729	0.1141		
4	1.00	Heuristic	75.222	64.093	72.406	77.431	81.101	80.589	79.707	4.5529	0.1163	1.928	22
		Closest	84.890	76.004	83.859	88.946	90.008	87.316	88.757	2.8823	0.1050		
5	1.25	Heuristic	84.580	76.115	83.661	87.225	90.070	89.081	88.792	3.1342	0.1069	1.810	19

2 ALS 3 BLS 12 Zones

ALS1: Station 4 and ALS2: Station 1

ID	Demand (calls/ hour)	Policy	Utilization						Mean BLS	Mean absolute deviation of BLS	Prob Survival	% Imp.	# of the imp. of lives saved /10,000 calls
			ALS1 :St4	ALS2 :St1	ALS3	BLS4 : St 4	BLS5 : St1	BLS6: St1					
		Closest	26.931	16.378	N/A	30.512	22.589	12.760	21.954	18.3876	0.1494		
1	0.25	Heuristic	26.263	16.299	N/A	23.783	23.418	19.366	22.189	5.6459	0.1519	1.673	25
		Closest	67.365	60.347	N/A	69.235	65.956	58.293	64.494	12.4033	0.1275		
2	0.50	Heuristic	67.490	59.504	N/A	63.693	64.032	65.339	64.355	1.9677	0.1307	2.510	32
		Closest	86.742	83.833	N/A	87.912	86.718	83.091	85.907	5.6315	0.1162		
3	0.75	Heuristic	85.753	82.718	N/A	84.994	85.984	84.395	85.125	1.7198	0.1188	2.238	26
		Closest	92.738	91.311	N/A	93.392	93.271	91.106	92.590	2.9681	0.1086		
4	1.00	Heuristic	92.546	91.064	N/A	92.160	92.956	92.503	92.540	0.8331	0.1106	1.842	20
		Closest	95.031	94.756	N/A	96.007	95.955	94.764	95.575	1.6233	0.1034		
5	1.25	Heuristic	95.417	94.834	N/A	95.468	96.101	95.478	95.682	0.8380	0.1049	1.451	15

2 ALS 3 BLS 12 Zones

ALS1: Station 4 and ALS2: Station 3

ID	Demand (calls/ hour)	Policy	Utilization						Mean BLS	Mean absolute deviation of BLS	Prob Survival	% Imp.	# of the imp. of lives saved /10,000 calls
			ALS1 :St4	ALS2 :St3	ALS3	BLS4 : St 4	BLS5 : St1	BLS6: St1					
		Closest	25.435	16.914	N/A	30.649	22.429	12.272	21.783	19.0231	0.1570		
1	0.25	Heuristic	25.967	17.424	N/A	24.491	23.549	20.124	22.721	5.1956	0.1589	1.210	19
		Closest	67.604	62.086	N/A	70.644	67.231	59.498	65.791	12.5863	0.1347		
2	0.50	Heuristic	67.950	62.404	N/A	65.697	66.291	67.171	66.386	1.5689	0.1382	2.598	35
		Closest	85.974	84.695	N/A	88.273	87.139	83.535	86.316	5.5612	0.1239		
3	0.75	Heuristic	85.763	84.093	N/A	85.972	86.999	85.280	86.083	1.8309	0.1257	1.453	18
		Closest	92.766	91.843	N/A	93.847	93.542	91.586	92.991	2.8111	0.1151		
4	1.00	Heuristic	92.318	91.912	N/A	92.715	93.411	93.086	93.070	0.7117	0.1180	2.520	29
		Closest	95.105	95.160	N/A	96.315	96.156	95.055	95.842	1.5743	0.1098		
5	1.25	Heuristic	95.087	95.050	N/A	95.690	96.204	95.605	95.833	0.7417	0.1109	1.002	11

2 ALS 3 BLS 12 Zones
ALS1: Station 3 and ALS2: Station 1

ID	Demand (calls/ hour)	Policy	Utilization						Mean BLS	Mean absolute deviation of BLS	Prob Survival	% Imp.	# of the imp. of lives saved /10,000 calls
			ALS1 :St3	ALS2 :St1	ALS3	BLS4 : St 4	BLS5 : St1	BLS6: St1					
		Closest	27.923	15.408	N/A	30.769	22.017	12.348	21.712	18.7269	0.1303		
1	0.25	Heuristic	27.797	15.208	N/A	23.796	23.044	19.337	22.059	5.4439	0.1340	2.840	37
		Closest	67.346	60.687	N/A	69.874	66.125	57.869	64.623	13.5071	0.1114		
2	0.50	Heuristic	66.737	59.242	N/A	63.477	63.774	64.389	63.880	1.0173	0.1163	4.399	49
		Closest	85.296	82.720	N/A	86.907	85.847	81.593	84.782	6.3778	0.1014		
3	0.75	Heuristic	84.964	82.425	N/A	84.623	85.796	83.659	84.692	2.2065	0.1047	3.254	33
		Closest	92.057	91.143	N/A	93.264	92.810	90.761	92.279	3.0342	0.0948		
4	1.00	Heuristic	91.967	90.856	N/A	91.810	92.732	91.918	92.153	1.1574	0.0982	3.586	34
		Closest	94.644	94.460	N/A	95.699	95.656	94.283	95.213	1.8590	0.0903		
5	1.25	Heuristic	94.692	94.589	N/A	95.234	95.762	94.988	95.328	0.8683	0.0924	2.326	21

1 ALS 3 BLS 12 Zones
ALS1: Station 4

ID	Demand (calls/ hour)	Policy	Utilization						Mean BLS	Mean absolute deviation of BLS	Prob Survival	% Imp.	# of the imp. of lives saved /10,000 calls
			ALS1 :St4	ALS2	ALS3	BLS4 : St 4	BLS5 : St1	BLS6: St1					
		Closest	69.275	N/A	N/A	70.302	66.066	61.373	65.914	9.0810	0.1357		
1	0.25	Heuristic	67.841	N/A	N/A	65.875	64.920	62.982	64.592	3.2207	0.1395	2.800	38
		Closest	91.567	N/A	N/A	91.715	90.276	87.687	89.892	4.4105	0.1180		
2	0.50	Heuristic	91.379	N/A	N/A	90.065	89.449	89.848	89.787	0.6764	0.1208	2.373	28
		Closest	96.304	N/A	N/A	96.231	95.604	94.144	95.326	2.3650	0.1065		
3	0.75	Heuristic	96.260	N/A	N/A	95.498	95.475	94.970	95.314	0.6887	0.1104	3.662	39
		Closest	98.002	N/A	N/A	97.890	97.540	96.647	97.359	1.4245	0.1013		
4	1.00	Heuristic	97.830	N/A	N/A	97.224	97.270	97.086	97.193	0.2144	0.1050	3.653	37
		Closest	98.625	N/A	N/A	98.491	98.268	97.634	98.131	0.9933	0.0958		
5	1.25	Heuristic	98.577	N/A	N/A	98.166	98.195	97.913	98.091	0.3570	0.0992	3.549	34

1 ALS 3 BLS 12 Zones
ALS1: Station 1

			ALS1 :St1	ALS2	ALS3	BLS4 : St 4	BLS5 : St1	BLS6: St1					
		Closest	61.254	N/A	N/A	62.072	56.415	50.273	56.253	11.9599	0.1098		
1	0.25	Heuristic	60.813	N/A	N/A	57.838	56.421	53.975	56.078	4.2067	0.1134	3.279	36
		Closest	89.427	N/A	N/A	89.180	87.246	83.808	86.744	5.8728	0.0957		
2	0.50	Heuristic	89.183	N/A	N/A	87.081	86.205	86.585	86.623	0.9145	0.1011	5.643	54
		Closest	95.404	N/A	N/A	95.053	94.148	92.188	93.796	3.2167	0.0877		
3	0.75	Heuristic	95.014	N/A	N/A	93.689	93.588	92.825	93.367	1.0841	0.0922	5.131	45
		Closest	97.437	N/A	N/A	97.090	96.598	95.370	96.353	1.9659	0.0818		
4	1.00	Heuristic	97.276	N/A	N/A	96.248	96.267	95.969	96.161	0.3840	0.0887	8.435	69
		Closest	98.305	N/A	N/A	97.945	97.625	96.774	97.448	1.3489	0.0794		
5	1.25	Heuristic	98.205	N/A	N/A	97.472	97.485	97.052	97.336	0.5693	0.0833	4.912	39

1 ALS 3 BLS 12 Zones ALS1: Station 3

ID	Demand (calls/ hour)	Policy	Utilization						Mean BLS	Mean absolute deviation of BLS	Prob Survival	% Imp.	# of the imp. of lives saved /10,000 calls
			ALS1: St3	ALS2	ALS3	BLS4 : St 4	BLS5 : St1	BLS6: St1					
		Closest	64.919	N/A	N/A	65.869	60.854	55.387	60.704	10.6329	0.1185		
1	0.25	Heuristic	65.274	N/A	N/A	62.838	61.604	59.477	61.306	3.6579	0.1216	2.616	31
		Closest	90.411	N/A	N/A	90.368	88.621	85.579	88.189	5.2199	0.1037		
2	0.50	Heuristic	89.947	N/A	N/A	88.210	87.444	87.819	87.824	0.7702	0.1077	3.857	40
		Closest	95.902	N/A	N/A	95.711	94.947	93.240	94.633	2.7849	0.0932		
3	0.75	Heuristic	95.524	N/A	N/A	94.469	94.434	93.767	94.223	0.9125	0.0972	4.292	40
		Closest	97.507	N/A	N/A	97.249	96.800	95.620	96.557	1.8726	0.0881		
4	1.00	Heuristic	97.639	N/A	N/A	96.854	96.866	96.633	96.784	0.3032	0.0923	4.767	42
		Closest	98.423	N/A	N/A	98.179	97.908	97.133	97.740	1.2143	0.0850		
5	1.25	Heuristic	98.371	N/A	N/A	97.821	97.842	97.478	97.714	0.4722	0.0879	3.412	29

REFERENCES

[1] Gibson, G. (1977). Emergency medical services. *Proceedings of the Academy of Political Science*, *32*(3), 121 – 135.

[2] Kuisma, M., Holmström, P., Repo, J., Määttä, T., Nousila-Wiik, M., and Boyd, J. (2004). Prehospital mortality in an EMS system using medical priority dispatching: a community based cohort study. *Resuscitation*, *61*(3), 297-302.

[3] Nicholl, J., Coleman, P., Parry, G., Turner, J., and Dixon, S. (1999). Emergency priority dispatch systems – a new era in the provision of ambulance services in the UK. *Pre-hospital Immediate Care*, *3*, 71-75.

[4] Clawson, J.J., Dernocoeur, K.B., and Rose, B. (2008). Principles of emergency medical dispatch. *National Academy of EMD*, 4th Edition, Utah, USA

[5] Alanis, R., Ingolfsson, A., and Kolfal, B. (2012). A markov chain model for an EMS system with repositioning. *Production and Operations Management*, *22*(1), 216 – 231.

[6] King, B.G., and Sox, E.D. (1967). An emergency medical service system: Analysis of workload: San Francisco Area. *Association of schools of public health*, *82*(11), 995 – 1008.

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

[8] Eaton, D.J., Daskin, M.S., Simmons, D., Bulloch, B., and Jansma, G. (1985). Determining emergency service vehicle deployment in Austin, Taxas. *Interfaces*, *15*(1), CPMS/TIMS Prize Papers, 96.

[9] Hougham, M. (1996). London ambulance service computer – aided dispatch system. *International Journal of Project Management*, *14*(2), 103 – 110.

[10] Palumbo, L., Kubincanek, J., Emerman, C., Jouriles, N., Cydulka, R., and Shade, B. (1996). Performance of a system to determine EMS dispatch priorities. *The American Journal of Emergency Medicine*, *14*(4), 388 - 390.

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

[12] Kolesar, P., and Walker, W.E. (1974). An algorithm for the dynamic relocation of fire companies. *Operations Research, 22*(2), 249 -274.

[13] Larson, R.C. (1974). A hypercube queuing model for facility location and redistricting in urban emergency services. *Computers and Operations Research, 1*(1), 67–95.

[14] Larson, R.C. (1975). Approximating the performance of urban emergency service system. *Operations Research, 23*(5), 845–868.

[15] Jarvis, J.P. (1985). Approximating the equilibrium behavior of multi-server loss systems. *Management Science, 31*(2), 235-239.

[16] Budge, S., Ingolfsson, A., and Erkut, E. (2009). Approximating vehicle dispatch probabilities for emergency service systems with location-specific service times and multiple units per location. *Operations Research*, *57*(1), 251–255.

[17] Chiyoshi, F., Iannoni, A.P., and Morabito, R. (2011). A tutorial on hypercube queuing models and some practical applications in emergency service systems. *Pesquisa Operacional (Impresso), 31*(2). 271-299.

[18] Church, R.L., and ReVelle, C. (1974). The maximal covering location problem. *Papers of the Regional Science Association, 32*(1), 101-118.

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

[20] Gendreau, M., Laporte, G., and Semet, F. (2006). The maximal expected coverage relocation problem for emergency vehicles. *Journal of the Operational Research Society*, *57*(1), 22–28.

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

[22] Marianov, V., and ReVelle, C. (1996). The queuing maximal availability location problem: A model for the siting of emergency vehicles. *European Journal Operational Research*, *93*(1), 110-120.

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

[24] Mendonça, F.C., and Morabito, R. (2001). Analyzing emergency medical service ambulance deployment on a Brazilian highway using the hypercube model. *Journal of the Operational Research Society*, *52*(3), 261-270.

[25] Atkinson, J.B., Kovalenko, I.N., Kuznetsov, N. Yu., and Mikhalevich, K.V. (2006). Heuristic methods for the analysis of a queuing system describing emergency medical service deployed along a highway. *Cybernetics and Systems Analysis*, *42*(3), 379-391.

[26] Iannoni, A.P., Morabito, R., and Saydam, C. (2007). A hypercube queuing model embedded into a genetic algorithm for ambulance deployment on highways. *Annals of Operations Research*, *157*(1), 207-224.

[27] McLay, L.A., and Mayorga, M.E. (2011) Evaluating the impact of performance goals on dispatching decisions in emergency medical service. *IIE Transactions on Healthcare Systems Engineering, 1*(3), 185-196.

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

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

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

[31] Chelst, K.R., and Barlach, Z. (1981). Multiple unit dispatches in emergency services: Models to estimate system performance. *Management Science*, *27*(12), 1390-1409.

[32] Gau, S.H., and Larson, R.C. (1988). Hypercube model with multiple-unit dispatches and police patrol-initiated activities. *MIT Operations Research Center Working Paper;* OR 188-88.

[33] Iannoni, A.P., and Morabito, R. (2007). A multiple dispatch and partial backup hypercube queuing model to analyze emergency medical system on highways. *Transportation Research Part E: Logistics and Transportation Review*, *43*(6), 755-771.

[34] Iannoni, A.P., and Morabito, R., and Saydam, C. (2011). Optimizing large-scale emergency medical system operations on highways using the hypercube queuing model. *Socio-Economic Planning Sciences*, *45*(3), 105-117.

[35] Andersson,T., and Värbrand, P. (2007). Decision support tools for ambulance dispatch and relocation. *Journal of the Operational Research Society*, *58*(2), 195-201.

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

[37] Bandara, D., Mayorga, M. E., and McLay, L. A. (2013). Priority dispatching strategies for EMS systems. *Journal of the Operational Research Society*.

[38] Larsen, M.P., Eisenberg, M.S., Cummins, R.O., and Hallstorm, A.P. (1993). Predicting survival from out-of-hospital cardiac arrest: a graphic model. *Annals of Emergency Medicine*, *22*(11), 1652–1658.

[39] McLay, L.A., and Mayorga, M.E. (2010). Evaluating emergency medical service performance measures. *Health Care Management Science*, *13*(2), 124–136.

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

[41] Marsh, M. T., and Schilling, D. A. (1994). Equity measurement in facility location analysis: a review and framework. *European Journal of Operational Research*, *74*(1), 1-17.

[42] Savas, E. S. (1978). On equity in providing public services. *Management Science*, *24*(8), 800-808.

[43] Bodily, S. E. (1978). Police sector design incorporating preferences of interest groups for equality and efficiency. *Management Science*, *24*(12), 1301-1313.

[44] Henderson, J. C., and Schilling, D. A. (1985). Design and Implementation of Decision Support Systems in the Public Sector. *MIS quarterly*, *9*(2), 157 – 161.

[45] Keeney, R. L., and Winkler, R. L. (1985). Evaluating decision strategies for equity of public risks. *Operations Research*, *33*(5), 955-970.

[46] Mandell, M.B. (1991). Modeling effectiveness – equity trade – offs in public service delivery systems. *Management Science*, *37*(4), 467 - 482.

[47] Ogryczak, W. (2000). Inequality measures and equitable approaches to location problem. *European Journal of Operational Research*, *122*(2), 374 – 391.

[48] Felder, S., and Brinkmann, H. (2002). Spatial allocation of emergency medical services: minimizing the death rate or providing equal access. *Regional Science and Urban Economics*, *32*(1), 27 – 45.

[49] Kostreva, M.M., Ogryczak, W., and Wierzbicki, A. (2004). Equitable aggregations and multiple criteria analysis. *European Journal of Operational Research*, *158*(2), 362 – 377.

[50] Heshmati, A. (2004). Inequalities and their measurement. *IZA Discussion Papers*, *1219*, 1 – 17.

[51] Hakimi, S. (1964). Optimal locations of switching centers and the absolute centers and medians of a graph. *Operations Research*, *12*(3), 450–459.

[52] Halpern, J. (1978). Finding minimal center-median convex combination (cent-dian) of a graph. *Management Science*, *24*(5), 353–544.

[53] Ogryczak, W. (2007). Inequality measures and equitable locations. *Annals of Operations Research*, *167*(1), 61 – 86.

[54] Lorenz, M.O. (1905). Methods for measuring concentration of wealth. *Journal of the American Statistical Association*, *9*(70), 209 – 219.

[55] Gini, C. (1921). Measurement of inequality of incomes. *The Economic Journal*, 124-126.

[56] Drezner, T., Zvi, D., and Guyse, J. (2009). Equitable service by a facility: minimizing the Gini coefficient. *Computers and Operations Research*, *36*(12), 3240 – 3246.

[57] Marín, A., Nickel, S., Puerto, J., and Velten, S. (2009). A flexible model and efficient solution strategies for discrete location problems. *Discrete Application Mathematics*, *157*(5), 1128–1145.

[58] Drezner, T., and Drezner, Z. (2011). A note on equity across groups in facility location. *Naval Research Logistics (NRL)*, *58*(7), 705 -711.

[59] Marín, A. (2011). The discrete facility location problem with balanced allocation of customers. *European Journal of Operational Research*, *210*(1), 27 – 38.

[60] Bertsimas, D., Farias, V.F., and Trichakis, N. (2011). The price of fairness. *Operations Research*, *59*(1), 17 – 31.

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

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

[63] Bertsimas, D., Farias, V.F., and Trichakis, N. (2011). Fairness, efficiency and flexibility in organ allocation for kidney transplantation. *Operations Research*, *61*(1), 73 – 87.

[64] Noyan, N. (2010). Alternate risk measures for emergency medical service system design. *Annals of Operations Research*, *181*(1), 559 – 589.

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

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

[67] Toregas, C., and ReVelle, C. (1972). Optimal location under time or distance constraints. *Papers in Regional Science*, *28*(1), 133-144.

[68] Aly, A.A., and White, J.A. (1978). Probabilistic Formulation of the Emergency Service Location Problem*. *Journal of the Operational Research Society*, *29*(12), 1167-1179.

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

[70] Tansel, B.C., Francis, R.L., and Lowe, T.J. (1983). State of the Art—Location on Networks: A Survey. Part I: The p-Center and p-Median Problems. *Management Science*, *29*(4), 482-497.

[71] Tansel, B.C., Francis, R.L., and Lowe, T.J. (1983). State of the Art—Location on Networks: A Survey. Part II: Exploiting Tree Network Structure. *Management Science*, *29*(4), 498-511.

[72] Krarup, J., and Pruzan, P.M. (1983). The simple plant location problem: survey and synthesis. *European Journal of Operational Research*, *12*(1), 36-81.

[73] Brandeau, M. L., and Chiu, S.S. (1989). An overview of representative problems in location research. *Management science*, *35*(6), 645-674.

[74] Beraldi, P., and Ruszczynski, A. (2002). The probabilistic set-covering problem. *Operations Research*, *50*(60), 956-967.

[75] Saxena, A., Goyal, V., and Lejeune, M.A. (2010). MIP reformulations of the probabilistic set covering problem. *Mathematical programming*, *121*(1), 1-31.

[76] Pirkul, H., and Schilling, D. (1989). The capacitated maximal covering location problem with backup service. *Annals of Operations Research*, *18*(1), 141 – 154.

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

[78] Galvão, R.D., Chiyoshi, F.Y., Morabito, R. (2005). Towards unified formulations and extensions of two classical probabilistic location models. *Computers & Operations Research*, *32*(1), 15-33.

[79] Galvão, R.D., and Morabito, R. (2008) Emergency service systems: The use of the hypercube queueing model in the solution of probabilistic location problems. *International Transactions in Operational Research*, *15*(5), 525-549.

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

[81] Erkut, E., Ingolfsson, A., and Budge, S. (2008). Maximum availability/reliability models for selecting ambulance station and vehicle locations: a critique. *Natural Sciences and Engineering Research Council of Canada*, 1-11.

[82] Kolesar, P., and Walker, W.E. (1972). An algorithm for the dynamic relocation of fire companies. *Operations Research*, *22*(2), 249 – 274.

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

[84] Rajagopalan, H.K., and Saydam, C., and Xiao, J. (2008). A multiperiod set covering location model for dynamic redeployment of ambulances. *Computers and Operations Research*, *35*(3), 814 – 826.

[85] Majzoubi, F., Bai, L., and Heragu, S.S. (2012). An optimization approach for dispatching and relocating EMS vehicles. *IIE transactions on Healthcare Systems Engineering*, *2*(3), 211 – 223.

[86] Berman, O. (1981). Dynamic repositioning of indistinguishable service units on transportation networks. *Transportation Science*, *15*(2), 115 – 136.

[87] Maxwell, M.S., Restrepo, M., Henderson, S.G., and Topaloglu, H. (2010). Approximate dynamic programming for ambulance redeployment. *Journal on Computing*, *22*(2), 266 – 281.

[88] Hess, S.W., Weaver, J.B., Siegfeldt, H.J., Whelan, J.N., and Zitlau, P.A. (1965). Nonpartisan political redistricting by computer. *Operations Research*, *13*(6), 998–1006.

[89] Gass, S.I. (1968). On the division of police districts into patrol beats. *Proceedings of ACM national conference 23rd*, 459-473.

[90] Bertolazzi, P., Bianco, L., and Ricciardelli, S. (1977) A method of determining the optimal districting in urban emergency service. *Computer & Operations Research*, *4*(1), 1-12.

[91] Marlin, P.G. (1981). Application of the transportation model to a large-scale districting problem. *Computer & Operations Research*, *8*(2), 83-96.

[92] Fleischmann, B., and Paraschis, J.N. (1988). Solving a large scale districting problem: a case report. *Computer & Operations Research*, *15*(6), 521-533.

[93] Schoepfle, O.B., and Church, R.L. (1991) A new network representation of a classic school districting problem. *Socio-Economic Planning Sciences*, *25*(3), 189-197.

[94] Hojati, M. (1996). Optimal political districting. *Computer & Operations Research*, *23*(12), 1147-1161.

[95] Geroliminis, N., Karlaftis, M.G., and Skabardonis, A. (2009). A spatial queuing model for the emergency vehicle districting and location problem. *Transportation Research Part B: Methodological*, *43*(7), 798-811.

[96] Larson, R.C., and Stevenson, K.A. (1972). On Insensitivities in urban redistricting and facility location. *Operations Research*, *20*(3), 595-612.

[97] Larson, R.C. (1974). A hypercube queuing model for facility location and redistricting in urban emergency services. *Computers & Operations Research*, *1*(1), 67-95.

[98] Plane, D.A. (1982). Redistricting reformulated: A maximum interaction/minimum separation objective. *Socio-Economic Planning Sciences*, *16*(6), 241-244.

[99] Williams, J.C. (1995) Political redistricting: a review. *Papers in Regional Science 74*(1), 13-40.

[100] Mehrotra, A., Johnson, E.L., and Nemhauser, G.L. (1998). An optimization based heuristic for political districting. *Management Science*, *44*(8), 1100-1114.

[101] Muyldermans, L., Cattrysse, D., Van Oudheusden, D., and Lotan, T. (2002). Districting for salt spreading operations. *European Journal of Operational Research*, *139*(3), 521-532.

[102] D'Amico, S.J., Wang, S.J., Batta, R., and Rump, C.M. (2002). A simulated annealing approach to police district design. *Computers & Operations Research*, *29*(6), 667-684.

[103] Bozkaya, B., Erkut, E., and Laporte, G. (2003). A tabu search heuristic and adaptive memory procedure for political districting. *European Journal of Operational Research*, *144*(1), 12-26.

[104] Ricca, F., and Simeone, B. (2008). Local search algorithms for political districting. *European Journal of Operational Research*, *189*(3), 1409-1426.

[105] Iannoni, A.P., Morabito, R., and 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.