8-2012

DISTRICTING AND DISPATCHING POLICIES TO IMPROVE THE EFFICIENCY OF EMERGENCY MEDICAL SERVICE (EMS) SYSTEMS

Damitha Bandara
Clemson University, dbandar@clemson.edu

DISTRICTING AND DISPATCHING POLICIES TO IMPROVE THE EFFICIENCY OF EMERGENCY MEDICAL SERVICE (EMS) SYSTEMS

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
Damitha Bandara
August 2012

Accepted by:
Dr. Maria E. Mayorga, Committee Chair
Dr. William G. Ferrell
Dr. Mary Elizabeth Kurz
Dr. Kevin M. Taaffe

Abstract

The major focus of Emergency Medical Service (EMS) systems is to save lives and to minimize the effects of emergency health incidents. The efficiency of the EMS systems is a major public concern. Thus, over the past three decades a significant amount of research studies have been conducted to improve the performance of EMS systems. The purpose of this study is also to improve the performance of EMS system. The contribution of this research towards improving the performance of EMS systems is twofold. One area is to implement optimal or near optimal dispatching strategies for EMS systems and the other is to determine the response boundaries for EMS vehicles.

Proposed dispatching strategies are implemented incorporating the degree of the urgency of the call. A Markov decision process (MDP) model is developed to obtain optimal dispatching strategies in less complex models. A heuristic algorithm is proposed to dispatch ambulances for more complex models. In this study, an integer programming formulation and a constructive heuristic are proposed to determine response areas or districts for each ambulance. Additionally, dispatching rules to dispatch paramedic units within districts and out of districts are examined.

Simulation is used to evaluate the performance of the EMS system after introducing proposed dispatching policies. Performance is measured in terms of patients' survival probability rather than measuring the response time thresholds, since survival probability reflects the patients' outcome directly. Results are illustrated using real-data collected from Hanover county Virginia. Results show that proposed dispatching rules are valuable in increasing patients' survivability.

Acknowledgments

I would like to express my heartiest gratitude to my advisor Dr. Maria E. Mayorga for her guidance and support. Her thoughtfulness and new ideas inspired me and helped me immensely in making this a success. My sincere thanks goes to Dr. William G. Ferrell, for sharing his experience with us about dynamic programming. Also my many thanks go to Dr. Mary Elizabeth Kurz for sharing her experience and ideas with us to complete this successfully. In addition I would like to thank Dr. Kevin M. Taaffe for sharing his simulation experience with us and to Dr. Laura A Mclay for providing data of Hanover county fire and EMS department to support this research effort. Finally but not last to my wife and all of my friends who helped me to complete this successfully.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The fundamental responsibilities of Emergency Medical Service (EMS) systems are to provide urgent medical care, such as pre-hospital care, and to transport the patient to the hospital if needed. The efficiency of EMS systems is a major public concern [35]. Problems, such as where to locate ambulances and how to dispatch ambulances, must be solved by EMS planners to provide effective and efficient service to the public.

Over the past three decades, a significant amount of research studies have been conducted to improve the performance of EMS systems. The major focus of these models is to reduce response time (the time between the receipt of a call at the dispatch center and the arrival of the first emergency response vehicle at the scene) by placing the ambulances in optimal locations. The focus has been on response time because EMS systems are designed to rapidly provide advance medical care to critical patients such as cardiac arrest or trauma. As a result, the common method to evaluate the EMS system is to measure *coverage level* that reflects the proportion of patients who experience a response time less than a given time standard. However, the focus of some recent research studies is on survival probability of patients in an emergency because survivability mirrors the patient outcome directly. As such, patients' survivability is used as our performance measure in this study.

Our contribution towards improving the performance of EMS systems is two fold. One area is to implement optimal or near optimal dispatching strategies for EMS systems and the other is to determine the response boundaries for EMS vehicles that maximize the patients' survivability. When implementing optimal dispatching strategies calls are prioritized considering the degree of the urgency of the call. Then, different dispatching strategies are implemented for each prioritized category. The proposed dispatching strategies are

developed for EMS systems that use fixed deployment, meaning paramedic units are located at specific locations, respond form those stations, and return to their home stations after providing the service. A Markov decision process (MDP) approach and a simulation-based approach are utilized when implementing optimal dispatching strategies. Since these approaches have the ability to address the stochastic and dynamic behavior of the EMS system. The model formulations and computational results are explained in detail later in this document. Computational results show that it is beneficial to implement prioritized dispatching strategies in EMS systems to maximize patients' survivability.

In another study, to determine the response areas or boundaries for each ambulance, an integer programing formulation and a constructive heuristic will be proposed. When operations of EMS vehicles are restricted to predetermined boundaries, it enables the EMS system to reduce the mean response time of paramedic units support to the scene [10]. Thus, we district the EMS service area into sub-regions in order to determine the emergency service vehicle response boundaries that increases patient survivability. After obtaining vehicles districts, we proposed intra-district (within the district) and inter-district (out of district) dispatching disciplines to improve the efficiency of the EMS system. These dispatching rules are developed incorporating the degree of the urgency of the call. The performance of integrated dispatching and districting policies are illustrated using real-world data collected from Hanover County, Virginia. Results show that operating ambulances according to boundaries given by our proposed methods and dispatching rules, can help to increase patients survivability.

The remainder of this doctoral thesis is organized as follows: In the next section relevant literature on models used for EMS systems is discussed. Chapter 2 presents "Optimal Dispatching Strategies for Emergency Vehicles to Increase Patient Survivability", a MDP approach and computational results, for obtaining optimal dispatching strategies that maximize patients survival probability. A simulation-based approach to determine optimal dispatching strategies, "Priority Dispatching Strategies for EMS systems," is presented in Chapter 3. Additionally, a heuristic approach for obtaining improved dispatching policies is discussed in Chapter 3. In Chapter 4, a mathematical model formulation and a constructive heuristic to determine geographical boundaries for paramedic units, "Districting and Dispatching Policies to Improve the Efficiency of Emergency Medical Service (EMS) Systems," is presented. Conclusions and future research suggestions are explained in Chapter 5.

1.1 Literature Review

The major focus of EMS system is to save lives and to minimize the effect of an emergency health incident. Thus, the decision-making process of an emergency medical services focused on effectiveness and efficiency, becomes a strategic challenge. EMS planners must solve problems, such as where to locate emergency service stations, how many ambulances to allocate to each station, and how to dispatch the appropriate paramedic unit to the emergency scene. A significant amount of research work has been done towards addressing these problems faced by EMS systems in order to improve the performance of the system. The relevant literature can be divided into which resource allocation decision is being made.

Most of the early models of the 1970's developed for EMS systems focused on placing ambulances at optimal locations to improve the EMS performances by providing better coverage for the population. In these covering location models, a demand point is said to be covered if there exists at least one vehicle within the distance or time standard. A few such static mathematical covering models are: the location set covering problem (LSCP) [49], the maximal covering location problem (MCLP) [11], the tandem equipment allocation model (TEAM) [43], and the double standard model (DSM) [16]. Probabilistic models are developed by researchcers to address the stochastic behavior of EMS systems, considering the fact that emergency vehicles are busy once they respond to a call. In these probabilistic models, the emergency medical units are considered as servers operating in a queuing system. The first such model is developed by Larson et al. [25]. Other well-known probabilistic models are the Maximum Expected Covering Location model (MEXCLP) and the Maximal Availability Location Problem (MALP). Additional extensions of the (MEXCLP) model can be found in EMS literature (e.g ReVelle et al. PLSCP [38], Marianov et al. Q-PLSCP [29], McLay [31]).

Ambulance relocation models found in EMS literature are developed to properly account for actual coverage level, since the dispatching of an ambulance in response to a call may leave a significant proportion of the population without sufficient coverage. In relocation models, the main focus is to relocate dynamically vehicles in real-time when vehicles are dispatched to the scene instead of looking for a single solution to a static or probabilistic model for providing proper coverage. The first such vehicle relocation model is developed by Koslea et al. [22] to relocate fire trucks. A more recent example for ambulance redeployment is the dynamic double standard model at time t (DDSMt) by Gendreau et al. [17]. The primary disadvantage of dynamic relocation models is the necessity of finding a new solution whenever a vehicle is dispatched. There are some other practical issues in implementing the relocation of ambulances in real time for EMS systems. For example, with an increase in the complexity of EMS systems, the number of relocations grows dramati-

cally; when calls come in quick succession, solutions may be infeasible. Additionally, relocating ambulance may change the route or destination of a vehicle frequently which can lead to a confusion of drivers and thereby causing mistakes. Current research focuses on developing more powerful solution methodologies to solve these models quickly. There are practical issues in implementing dynamic relocation models to real world EMS systems; in our study, an EMS system with fixed deployment (where paramedic units are located at specific stations, respond from stations, and return back to their station after serving a call) is considered instead of ambulances relocating. The main focus is to improve the EMS system performance by implementing efficient dispatching strategies, which includes an ordered preference list of ambulances to dispatch.

There are only a few research studies that have been conducted incorporating dispatching strategies to improve the performance of EMS systems. In many EMS systems the existing dispatching policies do not consider the degree of the urgency of the call. Most of the EMS models previously discussed follow the most common dispatching rule: sending the closest available unit to minimize the response time. Carter et al. [10] found that this rule is not always optimal in minimizing the average response time. Their goal was to determine the boundaries for each emergency unit (i.e. the area of each demand zone) in order to minimize the average response time. They proposed a queuing model to represent the emergency medical system with a continuous-time Markov process.

A few research studies have shown that dispatching emergency vehicles according to the degree of the urgency of the call helps to increase the survival probability of patients. For example, Nicholl et al. [35] conducted a case study with the objective of evaluating the safety and reliability of a two priority dispatch systems operated by ambulance service in the UK. They found that priority dispatching systems have the ability to respond quickly to life-threatening calls by focusing resources on these calls, thereby increasing the survival probability of the patients. In addition, they recommended that low priority calls, or non-life threatening calls, should be responded to as soon as possible rather than immediately. Another study was conducted in the Emergency Medical Service in Helsinki, Finland, by Kusima et al. [23] to record pre-hospital death rates in four medical priority categories (most severe to least severe) to evaluate if deaths in lower urgency categories could have been prevented by faster ambulances responses. This community-based cohort study showed that the four-category medical priority dispatching of ambulances helps to maintain a lower pre-hospital mortality in the two lower urgency categories. These studies suggest that priority dispatching plays a key role in saving lives.

EMS literature includes only a few studies incorporating dispatching policies or strategies considering the degree of the urgency of the call. Comparing the dispatching strategies of first-called first-served

(FCFS), nearest-origin assignment, and the flexible assignment strategy, Haghani et al. [19] discussed the benefits using priority dispatching in EMS systems to reduce response time. The objective was to evaluate the performance of these three dispatching strategies considering dynamic travel time information, vehicle diversion, and route changing. Henderson et al. [20] developed a decision support tool for St. John Ambulance Service in New Zealand considering two types of calls: Priority 1 and Priority 2. Priority 1 calls are considered to be the more severe incidents, while Priority 2 calls are less severe incidents. Andersson et al. [1] also investigated the advantage of priority dispatching considering the urgency of the call. Their goal was to relocate ambulances in order to minimize the response time. In addition to dynamic ambulance relocation, they also proposed an algorithm to dispatch ambulances automatically incorporating the severity of the call. They considered three types of priority calls: the most urgent life threatening calls categorized as Priority 1 calls, urgent but not life threatening calls as Priority 2, and non-urgent calls as Priority 3 calls.

Determining the response area for an ambulance or set of ambulances is another important resource allocation decision context in EMS systems [25]. Larson defined this resource allocation problem as a "districting" problem. The districting literature goes back to early 1960's. One of the first study was done by Smith [46] to redesign the police patrol response area in order to minimize the traveling time within districts. Later Gass [15] proposed a heuristic technique for police sectors in Cleveland. A network approach was implemented by Santone et al [41] to evaluate the alternative fire station sites and districts at the National Bureau of Standards. Larson ([25], [27], [26]) and Carter et al. [10] are the other researchers who drew significant amount of attention to the districting problem. A detailed discussion of their work will be given in Chapter 4.

The models discussed thus far, were developed using three approaches: mathematical models, queuing models, and simulation models. However, developing mathematical or queuing models for EMS systems to obtain optimal dispatching strategies using existing tools leads to several issues. Decision making in emergency vehicle dispatching process is a complex real world problem that is filled with uncertainty. Furthermore, the dynamic behavior of the EMS systems leads the dispatchers to make a series of random sequential decisions in each stage and ultimately operate the system optimally. Since the vehicle dispatching problem is dynamic and stochastic in behavior, the use of traditional deterministic optimization modeling to solve EMS dispatching problems is not applicable. Therefore, the Markov Decision Process (MDP) and simulation approaches, which are capable of addressing those issues (dynamic and stochastic behavior), can be used to determine the optimal dispatching decisions in EMS systems. McLay et al. [34] used the MDP approach to model the EMS system and obtain optimal dispatching rules. In our study we also used the MDP

approach to obtain the optimal dispatching strategies. Although the MDP is capable of addressing problems with stochastic behavior, simulation approach has some advantages over the MDP. When the MDP approach is used for real world EMS systems the model formulation becomes complicated; the state space grows dramatically with the complexity (number of ambulances and demand zones managed) of the EMS system.

In contrast, simulation models are easy to use for modeling EMS systems and measuring performances. They also allow us to compare systems under different sets of assumptions, providing the ability to test new operational strategies such as different ambulance locations or dispatching rules. EMS literature includes several simulation models that have been developed and used to evaluate the performance of EMS systems. One early example is that of Savas et al. [42] for ambulance operation in New York City. In this study, a computer-based simulation was developed to conduct a analysis of cost-effectiveness in New York's emergency ambulance service. The objective was to improve the service of the EMS at a low cost. A considerable improvement in average response time has been showed by redistributing the existing ambulances in the district by locating some of them at satellite garages rather than all of them at a hospital. Another simulation model was developed by Henderson et al. [20] as a decision support tool for ambulance dispatching in the Auckland region, New Zealand. Two other simulation models were developed by Andersson et al. [1] and Haghani et al. [19] to study the performance of the different dispatching rules of EMS systems. These examples show the usefulness of simulation in modeling EMS systems. In our research we are also considering a simulation approach to develop EMS systems and study the performance of the proposed dispatching strategies to the system.

The main goal of this study is to maximize the average survival probability of the patients by implementing optimal dispatching strategies and determining the response boundaries for EMS vehicles. In contrast, most EMS systems attempt to improve their performance by minimizing the average response time. To reduce this time, the rule followed by those systems is to dispatch the closest unit, known as the myopic policy, giving no reference to the severity of the emergency call. When attempting to maximize the survival probability of the patient, the myopic policy is not always optimal. For example, in a situation when the dispatching center receives two types of calls from the same region, an urgent call subsequent to a less urgent call, the victim involved in the second incident will be in jeopardy because the closest emergency unit may have been sent to the less serious first call. In such a situation the dispatcher has to send another unit to the more severe incident. This strategy may not be ideal, as the next closest available unit may take more time to arrive at the incident. This dispatching could lead to a decrease in the survival probability of the severe incident due to inability to minimize the response time. In order to address this situation, a system consider-

ing the severity of the emergency has to be developed. Therefore, one objective of this research is to *develop optimal dispatching strategies for EMS systems that will maximize the patient survivability by incorporating the degree of the urgency of the call*. The other objective is to *determine the response boundaries for EMS vehicles that maximize the patients survivability.*

In this study, EMS performance evaluation was conducted by measuring the patient survivability instead the number of calls covered within a given time standard, known as the response time threshold (RTT). Utilizing patient survivability to evaluate performance of EMS system is a more precise measure in terms of number of patients survive. In addition, measuring the patients survivability directly mirrors the patient outcome. The models developed incorporating survival probability that can be found in EMS literature are [13],[33],and [34]. Results of these research studies indicated that incorporating survival function in the models ultimately helped to increase of patients' survivability.

The MDP approach and simulation approaches are used to determine the optimal dispatching strategies for EMS systems. The results of these findings were used to develop a heuristic to determine the improved dispatching strategies in more complex models in order to maximize the patients survival probability. Computational examples were considered to study the performances of the proposed dispatching heuristic. The results indicated that it is not always optimal to dispatch the closest ambulance especially for low priority calls that are non-life threatening. In addition, in this study, an integer programing model and a constructive heuristic are proposed, to determine the response area (or boundaries) for each ambulance that increase the patients survivability in emergencies. Results show that operating ambulance according to boundaries given by the proposed model can help to increase patients' survivability.

Chapter 2

Optimal Dispatching Strategies for Emergency Vehicles to Increase Patient Survivability

2.1 Introduction

The goal of most emergency medical services is to provide medical care for urgent 911 calls, and/or to transport patients to a hospital, and to ultimately save patients' lives. In recent years, demands on EMS systems have increased due to population growth and scarcity of resources. Thus, EMS system planners are interested in improving the performance of EMS system by providing effective and efficient service to customers through improved operations, such as optimal allocation of resources. Locating ambulances is one class of resource allocation problem and a widely used method for improving the performance of EMS systems. Dispatching emergency vehicles is another class of resource allocation problem, which is less studied, for improving EMS system performance. In practice, the optimal strategies are not obvious and the medical literature highlights the need for finding effective ways to dispatch ambulances to patients [40]. In this study our focus is to improve the performance of EMS systems by implementing dispatching strategies to use with currently available resources. In particular, in this paper we present stationary dispatching rules for emergency vehicles. The dispatching rule provides an ordered preference list of ambulances (prioritized list of ambulances) to dispatch for each demand zone which depends on the degree of urgency of the call.

The dispatch center handles each 911 call, where a dispatcher determines the nature and priority of the call and dispatches the appropriate medical unit(s). The nature of the call reflects the type of call (such as motor vehicle accident, trauma, or difficulty breathing). The priority assigned to the call reflects the operator's perception of whether the call is an emergency. Here we assume that calls are classified into two priorities, Priority 1 or life-threatening calls, and Priority 2 or non-life threatening calls. Most frequently, EMS systems have no more than three priority levels, as these classifications need to be made in a matter of seconds. While in this paper we limit our analysis to two priority levels, we can extend the results beyond two levels easily with minor modifications to the model. In addition, survival probability is used as our performance measure of the EMS system as opposed to a response time threshold, since survival probability mirrors patient outcomes directly [13].

The vehicle dispatching problem is dynamic and stochastic in nature; thus the use of traditional deterministic optimization models to solve the EMS dispatching problem is not applicable. Therefore, a Markov decision process (MDP) approach, which is capable of addressing the dynamic and stochastic behavior of the EMS system, is used in our research to obtain the optimal dispatching strategies. We model this problem as a discounted, infinite horizon Markov decision process. The proposed model determines how to optimally dispatch paramedic units (ambulances) in response to emergency 911 calls in order to maximize the patients' survivability.

The optimal policy is compared with the myopic policy of always sending the closest server. This comparison is done using a hypothetical example using data similar to Hanover County, Virginia. Results show that dispatching the closest vehicle is not always optimal and dispatching vehicles considering the priority of the call leads to an increase in the average survival probability of patients. Moreover, it is observed that additional lives can be saved with no extra cost (in terms of paramedic units available) by implementing the optimal dispatching policy suggested by the MDP model. In other words, the optimal dispatching policy is economical.

The remainder of this Chapter is organized as follows. Section 2.2 reviews relevant literature on models used for EMS systems. The MDP model formulation is described in section 2.3. Section 2.4 presents a computational example using data collected from Hanover County, Virginia. Conclusions and future research directions are presented in section 2.5.

2.2 Related Work

The decision making process in EMS systems is a strategic and complex challenge, made even more difficult by the uncertain nature of the system dynamics (e.g. when and where calls originate, length of service, etc.). Several different approaches such as discrete optimization, queuing, and simulation models have been utilized in the context of EMS systems to help the decision maker. The models found in the literature for improving the performance of EMS systems using these approaches are briefly discussed below.

Most of the early models developed for EMS systems dealt with static and deterministic ambulance station location problem. The location set covering problem (LSCP) [49], the maximal covering location problem (MCLP) [11], the tandem equipment allocation model (TEAM) [43] and the double standard model (DSM) [16] are few examples. Probabilistic covering location models were developed considering the fact that emergency vehicles are busy once they respond to a call. A few such model are, the maximum expected covering location model (MEXCLP) [12], the maximal availability location problem (MALP) [38] and a more recent hybrid of these (LR-MEXCLP) [47]. [39, 29, 31] and [9] are the additional extensions of the (MEXCLP) model can be found in EMS literature. Ambulance relocation models were developed to explicitly account for busy ambulances when calculate coverage. In relocation (or redeployment) models, the main focus is to dynamically relocate vehicles in real-time when vehicles are dispatched to the scene. A few such vehicle relocation models are Kolesar et al. [22], Gendreau et al. [17], and Rajagopalan et al. [37]. In another study, Paul et al. [36], provide a case study of how ambulances (and other resources) can be reallocated during a disaster.

The models discussed thus far were developed to locate and relocate ambulances using two approaches, mathematical models and queuing models. There are only a few research studies have investigated dispatching strategies to improve the performance of EMS systems. Since the vehicle dispatching problem is dynamic and stochastic in nature; simulation and MDP approaches are most often used by researchers to address these issues. The EMS literature includes several simulation models that have been developed and used to evaluate the performance of EMS systems. Savas [42], Henderson et al. [20], Andersson et al.[1] and Haghani et al.[19] are researchers who developed simulation approaches for EMS models.

In many EMS systems the existing dispatching policies do not consider the degree of the urgency of the call. Most of the EMS models discussed previously follow the most common dispatching rule, sending the closest available unit. Carter [10] found that this rule is not always optimal in minimizing the average response time. Their goal was to determine the boundaries for each emergency unit (i.e. the area of

each demand zone). Carter developed a queuing model to represent the emergency medical system with a continuous-time Markov decision process. A few research studies have shown that dispatching emergency vehicles considering the degree of the urgency of the call will lead to an increase of survival probability of patients. For example, Nicholl et al. [35] and Kuisma et al. [23]. Thus, in this study we also consider the severity of the call when implementing optimal dispatching rules.

EMS system performance is commonly evaluated by measuring the response time threshold (RTT), meaning measuring the number of calls covered within a given time standard. However, recent evaluation is focused on measuring the survival probability of patients. Utilizing patient survivability can be a more precise measure in terms of number of patients that survive. Erkut [13] and McLay [33] developed models for locating EMS vehicles incorporating survival probability. A discussion of these and other EMS models that aim to improve patient survivability can be found in [32]. Results of these research studies specify that incorporating survival function in the models ultimately helped to increase patients' survivability.

Therefore, the objective of this study is to implement optimal dispatching strategies for EMS systems that maximize patient survival probability by incorporating the degree of the urgency of the call. These dispatching strategies are developed using the MDP approach. In EMS literature a few studies have been conducted using the MDP approach. McLay [34] used an MDP approach to optimally dispatch EMS vehicles. Maxwell et al. [30] proposed approximate dynamic programming for ambulance redeployment. The main differences between our model and Henderson's model can be described as follows: Henderson considered an EMS system with ambulance redeployment. However, we consider an EMS system with fixed deployment, meaning ambulances return back to their home station after serving the calls.

While redeployment has its advantages, there are some practical issues when implementing ambulance redeployment for use with some real world EMS systems. One disadvantage of dynamic relocation models is the necessity of finding a new solution whenever a vehicle is dispatched. This procedure is time consuming since it has to be done frequently. Furthermore, when calls come in quick succession solutions may be infeasible [18]. Another drawback associated with redeployment is, with an increase in the complexity of EMS systems, the number of relocations grows dramatically [1]. Additionally, frequent ambulance relocations will change the route or destination and can lead to a confusion of drivers thereby causing mistakes. Also, to implement redeployment in EMS systems advanced technologies such as CAD and GPS are required. Moreover, it is necessary to incorporate the relocation cost in EMS systems; because each time a relocation occurs the ambulances have to move from one location to another. The operating cost of EMS systems may increase due to relocation. There is a trade off between the benefits of the relocating and the

operational costs [19]. Since there are practical issues in implementing dynamic relocation models with some real world EMS systems, especially those with few or limited resources, in our study an EMS system with fixed deployment is considered instead (where paramedic units are located at specific stations, respond from stations, and return back to their home station after serving a call). The main focus is to improve the EMS system performance by implementing efficient dispatching strategies.

McLay [34] do consider an EMS system with fixed deployment in their model. The difference between this work and our work can be described in the following manner. The objective of their model was to maximize the coverage level (which reflects the proportion of calls covered within a given time standard). In our model, however, the objective is to maximize the patient survival probability. Furthermore, their optimal dispatching policy is not always a priority list. The optimal ambulance to dispatch depends on the locations to which the busy ambulances have been dispatched. In our study we propose an ordered list (priority list) of ambulances to dispatch depending on call severity. Our approach is easy to implement in EMS systems since we consider only the call severity not the location of the busy ambulances. Furthermore, McLay [34] showed that their dispatching policy in most cases does follow a priority list of ambulances; thus, implementing a priority list dispatching strategy in EMS system is reasonable and constraining the decision space ex-ante is computationally less expensive.

2.3 MDP Model Description and Parameters

Consider an EMS system with i demand zones $(i = 1, ..., n)$ and j ambulance stations $(j = 1, ..., m)$ with an ambulance at each station ($n \times m$ case). Here demand is considered to be the calls requesting paramedic units from the EMS system. In this model we assumed that locations of all demand zones and ambulance stations are known. A 911 emergency medical service starts with a call to the dispatching center requesting an ambulance. An arriving call is of one of two types: either Priority 1 or Priority 2. Priority 1 calls are considered to be life-threatening while Priority 2 calls are non-life threatening calls. The time between the arrival of the call and the time the first ambulance is dispatched to the scene is known as *preparation time*. The time between the receipt of a call at the dispatch center and the arrival of the first emergency response vehicle at the scene is known as the *response time*. *Service time* can be defined as the time required for an ambulance to return back to its original station after leaving the station to attend the call, which includes the transportation time of the patient to the hospital if needed. The assumptions we made when developing this model are explained below:

- We consider two types of calls: Priority 1 and Priority 2.
- Any type of emergency (either life-threatening or non-life threatening) can be handled by any ambulance.
- Only one paramedic unit is dispatched to each call.
- Inter-arrival times of calls and service times are exponentially distributed.
- Service time is independent of the call priority. This assumption can be easily relaxed.
- When a call arrives, if a paramedic unit is available, then it must be dispatched.
- When a call arrives if all paramedic units are busy, that calls is served by outside resources (e.g. an ambulance from a neighboring county).

Input parameters for the model are summarized below;

- λ = call arrival rate (to the entire system)
- n = total number of demand zones
- m = total number of paramedic units, each at a fixed location
- z_i= proportion of calls from i^{th} demand zone: such that $\sum_{i=1}^{n} z_i = 1$
- p_i^k= proportion of priority k calls from demand zone i (where priority denotes severity) : such that $\sum_{k=1}^{2} p_i^k = 1, \forall\ i$
- $F_{ij}(t_R)$= response time distribution for ambulance j for zone i with Mean Response Time (MRT) of x_{ij} and standard deviation of σ_{ij}.
- μ_{ij}= average service rate by ambulance j for zone i.
- λ_i= λz_i (call arrival rate from demand zone i)
- C_{ij}^k = Reward if ambulance j responds to zone i, for call of priority k. (This reward depends on the response time and priority of the call.)

EMS performance evaluation is commonly done by measuring the number of calls covered within a given time standard, known as the response time threshold (RTT). However, later research has suggested that using patient survivability is a more precise measure [13] in terms of the expected number of survivors in case of emergencies. In addition, measuring the patients survivability directly mirrors the patient outcome. Erkut [13] showed that incorporating a survival function in the EMS location problem helped to save more lives. Therefore, to get a precise estimate of the EMS system performance, survival probability will be used as the performance measure in this study. However, other performance measures can be used without changing the model by adjusting the reward parameter. Although several studies have been conducted to determine the relationship between response time and survival probability (eg. [7]), the one considered in this study is based on Larsen et al. [24] and subsequently simplified by McLay et al. [33]. We apply their survival function directly to calculate the survival probability of the patient since as a function of the response time. The survival function is explained below. Let S denote the patient survival probability, then

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

where t_R represents the response time. We used this survival function when calculating rewards (C_{ij}^k) in the MDP model. The reward calculation is discussed in detail in the next section, where the MDP Model formulation is presented.

2.3.1 MDP Model Formulation

Here we present the MDP model formulation to determine optimal dispatching strategies for EMS systems. The objective of the MDP model is to optimally determine which ambulance to dispatch for arriving calls in order to maximize the average reward of responding to life-threatening calls (Priority 1 calls). The reward is considered to be the survival probability of patients as mentioned earlier. In this MDP approach we assumed that calls arrive to the EMS system requesting paramedic units according to a Poisson process with rate λ. The MDP model formulation is described below:

States

The state $\mathbf{s}(t)$ describe the status of each ambulance in the system at time t. Therefore, the state $\mathbf{s}(t)$ is the vector, $\mathbf{s}(t) = \{s_1(t), s_2(t), s_3(t), \ldots, s_m(t)\}$, where $s_j(t)$ describes the status of ambulance j.

$$s_j(t) = \begin{cases} i & \text{if ambulance } j \text{ is serving a call originating at zone } i \text{ at time } t; \\ 0 & \text{if ambulance } j \text{ is idle at time } t. \end{cases}$$

The resulting state space of $(n+1)^m$ states, can be described as follows:

State space $\mathbf{S} = \{\mathbf{s}(t) : s_j(t) \in \{0, 1, 2, \ldots, n\},\ , j = 1, 2, \ldots, m\}$

Decisions

In the MDP model the decision is to determine which ambulance to dispatch at each state. We assume that when calls arrives to the system one of the available ambulances must be dispatched to the incident. Let $U(\mathbf{s}(t)) = \{j \in \{1, 2, ..., m\} : s_j = 0\}$, denote the set of available decisions when a customer arrives at time t and the state is $\mathbf{s}(t)$. We define $u(\mathbf{s}(t))$ as the optimal decision when the state of the EMS system is $\mathbf{s}(t)$, where $u(\mathbf{s}(t)) \in U(\mathbf{s}(t))$. In our model, since we are interested in obtaining a priority list of ambulances to dispatch, the set of available decisions ($U(\mathbf{s}(t))$) are restricted so that the decision does not depend on the location of the busy ambulance. For example, consider an EMS system with three ambulances and two demands zones. When ambulance 1 is busy, the optimal decision is restricted such that $u(\mathbf{s}(t)) = u(i, 0, 0); i = 1, 2$. In other words, when ambulance 1 is busy, whether ambulance 2 or 3 is assigned to the next incoming call is independent of which location ambulance 1 is busy serving.

Rewards(C_{ij}^k)

If ambulance j is dispatched to a call at zone i of Priority k, then a fixed reward C_{ij}^k is received. In this model, the reward is considered to be the survival probability of the patients. These rewards depend on the type of the call, the decision made in each stage (dispatching ambulance) and the location of the call. Reward calculation for Priority 1 calls is explained below.

Let the probability of survival be S_{ij}, if ambulance j responds to a call from demand zone i. S_{ij} can be obtained by solving the following equation.

$$S_{ij} = \int_0^\infty S(t_R) dF_{ij}(t_R) \tag{2.2}$$

where t_R denotes the response time, F_{ij} is the cumulative distribution function for the response times, and

$S(t_R)$ is the survival function explained in Equation 2.1. Using this calculation, the resulting reward for Priority 1 calls is $C_{ij}^1 = S_{ij}$. For Priority 2 calls reward is set to be zero (this assumption can be easily modified), because in this model our goal is to maximize the survival probability of life-threatening calls (Priority 1 calls). Therefore, setting the reward of Priority 2 calls to zero does not affect the decision made at each stage since those are non-life threatening calls. Thus, we can summarize the rewards as follows:

$$C_{ij}^k = \begin{cases} S_{ij} & \text{if } k = 1; \\ 0 & \text{if } k = 2\,. \end{cases}$$

Thus far we have presented thee EMS vehicle dispatching problem as a continuous-time MDP. Next we use the uniformization approach to convert this problem into a discrete time equivalent MDP (see for example [5] Chapter 5.1). We define the uniformization rate ν as follows:

$$\nu = \lambda + \textstyle\sum_{j=1}^{m} \delta_j, \text{ where } \delta_j = \max_{i=1,2,\ldots,n} \{\mu_{ij}\}.$$

Based on this uniformization rate the CTMC is equivalent to a DTMC with discount factor $\theta = \frac{\nu}{\nu+\beta}$, where β is the continuous rate of discount and $\beta \in (0, 1)$.

The value iteration algorithm was used to determine the optimal value function. Let the value function $J_n(\mathbf{s})$ be the optimal finite horizon discounted reward with n periods left to go starting in state $\mathbf{s} = (s_1, s_2, ..., s_m)$. The infinite horizon discounted reward can be approximated using the limit of the finite horizon problem starting with $J_0(\mathbf{s}) = 0,\ \forall\ s \in S$. We used the optimality equation (recursive equation) defined by McLay et al. [34] (equation 4, page 31) to obtain the optimal dispatching policy. We modified their optimality equation according to our notation and assumptions as follows:

$$\begin{aligned} J_{n+1}(\mathbf{s}) =& \frac{1}{\beta+\nu}\Bigg[\sum_{j=1}^{m} I_{\{s_j=i|i>0\}}\mu_{ij} J_n(s_1, s_2, \ldots, s_{j-1}, 0, s_{j+1}, \ldots, s_m) \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{2} \lambda_i p_i^k \max_{j\in U(s)} \left\{ I_{\{s_j=0\}} J_n(s_1, s_2, \ldots, s_{j-1}, i, s_{j+1}, \ldots, s_m) + (\beta+\nu) C_{ij}^k \right\} \\ &+ (\nu - \lambda - \sum_{j=1}^{m} I_{\{s_j=i|i>0\}}\mu_{ij}) J_n(\mathbf{s}) \Bigg] \end{aligned} \tag{2.3}$$

where,

$I_{\{s_j=i|i>0\}}$ = indicator variable that denotes ambulance j is serving a patient at zone i

$I_{\{s_j=0\}}$ = indicator variable that denotes if ambulance j is available.

$U(\mathbf{s})$ = the set of available decisions in state $\mathbf{s}$.

Note that in [34], $U(\mathbf{s})$ depends on the location of the busy ambulance. However, in our model $U(\mathbf{s})$ is restricted, as explained previously, to obtain a priority list of ambulances to dispatch. We can further express this decision difference as follows. Recall that the state is defined as $\mathbf{s} = (s_1, s_2, ..., s_m)$. Suppose $\mathbf{s}$ is such that $s_{j^*} = 0$ for some $j^* = 1, 2, ..., m$ and $s_j \geq 1$ for some $j = 1, 2, ..., m$. In our model the decision depends only on the location of free ambulances (location of j^* where $s_{j^*} = 0$). In the model presented by McLay et al. [34] the decision depends not only on the free ambulance locations but also on the location of busy ambulances (location of j where $s_j \geq 1$).

The first term of the Equation 2.3 describes busy ambulances becoming available. The second term describes new calls arriving to the system, where $U(\mathbf{s})$ denotes the available decisions in state $\mathbf{s}$. The third term describes the EMS system remaining in the same state (no new call arrives to the system and no ambulance becomes available). In the next section the MDP is applied to a scenario using data (such as response times and service times) similar to Hanover County, Virginia to obtain optimal dispatching policies for paramedic units.

2.4 Computational Examples

2.4.1 Two demand zones and two ambulances

An illustrative example is discussed to study the behavior of the optimal dispatching strategy given by the MDP model. This hypothetical example was constructed using data such as response time distributions and service time distributions, similar to that of Hanover County, Virginia Fire an EMS. For simplicity, this example (hereby referred to as the 2×2 case) assumed an EMS system with two demand zones (i.e. the zones which are requesting emergency vehicles) and two ambulance stations (i.e. the locations that the ambulances are sited) with one paramedic unit (ambulance) at each station. The paramedic units at each station were labeled Ambulance 1 and Ambulance 2, sited at station 1 and station 2 respectively. Calls arrive according to a Poisson process with rate $\lambda = 1$ per hour to the entire system. The probability of receiving a Priority 1 or

a Priority 2 call is equally likely from either of the demand zones. i.e. $p_1^1 = p_1^2 = p_2^1 = p_2^2 = 0.5$. Response times and service times are assumed to be lognormally and exponentially distributed respectively, since those are the best fit with data collected from Hanover County. Tables 2.1 and 2.2 summarize the response times and service times for each demand zone.

Zone	*Ambulance 1*	*Ambulance 2*
zone 1	logn(9.07,4.19)	logn(14.03,6.48)
zone 2	logn(14.03,6.48)	logn(9.02,6.48)

Table 2.1: Response Time Distributions-2×2 case

Zone	*Ambulance 1*	*Ambulance 2*
zone 1	expo(60)	expo(65)
zone 2	expo(75)	expo(65)

Table 2.2: Service Time Distributions-2×2 case

Position $(1, 1)$ of Table 2.1 shows that the response time (the time between the receipt of a call at the dispatch center and the arrival of the first emergency response vehicle at the scene) distribution for Ambulance 1 to demand zone 1 is lognormally distributed with mean of 9.07 minutes and standard deviation of 4.19 minutes. Similarly, other positions represent the corresponding response time distributions for each ambulance to each demand zone. Service times, the time between the receipt of a call at the dispatching center and the time vehicle returns to the ambulance station after serving the incident, are considered as in Table 2.2. For example, position (1,1) of Table 2.2 shows that the service time for Ambulance 1 to demand zone 1 is exponentially distributed with mean of 60 minutes. Modeling the hypothetical example using the MDP presented in Section 3.1 is seen below:

Since the state **s** describes the status of each ambulance in the system, the state **s** is the vector, $\mathbf{s} = (s_1, s_2)$, where s_1 is the status of ambulance 1 and s_2 is the status of ambulance 2. The status of an ambulance is 0 if it is idle at its home station, 1 if it is busy serving a call from zone 1, and 2 if it is busy serving a call from zone 2. The resulting state space of nine states, can be described as follows;

$$S = \{(0,0), (1,0), (0,1), (2,0), (0,2), (1,2), (2,1), (1,1), (2,2)\}$$

The state $(0,0)$ implies both ambulances are idle, $(1,0)$ implies Ambulance 1 is serving a call at zone 1 and Ambulance 2 is idle, and so on. In this example a decision occurs when both ambulances are available. Therefore, the decision is to either dispatch Ambulance 1 or Ambulance 2 when the state of the system is (0,0). In other words, decision $u((0,0))$ is either dispatch Ambulance 1 or Ambulance 2. Here we have only this decision because we assume that the optimal policy is a priority list of ambulance to dispatch.

If ambulance j responds to zone i, then a fixed reward (C_{ij}^k) is received. For Priority 1 ($k=1$) calls the reward is obtained using Equation 2.1 and Equation 2.2 as explained above. Resulting rewards (survival probabilities) for Priority 1 calls can be summarized as in Table 2.3. For example, position (1,1) of Table 2.3 indicates that, if Ambulance 1 responds to a call from zone 1, a fixed reward of 0.15 is received. As we mentioned earlier for Priority 2 calls reward is set to be zero.

Zone	*Ambulance 1*	*Ambulance 2*
zone 1	0.15	0.05
zone 2	0.05	0.10

Table 2.3: Rewards- 2×2 case

Denoting $J(\mathbf{s})$ as the reward function when the state of the system is $\mathbf{s}$, we can write the recursive formula as show in Equation 2.3.

$$\begin{aligned} J_n(0,0) =& \lambda_1 p_1^1 \ \max\left(\left[S_{11}+\frac{J_{n-1}(1,0)}{\beta+\nu}\right],\left[S_{12}+\frac{J_{n-1}(2,0)}{\beta+\nu}\right]\right) \\ &+\lambda_2 p_1^2 \ \max\left(\left[S_{22}+\frac{J_{n-1}(0,2)}{\beta+\nu}\right],\left[S_{21}+\frac{J_{n-1}(0,1)}{\beta+\nu}\right]\right) \\ &+\lambda_1 p_2^2 \max\left(\frac{J_{n-1}(1,0)}{\beta+\nu},\frac{J_{n-1}(0,1)}{\beta+\nu}\right)+ \ \lambda_2 p_2^2 \ \max\left(\frac{J_{n-1}(2,0)}{\beta+\nu},\frac{J_{n-1}(0,2)}{\beta+\nu}\right) \\ &+(\nu-\lambda_1-\lambda_2)\frac{J_{n-1}(0,0)}{\beta+\nu}; \end{aligned} \tag{2.4}$$

$$J_n(k,0) = \lambda_1 p_1^1 S_{12} + \lambda_1 \frac{J_{n-1}(k,1)}{\beta+\nu} + \lambda_2 p_1^2 S_{22} + \lambda_2 \frac{J_{n-1}(k,2)}{\beta+\nu} + \mu_{k1} \frac{J_{n-1}(0,0)}{\beta+\nu} + (\nu - \lambda_1 - \lambda_2 - \mu_{k1}) \frac{J_{n-1}(k,0)}{\beta+\nu} \quad ; \text{where } k = 1,2; \tag{2.5}$$

$$J_n(0,l) = \lambda_1 p_1^1 S_{11} + \lambda_1 \frac{J_{n-1}(1,l)}{\beta+\nu} + \lambda_2 p_1^2 S_{21} + \lambda_2 \frac{J_{n-1}(2,l)}{\beta+\nu} + \mu_{l2} \frac{J_{n-1}(0,0)}{\beta+\nu} + (\nu - \lambda_1 - \lambda_2 - \mu_{l2}) \frac{J_{n-1}(0,l)}{\beta+\nu} \quad ; \text{where } l = 1,2; \tag{2.6}$$

$$J_n(k,l) = \mu_{k1} \frac{J_{n-1}(0,l)}{\beta+\nu} + \mu_{k2} \frac{J_{n-1}(k,0)}{\beta+\nu} + (\nu - \mu_{k1} - \mu_{l2}) \frac{J_{n-1}(k,l)}{\beta+\nu} \quad ; \text{where } k = 1,2; \text{ and } l = 1,2. \tag{2.7}$$

The value iteration algorithm was implemented in MATLAB. In addition, a MATLAB program was developed to determine steady-state and survival probabilities. The computational time for each program was less than 1 second. All programs were executed on Dell Vostro 1400 computer with a Pentium-IV processor and 2 GB RAM.

To study the effect of the geographic dispersion of demand on the optimal dispatching strategy, the call volume between the two demand zones is varied such that $z_1 + z_2 = 1$, resulting in $1/10 \leq \lambda_1/\lambda_2 \leq 10$. The value iteration algorithm is applied to solve for the value function $J_n(s)$ and to obtain the optimal decision in the long run. The results indicated that it is always optimal to dispatch the closest unit for Priority 1 calls (not only for this example but also for all other examples tested). Interestingly, this result is in contrast to the findings of McLay et al. [34], who showed that under different settings it is not always best to dispatch the closest ambulance to a Priority 1 call. We conjecture tat the reason for the disparity in our results is that we restrict our attention to policies that follow a priority list; whereas in McLay et al. [34] the policy depends on the location of the busy ambulances.

The optimal order of dispatching ambulances for Priority 2 calls is shown in Table 2.4 for this hypothetical example. According to the table when $z_1 = 0.1$ (i.e., probability of requesting a paramedic vehicle for demand zone 1 is 0.1), the optimal order of dispatching paramedic vehicles to demand zone 1 is: Am-

bulance 1 is first choice and Ambulance 2 is the second choice. To demand zone 2, Ambulance 1 is the first choice and Ambulance 2 is second. Table 2.4 shows the optimal order of dispatching units for Priority 2 calls along with the closest dispatch order for each corresponding z_1 value.

		Closest Policy				Optimal Policy-by MDP			
		order zone1		order zone2		order zone1		order zone2	
z_1	$log_{10}(\lambda_1/\lambda_2)$	choi 1	choi 2	choi 1	choi 2	choi 1	choi 2	choi 1	choi 2
0.1	-0.95	1	2	2	1	1	2	1	2
0.2	-0.60	1	2	2	1	1	2	1	2
0.3	-0.36	1	2	2	1	1	2	1	2
o.4	-0.17	1	2	2	1	2	1	2	1
0.5	0	1	2	2	1	2	1	2	1
0.6	0.17	1	2	2	1	2	1	2	1
0.7	0.36	1	2	2	1	2	1	2	1
0.8	0.60	1	2	2	1	2	1	2	1
0.9	0.95	1	2	2	1	2	1	2	1

Table 2.4: Comparison of order of dispatching ambulances for Priority 2 calls- Closest Policy and Optimal Policy by MDP

The steady-state or long-run probabilities of each state were calculated to analytically determine the optimal dispatching polices and to calculate the long-run average survival probability of life-threatening patients in this example. This analytical approach allows us to enumerate all possible orders of dispatching ambulances and to study the performance of each dispatching order. However, this approach becomes unmanageable when the problem size increases (in terms of number of demand zones, number of ambulances, and number of call types) because there are $(m!)^{kn}$ possible dispatching orders, if we consider an EMS system with n demand zones, m ambulances, and k types of calls. This calculation can be described as fallows: There are $m!$ possible dispatching orders for each district when we consider only one type of call. Since there are k types of calls, there exists $(m!)^k$ possible dispatching orders for each district. Thus, for n districts, there are $(m!)^{kn}$ total possible dispatching orders. As such, for this hypothetical example there are 16 possible cases. We calculated the long-run average survival probability for these 16 cases and determined the optimal order of dispatching ambulances. The analytical results confirmed the MDP policy, and indicated that it is optimal to send the closest unit to the Priority 1 calls.

By examining the optimal dispatching order (see Table 2.4), it is suggested that the optimal policy is likely to reserve the closest ambulance to serve the zone with higher call arrival rate for Priority 1 calls. For example, Ambulance 2 will not be deployed when $z_1 = 0.1$ (i.e. the higher customer arrival rate comes from

zone 2, since $z_2 = 0.9$) because it is reserved for Priority 1 calls and it is the closest paramedic unit for zone 2. Ambulance 1, on the other hand, is dispatched for Priority 2 calls from demand zone 2 according to the optimal policy if both ambulances are idle. We obtained the long-run average survival probability of patients to compare the two dispatching strategies; myopic and optimal. The calculation of long-run average survival probabilities are described below. Let $\pi_{\mathbf{s}}$ be the steady-state probabilities of each state. Then the calculation of the survival probability of closest dispatching policy can be expressed as follows:

Long-run probability that an incoming Priority 1 call survives (Conditional survival probability) when the closest ambulance is dispatched =

$$\frac{\lambda_1 p_1^1[S_{11}\pi_{(0,0)} + S_{12}\sum_{k=1}^{2}\pi_{(k,0)} + S_{11}\sum_{l=1}^{2}\pi_{(0,l)}] + \lambda_2 p_2^1[S_{22}\pi_{(0,0)} + S_{22}\sum_{k=1}^{2}\pi_{(k,0)} + S_{21}\sum_{l=1}^{2}\pi_{(0,l)}]}{\lambda_1 p_1^1 + \lambda_2 p_2^1} \tag{2.8}$$

To compare the performance of two dispatching strategies, closest and optimal, the proportion of call from demand zone 1 is varied from 0 to 1 while maintaining $z_1 + z_2 = 1$. Figure 2.1 graphs $log_{10}(\frac{\lambda_1}{\lambda_2})$ against $P(survival)$. As this figure shows, the conditional survival probability increases when the ambulances are dispatched according to the optimal policy rather than always sending the closest unit. The objective value (i.e. the survival probability) difference between the two dispatching strategies is greatest when call volume is not balanced between two demand zones. In other words, it is suggested that dispatching ambulances under the optimal policy as opposed to a myopic policy is most beneficial, in terms of patients' survival, when call arrival rate is unbalanced between demand zones. Although the absolute difference in average survival probabilities between the two policies seems low, this translates to a large increase in the expected number of patients who can survive with no additional cost (in terms of available paramedic units). For example, when $z_1 = 0.9$, the survival probability of patients can increase by 5.63 % compared to the myopic policy of sending the closest ambulance. Thus, if it is assumed that this system receives 1000 Priority 1 calls per year; approximately an additional 56 lives can be saved by following the optimal dispatching rule with available resources.

We would also like to study the impact of the optimal dispatching rule on other performance measures of interest. For example, Carter [10] showed that work load of paramedic units is a key factor associated with EMS system performance. Thus, we studied the work load of each ambulance in this hypothetical example. Figure 2.2 compares the proportion of time that the ambulances are busy for the two dispatching

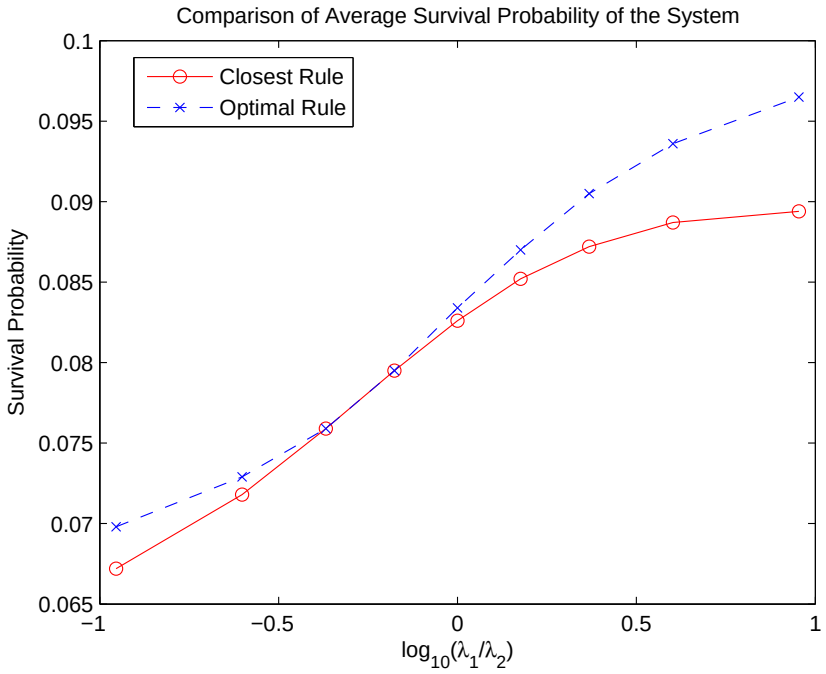


Figure 2.1: Comparison of Survival Probability for Two dispatching strategies

policies for the hypothetical example (2×2 case). As this figure illustrates, when ambulances are dispatched according to the closest dispatching rule, the ambulance busy probability continuously decreases or increases with respect to the call volume at the zones they are closest to. That is, in this example, for the myopic policy Ambulance 1, which is closest to zone 1, become more busy as z_1 increases, while Ambulance 2, which is closest to zone 2, becomes more busy as z_2 increase. Thus, by following a myopic policy, ambulance utilization is unbalanced when demand call volume is unbalanced. For the optimal dispatching strategy, the utilization of ambulances was smoothly distributed between two ambulances. In other words, the optimal policy tends to balance the work load between the two ambulances. In addition, it was observed that the order of sending paramedic units to each demand zone depended on the ambulance busy proportion. According to the MDP model, the optimal dispatching rule tends to send the less busy unit for Priority 2 calls.

2.4.2 Three demand zones and two ambulances

An EMS system with three demand zones and two ambulances (3×2 case) is considered to further illustrate the nature of the optimal dispatching policy. The input data for this hypothetical example is consid-

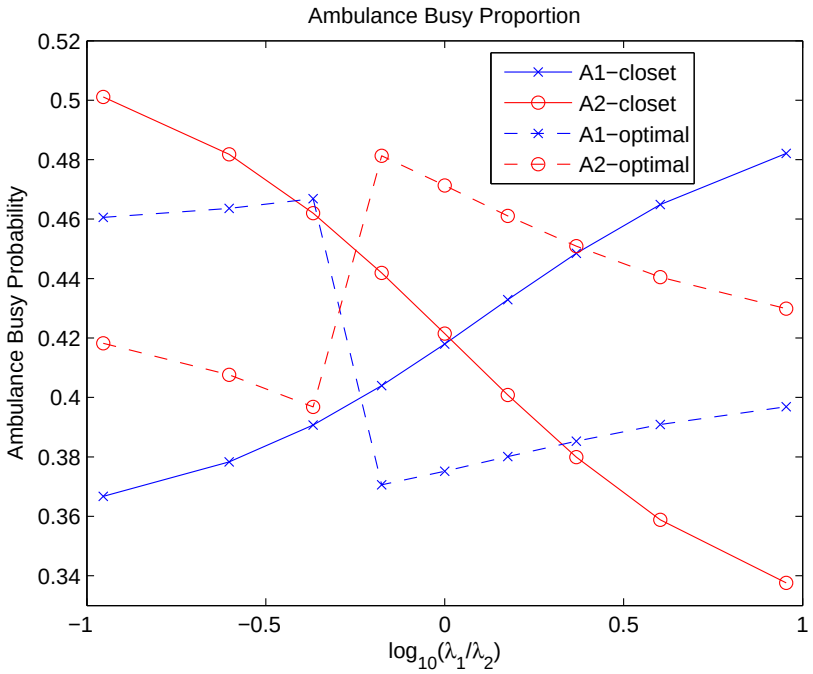


Figure 2.2: Comparison of Busy Probability of Each Ambulance for Two Dispatching Strategies - 2×2 case

ered as follows, Table 2.5 summarizes the response times while Table 2.6 summarizes service times for each demand zone.

Zone	*Ambulance 1*	*Ambulance 2*
zone 1	logn(9.07,4.19)	logn(14.03,6.48)
zone 2	logn(14.03,6.48)	logn(10.92,5.05)
zone 3	logn(8.03,4.19)	logn(9.07,6.48)

Table 2.5: Response Time Distributions- 3×2 case

We assumed that calls arrive according to a Poisson process with rate $\lambda = 1$ per hour to the entire system and probability of receiving a Priority 1 or a Priority 2 call is equally likely from any demand zone. The corresponding rewards for this example are given in Table 2.7.

Table 2.8 provides the order of dispatching ambulances according to the closest dispatching policy for this EMS system. For example, according to the table, the closest order of dispatching paramedic vehicles to demand zone 1 is: Ambulance 1 is first choice and Ambulance 2 is the second choice regardless of the call

Zone	*Ambulance 1*	*Ambulance 2*
Zone 1	expo(60)	expo(65)
Zone 2	expo(75)	expo(65)
Zone 3	expo(50)	expo(55)

Table 2.6: Service Time Distributions- 3×2 case

Zone	*Ambulance 1*	*Ambulance 2*
Zone 1	0.15	0.05
Zone 2	0.05	0.10
Zone 3	0.20	0.15

Table 2.7: Rewards- 3×2 case

arrival rate from any demand zone.

Order zone1		Order zone2		Order zone3	
choice 1	choice 2	choice 1	choice 2	choice 1	choice 2
1	2	2	1	1	2

Table 2.8: Closest Policy - 3×2 case

In this example, first we set the proportion of calls from demand zone 1 to 0.1 ($z_1 = 0.1$) and varied the call volume between demand zone 2 and demand zone 3 such that $z_1 + z_2 + z_3 = 1$, resulting $1/8 \leq \lambda_2/\lambda_3 \leq 8$ to study the effect of the geographic dispersion of demands. MDP model suggested that, it is optimal to send the closet unit for Priority 1 calls always. The optimal order of dispatching ambulances for Priority 2 calls is given in Table 2.9. Comparing Table 2.8 and Table 2.9, we can conclude that it is not optimal to dispatch the closet unit always for Priority 2 calls. The comparison of survival probability for two dispatching strategies, closest and optimal is depicted in Figure 2.3. A comparison of ambulance busy probabilities is depicted in Figure 2.4. Results of these comparisons are similar to the observations we obtained previously in 2×2 case example.

By observing the optimal dispatching order for Priority 2 calls, we can say that the optimal policy tends to reserve the closest ambulance to serve the demand zone with higher rewards and arrival rate. For instance, consider the case when $z_1 = 0.1$, $z_2 = 0.1$ and $z_3 = 0.8$. Demand zone 3 has the higher arrival rate and higher reward. Ambulance 1 is the closest paramedic unit for demand zone 3. Thus, Ambulance 1 is reserved to respond to Priority 1 calls from demand zone 3 and dispatch the Ambulance 2 for Priority 2 calls

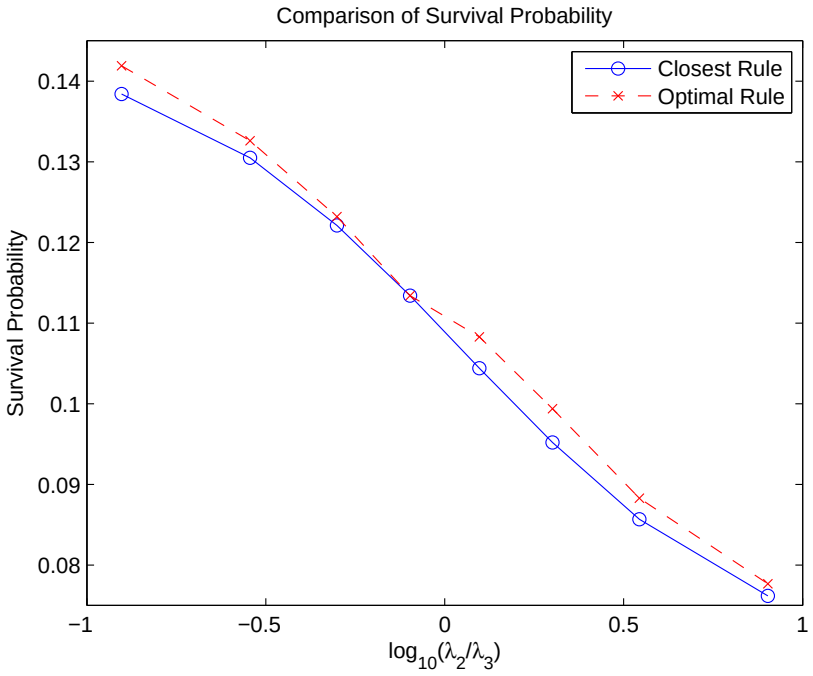


Figure 2.3: Comparison of Survival Probability of Two Dispatching Strategies - 3×2 case

instead (if both ambulances are available to dispatch). This dispatching order is vise versa when $z_1 = 0.1$, $z_2 = 0.8$ and $z_3 = 0.1$. Because in this instance, higher call arrival rate occurs from demand zone 2 and higher reward is given by Ambulance 2. Thus, Ambulance 2 is reserved to respond to Priority 1 calls from zone 3 and Ambulance 1 is dispatched to Priority 2 calls (if both ambulances are available).

		Optimal Policy-by MDP					
		Order zone1		Order zone2		Order zone3	
z_2	$log_{10}(\lambda_2/\lambda_3)$	choice 1	choice 2	choice 1	choice 2	choice 1	choice 2
0.1	-0.90	2	1	2	1	2	1
0.2	-0.54	2	1	2	1	2	1
0.3	-0.30	2	1	2	1	2	1
0.4	-0.09	1	2	2	1	1	2
0.5	0.09	1	2	1	2	1	2
0.6	0.30	1	2	1	2	1	2
0.7	0.54	1	2	1	2	1	2
0.8	0.90	1	2	1	2	1	2

Table 2.9: Optimal Policy for Priority 2 Calls - 3×2 case, when $z_1 = 0.1$

This observation is also confirmed by the next hypothetical example. In this instance we set the

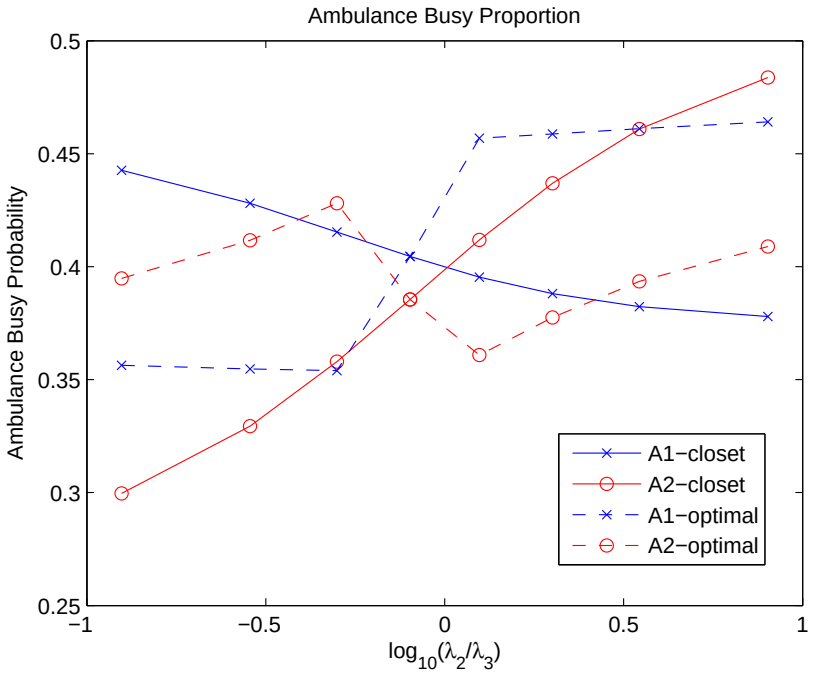


Figure 2.4: Comparison of Busy Probability of Each Ambulance for Two Dispatching Strategies - 3×2 case

proportion of calls from demand zone 2 to 0.1 ($z_2 = 0.1$) and varied the call volume between demand zone 1 and demand zone 3 such that $z_1 + z_2 + z_3 = 1$, resulting $1/8 \leq \lambda_1/\lambda_3 \leq 8$. The MDP results show that it is optimal to dispatch the closest available unit for Priority 1 calls. The optimal order for Priority 2 calls is given in Table 2.10. Results confirm that the optimal policy is likely to reserve the closest ambulance to serve the zone with higher call arrival rate and higher reward.

		Optimal Policy-by MDP					
		Order zone1		Order zone2		Order zone3	
z_1	$log_{10}(\lambda_1/\lambda_3)$	choice 1	choice 2	choice 1	choice 2	choice 1	choice 2
0.1	-0.90	2	1	2	1	2	1
0.2	-0.54	2	1	2	1	2	1
0.3	-0.30	2	1	2	1	2	1
0.4	-0.09	2	1	2	1	2	1
0.5	0.09	2	1	2	1	2	1
0.6	0.30	2	1	2	1	2	1
0.7	0.54	2	1	2	1	2	1
0.8	0.90	2	1	2	1	2	1

Table 2.10: Optimal Policy for Priority 2 Calls - 3×2 case, when $z_2 = 0.1$

2.5 Conclusions and Future Research

This Chapter explains the use of a Markov decision process approach for determining optimal dispatching strategies for EMS systems. A discounted, infinite horizon Markov decision process model is developed and analyzed to obtain optimal dispatching policies. In this model calls are prioritized according to the severity of the call. The results show that dispatching ambulances considering the degree of urgency of the call will lead to an increase in the average survival probability of patients. It is also observed that many lives can be saved at no additional cost (in terms of paramedic units available) by following the optimal policy. Further, the results show that the optimal policy is likely to balance the work load between paramedic units.

We compared the myopic policy of always sending the closest ambulance with the optimal policy given by the MDP model. Results indicate that it is always optimal to dispatch the closest ambulance for Priority 1 patients. The optimal policy for Priority 1 calls is intuitive, since faster response times increase patient survival probability. For Priority 2 calls, a priority list of ambulances to dispatch is obtained using the model. The proposed dispatching rule is easy to implement in EMS systems since a priority list of ambulances to dispatch depends only on the location and degree of urgency of the call and not on the location of all busy ambulances. Thus, obtaining priority list heuristic policies for EMS systems will be a vital area for future research.

This MDP approach allows us to address the stochastic behavior of the EMS system. Additionally, the running time for our MDP formulation in MATLAB is not significant. One potential drawback is that the formulation of the dynamic programming model is complicated when problem size increases. For example, if we consider an EMS system with m ambulances and n demand zones, then the total number of states become $(n+1)^m$. As such, if we try to apply this dynamic programming approach for larger problems we will face the curse of dimensionality. However, in the future a simulation approach can be used to overcome this drawback of the MDP approach, and this is an area of future research that we are currently pursuing. Simulation can also help to alleviate some other potential drawbacks associated with an MDP approach, namely the assumption of exponential service times and of zero-length queue.

We believe the parsimonious model we presented here provides unique contribution in that it shows that it is possible to achieve significant improvements, in terms of lives saved, at little cost by considering the degree of urgency of the call. Furthermore, this can be achieved even with a simple policy: send the closest ambulance to Priority 1 calls, follow an ordered preference list for Priority 2 calls. This should be easy to implement in practice as ordered preference lists are already widely accepted policy types in

EMS systems. Lastly, we observe that applying the optimal policy is most beneficial when the demand is imbalanced between zones. Interestingly, the optimal policy tends to balance the workload as compared to the myopic policy of always sending the closets unit. This is an important observation as workload imbalance resulting from a myopic policy is also the greatest when demand is imbalanced between zones.

Finally, the methodology presented in this Chapter can be extended to consider multiple types of vehicles, multiple patient categories etc. In addition, this methodology can be applied to other problems such as dispatching police cars and fire engines and military deployment.

Chapter 3

Priority Dispatching Strategies for EMS Systems

3.1 Introduction

The primary objective of an emergency medical service (EMS) system is to save lives and to minimize the effect of an emergency health incident. This goal can be achieved by providing adequate and timely paramedic support to the scene. There is a well documented correlation between the response time, the time between the receipt of a call at the dispatch center and the arrival of the first emergency response vehicle at the scene, and the survival probability of the patient ([48], [13]). Hence, the response time is vital in minimizing the impact of the incident. Thus, the dispatcher must consider the location of the available vehicles. In addition, the degree of the urgency of the call is an important factor in dispatching paramedic units, as some types of emergencies may be more time critical (e.g. heart attack), than others (e.g. broken leg).

As this discussion suggests, it is vital to use an efficient dispatching strategy to increase the survivability of patients in an emergency and thereby improve the performance of the EMS system. While a significant amount of research has been conducted related to improving the efficiency of the EMS system, most models focus on the decision context of locating vehicles where the outcome of interest is coverage to an area (e.g. [49], [11], [16], [8]), where coverage is defined as the number of demand points that can be reached by paramedic units within a given time standard. In most of these models, the implicit dispatching rule is to send the closest unit to the scene regardless of the severity of the call. However, Carter et al. [10]

showed that this dispatching rule is not always optimal in minimizing the average response time.

In this chapter we focus on the complementary decision context of which ambulance to dispatch to an incident. Ambulance dispatching decisions allocate appropriate paramedic units considering the nature and risk level of the call. As such, different types of dispatching strategies incorporating the degree of the urgency of the call will be studied to ascertain the impact on the response time and ultimately patients' survival probability. We examine the behavior of optimal dispatching strategies through several examples. Results are used to guide us in creating a useful heuristic to dispatch ambulances that increase the average survival probability of patients and thereby improve the efficiency of the EMS systems.

3.2 Related work

The overall goal of emergency medical service (EMS) systems is to prevent life-threatening and disabling injuries by providing care in timely manner. Thus, the decision-making process for EMS, which focuses on providing effective and efficient services, is a strategic challenge. EMS planners must solve problems such as where to locate emergency service stations, the number of ambulances to allocate to each station, and how to dispatch the appropriate paramedic unit to the emergency scene. A significant amount of research has been done towards addressing these. Much of the relevant literature can be divided into which resource allocation decision is being made.

While much work has focused on the optimal location of vehicles ([49], [11], [12]), these assume that the closest paramedic unit is dispatched. We limit the remainder of our discussion to research that in some way directly addresses the dispatching problem. Ambulance relocation models found in EMS literature are developed to explicitly account for busy ambulances when calculating coverage, since dispatching of an ambulance in response to a call may leave a significant proportion of the population without sufficient coverage. In this approach, the main focus is to dynamically relocate vehicles in real-time when vehicles are dispatched to the scene. The first such vehicle relocation model was developed by Kolesar et al. [22] to relocate fire trucks. A more recent example for ambulance redeployment is the dynamic double standard model at time t (DDSMt) by Gendreau et al. [17]. The primary disadvantage of dynamic relocation models is the necessity of finding a new solution whenever a vehicle is dispatched. This procedure is time consuming since it has to be done frequently [17]. There are some other practical issues in implementing the relocation of ambulances in real time for EMS systems. For example, with an increase in the complexity of EMS systems,

the number of relocations grows dramatically [1]; when calls come in quick succession solutions may be infeasible. From a practical standpoint, relocating ambulances may change the route or destination of vehicles frequently which can lead to a confusion of drivers and thereby causing mistakes [19]. Current research focuses on developing more powerful solution methodologies to solve these models quickly. However, since there are practical issues in implementing dynamic relocation models to real world EMS systems, our study focuses on the complementary problem of fixed deployment (paramedic units are located at specific stations, respond from stations, and return back to their station after serving a call) instead of ambulances relocating while en-route. The main focus is to improve the EMS system performance by implementing efficient dispatching strategies, which includes an ordered preference list of ambulances to dispatch. Such dispatching strategies are easy to implement in practice as ordered preference lists are already widely accepted policy types in EMS systems.

The most commonly used dispatching rule is to send the closest available unit without considering the severity of the call. Carter et al. [10] found that this rule is not always optimal in minimizing the average response time. Their goal was to determine the boundaries for each emergency unit (i.e. the response area of ambulance) in order to minimize the average response time. A few research studies have shown that dispatching emergency vehicles according to the degree of the urgency of the call helps to increase the survival probability of patients. For example, Nicholl et al. [35] conducted a case study with the objective of evaluating the safety and reliability of a two-priority dispatch system operated by an ambulance service in the UK. They found that priority dispatching systems have the ability to respond quickly to life-threatening calls by focusing resources on these calls, thereby increasing the survival probability of patients. In addition, they recommended that low priority calls, or non-life threatening calls, be responded to as soon as possible rather than immediately. Another study was conducted for an EMS system in Helsinki, Finland, by Kuisma et al. [23] to record pre-hospital death rates in four medical priority categories (most severe to least severe) to evaluate if deaths in lower urgency categories could have been prevented by faster ambulances responses. This community-based cohort study showed that the four-category medical priority dispatching of ambulances helps to maintain a lower pre-hospital mortality in the two lower urgency categories. These studies suggest that priority dispatching can play a key role in saving lives.

EMS literature includes only a few studies incorporating dispatching policies or strategies considering the degree of the urgency of the call. Comparing the dispatching strategies of First Come First Serve (FCFS), nearest origin assignment, and the flexible assignment strategy, Haghani et al. [19] discussed the benefits of using priority dispatching in EMS systems to reduce response time. The objective was to evaluate

the performance of these three dispatching strategies considering dynamic travel time information, vehicle diversion, and route changing. Henderson et al. [20] developed a decision support tool for St. John Ambulance Service in New Zealand considering two types of calls: Priority 1 calls are considered to be the more severe incidents, while Priority 2 calls are less severe incidents. Andersson et al. [1] also investigated the advantage of priority dispatching considering the urgency of the call. Their goal was to relocate ambulances in order to minimize the response time. In addition to dynamic ambulance relocation, they also proposed an algorithm to dispatch ambulances automatically incorporating the severity of the call. They considered three types of priority calls: the most urgent life-threatening calls (Priority 1), urgent but not life-threatening (Priority 2), and non-urgent calls (Priority 3).

Other authors use Markov Decision Process (MDP) approach to model EMS systems. McLay et al. [34] and Bandara et al. [3] used an MDP approach to model the EMS system and obtain optimal dispatching rules. The objective of the model in McLay et al. [34] is to maximize the coverage level while the objective of Bandara et al. [3] is to maximize patient survival probability. Although MDPs are capable of addressing problems with stochastic behavior and finding optimal solutions, a simulation approach has some advantages. When the MDP approach is used for real world EMS systems the model formulation becomes complicated and the state space grows dramatically with the complexity (number of ambulances and demand zones managed) of the EMS system. Furthermore, MDP models assume that service times are exponentially distributed.

In contrast, simulation models are easy to use for modeling EMS systems and measuring performance. They also allow us to compare systems under different sets of assumptions, providing the ability to test new operational strategies such as different ambulance locations or dispatching rules. The literature includes several simulation models that have been developed and used to evaluate the performance of EMS systems. One early example is that of Savas [42] for ambulance operations in New York City. In this study, a simulation was developed to conduct a cost-effectiveness analysis of New York's emergency ambulance service. The objective was to improve the service of the EMS system at a low cost. This study showed a considerable improvement in average response time by redistributing the existing ambulances in the district by locating some of them at satellite garages rather than all of them at a hospital. Henderson et al. [20] also developed a simulation model (BartSim) as a decision support tool for ambulance dispatching in Auckland, New Zealand. Andersson et al.[1] and Haghani et al. [19] also developed simulation approaches for the EMS models described earlier. These examples show the usefulness of simulation in modeling EMS systems. In our research we are also considering a simulation approach to model EMS systems and study the performance

of the proposed dispatching strategies.

The main goal of this study is to increase the average survival probability of the patients by implementing dispatching strategies that incorporate the degree of urgency of the call. In contrast, most of the models developed for EMS systems attempt to improve their performance by maximizing the coverage while operating according to the myopic dispatching policy, giving no reference to the severity of the emergency call. When attempting to maximize the survival probability of the patient, the myopic policy is not always optimal. For example, in a situation when the dispatching center receives two types of calls from the same region, an urgent call subsequent to a less urgent call, the victim involved in the second incident will be in jeopardy because the closest emergency unit may have been sent to the first but less serious call. In such a situation the dispatcher has to send another unit to the more severe incident. This strategy may not be ideal, as the next closest available unit may take more time to arrive at the incident. This dispatching could lead to a decrease in the survival probability of the severe incident due to inability to minimize the response time. In order to address this situation, a system considering the severity of the emergency has to be developed. Therefore, the objective of this article is to develop, easy to implement, dispatching strategies for EMS systems that will increase patient survivability while incorporating the degree of the urgency of the call.

A simulation model was developed to represent the EMS system. To obtain the optimal dispatching strategies in less complex models either full enumeration or a commercial optimizer was used. The results of these findings were used to develop a heuristic to determine the improved dispatching strategies in more complex models in order to maximize patient survival probability. Andersson [1] and Lee [28] also implemented heuristics for dispatching ambulances. Finally, we discuss computational examples to study the performance of the proposed dispatching heuristic. EMS performance evaluation was conducted by measuring the patient survivability instead the number of calls covered within a given time standard, known as the response time threshold (RTT). The results indicated that it is not always optimal to dispatch the closest ambulance especially for low priority calls that are non-life threatening. Moreover, it is found that dispatching vehicles considering priority of the call leads to an increase in the expected average survival probability of the patients and to a decrease in the average response time for life-threatening calls.

3.3 Model Description

An EMS system with i demand zones $(i = 1, ..., n)$ and j ambulance stations $(j = 1, ..., m)$ with an ambulance at each station ($n \times m$ case) is considered. The emergency vehicle dispatching process is depicted in Figure 3.1. A 911 emergency medical service starts with a call to the dispatching center requesting an ambulance. The time between the arrival of the call and the time the first ambulance is dispatched to the scene is known as the "preparation time". The time between the receipt of a call at the dispatch center and the arrival of the first emergency response vehicle at the scene is known as the "response time". "Turn around" time can be defined as the time required for an ambulance to return back to its original station after serving the patient, which includes the transportation time of the patient to the hospital if needed. Not every emergency call is life-threatening. Therefore, the dispatcher has to select and assign the appropriate ambulance according to the severity of the call. In this study we consider two types of calls: Priority 1 calls are considered to be life-threatening and Priority 2 calls are non-life threatening. Input parameters for the model are summarized below.

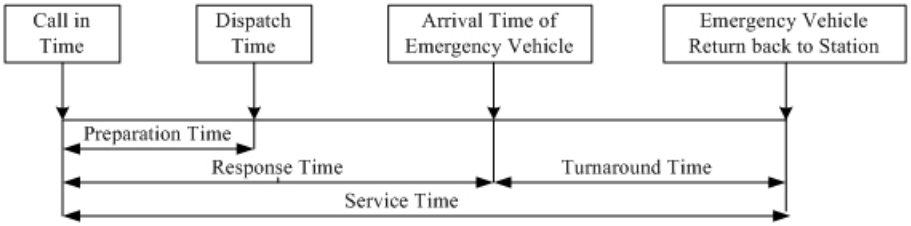


Figure 3.1: Emergency vehicle dispatching process

- λ = call arrival rate (to the entire system)
- n = total number of demand zones
- m = total number of paramedic units, each at a fixed location
- z_i= proportion of calls from i^{th} demand zone: such that $\sum_i z_i = 1. \forall\, i = 1, ..., n$
- p_{k_i}= probability of Priority k calls from demand zone i : such that $\sum_k p_{k_i} = 1. \forall\, i = 1, ..., n$
- R_{ij}= response time distribution for ambulance j for zone i with Mean Response Time (MRT) of $\mu_{R_{ij}}$ and standard deviation of $\sigma_{R_{ij}}$

- A_{ij}= turn around time distribution for ambulance j for zone i with mean of $\mu_{A_{ij}}$ and standard deviation of $\sigma_{A_{ij}}$
- λ_i= λz_i , call arrival rate in demand zone i

EMS performance evaluation is commonly done by measuring the number of calls covered within a given time standard, known as the "response time threshold" (RTT). Later research has suggested that using patient survivability is a more precise measure in terms of the expected number of survivors in case of emergencies [13]. Erkut et al. [13] showed that incorporating a survival function in EMS location problems helped to save more lives. In order to get a precise estimate of the EMS system performances, survival probability will be used as the performance measure of this study. Although several studies have been conducted to determine the relationship between response time and survival probability, the one utilized in this study is based on Larsen et al. [24] and subsequently simplified by McLay et al. [33]. This survival function is is given in equation (4.1), where S denotes the patient survival probability and the realized response time is t_R..

$$S(t_R) = \max[(0.594 - 0.055 * t_R); 0] \tag{3.1}$$

3.3.1 Simulation Model

Arena software is utilized to develop a simulation model to represent an EMS system. In the model we assumed that calls arrive according to a Poisson process with rate λ, which is consistent with the call arrival process of Hanover county Fire and EMS, the EMS system used in our case study. The status of an ambulance is either "busy" serving a call from demand zone i or "idle" at station j. The status of all ambulances is used to model the state of the EMS system. This simulation model is designed to identify the call location and its severity as soon as a call arrives to the system and to dispatch one of the available ambulances. Once the ambulance is dispatched, the ambulance status is set to "busy" and generates the response time according to a given distribution, R_{ij}, which depends on the call location and the responding paramedic unit. Then, the survival probability is calculated according to the response time using equation (4.1) for Priority 1 patients. (Priority 2 patients add zero value to the objective function because their calls are non-life-threatening.) After calculating the survival probability, the ambulance status is reset to "busy" for a time until it returns to the original station. We considered this time as "Turn around" time as explained previously, and again is drawn from a known distribution, A_{ij}, which depends on the location of the call and the responding vehicle. If all ambulances are busy then it is assumed calls are served by outside resources (by

fire engines or ambulances from neighboring county). Such arrangements are in place in Hanover county. In addition, we assumed that the response time for outside unit is greater than 11 minutes. Thus, zero survival probability is given to the calls that are served by the outside paramedic unit when calculating the average survival probability. The zero-queue assumption can be relaxed by allowing those calls to be queued in the system (we explore the results of lifting this assumption in a later section). However, we believe the zero-queue assumption is reasonable for the following reasons. (1) The objective of the model is to maximize the survival probability of Priority 1 calls and this probability is zero for response times greater than 11 minutes (according to equation (4.1)). Therefore, the contribution of the calls waiting in the queue, when calculating the objective function is negligible. (2) The examples we have studied with non-zero queue in the simulation model showed that the zero-length queue assumption does not significantly impact the steady-state results such as average survival probabilities and dispatching strategies. Therefore, the EMS system was modeled with zero queue when investigating the structure of the optimal policy, as well as developing the heuristic. The effect of the zero-queue length assumption on the performance of the system is studied in detail in a later section. Figure 3.2 illustrates the flow chart of this simulation model. This simulation was modeled assuming that the EMS system operates 24 hours per day and seven days per week.

Dispatching strategies can be characterized into static and dynamic. In a static policy, fixed deployment (ambulances return to their home station after serving a call) is considered; while in a dynamic policy, ambulances relocate based on real-time information. In this study a static dispatching rule for EMS systems is proposed in order to maximize the patients survivability. In particular, we restrict our attention to dispatching rules that provide an ordered preference list of ambulances to dispatch for each demand zone depending on the priority of the call. For example, if the preference order of dispatching paramedic units is 2, 3, 1 then Ambulance 2 is the first choice to dispatch, if it is busy Ambulance 3 is the second choice and, if both 2 and 3 are busy Ambulance 1 is the third choice. From here on we refer to a preference order list as a "contingency table" for short, a name commonly used by EMS system administrators. OptQuest of Arena (or the optimal simulator in Arena) is used to determine the optimal dispatching policies (the optimal order of dispatching ambulances) in less complex models and improved dispatching policies in more complex models. The results are used to develop a heuristic approach for implementing improved dispatching policies. Once the dispatching strategies are determined, the simulation model will be used to evaluate the performance of the EMS system. This simulation model is developed to incorporate different dispatching strategies and evaluate the performance of each dispatching strategy.

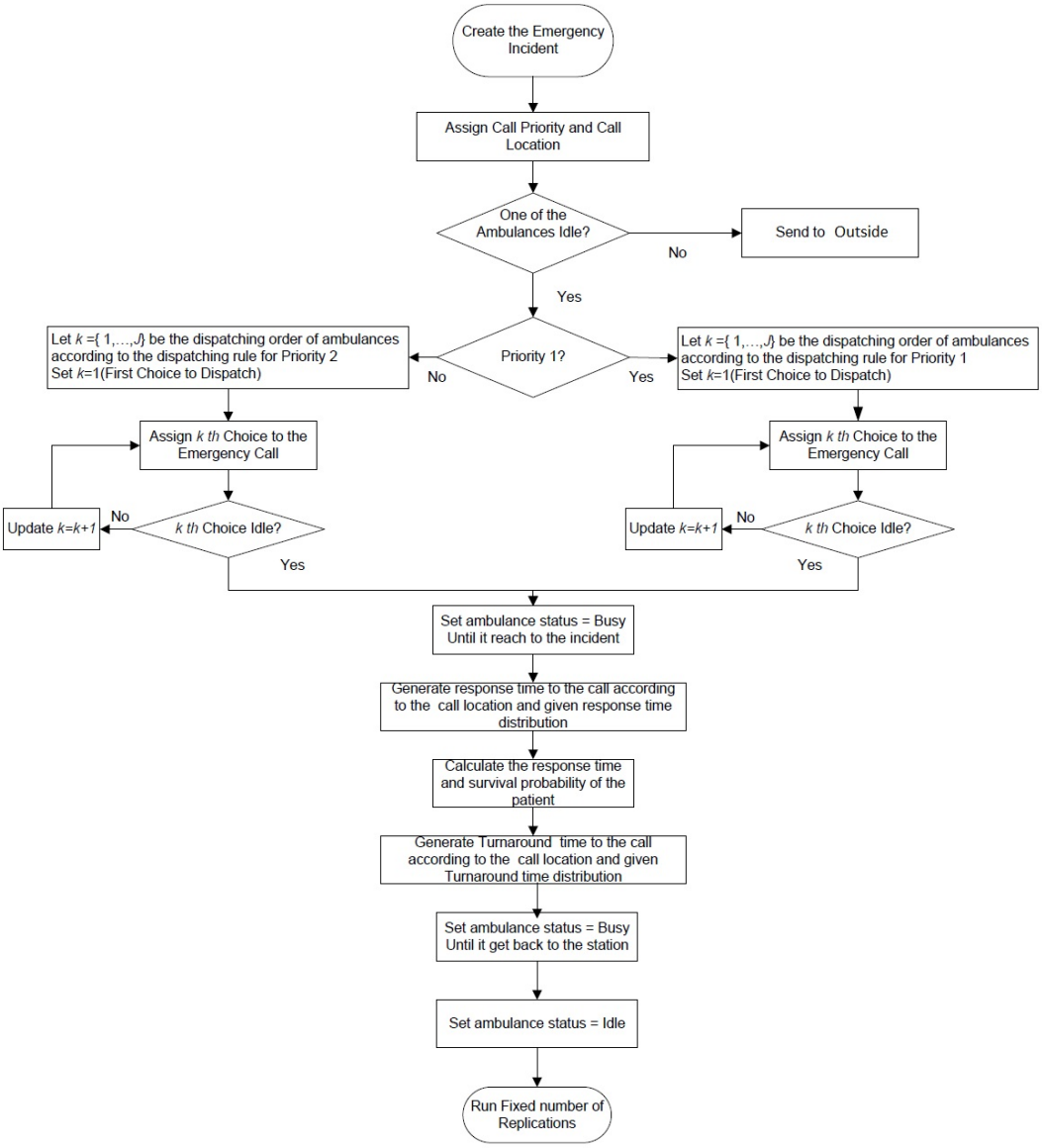


Figure 3.2: Simulation flow Chart

An illustrative example is discussed to study the behavior of the optimal dispatching strategy. This hypothetical example is constructed using data such as response times and turn around times similar to Hanover County, Virginia. A detailed description of Hanover County Fire an EMS is provided in a later section. For illustrative purposes, this example (2×2 case) assumes an EMS system with two demand zones (i.e. the zones that are requesting emergency vehicles) and two ambulance stations (i.e. the locations of the ambulances) with one paramedic unit (ambulance) at each station. The paramedic units at each station are

labeled Ambulance 1 and Ambulance 2, sited at station 1 and station 2 respectively. Calls arrive according to a Poisson process with rate $\lambda = 1$ per hour to the entire system. Both demand zones are equally likely to receive Priority 1 and Priority 2 calls , i.e. $p_{11} = p_{12} = p_{21} = p_{22} = 0.5$. Response times and turn around times are assumed to be lognormally and exponentially distributed respectively, since those are the best fit with data collected from Hanover County. Tables 3.1 and 3.2summarizes the response times and turn around times for each demand zone:

Zone	*Ambulance 1*	*Ambulance 2*
zone 1	logn(9.07,4.19)	logn(14.03,6.48)
zone 2	logn(14.03,6.48)	logn(9.02,6.48)

Table 3.1: Response Time Distributions-2 $\times$ 2 case

Zone	*Ambulance 1*	*Ambulance 2*
zone 1	expo(50)	expo(54)
zone 2	expo(60)	expo(50)

Table 3.2: Turn around time Distributions-2 $\times$ 2 case

Position $(1, 1)$ of Table 3.1 shows that the response time for Ambulance 1 to demand zone 1 is lognormally distributed with mean of 9.07 minutes and standard deviation of 4.19 minutes. Position $(1, 1)$ of Table 3.2 shows that the turn around time for Ambulance 1 to demand zone 1 is exponentially distributed with a mean of 50 minutes. We vary the call volume between the two demand zones such that $z_1 + z_2 = 1$, resulting $1/10 \leq \lambda_1/\lambda_2 \leq 10$ to study the effect of the geographic dispersion of demand on the optimal dispatching strategy.

To study the structure of the optimal policy on this small example, the optimal simulator in Arena (OptQuest) is used to obtain the preference order list for dispatching ambulances which maximizes patient survivability. This simulator ran for 200 replications per simulation with tolerance of 0.0001 until it obtains the optimal solution (optimal dispatching order); each replication ran for 336 simulated hours to obtain steady-state results with the half width of a 95% confidence interval. The performance, i.e. the average survival probability, of the optimal dispatching rule is compared to the myopic policy of always sending the closest ambulance. Comparison of the survival probability of two dispatching policies is depicted in Figure 3.3. As Figure 3.3 shows, the simulation indicates that $P(survival)$ increases when ambulances are

dispatched based on the optimal policy rather than sending the closest unit for every call. In some cases, however, the optimal rule is to send the closest unit. This occurs when the call volume is almost balanced between two demand zones. (i.e. $log_{10}(\lambda_1/\lambda_2) = 0, log_{10}(\lambda_1/\lambda_2) = -.17$ and $log_{10}(\lambda_1/\lambda_2) = 0.17$). The objective value difference (i.e. the survival probability) of the two dispatching strategies is greatest when the call volume is not balanced between the two demand zones. Since the objective difference is small (in absolute terms), we did a paired t-test to determine whether the objective improvement is statistically significant. We found that objective improvement is statistically significant when $z_1 = 0.1, 0.8$ and 0.9 (i.e. $log_{10}(\lambda_1/\lambda_2) = -0.95, 0.6$ and 0.95) at the 95% confidence level. Although the objective difference is low, the number of patients who can survive will increase with no additional cost by following the optimal dispatching rule. For example, when $z_1 = 0.9$, the survival probability can increase by 5.63% compared to the send-the-closest rule. If it is assumed that this system serves 1000 Priority 1 calls per year, approximately an additional 17 lives can be saved by following the optimal dispatching rule with available resources. To get a sense for this, Hanover county serves around 4760 Priority 1 calls per year. In addition, the average response time for Priority 1 calls is reduced by dispatching ambulances according to the optimal dispatching strategy. Figure 3.3 compares the average response time for Priority 1 calls. This figure indicates that the optimal dispatching strategy helps to decrease the average response time for Priority 1 calls. Two additional hypothetical examples are constructed in this study using data similar to Hanover County. Example 2 is an EMS system with three demand zones and three ambulances (3×3 case), while Example 3 is an EMS system with five demand zones and three ambulances (5×3 case). These two examples are summarized in Appendix A.

For this example (2×2 case) OptQuest of Arena determined that it is optimal (OptQuest enumerate all possibilities in this case) to dispatch the closest paramedic unit for Priority 1 calls but that policy is not always optimal for Priority 2 calls. Similar results have been obtained by Bandara et al. [3] for low priority calls in a system with exponential response times. In Table 3.3, the contingency table for Priority 2 calls provided by the OptQuest of Arena is compared with the closest dispatching strategy for the (2×2) case. Table 3.3 shows that when probability of requesting an ambulance for demand zone 1 is 0.1 (i.e. $z_1 = 0.1$), the optimal order of dispatching paramedic units to demand zone 1 is that Ambulance 1 is first choice and Ambulance 2 is second choice. For demand zone 2 optimal order is Ambulance 1 first choice and Ambulance 2 second choice. Similarly we can obtain the preference order of dispatching ambulances for both policies using Table 3.3. By observing the ambulance dispatching orders, it is suggested that the optimal policy is likely to reserve the closest ambulance to serve the zone with higher customer arrival rate for Priority 1 calls.

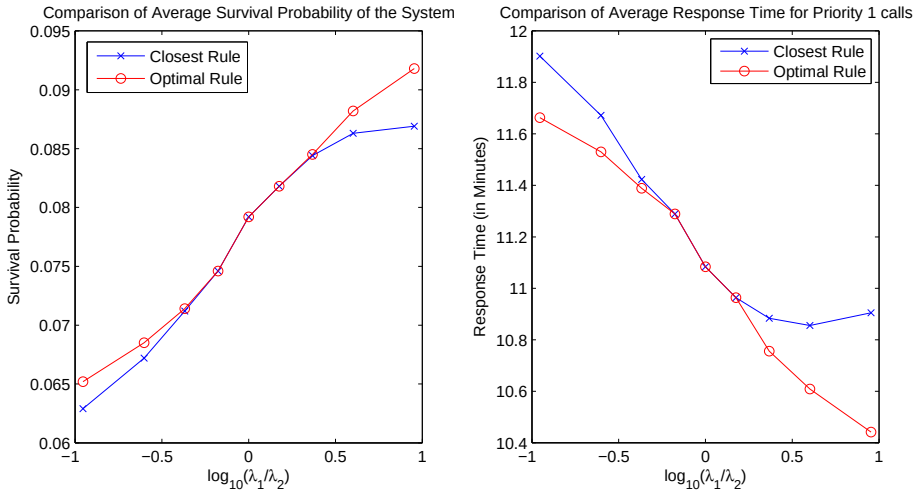


Figure 3.3: Comparison of Survival Probability and Average Response Time

For example, Ambulance 2 will not be deployed for Priority 2 calls (if both ambulances available) from demand zone 2 according to the optimal policy when $z_1 = 0.1$ (i.e. higher customer arrival rate occur from zone 2, since $z_2 = 0.9$) because it is reserved for Priority 1 calls and it is the closest paramedic unit for zone 2. Ambulance 1 on the other hand, is dispatched for Priority 2 calls from demand zone 2 under the optimal policy if both ambulances are available.

		Closest Policy				Optimal Policy-by OptQuest			
		order zone1		order zone2		order zone1		order zone2	
z_1	$log_{10}(\lambda_1/\lambda_2)$	choi 1	choi 2	choi 1	choi 2	choi 1	choi 2	choi 1	choi 2
0.1	-0.95	1	2	2	1	1	2	1	2
0.2	-0.60	1	2	2	1	1	2	1	2
0.3	-0.36	1	2	2	1	1	2	1	2
o.4	-0.17	1	2	2	1	1	2	2	1
0.5	0	1	2	2	1	1	2	2	1
0.6	0.17	1	2	2	1	1	2	2	1
0.7	0.36	1	2	2	1	2	1	2	1
0.8	0.60	1	2	2	1	2	1	2	1
0.9	0.95	1	2	2	1	2	1	2	1

Table 3.3: Comparison contingency tables (closest versus optimal policies) for Priority 2 calls

To study the effect of the response time on the survival probability of the 2×2 example, the Mean Response time (MRT) is varied from three minutes to thirteen minutes in one minute increments and we obtain the survival probability for the two cases with the most dispersed call volume, when $z_1 = 0.1$ and $z_1 = 0.9$. Additionally, the objective differences between optimal and myopic policies are observed. See

Figure 3.4. This figure shows that when MRT increases survival probability decreases gradually, as expected. As Figure 3.4 shows, when the probability of requesting a paramedic unit to demand zone 1 is low (i.e.$z_1 = 0.1$) compared to zone 2, the objective difference between the two dispatching strategies is not significant for smaller response times (or MRT). However, objective differences become gradually significant as the response time increases (see Figure 3.4 Case 1). This relationship between MRT and survival probability is reversed when $z_1 = 0.9$ (see Figure 3.4 Case 2). This result can be confirmed by observing the objective difference graph (rightmost graph in Figure 3.4). The MRT 3 mins curve shows that the objective difference is not significant when demand zone 1 has smaller call arrival rate. This objective difference becomes significant when call arrival rate increases for zone 1. This result is reversed when MRT is 13 minutes.

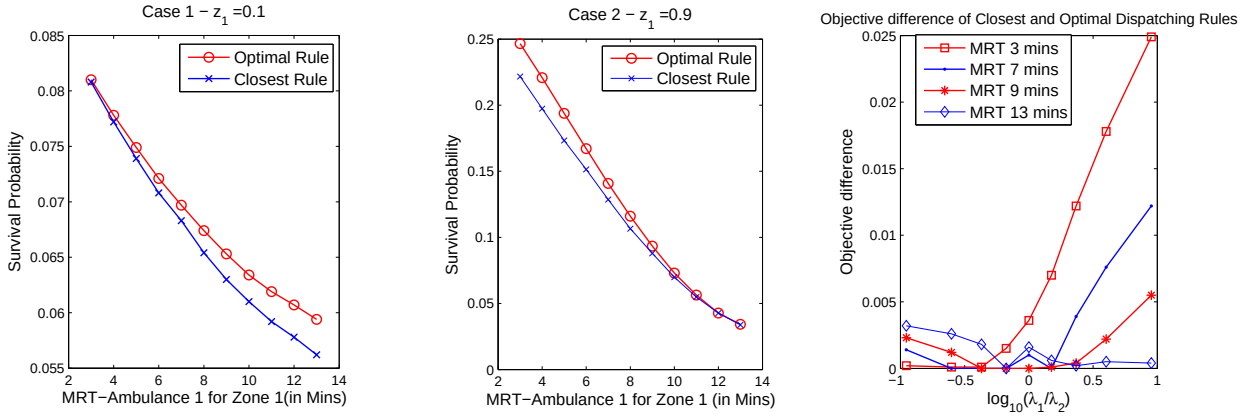


Figure 3.4: Sensitivity analysis-Vary Mean Response Time (MRT) for demand zone 1 by Ambulance 1

While we chose a very small example to illustrate the structure of the optimal policy, the discussion and observations presented for this example are consistent with many other small (up to 5 demand zones and 3 ambulances) examples that were tested. For larger examples, it becomes impractical to obtain the optimal policy via simulation.

3.4 Improved Dispatching Strategies- Heuristic Approach

The running time for simulation model, when analyzing the performance of the system, was not significant (less than two minutes) for most of the examples considered. However, the running time of the commercial simulation optimizer to obtain an improved solution is significant and grows with problem size.

For example, OptQuest running time for an example with three demand zones and three ambulances is 38 minutes while it is 9292 minutes for an example with five demand zones and three ambulances. (OptQuest running times for computational examples considered in this study are summarized in Table 3.4). In addition, full enumeration is not practical for obtaining the optimal dispatching order, because there are $(m!)^{2n}$ possible dispatching orders.

Example	*Size*	*OpQuest Running Time(mins)*
Example 1	2×2 case	4
Example 2	3×3 case	38
Example 3	5×3 case	9292
Example 4	5×5 case	> 10000

Table 3.4: OptQuest running Times

Since the running times increase dramatically with the problem size and full enumeration is not practical, a heuristic was developed to determine improved dispatching strategies for real-world EMS systems. While studying the optimal policy, we observed that there was a marked difference in the ambulance busy probabilities between the optimal and myopic policies. Figure 3.5 compares the proportion of time during a day that the ambulances are busy for the myopic and optimal dispatching policies for the hypothetical Example 1 (2×2 case). As this figure illustrates, when ambulances are dispatched according to the closest dispatching rule, the ambulance busy probability continuously decreases or increases with respect to z_1 and z_2 values. For instance, Ambulance 1 busy probability increases when z_1 increases and Ambulance 2 busy probability decreases when z_2 decreases. In this example we find that ambulance utilization is unbalanced when z_1 and z_2 values differ substantially and the closest dispatching rule is followed. For the optimal dispatching strategy, the utilization of ambulances was smoothly distributed between two ambulances. Also, it was observed that the order of sending paramedic units to each demand zone is linked to the ambulance busy proportion. According to the simulation model, the optimal dispatching rule tends to send the less busy unit (in terms of what unit would be busy under a myopic policy) for Priority 2 calls.

As observed in earlier examples the optimal dispatching rule did not always send the closest unit for Priority 2 calls. While for Priority 1 calls the optimal rule was to send the closest unit if it is available. Hence we studied the probability that the closest ambulance is dispatched to Priority 1 and Priority 2 calls for these examples. Figure 3.6 indicates the probability that the closest (absolute closest, not closest among

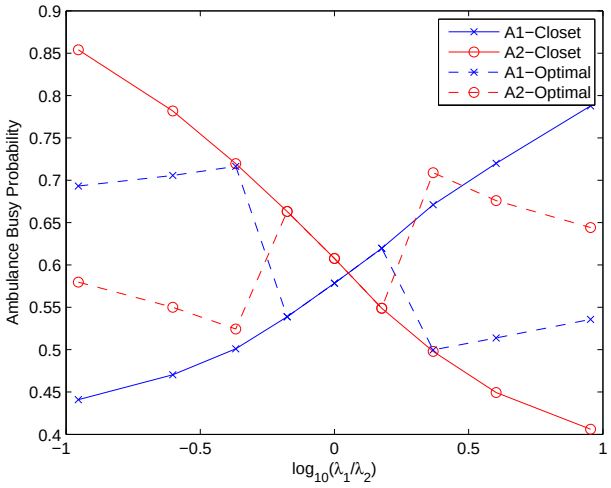


Figure 3.5: Comparison of Busy Probability of Each Ambulance for Two Dispatching Strategies - 2×2 case

those available) server is dispatched to each Priority call for Example 1. This figure indicates that the optimal policy tends to serve the Priority 1 calls from the location (zone) with higher arrival rate by the closest unit more often, while Priority 2 calls from either of the demand zones are served by the unit which is closest to the zone with lower arrival rate (by less busy ambulance). Consequently, Priority 2 calls from the demand zone with higher arrival rates are not likely to be served by the closest ambulance according to the optimal policy (see Figure 3.6 Priority 2). Figure 3.6 (Overall) shows that the proportion of calls served by the closest ambulance has increased by following the optimal dispatching compared to closest rule for Priority 1 calls. For Priority 2 calls, however, the proportion of calls served by the closest ambulance has decreased. These results also are confirmed by the examples 3×3 case and 5×3 case. Figure 3.7 and Figure 3.8 show the probability that the closest server is dispatched to Priority 1 and Priority 2 calls for each example. These figures illustrate that the proportion of Priority 1 calls served by the closest ambulance can be increased by following the optimal rule in comparison to the closest rule. It is also observed that the Priority 2 calls are unlikely to be served by the closest unit according to the optimal rule.

Based on these observations, a heuristic rule is developed to dispatch the ambulances for emergencies considering the priority or the severity of the call. This heuristic provides an ordered preference list of ambulances to dispatch (contingency table) for each demand zone depending on the priority of the call. According to the the heuristic rule for Priority 1 calls, the closest available unit is dispatched. For Priority 2 calls, however the contingency table does not depend on the distance from the ambulance station to the demand zone. This result is also confirmed by the findings of [34]. The ordered list for dispatching ambulances to

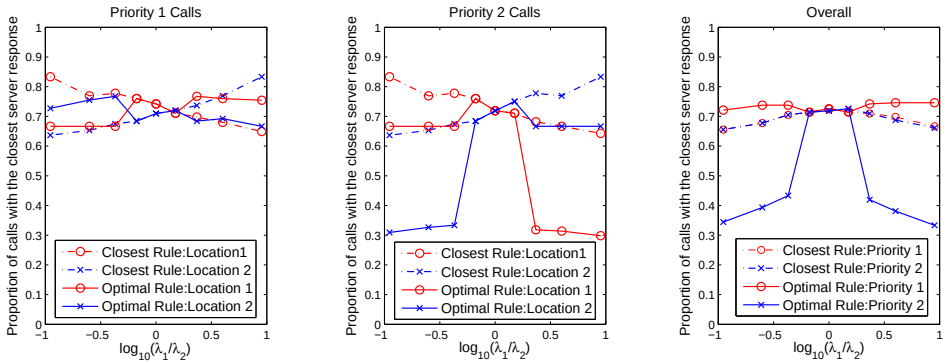


Figure 3.6: Probability that the closest server is dispatched to Priority 1 and Priority 2 calls for Optimal Rule and Closest Rule - (2×2 case)

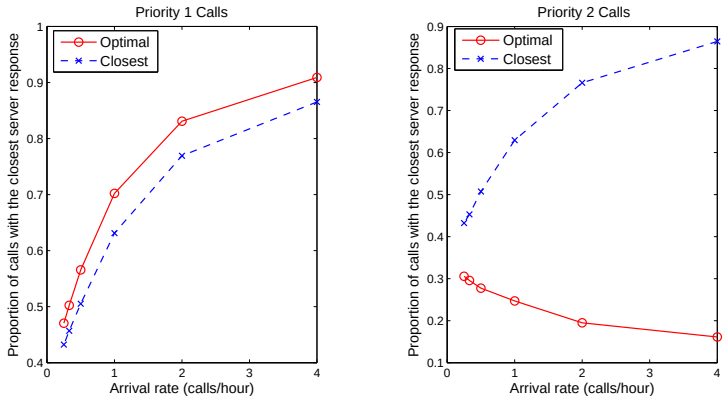


Figure 3.7: Probability that the closest server is dispatched to Priority 1 and Priority 2 calls for Optimal Rule and Closest Rule - (3×3 case)

Priority 2 calls is constructed considering the busy probability of each paramedic unit. Previous results show that optimal policy tends to dispatch the less busy ambulance to Priority 2 calls. Therefore, this heuristic is also developed to dispatch the less busy ambulance for Priority 2 calls. The heuristic (H_1) algorithm used to obtain the contingency table for Priority 2 calls is outlined below.

Let n be the total number of demand zones and m be the total number of ambulances.

Step 1:

Let $r_1, r_2, r_3, \ldots, r_k \ldots, r_m$ be a permutation of $(1, 2, 3, \ldots, m)$ and $t_i^{r_k}$ be the response time of the ambulance r_k to zone i.

for each i where $i = 1, 2, 3, \ldots, n$

rank response time $t_i^{r_k}$ as $t_i^{r_1} \leq t_i^{r_2} \leq t_i^{r_3} \leq \ldots \leq t_i^{r_k} \ldots \leq t_i^{r_m}$

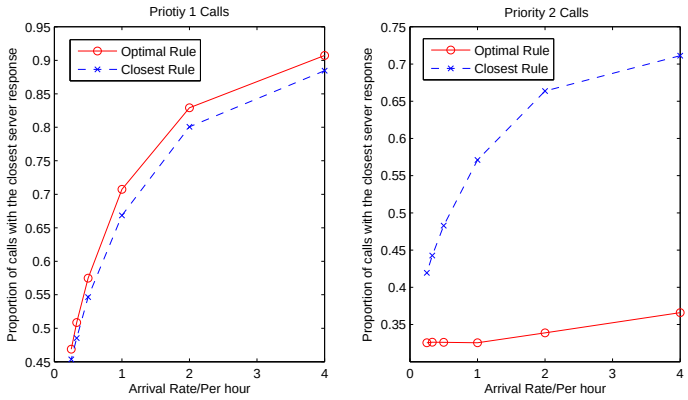


Figure 3.8: Probability that the closet server is dispatched to Priority 1 and Priority 2 calls for Optimal Rule and Closest Rule - (5×3 case)

Let $a_{ij} \in A$ be an $(n \times m)$ matrix, where i^{th} row of A denotes $r_1, r_2, r_3, \ldots, r_k \ldots, r_m$

Step 2:

Consider the row vector z where $z(i)$ denotes the proportion of calls from demand zone (i), $i = 1, 2, 3, \ldots, n$

Step 3:

Let $b_{kj} \in B$ be the Heuristic proportion matrix of size $(m \times m)$

for $j = 1, 2, 3, \ldots, m$ Do

for $k = 1, 2, 3, \ldots, m$ Do

$sum = 0$;

for $i = 1, 2, 3, \ldots, n$ Do

if $a_{ij} == k$

$sum = sum + z(i)$

$b_{kj} = sum$

Step 4:

Let $seq = 1, 2, 3, \ldots, m$ be a column vector and

$B^{'} = adjoin(B, seq)$

Do priority sort on $B^{'}$ in sequential order to permute the last column of $B^{'}$.

The last column of $B^{'}$ gives the dispatching order of ambulances for non-life threatening calls (For Priority 2 calls).

Using this heuristic we obtained the order of dispatching ambulances for the (3×3 case) shown

in Table 3.5. To interpret this table, consider the Heuristic policy portion of the table. Column "zone 1" under the Priority 1 heading in Table 3.5 indicates that Ambulance 1 is the first choice to dispatch, followed by Ambulance 2 (if Ambulance 1 is busy) then Ambulance 3 (if both Ambulance 1 and Ambulance 2 are busy) for Priority 1 calls. We observed that the dispatching policy given by the heuristic (H_1) is similar to the order given by the optimal rule (compare dispatching policies given in Table 3.5). Dispatching order for zone 3 given by the H_1 is slightly different in comparison to the optimal order, since H_1 provides the same dispatching order for all Priority 2 calls.

	Heuristic Policy						Optimal Policy					
	Priority 1 calls			Priority 2 calls			Priority 1 calls			Priority 2 calls		
Dispatch order	zone 1	zone 2	zone 3	zone 1	zone 2	zone 3	zone 1	zone 2	zone 3	zone 1	zone 2	zone 3
1^{st} Choice	1	2	3	1	1	1	1	2	3	1	1	1
2^{nd} Choice	2	3	2	2	2	2	2	3	2	2	2	3
3^{rd} Choice	3	1	1	3	3	3	3	1	1	3	3	2

Table 3.5: Order of dispatching ambulances for Heuristic and Optimal Policies - (3×3 case)

3.4.1 Performance of Heuristic Policy

The survival probability of the dispatching rules of closest dispatching (myopic policy), OptQuest dispatching, Andersson heuristic dispatching [1], and the dispatching rule proposed by the heuristic H_1 are compared in Figure 3.9 for the (3×3) example. The Andersson dispatching rule is developed based on the preparedness function which describes the ability to cover each demand zone. According to this algorithm, the ambulance whose unavailability causes the least drop in the preparedness value, is dispatched for non life-threatening calls, while the closest available unit is dispatched for life-threatening calls. [28] illustrated the role of preparedness in ambulance dispatching and also provide a dispatching heuristic. However, we did not consider the heuristic proposed by [28] when comparing different dispatching strategies in our study because their study does not consider the degree of the urgency of the call when dispatching paramedic units. Furthermore, initial testing showed that the [28] heuristic performs similar to the myopic policy.

For this small example OptQuest provides the optimal dispatching order for each demand zone. As we can observe from Figure 3.9, H_1 performs better than Andersson's heuristic in all aspects.(e.g. increase patient survivability, decrease response time). In addition, H_1 is easy to implement in EMS systems since it provides a static dispatching rule. On the other hand, the Andersson heuristic provides a dynamic dispatching rule and it needs to be executed every time an emergency call arrives. The improvement of the objective (survival probability) value by following the heuristic (H_1) dispatching rule seems low (in absolute terms)

when compared to the myopic policy (see Survival rate comparison of Figure 3.9). However this corresponds to a potential large number of Priority 1 calls by following the proposed heuristic (H_1) rule. For example, the average number of lives saved per 1000 Priority 1 calls in the (3×3 case) is 76 while it is 135 for an example with 5 demand zones and 3 ambulances.

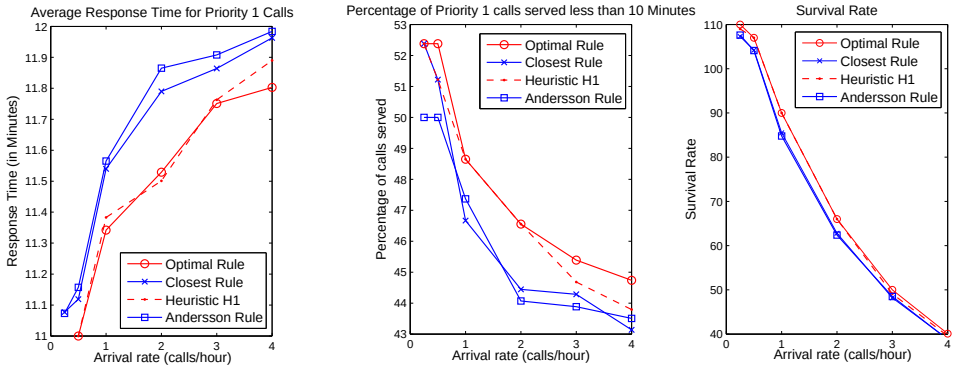


Figure 3.9: Comparison of Dispatching Strategies -3×3 case

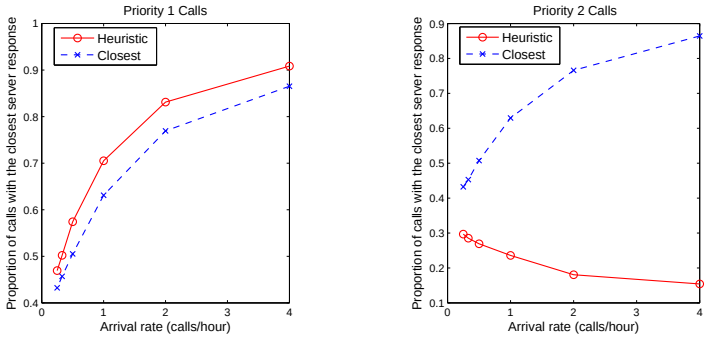


Figure 3.10: Probability that the closest server is dispatched to Priority 1 and Priority 2 calls for Heuristic Rule and Closest Rule -3×3 case

Figures 3.9 compares the response time for Priority 1 calls and the percentage of calls responded within 10 minutes for each dispatching policy. These Figures show that the results obtained (survival rate, average response time, and percentage of calls covered within 10 mins) by following the proposed dispatching heuristic for those EMS systems are similar to the results obtained by following the rule given by OptQuest. Hence it can be said that the proposed heuristic performs as well as optimal dispatching given by OptQuest for these two examples, and this observation is consistent with other cases tested. Dispatching ambulances according to the proposed heuristic (H_1) will lead to an increase of patient survivability and thereby increase the EMS systems performance. Further, average response time for Priority 1 calls decreases while percentage

of calls covered within 10 minutes increase for Priority 1 calls increases. Figure 3.10 compares the proportion of calls served by the (absolute) closest paramedic unit by following the dispatching strategies closest and heuristic for Example 2, (3×3) case. As this figure shows, more Priority 1 calls can be served with the closest unit by following the heuristic rule rather than sending the closest unit for every call (because the closest unit is not occupied serving less severe calls). The performance heuristic policy of (H_1) for Example 2, (5×3) case is summarized in Appendix B . In a later section the proposed heuristic is applied to a scenario using real-world data from Hanover County, Virginia.

3.4.2 Exploring the Zero-queue Assumption

This section illustrates the effect of the zero-queue length assumption on the performance of the EMS system. We modified our zero-queue length EMS system to a system with a queue by allowing the calls to be queued if all ambulances are busy. Then we compared the performance of the EMS system with queue to the EMS system with zero-queue. Though there are several queuing disciplines to serve customers in the queue, we assumed that customers in the queue are served according to the FCFS (First Come First Serve) discipline, because most of the EMS systems follow this rule. Studying the best queuing discipline to serve customers in the EMS system queue would be another interesting research topic.

We used several examples to study the effect of the queue on the optimal dispatching policy. However, we illustrate the results here via Example 1 (2×2 case) to compare the performance of the EMS system with a queue (in which calls must wait until one of the ambulances in the system is idle to receive service) to the EMS system with zero-queue. First we studied the optimal order of dispatching ambulances for each demand zone. The optimal order remains the same even if we consider a waiting queue in the EMS system. We did full enumeration to obtain the optimal dispatching order. Next we compared the survival probability of these two EMS systems (see Figure 3.11). As can be observed from Figure 3.11 the EMS system with queue has similar behavior to the EMS system with no queue. The only difference is the survival probability of the EMS system with queue decreases compared to the EMS system with zero-queue. This decrease is due to the calls waiting in line, which increase the busy probability of the ambulances which in turn decreases the survival probability of future calls.

Figure 3.12 compares the ambulance busy probabilities of the two EMS systems, with queue and zero-queue. As we can observe from the figure, the busy probabilities of ambulances follow similar trends in both systems. However, the busy probabilities for the EMS system with queue increase slightly compared

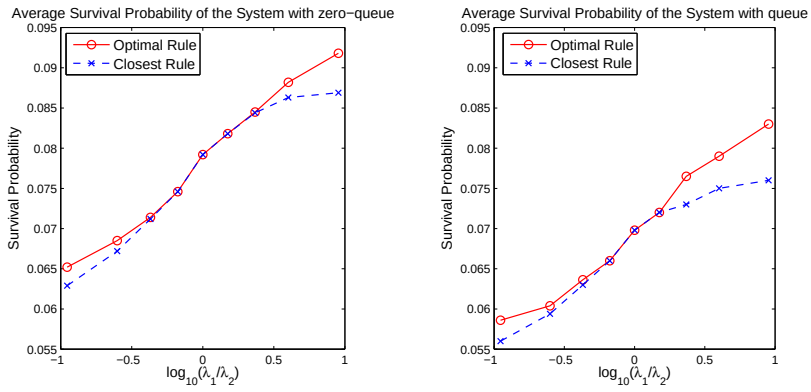


Figure 3.11: Comparison of Survival probability of EMS systems

to the EMS system with zero-queue, as can be expected. Thus we studied the average waiting time and the average number waiting in the queue for this example. We observed that the number waiting in the system was less than 0.4 calls on average for all scenarios. Additionally, we found that average waiting time in the queue varied from 18 minutes to 22 minutes. These last two observations imply that while the number of calls waiting seems negligible, if a call waits, it has to wait for an exceptional amount of time. This helps to explain why many EMS systems employ alternative strategies for dealing with calls when all ambulances are busy. Also it was observed that in the zero-queue system only 9% of the calls are served by the outside paramedic unit.

Interestingly, these differences do not affect the optimal dispatching strategy, which happens to be the same for both systems. Since the optimal policy structure informed our heuristic, we must look at the performance of the heuristic under a non-zero queue. In fact, we find that the survival probability improvement by following heuristic policy is greater in a EMS system with a queue compared to a EMS system with zero-queue, meaning that the dispatching rule is more beneficial (in terms of patients survivability) when EMS systems operate with a queue. For example, when $z_1 = 0.9$ (i.e. $log_{10}(\lambda_1/\lambda_2) = 0.95$), the heuristic results in 9.21% higher survival probability than the myopic policy compared to a 5.63% difference in survival probability between the heuristic policy and the myopic policy in the zero-queue system. We observed similar results for other examples tested. In summary, the queue has a negligible affect on the optimal policy and the heuristic (H_1) rule is more beneficial when there is a queue in the EMS system. Therefore, we believe our zero-queue assumption is justified.

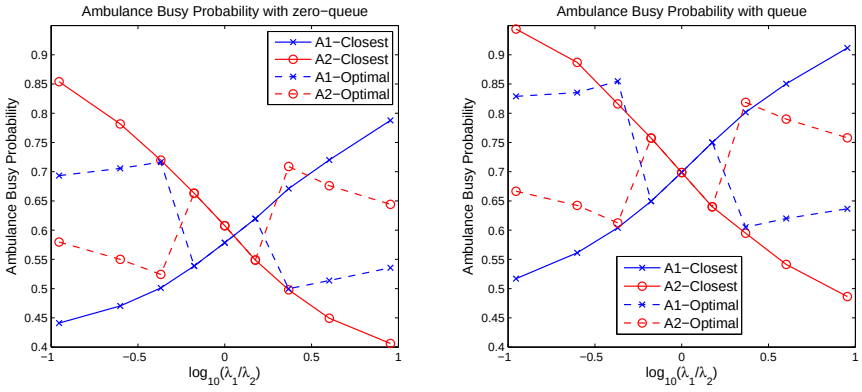


Figure 3.12: Comparison of busy probability of each ambulance for two EMS systems

3.5 Case Study

Hanover County, Virginia was used as the study area in our simulation model. The Hanover County EMS department responds to 911 calls 24 hours a day with a population of approximately 100,000 and an area of 471 square miles. Based on data collected in 2007, the average number of calls in Hanover is 1.2 calls/hour (peak rate) with 9521 total calls through out the year. An instance with twelve demand zones, four ambulance stations and five paramedic units was considered in this EMS simulation model. All demand zones and station locations are shown in Figure 3.13. Rescue stations 1, 2, 3 and 4 are the four fire stations with a paramedic unit at each station. The fifth paramedic unit is sited in the fourth station. Response time and turn around time distributions for this example are summarized in Appendix C.

	Priority 1 calls											
Dispatch order	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	z_{10}	z_{11}	z_{12}
1^{st} Choice	4	3	2	4	1	1	2	3	4	1	2	2
2^{nd} Choice	5	4	3	5	3	4	3	1	5	4	4	4
3^{rd} Choice	3	5	4	3	2	5	4	4	3	5	5	5
4^{th} Choice	2	2	5	1	4	2	5	5	1	3	3	3
5^{th} Choice	1	1	1	2	5	3	1	2	2	2	1	1

Table 3.6: Heuristic (H_1) order of dispatching ambulances for Priority 1

	Priority 2 calls											
Dispatch order	z 1	z 2	z 3	z 4	z 5	z 6	z 7	z 8	z 9	z 10	z 11	z 12
1^{st} Choice	5	5	5	5	5	5	5	5	5	5	5	5
2^{nd} Choice	3	3	3	3	3	3	3	3	3	3	3	3
3^{rd} Choice	1	1	1	1	1	1	1	1	1	1	1	1
4^{th} Choice	4	4	4	4	4	4	4	4	4	4	4	4
5^{th} Choice	2	2	2	2	2	2	2	2	2	2	2	2

Table 3.7: Heuristic (H_1) order of dispatching ambulances for Priority 2

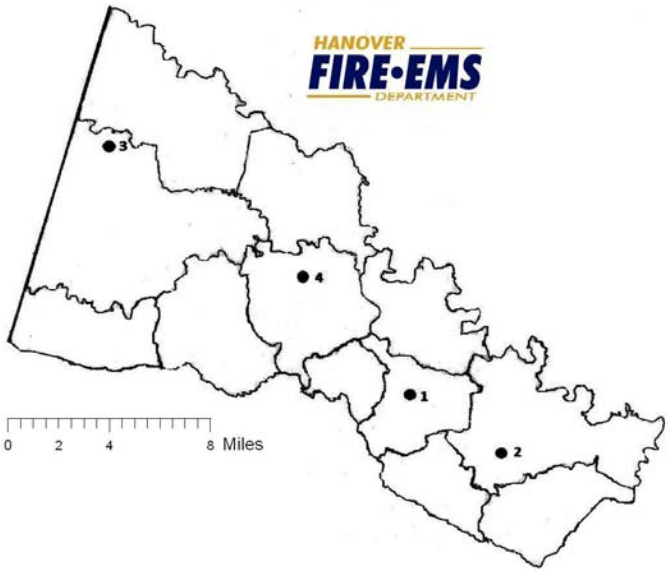


Figure 3.13: Hanover County Map, dots represent station locations

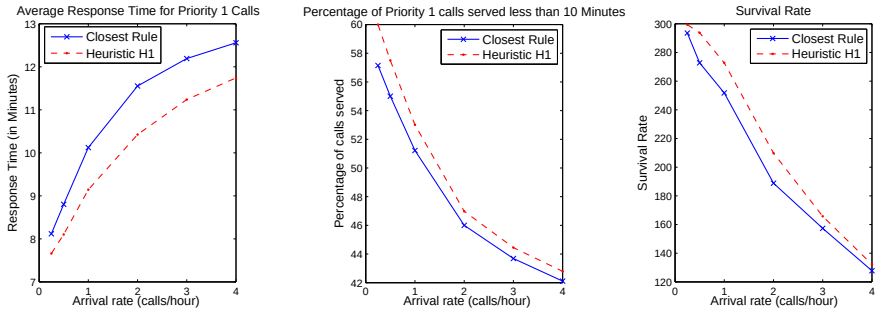


Figure 3.14: Comparison of dispatching Strategies - 12×5 case

Table 3.6 provides the contingency for Priority 1 calls according to the heuristic (H_1) rule. The order of dispatching ambulances for Priority 2 calls is given in Table 3.7. The survival rate per 1000 Priority 1 calls, when using the dispatching heuristic (H_1) is compared to the policy of always sending closest paramedic unit in Figure 3.14 as the total arrival rate to the system is varied. In addition, Figure 3.14 compares the average response time and percentage of calls covered within 10 minutes for each dispatching strategy. As the figure indicates, when call rate increases, survival rate and percentage of calls covered within 10 minutes decreases while the average response time increases. When dispatching paramedic units according to H_1 rule, survival rate and coverage can increase while average response time decreases in comparison to the myopic policy. Although the differences between survival probability for two dispatching strategies is low, many lives can be

saved at no additional cost (in terms of servers used) by following the proposed dispatching heuristic (H_1). For example, there is an 8.33% increase in survival probability compared to the closest rule when call arrival rate equals to 1 per hour. Since there are 9521 total calls to the system during a year, assuming half of them are life-threatening, approximately an additional 28 lives can be saved per year with available resources by following the dispatching heuristic proposed in this study. Hence we can say that it is beneficial to implement dispatching heuristic H_1 for EMS systems such as the one presented in this study.

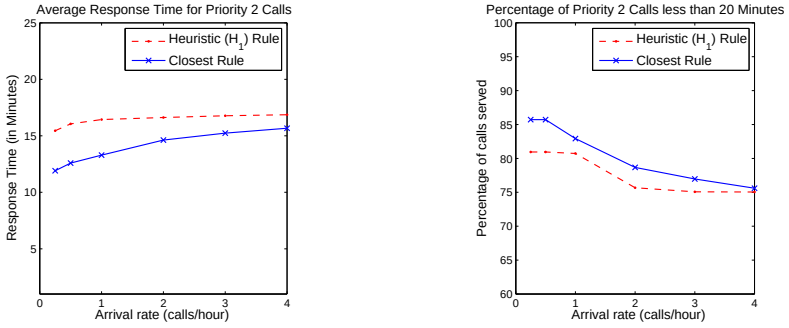


Figure 3.15: Comparison of dispatching Strategies for Priority 2 calls - 12×5 case

So far we have discussed the impact of proposed dispatching rule on Priority 1 calls. The effect of this dispatching rule on Priority 2 calls is depicted in Figure 3.15. As the figure indicates, average response time for Priority 2 calls increases. Though, proportion of calls covered within 20 minutes decreases slightly. This impact did not affect the average survival probability since Priority 2 calls are non-life threatening.

Finally, we believe that the heuristic we presented here provides significant improvements for EMS systems in terms of lives saved at no additional cost. Moreover, this improvement can be achieved by a simple policy of dispatching the closet available paramedic unit for Priority 1 calls while Priority 2 calls are served according to a pre-determined ordered preference list. This policy is applicable for existing EMS systems since it is easy to implement.

3.6 Conclusion and Future Research

Implementing optimal dispatching strategies for EMS systems to increase patient survivability is a challenging problem. In this study a heuristic algorithm was proposed for dispatching ambulances incorporating the degree of the urgency of the call to maximize patient survivability. This dispatching heuristic is

developed to send the closest available ambulance for Priority 1 calls and the less busy ambulance for Priority 2 calls. The proposed rule provides an order of dispatching ambulances for each demand zone depending on the priority of the call. However, the heuristic algorithm (H_1) provides the same dispatching order for Priority 2 calls for every demand zone. Future research can be conducted to obtain the order of dispatching ambulances for Priority 2 calls depending on the demand zone.

Computational examples showed that the proposed heuristic is beneficial in increasing patient survivability with no extra cost in terms of the number of paramedic units; meaning that patient survivability can be increased using the available resources simply by implementing the proposed dispatching rule. Even though this heuristic was developed to maximize the patient survivability, it helps to decrease the average response time and increase the percentage calls served within 10 minutes for Priority 1 calls. The average response time for Priority 2 calls increased slightly by following the proposed dispatching rule. Although the average response time increased it did not affect the average survivability of patients since Priority 2 calls are non-life threatening. Future research can concentrate towards obtaining dispatching rules to maximize patient survival probability of life-threatening calls while minimizing the effect on the average response time of Priority 2 calls.

The simulation model was developed enabling all the ambulances located at each station to serve the calls originating from any demand zone, meaning there is no restriction on ambulance response area. [10] showed that defining response areas (or boundaries) for each ambulance will lead to a decrease the average response time. Determining response areas for each ambulances while incorporating the proposed dispatching rule for dispatching ambulances in EMS systems, is another vital area of future research.

Chapter 4

Districting and Dispatching Policies to Improve the Efficiency of Emergency Medical Service (EMS) Systems

4.1 Introduction

The fundamental responsibilities of Emergency Medical Service (EMS) systems are to provide urgent medical care, such as pre-hospital care, and to transport the patient to the hospital if needed. The efficiency of EMS systems is a major public concern, because providing urgent medical care can literally mean the difference between life and death. Over the past thirty years a significant amount of research studies have been conducted to improve the performance of EMS systems by providing effective and efficient service to the public.

The response time, the time elapsed from an emergency call arriving at the dispatch center until the time an emergency vehicle arrives at the scene, is vital in minimizing the impact of the incident. Rapid response times by EMS systems can reduce the fatality of an emergency incident [7]. For example, the survival probability in critical emergency incidents such as a trauma, can be expressed as a function of the time to treatment [14]. Therefore, much of the research focus has been on reducing response time and thereby improving the performance of such EMS systems. There are three main resource allocation decisions in emergency medical service that can be used to reduce the response time. Locating ambulances is one class of

resource allocation problem and a widely used method in EMS systems. Relocating ambulances is another category of resource allocation problem which focused on dynamically relocating vehicles in real-time in order to increase the coverage based on demand patterns. Dispatching emergency vehicles is the third category of resource allocation problem, which is less studied, for improving EMS system performance. In this study our focus is to improve the performance of EMS systems by implementing dispatching strategies to use with currently available resources.

Our contribution towards improving the performance by implementing dispatching strategies is two fold. One area is to determine the response boundaries for each EMS vehicle, because when operations of EMS vehicles are restricted to predetermined response areas (or boundaries), it enables the EMS system to decrease the average response time of paramedic support to the scene [10]. In addition, Larson et al. [25] showed that determining districts (the region of primary responsibility of a response unit) of each unit is an important decision for EMS systems in order to balance the workload between paramedic units. Therefore, in this study one objective is to determine a response area for each ambulance by partitioning the service region of the EMS system into districts. The second area is to propose intra-district (within the district) and inter-district (out of district) dispatching discipline to improve the performance of the EMS system. The importance of implementing better intra-district and inter-district dispatching policies, is well documented by Larson et al. [27]. Thus, the second objective of this work is to implement dispatching rules in order to improve the performance of the EMS system. These dispatching strategies are developed incorporating the degree of the urgency of the call, because priority dispatching strategies can improve the survival probability of patients [3]. We used the survival probability as the performance measure to study the impact of the partitioning on the overall performance of the EMS system, since measuring the patients' survivability reflect the patient outcome directly (Erkut et al. [13]). This study presents a constructive heuristic for dividing the service region into sub-regions (or districts) to determine the response boundaries of each paramedic unit (ambulance). In addition, a simulation approach is used to study the performance of the EMS system by introducing integrated dispatching and districting policies to the system.

4.2 Related work

In the past three decades a significant amount of research work had been conducted to improve the performance of emergency services such as Emergency Medical Services (EMS), fire companies and police emergency services. The models developed for EMS systems address three vital decisions; location,

relocation and dispatch of paramedic units. However, most of the models focused on decision context of locating paramedic units at optimal stations in order to provide better coverage for the service population. A few such static mathematical covering models are LSCP [49], MCLP [11], TEAM [43] and DSM [16]. Probabilistic covering location models were developed incorporating the ambulance busy probabilities in the model. MEXCLP [12], MALP [38], AMEXCLP [4], PLSCP [39], Q-PLSCP [29] and McLay [31] are few examples for probabilistic covering models. The focus of ambulance relocation models is to dynamically relocate ambulances in real-time when dispatching ambulances to provide better coverage for population, since dispatching of an ambulance may leave a significant proportion of population without coverage (e.g. [22],[17]). The literature related to the paramedic unit dispatching problem will be discussed next in this section, since our main focus of this work is to study dispatching strategies to improve the performance of the EMS system.

Most of the existing models we discussed above mainly focused on the optimal location of vehicles to provide better coverage with different objective functions. However, determining response areas for each ambulance and dispatching appropriate paramedic unit to the scene are less studied but important decision contexts in EMS systems. The importance of determining response areas for emergency vehicles is clearly explained by Larson ([25], [27], [26]). Larson introduced this problem (determining response area) as a "districting" problem. Larson stated the districting problem as follows: "How should the service region be partitioned into areas of primary responsibility (districts) so that a level or combination of levels of service is best achieved?" [25].

One early example for vehicle districting is Smith [46]. A gradient search technique was used in his study to redesign police patrol response areas so as to minimize the traveling time within the district. Later Gass [15] developed a heuristic technique to determine police patrol districts that balance the call rate from each district. However, the work done by Larson ([25], [27], [26]) and Carter et al. [10] towards the districting problem influenced the later research significantly. Larson [26] developed a hypercube queuing model to represent the EMS system and to obtain several performance measures. In later research [25] he discussed the importance of intra-district (within districts) dispatching and inter-district (out of district) when designing EMS systems. Though, in few earlier models ([41], [44], [45]) all responses are assumed to be intra-district, and a district is assumed to be all elements closet to the each facility location. Carter et al. [10] also studied the effect of districting on the performance of the EMS system. They used average response time and workload imbalance as performance measures. Carter found that dispatching the closest paramedic unit is not always optimal and dispatching units according to pre-determined boundaries can decrease the average

response time.

Thus, in our study one goal is to determine response area (or district) for each ambulance while the second goal is to obtain dispatching policies for dispatching units within districts and out of districts in order to increase survival probability. In this study, first we propose a computational time efficient, constructive heuristic to obtain districts. The second goal of this study is to determine intra-district and inter-district dispatching policies. These dispatching policies are developed considering the degree of the urgency of the call. The benefits of priority dispatching strategies, are well documented by [3], [34],[19] and [35]. A simulation approach is used to study the performance of the proposed integrated districting and dispatching polices for EMS systems.

4.3 Methodology

Given an emergency service area of an EMS system and its demands (number of calls requesting emergency vehicles), we wish to divide the service area into K districts, such that each district determines the response area for an ambulance or a set of ambulances. In addition, it is desirable that a district consists of contiguous demand zones. For example, Figure 4.1 shows a geographical region partitioned into three sub-regions. A1, A2, A3 and A4 are the positions of each of the ambulance stations (with a paramedic unit at each station) in the service region. According to the partitioning, ambulance A2 and A3 serve calls arriving from district 2 while ambulance A4 serves for district 3 and ambulance A1 serves for district 1. Determining response districts for each paramedic unit such as this is done in order to minimize the average response time or to balance the work load among ambulances (Carter et al. [10]). In this study we proposed two methods to obtain districts. First we proposed an integer programming model to divide the service area in to districts, because the integer programming approach is a widely used method in districting, especially in political districting [21], [6], [2]. However, those work show that when the problem size increases the computational time to obtain a solution grows significantly. Thus we proposed a constructive heuristic to determine vehicle districts within less computational time, as our second method.

After determining vehicles districts, our next goal is to develop inter-district and intra-district dispatching policies for paramedic units. We considered dispatching rules that are static, meaning that we considered an EMS system with fixed deployment (paramedic units are located at specific stations, respond from stations, and return back to their station after serving the call). In addition, we considered stationary dispatching rules; in other words, we assumed that EMS systems followed an ordered preference list of ambu-

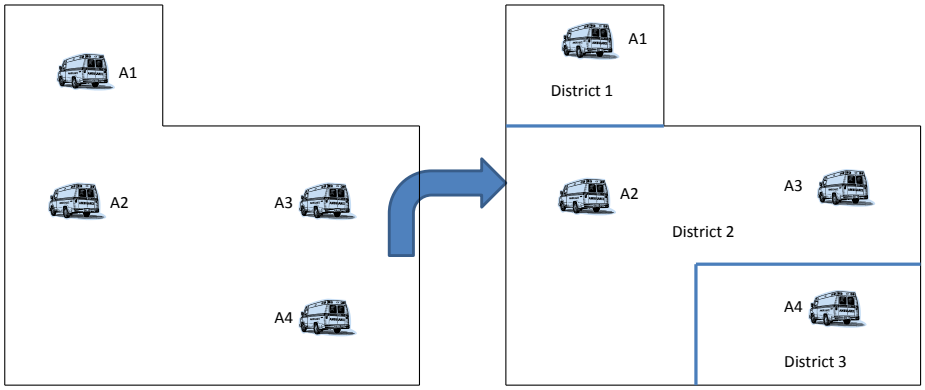


Figure 4.1: Example for partitioning

lances, when dispatching paramedic units. We studied several within-district and out of-district dispatching policies as follows:

Intra-district policies

To dispatch ambulances within districts we considered two polices. The first policy is the myopic policy of sending the closest available unit, giving no reference to the degree of the urgency of the call. Myopic rule (closest rule) is the most common dispatching rule existing in EMS systems. The second policy is the heuristic policy, developed by Bandara et al. [3] to dispatch paramedic units. The key differences between this dispatching heuristic and the myopic policy can be expressed as follows: One is that the dispatching heuristic was developed considering the severity of the call, the other is that the heuristic policy helps to balance the work load between units.

Inter-district policies

Here we considered two dispatching polices to cross district boundaries when all ambulances within the district are busy. In the first policy, we assumed that other emergency services such as fire engines or ambulances from other counties will assist the calls, when all available ambulances within the district are busy . These

kind of dispatching rules are common for EMS systems. For example, in Hanover County, Virginia EMS department, these types of arrangements are in place. An ordered preference list of ambulances is used, to dispatch paramedic units within the district. We refer this dispatching policy as "Nocross rule" throughout this paper from here onwards. The second policy is to dispatch ambulances from other districts to assist calls if all ambulances within the district are busy. We refer to this policy as "Cross rule" in this study. In this rule we also used an order preference list of ambulances to cross boundaries. We utilized the heuristic policy, developed by Bandara et al. [3] to obtain this preference list.

These intra-district and inter-district dispatching policies result in four different dispatching rules, which can be summarized as follows.

1. Closest-Nocross
2. Closest-Cross
3. Heuristic-Nocross
4. Heuristic-Cross

We used the survival probability as the performances measure to study the impact of the integrated districting and dispatching polices on the overall performance of the EMS system, because measuring the patients' survivability mirrors the patient outcome directly (Erkut et al [13]). To obtain the patient survivability we used the survival function developed by Larsen et al. [24] and subsequently simplified by McLay et al. [33]. The survival function is explained below. Let S denote the patient survival probability,

$$S(t_R) = \max[(0.594 - 0.055 * t_R); 0], \tag{4.1}$$

where t_R represents the response time. A simulation model is developed using Arena software to represent the EMS system and to obtain the patient survival probability.

4.4 Mathematical Model to Determine Response Boundaries

This section presents the procedure of determining response boundaries for emergency vehicles using a mathematical programming approach. First we used an integer programing model to partition the service region into a desired number of districts K (partition size). Then we apply different dispatching policies as introduced in the previous section. Finally we used a simulation-based approach to ascertain the

performance (survival probability of patients) of the partitioned EMS system to pick the best partitioning size and dispatching policy combination. Figure 4.2 depicts the procedure of obtaining the best number of districts (best K value) to operate for an EMS system, where $K = 1, \ldots, m$ (number of ambulance).

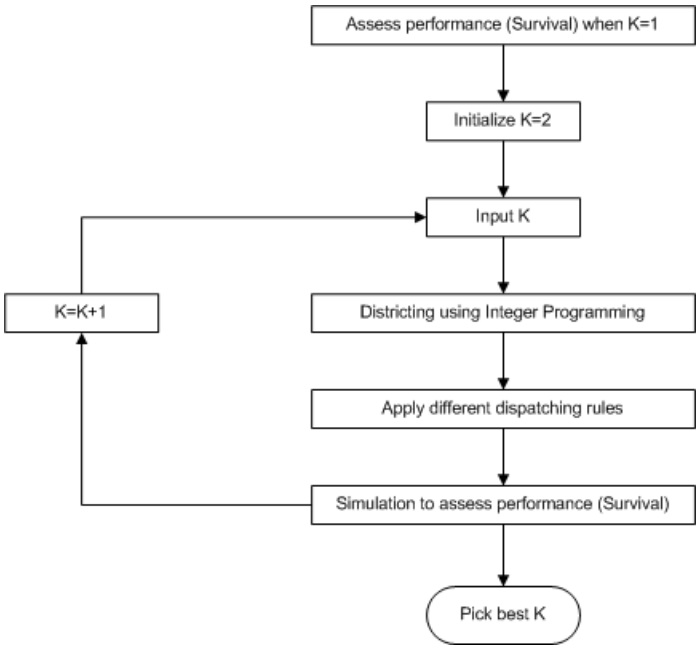


Figure 4.2: Procedure-Integer Programming

4.4.1 Mathematical Model Formulation

For a given geographical region of an EMS department, we are interested in partitioning this region into sub-regions (or districts) which determines the response area for an ambulance or a set of ambulances. Partitioning such as this may be done in order to decrease the average response time for emergency calls. Since the response time mainly depends on the traveling distance of ambulances, our objective is to minimize the weighted traveling distance of each ambulance. Therefore, districting is done so as to minimize the distance between demand zones and corresponding ambulance stations while minimizing the travel distance within a sub-region. Euclidean distance is used as the traveling distance between any two points within the service region. We used an integer programming approach to model this problem.

The model developed below assumes that an EMS system is partitioned in to square cells to represent demand zones. For example, Figure 4.3 shows a service region is partitioned into 28 demand zones. Further, assume that the EMS system has m ambulance stations, the positions of ambulances are known and each

ambulance station is located at the middle of the corresponding demand zone. For example, the EMS system shown in Figure 4.3 has 4 ambulance stations and positions are named as $A_1, A_2, \ldots, A_4$.

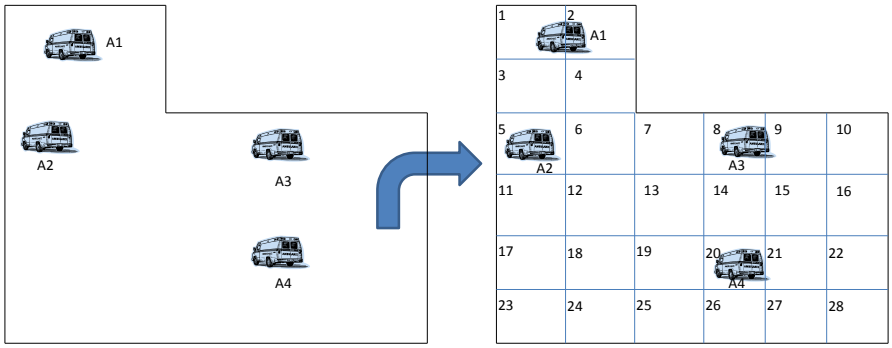


Figure 4.3: Square cells of demands

The notation used in the mathematical model is described below.

n = Number of demand zones

m = Number of ambulance stations with an ambulance in each station

K = Number of possible sub-regions or districts

d_{ij} = Euclidean distance from demand zone i to ambulance station j

a_{il} = Euclidean distance between demand zone i and demand zone l

h_i= Demand of zone i (number of calls)

Let $N = \{i, l : i, l = 1, 2, \ldots, n\}$, $M = \{j : j = 1, 2, \ldots, m\}$ and $\bar{K} = \{k : k = 1, 2, \ldots\ K\}$ be the index sets. The decision variable X_{jk}, indicates the allocation of ambulances to sub-regions.

$$X_{jk} = \begin{cases} 1 & \text{if ambulance } j \text{ is assigned to sub-region } k; \\ 0 & \text{otherwise.} \end{cases}$$

The decision variable Y_{ijk}, indicates the allocation of ambulances in a sub-region to a demand zone.

$$Y_{ijk} = \begin{cases} 1 & \text{if demand zone } i \text{ is covered by ambulance } j \text{ in sub-region } k; \\ 0 & \text{otherwise.} \end{cases}$$

The decision variable Z_{ik}, determines the allocation of demand zones to sub-regions.

$$Z_{ik} = \begin{cases} 1 & \text{if demand zone } i \text{ is assigned to sub-region } k; \\ 0 & \text{otherwise.} \end{cases}$$

The decision variable D_{ilk}, indicates which demand zones are included in the same sub-region.

$$D_{ilk} = \begin{cases} 1 & \text{if both demand zones } i \text{ and } l \text{ is assigned to sub-region } k; \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, we can formulate the linear integer programming model as follows.

$$\text{Minimize} \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{m} h_i d_{ij} Y_{ijk} + \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{l \geq i}^{n} a_{il} D_{ilk}$$

$$\text{subject to } \sum_{j=1}^{m} \sum_{k=1}^{K} Y_{ijk} \geq 1 \; \forall \, i \in I \tag{4.2}$$

$$\sum_{k=1}^{K} X_{jk} = 1 \;\; \forall \, j \in J \tag{4.3}$$

$$\sum_{j=1}^{m} X_{jk} \geq 2 \;\; \forall \, k \in \bar{K} \tag{4.4}$$

$$Y_{ijk} \leq X_{jk} \;\; \forall \, k \in \bar{K}, \; \forall \, j \in J, \; \forall \, i \in I \tag{4.5}$$

$$\sum_{i=1}^{n} Z_{ik} = 1 \;\; \forall \, k \in \bar{K} \tag{4.6}$$

$$Z_{ik} + X_{jk} - Y_{ijk} \leq 1 \;\; \forall \, k \in \bar{K}, \; \forall \, j \in J, \; \forall \, i \in I \tag{4.7}$$

$$D_{ilk} \leq Z_{ik} \quad \forall\, k \in \bar{K},\ \forall\, i,l \in I \ \ and \ \ i \neq l \tag{4.8}$$

$$Z_{ik} + Z_{lk} - D_{ilk} \leq 1 \ \ \forall\, k \in \bar{K},\ \forall i,l \in I \ \ and \ \ i \neq l \tag{4.9}$$

$$X_{jk}, Z_{ik}, Y_{ijk}, D_{ilk} \in \{0,1\}, \ \ \forall\, i \in I, \ \ \forall\, j \in J,\ \forall\, k \in \bar{K} \tag{4.10}$$

The objective function consists of two parts. It is the sum of the weighted distances from demand zone i to ambulance station j and the distance between two demand zones in the same sub-region. The first term of the objective function computes the distance from each demand zone to each ambulance station within the district. This term is used to minimize the weighted traveling distance from ambulance station to demand zones. As we mentioned earlier, we used Euclidean distance between demand zone and ambulance station as the traveling distance. The second term of the objective function calculates the total Euclidean distance between each demand zone within a particular sub-region in the manner such that $i, l \in S$ for $l \geq i$. (Selecting i and l according to this manner, helps us to reduce the running time of the model to some extent, because the distance between i and l is equal to the distance between l and i, it is not necessary to consider the same distance two times when a term is minimized.) Therefore, the second term of the objective function helps to minimize the traveling distance inside a sub-region by providing a compact sub-region.

Constraint (4.2) ensures that every demand zone is covered by at least one paramedic unit. Constraint (4.3) ensures that each ambulance is allocated to only one sub-region. Constraint (4.4) ensures that there are at least two ambulances in each sub-region. This constraint is used to ensure that every demand zone has a back up coverage when one ambulance is busy. We can relax this constraint according to the back up coverage we desire. For example, if we are interested in three ambulances for back up coverage R.H.S of this constraint should equal to three. If we are not interested in back up coverage then R.H.S of this inequality becomes one. Constraint (4.5) ensures that if demand zone i is covered by ambulance j in sub-region k then that ambulance is assigned to sub-region k. Constraint (4.6) ensures that demand zone i is only assigned to one sub-region. Constraint (4.7) ensures that if both demand zone i and station j are in the same sub-region k (i.e., $Z_{ik} = 1$ and $X_{jk} = 1$) then that demand zone i, can be covered by ambulance j in sub-region k (i.e., $Y_{ijk} = 1$). Consider constraints (4.8) and (4.9). These two constraints ensure that D_{ilk} must equal to zero unless both

Z_{ik} and Z_{lk} equal 1. In addition, constraint (4.9) ensures that if both demand zone i and l are selected for sub-region k then D_{ilk} have value 1. The constraint set (4.5) is summarized in the first two columns of Table 4.4.1 which shows the possible values for Y, X variables. Similarly the constraint set (4.7) is summarized in column 3 to column 5 of Table 4.4.1 which shows the possible values for Z, X, Y variables and the constrain set (4.8)-(4.9) is summarized in column 6 to column 8 of Table 4.4.1 which shows the possible values for Z, D variables. A few computational examples are discussed in the next section to obtain the performance of this proposed mathematical model.

Y_{ijk}	X_{jk}	Z_{ik}	X_{jk}	Y_{ijk}	Z_{ik}	Z_{lk}	D_{ilk}
0	0 or 1	0	0	0 or 1	0	0	0
1	1	1	0	0 or 1	1	0	0
		0	1	0 or 1	0	1	0
		1	1	1	1	1	1

Table 4.1: Explanation of the constraints (4.5), (4.7)–(4.9)

4.5 Computational Results Using Mathematical Model

In this section we present computational results for two examples using the mathematical model developed. Each optimization problem was formulated using Optimization Programming Language (OPL) which uses a solver called CPLEX to solve the mathematical models and Excel was used to input the data. Hanover County, Virginia is used as our study region and we considered the urban area of this county in our mathematical model. The Hanover County EMS department responds to 911 calls 24 hours a day with a population of approximately 100,000 and an area of 471 square miles. Based on data collected in 2007, the average number of calls in Hanover is 1.2 calls/hour (peak rate) with 9521 total calls through out the year. An instance with forty one demand zones, five ambulance stations with a paramedic units at each station was considered as our first example. All demand zones and station locations are shown in Figure 4.4. Rescue stations A1, A2, A3, A4 and A5 are the five stations with a paramedic unit at each station.

The solution for this example for different K (number of sub-regions) values given by the mathematical model is depicted in Figure 4.5 and 4.6. When $K > 2$ to obtain a feasible solution we need to change the right hand side of constraint set (4.4) to 1 because with five ambulances it would be impossible to assign at least two ambulances for more than two districts. When $k = 1$, the model does not partition the region in to subregions or districts (Figure 4.5-a). When $K = 2$, the optimal sub-regions given by the model is shown in Figure 4.5-b. The demand zones that belong to different sub-regions are shaded using different

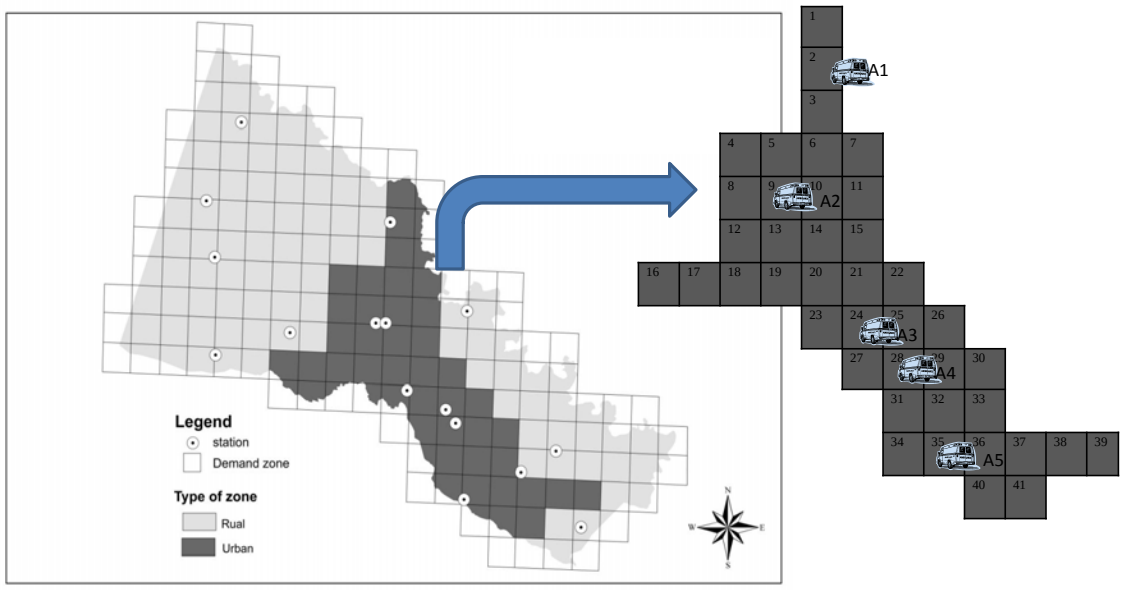


Figure 4.4: Study Area- Hanover County EMS department in Virginia

colors. According to this optimal solution ambulance $A1$ and ambulance $A2$ are allocated to one sub-region and ambulances $A3$, $A4$ and $A5$ are allocated to another sub-region. When $K = 3$, $K = 4$ and $K = 5$, the optimal districts are shown in Figure 4.6.

After determining the districts, to obtain the best K, the next step is to apply different dispatching policies to the EMS system (see Figure 4.2). Then we did a simulation analysis to determine the overall performance of the partitioned EMS system. As mentioned earlier, we used the survival probability as the performances measure to study the impact of the partitioning on the overall performance of the EMS system.

4.5.1 Performance of the Partitioned EMS System

A simulation model is developed using Arena software to represent the EMS system and to obtain the patient survival probability. We assumed that this EMS system operates according to the dispatching rules

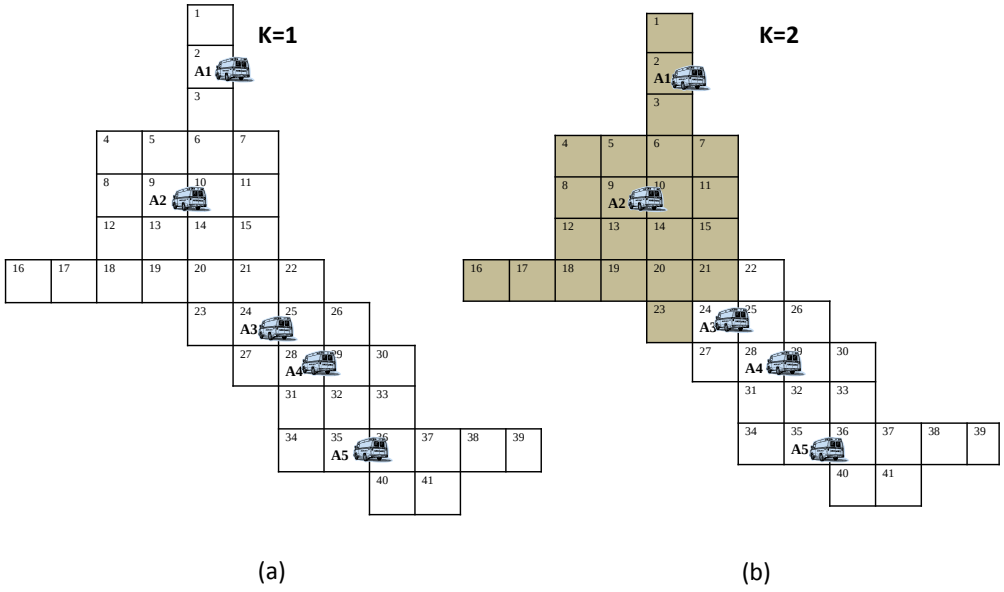


Figure 4.5: Districts given by mathematical model when $K = 1$ and $K = 2$

we introduced in section 4.2. However, when comparing the performance of dispatching policies for $K > 1$, we only illustrate the performance of Heuristic-Nocross rule and Heuristic-Cross rule. This is because, we found that the myopic policies, e.g Closest-Nocross rule and Closest-Cross rule, always perform worse than the heuristic policies. We defined Heuristic-Nocross rule as Nocross rule and Heuristic-Cross rule as Cross rule for convenience when comparing performance of those policies. We leave the $K = 1$ case (heuristic and closest) as a baseline reference.

We considered the previous example with different K values to compare the performance of partitioning. Figure 4.7 compares the survival rate for two dispatching rules (Nocross rule and Cross rule) with Closest policy and Heuristic policy when $K = 2$ and $K = 3$. Figure 4.8 compares the performance when $K = 4$ and $K = 5$.

According to those graphs when call arrival rate increases, survival rate decreases gradually for every policy as expected. In addition, we observed that the Heuristic rule performs well in comparison to Closest rule always, when the system operates without partitioning (when $K = 1$). However, the partitioning approach with Nocross rule is better than no partitioning in terms of patients survivability (see $K = 2$). This

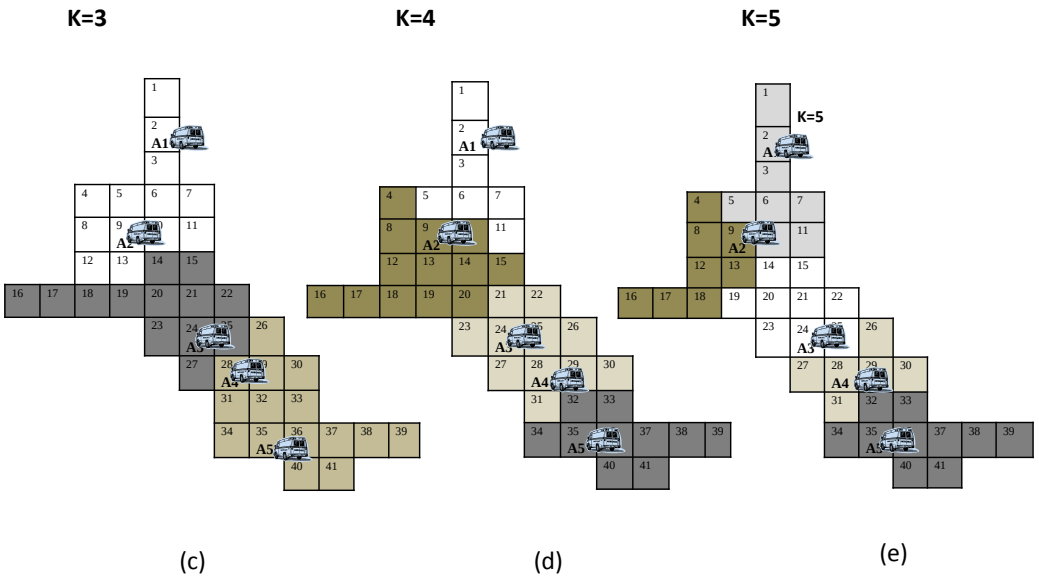


Figure 4.6: Districts given by mathematical model when $K = 3$, $K = 4$ and $K = 5$

occurs because in this instance partitioning allows the Priority 1 calls to be more likely to be served by the closest ambulance (instead of that ambulance being busy serving a call that is farther away). However, this observation is not true when $K \geq 3$. In these cases some of the subregions do not have backup coverage when the ambulances are busy. For example, when $K = 4$ (see Figure 4.6) , only one district operates with two ambulances (A3 and A4), while other districts have one ambulance (no backup coverage) each. Thus we can conclude that "over-partitioning" can reduce the effectiveness of an EMS system.

The performance of Nocross rule and Cross rule is worse than Closest rule performance when $K = 4$ and $K = 5$ (see Figure 4.8). This observation is also justified by Figure 4.9. Therefore, we can conclude that the optimal partitioning size is two for this example with Nocross rule, since operating the EMS system with two districts ($K = 2$) helps to save more lives than operating without partitioning. For example, there is a 14% increase in the average survival rate of Priority 1 patients with Nocross rule when call arrival rate is 3/hour and $K = 2$, compared to no partitioning Closest rule. If we assume that this EMS system receives 1000 life-threatening (Priority 1) calls during a year this corresponds to approximately an additional 39 lives saved without utilizing any additional resources. To get a sense for this, Hanover county serves around 4760

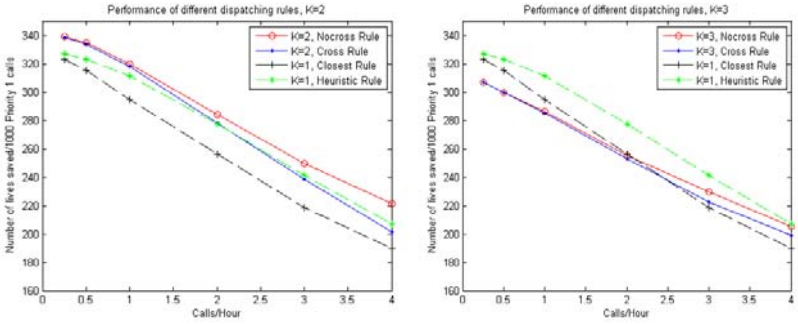


Figure 4.7: Comparison of survival rate when $K = 2$ and $K = 3$

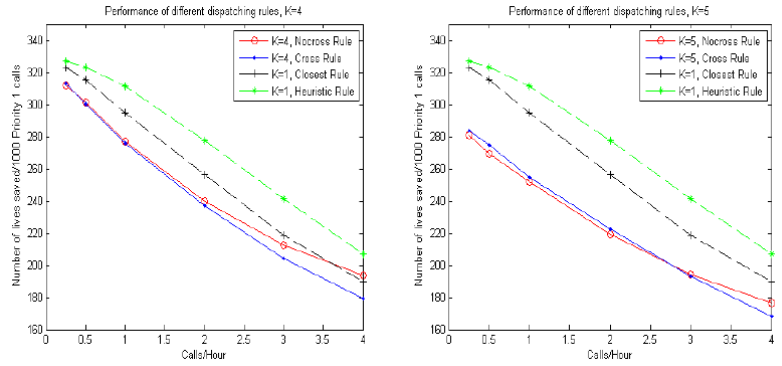


Figure 4.8: Comparison of survival rate when $K = 4$ and $K = 5$

Priority 1 calls per year. After studying several computational examples, it is observed that it may be beneficial to implement districting when operating EMS systems; but that one must be careful in selecting the number of sub-regions so as not to reduce backup coverage. To study this further we conider an EMS system with 6 ambulances, in which case $K = 3$ can still result in backup coverage for all districts. However, in this instance we assign demands to each zone randomly in order to have evenly distributed deamnds throught the service region. Solutions to this instance for different K values given by the mathematical model is shown in Figure 4.10 and Figure 4.11.

After obtaining districts, we did a simulation to asses the performance of partitioned EMS system in order to find the best K. We studied the performance of Nocross rule and performance of Cross rule for different K values. Figure 4.12 illustrates the performances of Nocross rule and Cross rule. By observing this graph we can conclude that best number of districts for this example is three ($K = 3$) with Nocross rule. Operating this EMS system with three districts is better than operating without districts ($K = 1$) in terms of

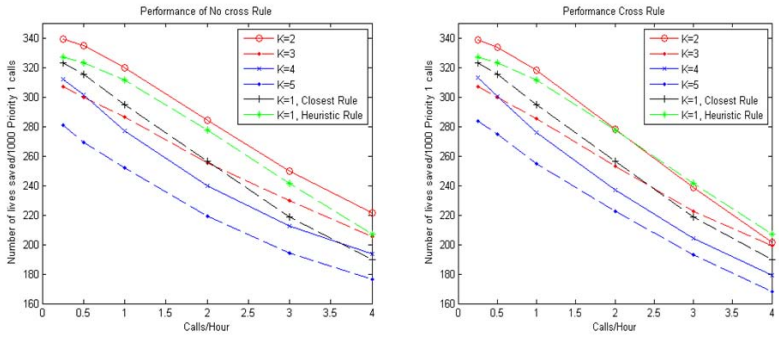


Figure 4.9: Comparison-performance of Nocross rule and Cross rule

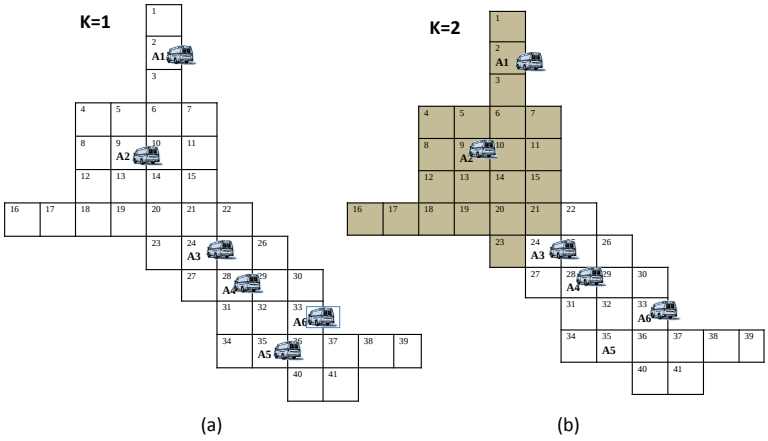


Figure 4.10: Vehicle districts when $K = 1$ and $K = 2$

patients' survivability. There is a 20% increase in the average survival probability of patients with Nocross rule when call arrival rate is 1/hour and $K = 3$.

In this study, we formulated a mathematical programming model to partition an EMS service area in to sub-regions to determine the service boundaries of the ambulances. Results show that it is beneficial to operate the EMS system with partitioning as long as backup coverage can be maintained. Though, the proposed integer model provides the vehicle districts, the execution time to obtain a solution is significant and grows with problem size. OPL running times for computational examples considered in this study are summarized in Table 4.2). Thus we proposed a constructive heuristic to determine vehicle districts in the next section.

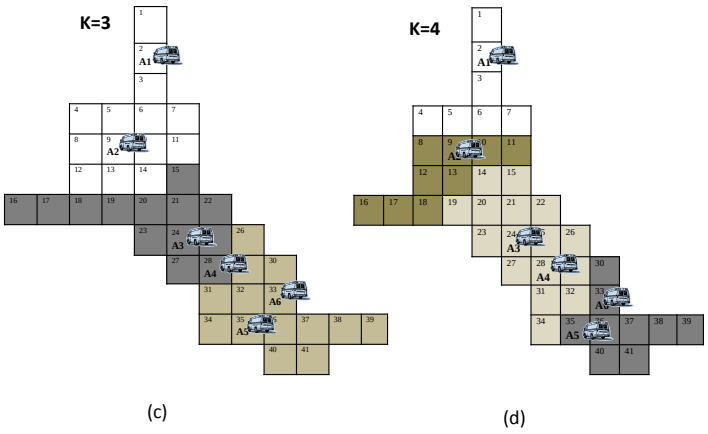


Figure 4.11: Vehicle districts when $K = 3$ and $K = 4$

Example	*K value*	*OPL Running Time(mins)*
41 zones and 5 ambulances	2	4
	3	43
	4	140
	5	272
41 zones and 6 ambulances	2	4
	3	53
	4	1110

Table 4.2: OPL running Times

4.6 A Constructive Heuristic to Determine Vehicle Districts

The districting problem is not only found in EMS literature but also in political districting literature. The most popular way to obtain district is to use mathematical programming such as integer programming or mixed integer programming (e.g. [21], [6], [2]). However, in these models execution time is significantly large. Thus in this study we proposed a constructive heuristic to obtain emergency vehicle boundaries within less computational time. In this heuristic we utilize the Adjusted Expected Coverage (AEXC) [4] concept for obtaining districts. The AEXC describes the ability to cover the demand zones taking in account ambulance busy probabilities and their dependencies. This AEXC concept is used in many mathematical models in order to find the location of ambulance stations. We used the objective function proposed by R.Batta et al. [4] to

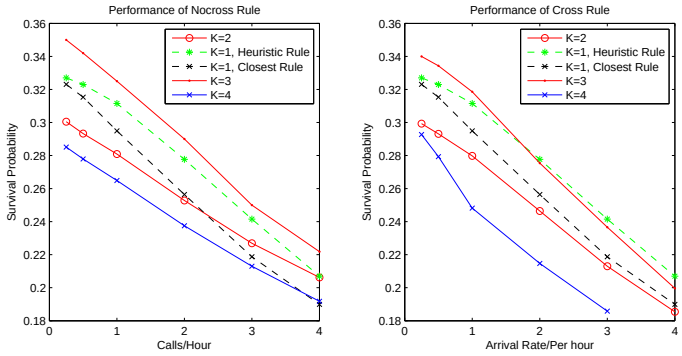


Figure 4.12: performance of Nocross rule and Cross rule

calculate the AEXC. In that function they relaxed the independence assumption when calculating server busy probability using the correction factor $Q(M, \rho, j)$, which is proposed by Larson [27]. The objective function can be summarized as follows:

Maximize $\sum_{i=1}^{N} \sum_{j=1}^{M-1} (1-\rho)\rho^j h_i y_{j+1,i} Q(M, \rho, j)$ where

$$y_{ji} = \begin{cases} 1 & \text{if zone } i \text{ is covered by at least } j \text{ servers;} \\ 0 & \text{otherwise.} \end{cases}$$

N = Number of demand zones

M = Number of ambulance stations with an ambulance at each station

h_i = Demand (number of calls) from zone i

ρ = Utilization factor

They used this objective function to determine ambulance station locations with respect to several constraints so as to maximize the AEXC. However, we can use the same function to calculate the AEXC since we already know the ambulance station locations and the value of y_{ji} with respect to a given distance standard. Thus, we can define AEXC as below:

$$\text{AEXC} = \sum_{i=1}^{N} \sum_{j=1}^{M-1} (1-\rho)\rho^j h_i y_{j+1,i} Q(M, \rho, j) \tag{4.11}$$

As mentioned earlier, AEXC level is used in our heuristic when determining the best number of districts to operate in the EMS system. The procedure for obtaining response boundaries can be summarized as fallows: First we calculate the AEXC for K=1 (without partitioning the service region in to districts). Then we partition the service region in to districts according to the constructive heuristic and calculate the corresponding AEXC level. This procedure is depicted in Figure 4.13. Once we cannot improve the AEXC level by partitioning into districts we discontinue the procedure and pick the districts that maximize the AEXC level as our solution. Simulation results also confirmed that maximum AEXC level maximizes the average survivability of the patients'. We will elaborate these findings in results section. After determining vehicles districts our next goal is to develop inter-district and intra-district dispatching policies for paramedic units in order to improve the performance of the EMS system using the dispatching rules we introduced in section 4.2.

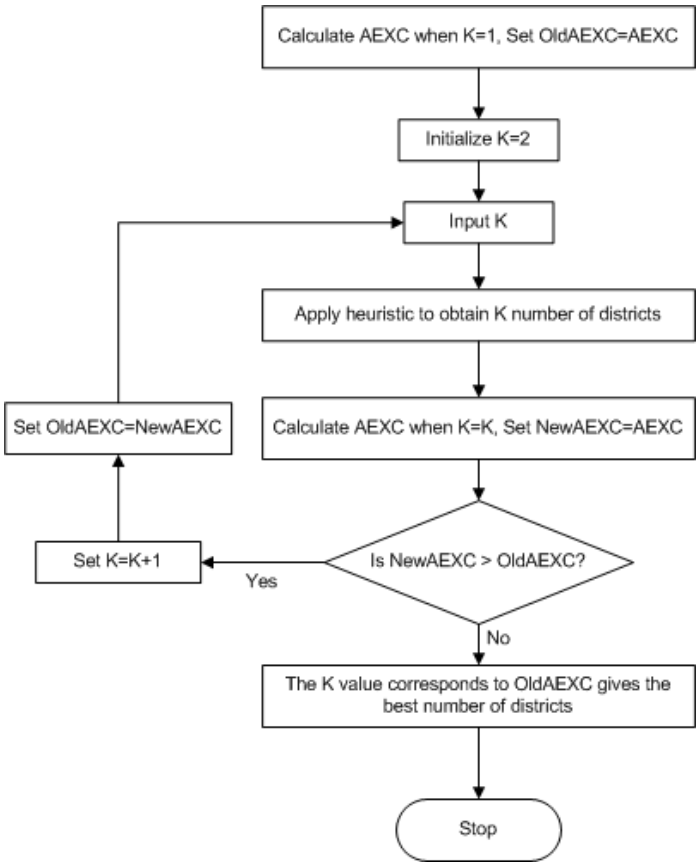


Figure 4.13: Procedure-Constructive Heuristic

4.7 The Heuristic Development

The heuristic algorithm developed below assumes that the response area of the EMS is partitioned into square cells of demand zones. Let N denote the set of demand zones (or nodes) and let E be the set of edges (adjacencies) in the system. According to this partition, the response area can be represented as a $p \times q$ grid. For example, Figure 4.14 depicts that an EMS system is partitioned into 6×6 grid system meaning that an EMS system with 36 demand zones. We included the shaded demand zones in to the EMS service region with zero demand to obtain a $p \times q$ grid system, which helps to construct a heuristic algorithm that can apply for EMS systems have different geographical shapes. In the $p \times q$ grid, any demand zone $i \in N$ can be represented as (x, y) where $x = 1, 2, , p$ and $y = 1, 2, , p$. If two demand zones $i_1 = (a, b), i_2 = (c, d) \in N$ are adjacent, then the ordered pair $(i_1, i_2) = ((a, b), (c, d)) \in E$.

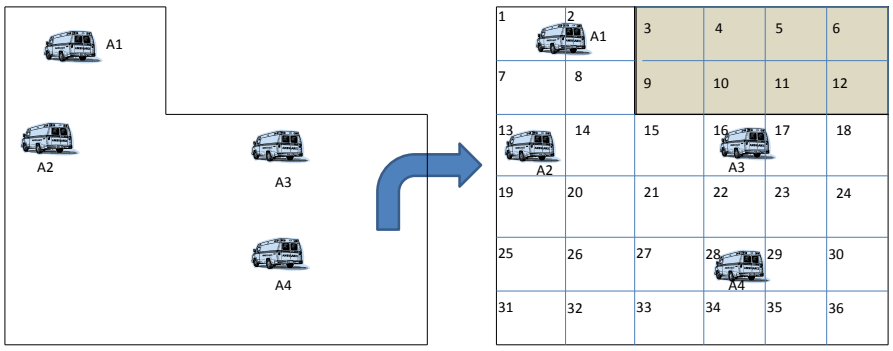


Figure 4.14: Service region is partitioned into square cells of demand zones

We adopt the following notation to represent the demand zones in the grid system. In $p \times q$ grid any demand zone $i \in N$ can be written as $i = (x - 1) * q + y$ where $x = 1, 2, \ldots, p$ and $y = 1, 2, \ldots, q$. For example, the demand zone $(1, 3)$ in Figure 4.14, corresponds to zone 3 which can be obtained as $(1 - 1) * 6 + 3 = 3$. Further, we assumed that the EMS system has m ambulance stations and the positions of all ambulances are known.

The notation used in the heuristic algorithm is summarized below:

i = demand zone

j = ambulance station

N = Number of demand zones

m = Number of ambulance stations with an ambulance at each station

K = Number of possible sub-regions or districts

$A(a,b) = \{(c,d) \in N : ((a,b),(c,d)) \in E\}$= the set of demand zones adjacent to zone (a,b)

h_i = demand of zone i

z_i = proportion of calls from i^{th} demand zone: such that $\sum_{i=1}^{n} z_i = 1$

λ = call arrival rate to the entire system

λ_i= λz_i (call arrival rate from demand zone i)

μ = service rate

M_K = number of servers within the district when there are K districts

ρ = utilization factor for infinite-capacity system; $\rho = \lambda / M_K \mu$

$Q(M, \rho, j)$= "Correction factor" for computing that the $(j+1)$st selected server is the first available server:given that there are total of M servers(ambulances) within the district

ρ_j = fraction of time that unit j is busy serving calls

This heuristic algorithm is developed in order to determine districts that maximize the AEXC level of the EMS system and to balance the workload among ambulances. According to the heuristic, first we obtain m districts assigning each demand zone to its' closest ambulance station. Then, we balance the work load among ambulances by swapping adjacent demand zones between districts. Next, we merge adjacent districts for a given K value to get new districts so as to maximize the AEXC level of the EMS system. The heuristic steps can be summarized as follows:

Step 1 Obtain m districts assigning each demand zone to its' closest ambulance station as described below;

Number the ambulances 1 through m. Let $c_{xy} \in C$ be an $(p \times q)$ matrix, where c_{xy}= number of calls from zone i. Let $d_{xy} \in D$ be an $(p \times q)$ matrix, where d_{xy}= district assignment for each demand zone (x, y). (Note that;$i = (x-1) * q + y$ where $x = 1, 2, \ldots, p$ and $y = 1, 2, \ldots, q$). The matrix D provides the district assignment for each demand zone.
for $x = 1 \rightarrow p$ **do**
  for $y = 1 \rightarrow q$ **do**
    if $c_{xy} \neq 0$ **then**
      calculate distance from zone (x, y) to each ambulance station
      assign $d_{xy} =$ closest ambulance number
    else
      assign $d_{xy} = 0$
    end if
  end for
end for
Return: D

Step 2 Procedure 1-Calculate the work load of each ambulance using the iterative procedure described by Larson [3] in page 859. We modified their procedure according to our notation and assumptions as follows

Let W be a column vector of size $(1 \times m)$, where $w(j)$ denotes the work load of each ambulance and $Var(W)$ be the variance of W.
Step 2.0 Input D matrix
Step 2.1 Initialization

1. Compute from the $M/M/m$ queuing model the exact value for $r \equiv average\ utilization\ factor$, where $r = \lambda/\mu$ for $M/M/m/\infty$ system
2. Set $n = 0$.
3. Define $\hat{\rho_j}(n) \equiv estimate\ of\ \rho_j\ at\ the\ nth\ iteration.$ Set $\hat{\rho_j}(n) = r, j = 1, 2, \ldots, m$

Step 2.2 Iteration
set $n \leftarrow n + 1$
for $j = 1 \rightarrow m$ **do**
  compute $\hat{\rho_j}(n)$ from equation given below.

$$\rho_j = \left[1 + \sum_{j \in G_j^1} \lambda_j + \sum_{j \in G_j^2} \lambda_j Q(m, \rho, 1) r + \sum_{j \in G_j^3} \lambda_j Q(m, \rho, 1) r^2 + \cdots + \sum_{j \in G_j^m} \lambda_j Q(m, \rho, 1) r^{m-1}\right]$$

end for
Step 2.3 Normalize [so that $N^{-1} \sum_{j=1}^{m} \hat{\rho_j}(n) = r$]

1. Compute $\Gamma \equiv \left[N^{-1} \sum_{j=1}^{m} \hat{\rho_j}(n)/r\right]$
2. $\hat{\rho_j}(n) \leftarrow \Gamma \hat{\rho_j}(n)$

Step 2.4 Convergence Test
if $\max |\hat{\rho_j}(n) - \Gamma \hat{\rho_j}(n)| > \epsilon$ **then**
  Goto Step 2.2
else
  STOP
end if
Return: W and $Var(W)$, where $W = [\hat{\rho_1}, \hat{\rho_2}, \ldots, \hat{\rho_m}]$

Step 3 Swap the adjacent demand zones between adjacent districts until workload is balanced among ambulances while remaining the contiguous within districts.

Initialization
$OldW = W$ and $OldVar(W) = Var(W)$
for $k = 1 \rightarrow m$ **do**
  for $i = 1 \rightarrow p$ **do**
    for $j = 1 \rightarrow q$ **do**
      if $d(i, j) = 1$ **then**
        $\forall A(i, j)$
        if $A(i, j) \neq k$ and $A(i, j) \neq 0$ **then**
          $OldA(i, j) = A(i, j)$
          $A(i, j) = k$
          Update D
          $[W, Var(W)] = Callprocedure1(D)$
          if Var(W) $\leq$OldVar(W) **then**
            exit
          else
            $A(i, j) = OldA(i, j)$
            Reset D
          end if
        end if
      end if
    end for
  end for
end for
Return: D

Step 4 Merge adjacent districts according to the number of districts (K) desired according to the following procedure.

Step 4.1. Initialize $K = 1$
Set $OldAEXC = AEXC$
Step 4.2 Set $K = K + 1$
Merge adjacent districts as fallows.
Let P be a list with K rows, where a row of P denotes the possible adjacent districts to merge and obtain a new district. i.e., $p(x, y) \subset m$
for $k = 1 \rightarrow K$ **do**
  for $i = 1 \rightarrow p$ **do**
    for $j = 1 \rightarrow q$ **do**
      for $l = 2 \rightarrow length(P(k, :))$ **do**
        if $d(i, j) = P(k, l)$ **then**
          set $d(i, j) = P(k, 1)$
        end if
      end for
    end for
  end for
end for
Return: D

Step 5

1: Input matrix D
2: Calculate AEXC using equation (4.12)
3: Set $NewAEXC = AEXC$
4: **if** $NewAEXC > OldAEXC$ **then**
5: Set $OldAEXC = NewAEXC$
6: Goto step 4
7: **else**
Exit
8: **end if**
9: K value provides the best number of districts to operate in EMS system and matrix D provides the allocation of demand zones to each district.
10: *Return:* D and K

4.8 Computational Results Using Constructive Heuristic

This section provides the computational results for several examples using the constructive heuristic. Hanover County, Virginia is used as the study area of these examples. Hanover County is a semi-rural, semi-suburban county in the metropolitan Richmond area. The study area is depicted in Figure 4.15. As it is observed the service region is divided into square cells of demand zones. First, we illustrate an instance with 41 demand zones and 6 ambulances (Example 1) which is the urban area of the county. Then, an example with 137 demand zones and 6 ambulances (Example 2) is discussed.

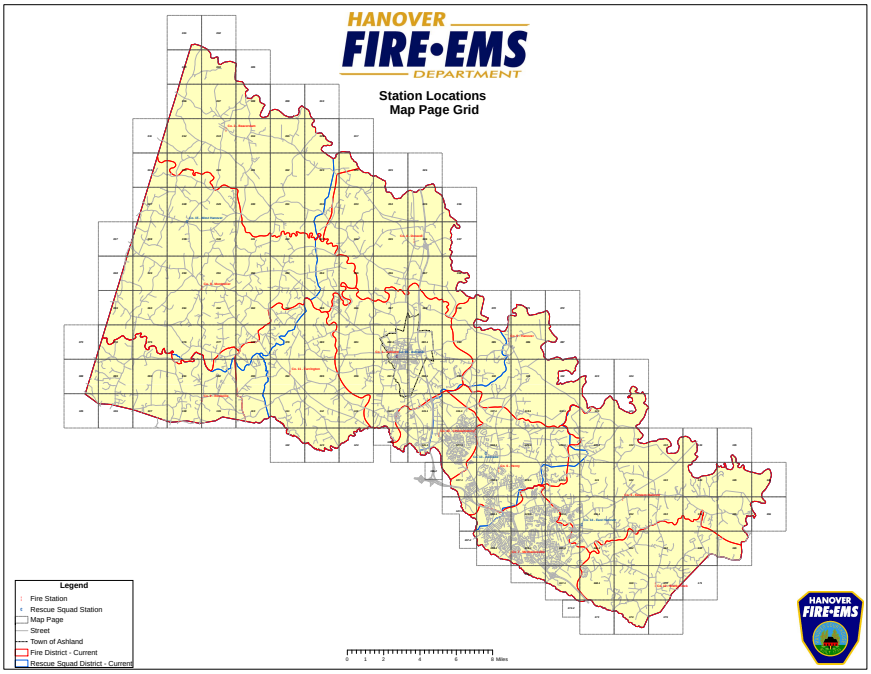


Figure 4.15: Service Region-Hanover County

4.8.1 Example 1

The districts given by the heuristic procedure for this example are illustrated in Figure 4.16 and Figure 4.17. We calculated the AEXC level for each K value to get the best number of districts to operate. Figure 4.18 (left most graph) depicts the AEXC for different K values while varying distance standard. The "distance standard" determines the ability to cover a demand zone by an ambulance. If the distance between a demand zone and an ambulance station is greater than "distance standard" then that node is considered not covered by that ambulance. According to the AEXC graph in Figure 4.18, the AEXC level increases for higher "distance standard" values as expected. In addition, it is observed that once AEXC level start to go down it decreases continuously. For example, the AEXC level when $K = 2$ is greater than $K = 1$. However, it started to go down when $K = 3$ and it remains for $K = 4$. We observed similar results, for other cases we tested. In this example, the AEXC level is greatest when $K = 2$. Thus we can conclude that operating this EMS system with two districts maximize the AEXC level. In addition, we did a simulation analysis to ensure these findings.

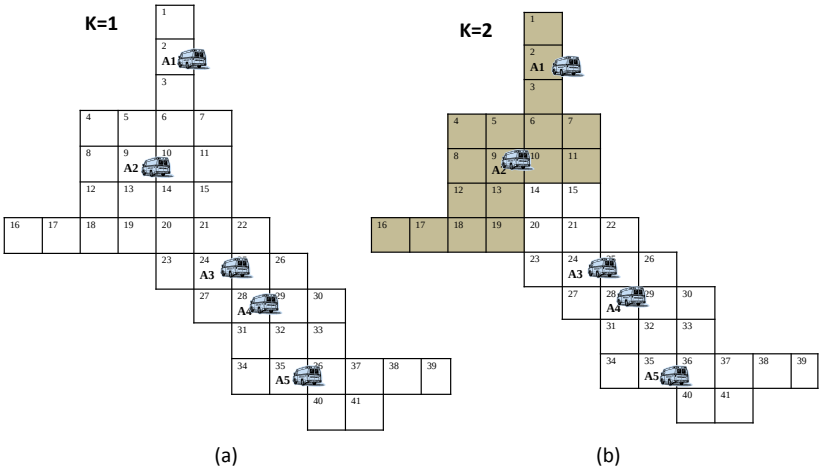


Figure 4.16: Districts given by Constructive Heuristic when $K = 1$ and $K = 2$-Example 1

A simulation model is developed using Arena software to represent EMS system and to obtain survival probability. In this comparison we present only the performance of Nocross rule, because Nocross rule performed better than Cross rule for this instance. Figure 4.18 (right most graph) compares the performance

of Nocross rule for different K values. Figure 4.18 illustrates that operating the EMS system with two districts ($K = 2$) performs better than any other case in terms of patients' survivability. Therefore, best K is two for this example. This observation confirms the previous finding, that the K value corresponds to maximum AEXC level provide the best district size. Additionally, we can conclude that the K value corresponds to maximum AEXC level provides the maximum survival probability. This result is also confirmed by the next example.

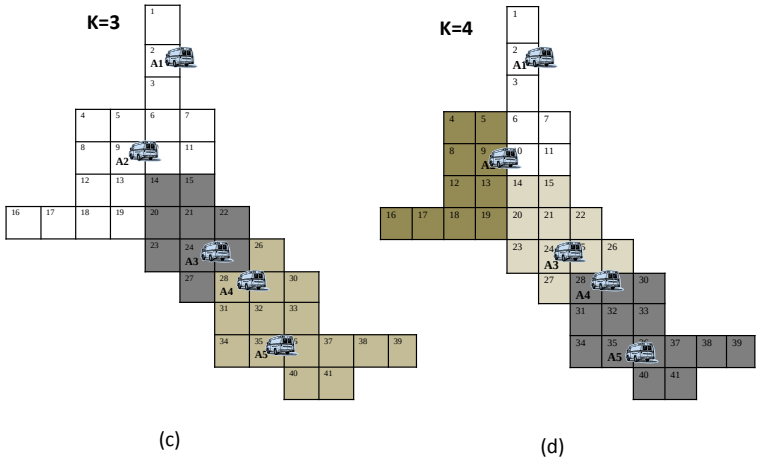


Figure 4.17: Districts given by Constructive Heuristic when $K = 3$ and $K = 4$-Example 1

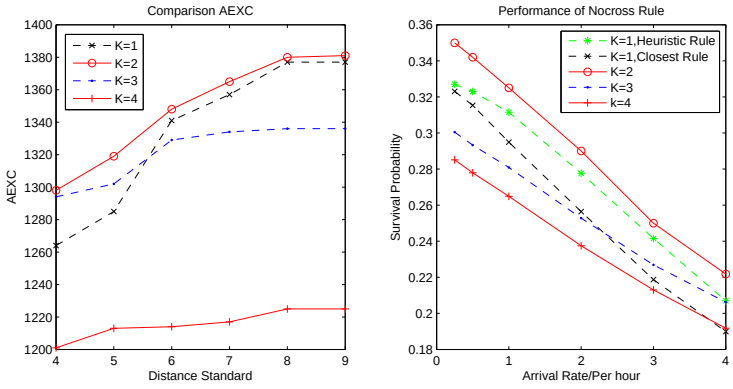


Figure 4.18: Performance of Districting-Example 1

4.8.2 Example 2

The second example is an EMS system with 137 demand zones and 6 ambulances (extracted from Hanover County). The study region of this example is shown in Figure 4.19. The locations of each ambulance station and all demand zones are depicted in Figure 4.19.

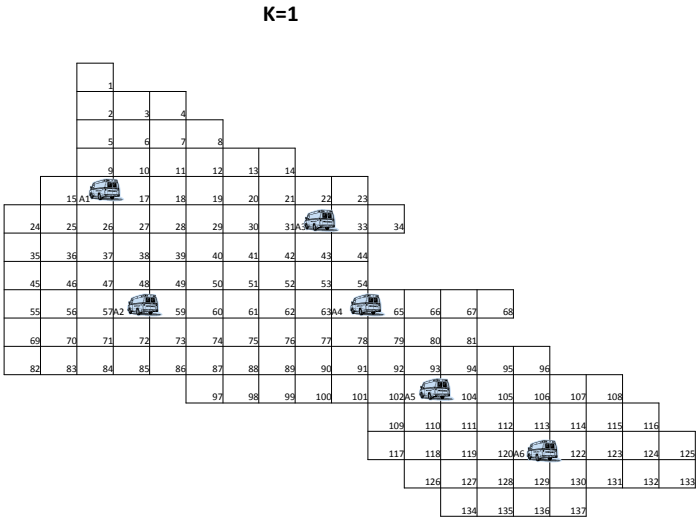


Figure 4.19: Study region- Example 2 ($K = 1$)

The districts given by the heuristic for this example are shown in Figure 4.20. The corresponding AEXC level for these solutions is illustrated in Figure 4.21 (left most graph). When $K > 3$, the heuristic does not provide districts and corresponding AEXC levels, since it cannot be increased the AEXC further. $K = 2$ provides the maximum AEXC level for this example. For $K \geq 3$, the AEXC level likely to decrease. Thus, operating this EMS system with two districts ($K = 2$) is better compared to $K = 1$. Simulation analysis also confirmed that operating this EMS system with two districts help to increase patients' survivability. Figure 4.21 (right most graph) compares the performance of Nocross rule. According to the graph, $K = 2$ performs better than any other k value in terms of survivability. In addition, we observed that the K value corresponds to maximum AEXC level, also maximizes the survival probability of patients for every case we tested.

In the next section, we compare the performance of the mathematical model and the constructive heuristic.

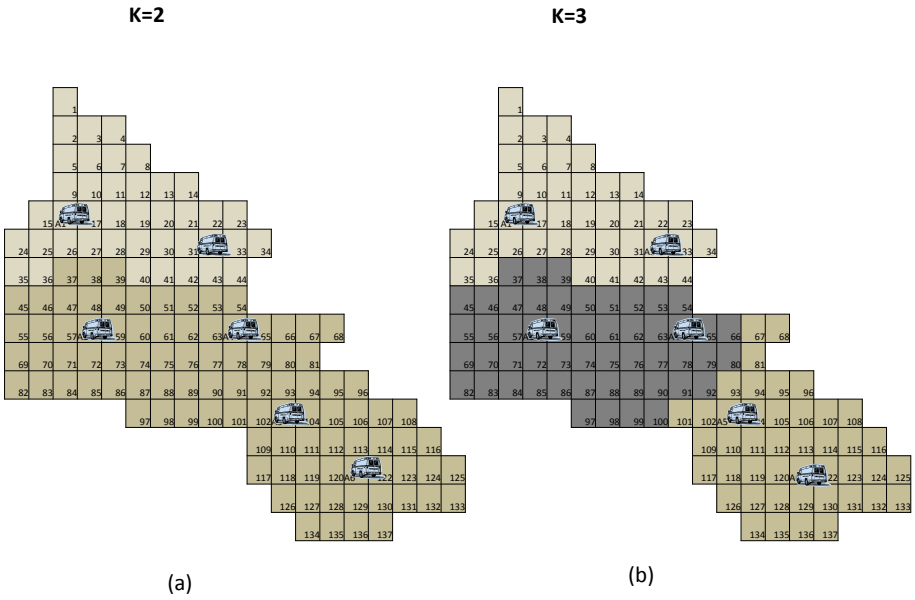


Figure 4.20: Districts given by the Constructive Heuristic when $K = 2$ and $K = 3$-Example 2

4.9 Comparison- Mathematical Model and Constructive Heuristic

This section illustrates the performance of two districting methods (mathematical model, constructive heuristic). Performance measures such as running time, survival probability are used to compare these two methods. In addition, we consider the 41 demand zones and 5 ambulances as an instance for the comparison. We defined the mathematical model as the "Integer model" and constructive heuristic as the "AEXC method" for convenience when comparing performance of those methods.

Example	*K value*	*Integer Model Time (mins)*	*AEXC Method Time (mins)*
41 zones and	2	4	0.05
5 ambulances	3	43	0.05
	4	140	0.1
	5	272	0.15

Table 4.3: Comparison- Running Times

Table 4.3 illustrates the execution time for both methods. The mathematical model was implemented in OPL while the AEXC method was implemented in MATLAB. All programs were executed on a Dell Vostro 1400 computer with a Pentium-IV processor and 2 GB RAM. According to the Table 4.3, we can

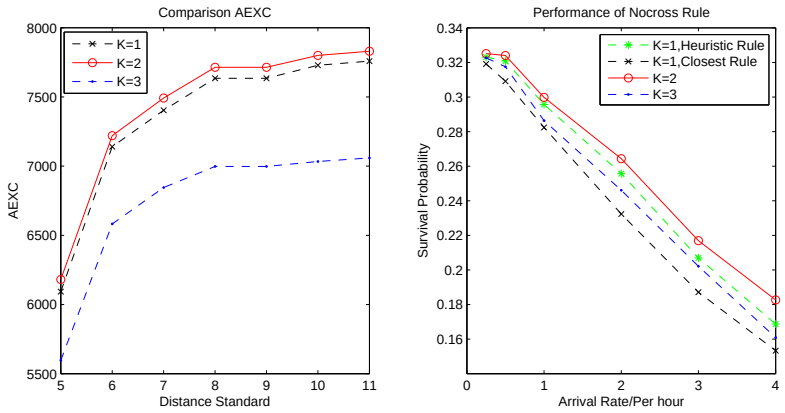


Figure 4.21: Performance of Districting-Example 2

observe that running time of the AEXC method is significantly smaller compared to the Integer model. Thus we can conclude that the AEXC method is computationally time efficient. Next we compared the survival probability of patients according to the solutions given by those two methods. Figure 4.22 compares the survival probability for the Integer method and for the AEXC method using Nocross rule and Cross rule. As the Nocross rule graph (Right most graph) shows, the districts given by the AEXC model increases the patients survivability compared to the districts given by the Integer model. However, in the Cross rule the survival probability difference is not significant for the two methods. Finally, we can conclude that the districts provided by the AEXC method is better compared to the Integer model in term of patients survival probability with the Nocross rule. Operating this EMS system according to the districts given by the AEXC method is better than operating with the solution given by Integer model in terms of patients' survivability. There is a 10% increase in the average survival probability of patients with Nocross rule when call arrival rate is 1/hour.

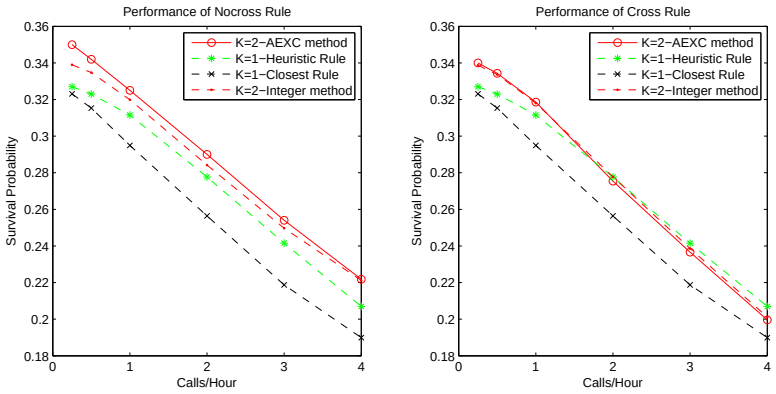


Figure 4.22: Survival Probability Comparison- Integer method and AEXC method

4.10 Sensitivity Analysis

We now explore the effect of response time (the time between the receipt of a call at the dispatch center and the arrival of the first emergency response vehicle at the scene) and turn around time (the time required for an ambulance to return back to its original station after serving the patient) on the performance of dispatching strategies to study the robustness of our previous findings. The EMS system with 6 ambulances and 41 demand zones case (Example 1) is considered in this sensitivity analysis. First we vary the turn around time from 30 minutes to 65 minutes range while holding the call arrival rate constant ($\lambda = 1$ call/ hour). The Figure 4.23 compares the survival rate for different dispatching rules for distinct partition sizes. As we expected, the Figure 4.23 shows that survival rate decreases when turn around time increases. However, our previous findings (e.g. best K is 2 and best dispatching rule is Nocross rule) remains the same in this instance.Thus, we can say that operating according to pre-determined boundaries an EMS system is able to increase patient survivability. In addition, priority dispatching strategies leads to increase the survival probability of patients.

Next we vary the response time from 4 minutes to 14 minutes range while holding the call arrival rate constant ($\lambda = 1$ call/ hour). The Figure 4.24 compares the survival rate for different dispatching policies. As the graph depicts, the survival rate decreases when response time increases. In this case also our previous findings remain the same, that is best K is 2 and best dispatching rule is Nocross rule.

Finally we vary the response time of outside ambulance from 4 minutes to 20 minutes range for two different λ values. The Figure 4.25 compares the performance of different dispatching policies while varying the response time of outside ambulance. The left most graph shows the performance when call arrival rate is

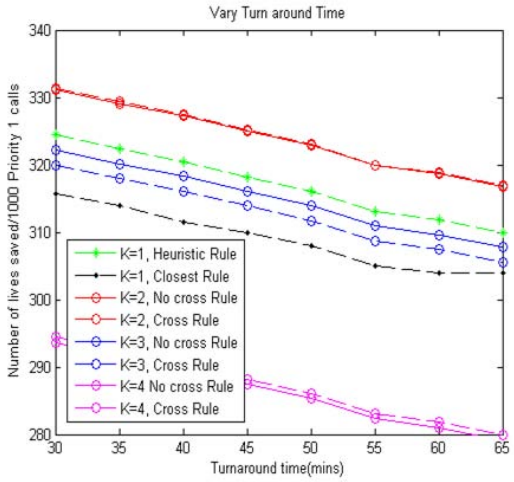


Figure 4.23: Survival rate comparison for different K values while varying the turn around time

one per hour while the right most graph shows the performance when call arrival rate is 3 per hour. When call arrival rate is low (see left most graph), there is no significant different between performance of No cross rule and Cross Rule for higher response time of outside ambulance. However, when call arrival rate is high (see right most graph), No cross Rule perform better than Cross Rule in terms of patients survival rate. This result also indicates that operating EMS systems according to pre-determined districts (or boundaries) and dispatching ambulances considering the severity of the call leads to an increase of patient survival probability.

In this section, we conducted a sensitivity analysis to see our findings remain the same while varying several parameters. Results show that our findings are robust. Thus we can conclude that it is beneficial to operate the EMS system with districts in terms of patients survivability and dispatching ambulances considering the degree of the urgency of the call also leads to increase the patient survival probability.

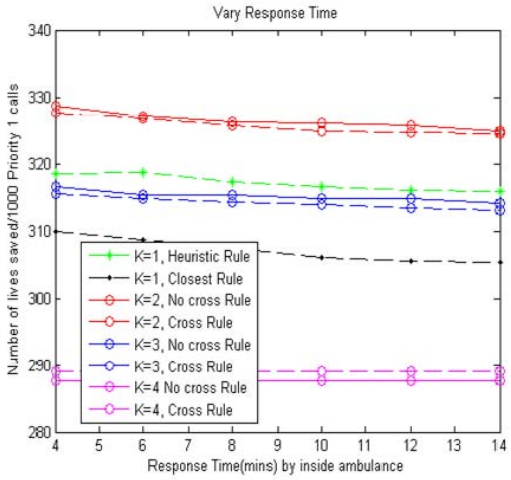
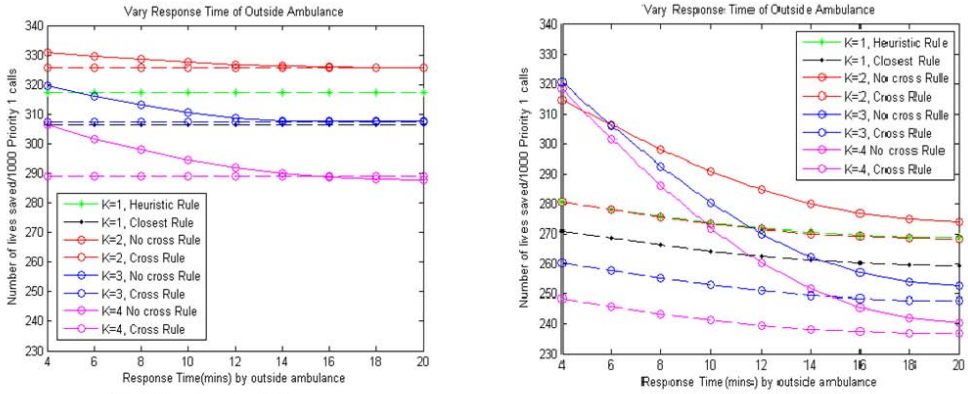


Figure 4.24: Survival rate comparison for different K values while varying the response time



(a) Call rate = 1 call/hour (b) Call rate = 3 calls/hour

Figure 4.25: Survival rate comparison for different K values while varying the response time of Outside Ambulance

4.11 Conclusion and Future Research

In this chapter we proposed a mathematical model and a constructive heuristic to determine response boundaries of emergency service vehicles. Computational results show that operating according to predetermined boundaries is beneficial for EMS systems in terms of patients survivability; however that one must be careful in selecting the number of districts so as not to reduce backup coverage. In addition, it is observed that districting helps to increase the coverage level of EMS systems. After determining districts we proposed intra-district and inter-district dispatching policies for EMS systems. Results show that integrated dispatching and districting policies leads to increase patient survival probability. The methodology in this chapter

can be applied to public services such as police patrol services. In addition, the methodology we proposed here can be extended to consider other objectives when determining response boundaries. For an example, we can district the service region into sub-regions so as to maximize the patient survival probability. Finally, it is concluded that the districts provided by AEXC method is better when compared to districts provided by Integer method in terms of computational time and survival probability.

Chapter 5

Conclusions and Discussion

Optimal dispatching rules potentially have significant effect on patient survivability during an emergency incident in addition to locating medical units. Thus this research studies and proposes dispatching and districting policies for EMS systems in order to improve the performance of EMS systems. Performance is measured in terms of patients' survivability as opposed to measuring the response time threshold, since survival probability reflects the patient outcome directly. These dispatching policies are implemented considering the degree of the urgency of the call. The performance of proposed dispatching and districting policies are illustrated using real-world data collected from Hanover County, Virginia.

Throughout this study, we assumed that calls arrives according to Poisson process with rate λ to the entire system and we consider an EMS system with fixed deployment. In addition, static dispatching rules were considered when developing dispatching strategies. Further, in this study calls are prioritized according to the severity of the call. All observations and conclusions are made under those assumptions.

First a discounted, infinite horizon, Markov decision process model is developed and analyzed to obtain optimal dispatching strategies for less complex EMS systems. In the MDP model we assumed that service times and turn around times are exponentially distributed. Computational results show that a myopic policy is not always optimal and dispatching ambulances considering the severity of the call leads to increase the patient survival probability. It is also observed that many lives can be saved at no additional (in terms of paramedic units available) cost. Further, the results show that the optimal policy given by the MDP model is likely to balance the work load between ambulances. We compared the myopic policy of always sending the closest ambulance with the optimal policy given by the MDP model. Results indicate that it is always optimal to dispatch the closest ambulance for Priority 1 patients. The optimal policy for Priority 1 calls is

intuitive, since faster response times increase patient survival probability. For Priority 2 calls, a priority list of ambulances to dispatch is obtained using the model. The proposed dispatching rule is easy to implement in EMS systems since a priority list of ambulances to dispatch depends only on the location and degree of urgency of the call and not on the location of all busy ambulances. This MDP approach allows us to address the stochastic behavior of the EMS system. Additionally, the running time for our MDP formulation in MATLAB is not significant. One potential drawback is that the formulation of the dynamic programming model is complicated when problem size increases. However, a simulation approach can be used to overcome this drawback of the MDP approach. Simulation can also help to alleviate some other potential drawbacks associated with an MDP approach, namely the assumption of exponential service times and of zero-length queue.

Next a simulation based approach is used to study the nature of the optimal dispatching policy and to develop a heuristic dispatching rule for complex EMS systems. In the simulation model we assumed that response times and turn around times are lognormally and exponentially distributed respectively. Simulation results also showed that it is better to dispatch the closest ambulance for Priority 1 patients. For Priority 2 calls, a priority list of ambulances to dispatch is obtained using a heuristic. According to this heuristic rule consider the ambulance busy probabilities when dispatching ambulances to Priority 2 calls. We calculated the ambulance busy probability by considering the demand of each zone. Computational results show that the heuristic rule is vital in increasing patient survivability at no extra cost when compared to myopic policy of always sending the closet unit without considering the degree of the urgency of the call. We believe the heuristic we presented here provides unique contribution in that it shows that it is possible to achieve significant improvements, in terms of lives saved, at little cost by considering the degree of urgency of the call. Furthermore, this can be achieved even with a simple heuristic: send the closest ambulance to Priority 1 calls, follow an ordered preference list for Priority 2 calls. This should be easy to implement in practice as ordered preference lists are already widely accepted policy types in EMS systems. However, the heuristic algorithm (H_1) provides the same dispatching order for Priority 2 calls for every demand zone. Future research can be conducted to obtain the order of dispatching ambulances for Priority 2 calls depending on the demand zone. Even though this heuristic was developed to maximize the patient survivability, it helps to decrease the average response time and increase the percentage calls served within 10 minutes for Priority 1 calls. The average response time for Priority 2 calls increased slightly by following the proposed dispatching rule. Although the average response time increased it did not affect the average survivability of patients since Priority 2 calls are non-life threatening. Future research can concentrate towards obtaining dispatching rules to maximize

patient survival probability of life-threatening calls while minimizing the effect on the average response time of Priority 2 calls.

Finally, we proposed districting techniques in order to determine the emergency service vehicles response boundaries. An integer mathematical model and a constructive heuristic are proposed to determine emergency vehicles response boundaries. We observed that the constructive heuristic is efficient compared to integer mathematical model in terms of computational time. After determining vehicle boundaries, we proposed intra-district and inter-district dispatching policies. Results show that these districting and dispatching policies are valuable in increasing patients' survival. Overall, we can conclude that integrated dispatching and districting policies proposed in this study can be used to improve the performance of EMS systems at no extra cost in terms of the number of paramedic units.

In this study only two types of calls were considered to address the severity of the call. Also, it was assumed that upon arrival we know the priority of the call exactly. In reality there are classification errors associated with call priority. Future research can focus on studying the classification errors of calls that impact patient survivability while incorporating few other call categories to address the severity of the call. In addition, the methodologies proposed in this study can be applied to other problems such as dispatching police cars and fire engines and military deployment.

EMS administrators and managers continually seek innovative methods to enhance system performance [7]. Although numerous studies have sought answers to such EMS related issues, certain unaddressed issues still exist. The integrated districting and dispatching policies we proposed can be used to address some of these issues such as enhancing the survival probability of patients in comparison to that of the existing methods in the EMS systems. We found that there is an 8% increase in the average survival probability by implementing our proposed rules in place of the existing EMS system in Hanover County, Virginia.

Our proposed integrated districting and dispatching policies can be easily implemented to EMS systems that follow fixed deployment with known station locations. Our proposed methods allow re-allocating of available paramedic units with no extra financial costs and without jeopardizing the patient survival while increasing the overall efficiency of the EMS system. Since our model was developed based on emergency calls received peak hours (12pm-6pm time period) it can be successfully applied to a region with higher 911 call rate. Furthermore, based on sensitivity analysis, the efficiency of our proposed system will be more pronounced with increasing call rate.

The most widely used dispatching methods have following draw backs. According to the myopic dispatching policy, the ambulances are dispatched based on proximity instead of the degree of the urgency of

the call [3], [34], [1]. We proposed Priority dispatching strategies instead of a myopic policy. Furthermore, most nationwide dispatch methods do not include the response boundary (or district) concept. EMS vehicles operating according to our dispatching policy can reduce the response time substantially for Priority 1 calls. However, the average response time for Priority 2 calls increased slightly. The standard common EMS performance is to respond to 80% of Priority 2 calls in less than 30 minutes for urban and less than 45 minutes for rural areas [20]. In our simulation we found that 80% of the Priority 2 calls can be responded in less than 20 minutes. We recommend our dispatching methods for EMS systems where the small increment in the average response time to priority 2 calls can be afforded without serious consequences in terms of the overall system efficiency. For other EMS systems (response time standard is less than 20 minute for Priority 2 calls) we need to study the EMS system carefully before recommending our dispatching methods.

Finally, we recommend these policies for EMS system given economical and technical feasibility of our proposed policies. No extra training is needed for dispatchers since our method produces an ordered preference list of ambulances to be dispatched considering the priority of the call. Since such preference lists (known as contingency tables) are already used in EMS system, no additional training is required for the EMS staff regarding preference lists. Thus we can recommend integrated districting and dispatching policies to improve the performance of EMS systems that follow a fixed deployment and myopic policy with scarce resources. In addition, we recommend these rules for EMS system with limited number of ambulances, without addition of extra paramedic units.

Appendices

Appendix A Additional Examples

A.1 Parameters for Examples

			Response Times				Turn Around Times		
Example	Size	Call proportion (z_i)		Amb 1	Amb 2	Amb 3	Amb 1	Amb 2	Amb 3
Example 2	3×3 case	$z_1 = 0.1$	zone 1	(9.07,4.19)	(10.92,5.05)	((14.03,6.48)	(60)	(65)	(50)
		$z_2 = 0.4$	zone 2	(14.03,6.48)	(11.05,6.48)	(12.03,5.05)	(75)	(65)	(60)
		$z_3 = 0.5$	zone 3	(14.03,6.48)	(12.03,5.05)	(10.92,5.05)	(65)	(65)	(75)
Example 3	5×3 case	$z_1 = 0.1$	zone 1	(9.07,4.19)	(12.03,5.05)	(10.07,4.19)	(60)	(65)	(60)
		$z_2 = 0.3$	zone 2	(8.03,6.48)	(11.03,6.48)	(13.03,6.48)	(75)	(50)	(75)
		$z_3 = 0.2$	zone 3	(14.03,6.48)	(12.03,5.05)	(9.03,6.48)	(65)	(60)	(65)
		$z_4 = 0.2$	zone 4	(10.92,5.05)	(9.07,4.19)	(11.05,6.48)	(65)	(75)	(65)
		$z_5 = 0.2$	zone 5	(11.05,6.48)	(9.07,4.19)	(8.03,6.48)	(65)	(65)	(65)

Table 1: Parameters for examples

- In table 1 under Response Times column (μ_1, σ) implies that response times are distributed lognormally with mean of μ_1 and standard deviation of σ. Also under Turn Around Times column (μ_2) implies that turn around times are exponentially distributed with mean of μ_2.
- Assumed that calls arrived according to a poison process with rate $\lambda = 1$ per hour to the entire system for both examples.
- For these two examples we assumed that the probability of receiving a Priority 1 or a Priority 2 call is equally likely from any of the demand zones. i.e.
  1. for 3×3 case - $p_{11} = p_{12} = p_{13} = p_{21} = p_{22} = p_{23} = 0.5$
  2. for 5×3 case - $p_{11} = p_{12} = p_{13} = p_{14} = p_{15} = p_{12} = p_{22} = p_{32} = p_{42} = p_{52} = 0.5$

A.2 Comparison of Myopic and Optimal Dispatching Rules

	Priority 1 Calls			Priority 2 Calls		
Dispatch Order	Zone 1	Zone 2	Zone 3	Zone 1	Zone 2	Zone 3
1^{st} Choice	1	2	3	1	1	1
2^{nd} Choice	2	3	2	2	2	3
3^{rd} Choice	3	1	1	3	3	2

Table 2: Optimal order of dispatching ambulances - (3×3 case)

Tables 2 and 3 show the optimal order of dispatching ambulances to Priority 1 and Priority 2 calls for each demand zone.For an example, the column zone 1 under Priority 1 heading in Table 2 indicates that

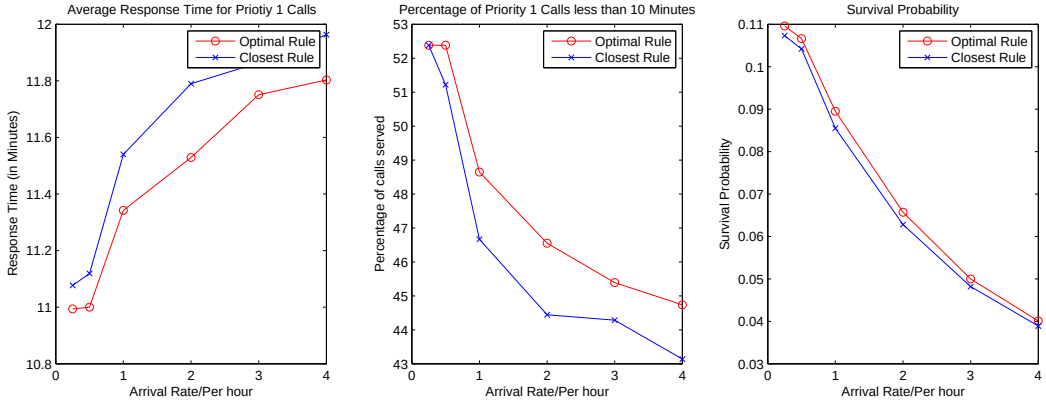


Figure 1: Comparison of Two dispatching Strategies - 3×3 case

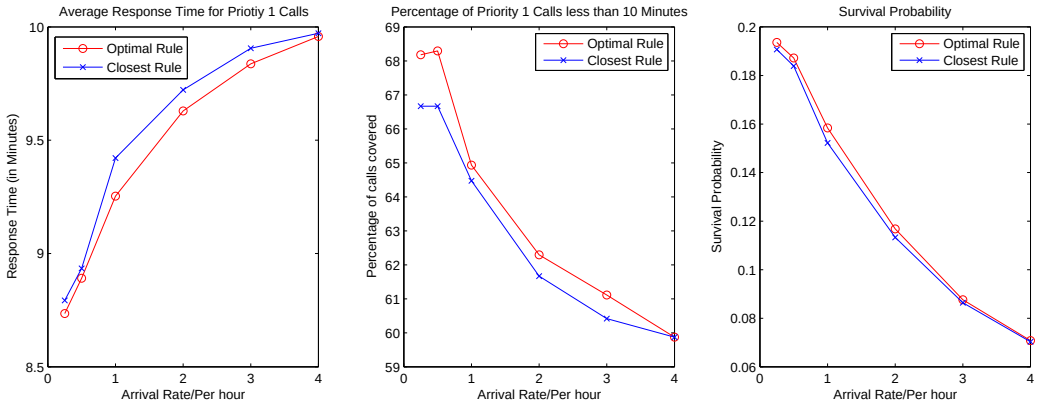


Figure 2: Comparison of Two dispatching Strategies - 5×3 case

Ambulance 1 is 1^{st} choice to dispatch, followed by Ambulance 2 (if Ambulance 1 is busy) then Ambulance 3 (if both Ambulance 1 and Ambulance 2 are busy).

	Priority 1 Calls					Priority 2 Calls				
Dispatch Order	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
1^{st} Choice	1	1	3	2	3	3	2	2	2	2
2^{nd} Choice	3	2	2	1	2	2	3	3	3	3
3^{rd} Choice	2	3	1	3	1	1	1	1	1	1

Table 3: Optimal order of dispatching ambulances - (5×3 case)

Appendix B Performance of heuristic policy (5×3) case

	Priority 1 calls					Priority 2 calls				
Dispatch order	zone 1	zone 2	zone 3	zone 4	zone 5	zone 1	zone 2	zone 3	zone 4	zone 5
1^{st} Choice	1	1	3	2	3	2	2	2	2	2
2^{nd} Choice	3	2	2	1	2	3	3	3	3	3
3^{rd} Choice	2	3	1	3	1	1	1	1	1	1

Table 4: Order of dispatching ambulances according to Heuristic Rule - (5×3 case)

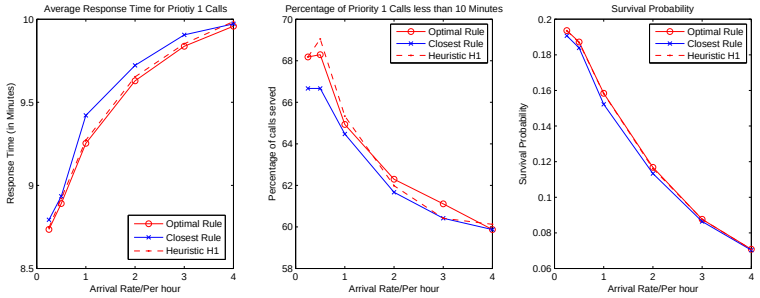


Figure 3: Comparison of dispatching Strategies -5×3 case

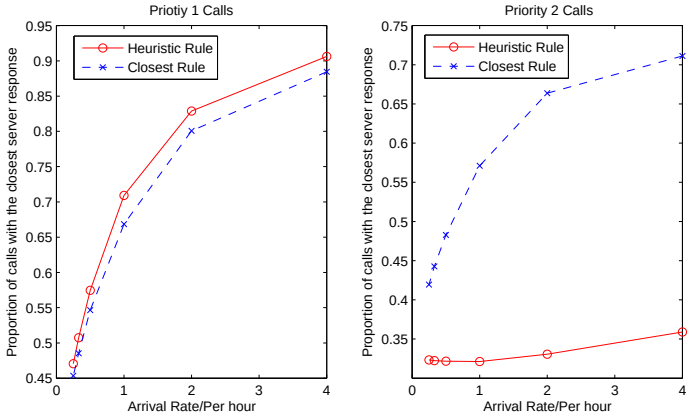


Figure 4: Probability that the closest server is dispatched to Priority 1 and Priority 2 calls for Heuristic Rule and Closest Rule -5×3 case

Appendix C Hanover County Fire and EMS department Data-Case Study

In table 5 under Response Times (μ_1, σ) implies that response times are distributed lognormally with mean of μ_1 and standard deviation of σ. Also In table 6 under Turn Around Times (μ_2) implies that turn around times are exponentially distributed with mean of μ_2. We assumed that the probability of receiving a Priority 1 or a Priority 2 call is equally likely from any of the demand zones. i.e. $p_{11} = p_{12} = p_{13} = p_{14} = p_{15} = p_{16} = p_{17} = p_{18} = p_{19} = p_{1,10} = p_{1,11} = p_{1,12} = p_{21} = p_{22} = p_{23} = p_{24} = p_{25} = p_{26} = p_{27} = p_{28}p_{29} = p_{2,10} = p_{2,11} = p_{2,12} = 0.5$.

Call proportion	Response Times					
(z_i)	Demand zone	Amb 1	Amb 2	Amb 3	Amb 4	Amb 5
$z_1 = 0.226034$	zone 1	(16.77,12.47)	(15.43,11.47)	(13.38,9.95)	(8.03,5.97)	(8.03,5.97)
$z_2 = 0.019513$	zone 2	(32.14,23.89)	(32.14,23.89)	(19.87,14.77)	(32.14,23.89)	(32.14,23.89)
$z_3 = 0.060281$	zone 3	(23.72,17.64)	(9.92,7.38)	(13.42,9.97)	(18.84,14.01)	(18.84,14.01)
$z_4 = 0.043914$	zone 4	(26.26,19.52)	(32.14,23.89)	(26.07,19.38)	(15.39,11.44)	(15.39,11.44)
$z_5 = 0.02657$	zone 5	(16.89,12.56)	(24.56,18.26)	(17.16,12.76)	(28.44,21.14)	(28.44,21.14)
$z_6 = 0.09327$	zone 6	(10.07,7.48)	(16.32,12.13)	(32.14,23.89)	(15.59,11.59)	(15.59,11.59)
$z_7 = 0.326744$	zone 7	(25.03,18.61)	(9.85,7.32)	(14.18,10.54)	(15.04,11.18)	(15.04,11.18)
$z_8 = 0.065128$	zone 8	(18.82,13.99)	(32.14,23.89)	(13.74,10.21)	(25.79,19.17)	(25.79,19.17)
$z_9 = 0.007525$	zone 9	(32.14,23.89)	(32.14,23.89)	(27.34,20.32)	(20.9,15.53)	(20.9,15.53)
$z_{10} = 0.077626$	zone 10	(12.6,9.36)	(19.62,14.59)	(14.63,10.87)	(12.7,9.44)	(12.7,9.44)
$z_{11} = 0.029886$	zone 11	(22.98,17.08)	(18.28,13.59)	(19.77,14.7)	(19.69,14.63)	(19.69,14.63)
$z_{12} = 0.023509$	zone 12	(32.14,23.89)	(18.63,13.85)	(32.14,23.89)	(19.72,14.66)	(19.72,14.66)

Table 5: Response times and proportion of calls for 12×5 case

	Turn Around Times				
Demand zone	Amb 1	Amb 2	Amb 3	Amb 4	Amb 5
zone 1	(75.61)	(77.68)	(91.86)	(69.87)	(69.87)
zone 2	(148.13)	(130.08)	(104.53)	(122.64)	(122.64)
zone 3	(84.56)	(63.10)	(63.25)	(95.30)	(95.30)
zone 4	(93.42)	(108.45)	(119.19)	(83.9)	(83.9)
zone 5	(74.9)	(75.39)	(76.58)	(95.01)	(95.01)
zone 6	(64.49)	(65.35)	(103.12)	(86.03)	(86.03)
zone 7	(67.63)	(58.47)	(83.59)	(77.40)	(77.40)
zone 8	(97.9)	(108.57)	(103.34)	(98.87)	(98.87)
zone 9	(129.59)	(108.57)	(103.34)	(113.6)	(113.6)
zone 10	(65.11)	(76.15)	(88.01)	(73.96)	(73.96)
zone 11	(87.25)	(101.74)	(99.14)	(82.96)	(82.96)
zone 12	(90.79)	(73.28)	(91.86)	(87.26)	(87.26)

Table 6: Turn Around Times for 12×5 case

Bibliography

[1] T. Andersson and P. Väbrand. Decision support tool for ambulance dispatch and relocation. *Journal of the Operational Research Society*, 58(2):195–201, 2008.

[2] M.A.G. Andrade and E.A.R. Garcia. Redistricting by square cells. *Lecture Notes in Computer Science*, 5845:669–679, 2009.

[3] D. Bandara, M.E. Mayorga, and L.A. McLay. Optimal dispatching strategies for emergency vehicles to increase patient survivability. *International Journal of Operational Research-In press*.

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

[5] D.P. Bertsekas. Dynamic programming and optimal contro. *Massachussets: Athena Scientific*, 2, 2001.

[6] J.R. Birge. Redistricting to maximize the preservation of political boundaries. *Social Science Research*, 12:205–214, 1983.

[7] T. H. Blackwellr and J. S. Kaufman. Response time effectiveness: Comparison of response time and survival in an urban emergency medical services system. *Academic Emergency Medicine*, 9(4):288–295, 1991.

[8] L. Brotcorne, G. Laporte, and F. Semet. Ambulance location and relocation models. *European Journal of Operational Research*, 147(3):451–463, 2003.

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

[10] G. Carter, M. Jan, and E. Ignall. Response area for two emergency units. *Operations Research*, 20(3):571–594, 1972.

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

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

[13] E. Erkut and A. Ingolfssonl. Ambulance location for maximum survival. *Naval Research Logistics*, 55(1):42–55, 2008.

[14] J. A. Fitzsimmons. A methodology for emergency ambulance deployment. *Management Science*, 19(6):627–636, 1973.

[15] S. Gass. On the division of police districts into patrol beats. *Proceedings of the 23rd National Conference of the Association for Computing Machinery*, 1968.

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

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

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

[19] A. Haghani, Q. Tian, and H. Huijun. Simulation model for real-time emergency vehicle dispatching and routing. *Journal of the Transportation Research Board*, 1882:176–183, 2004.

[20] S.G. Henderson and A.J. Mason. Ambulance service planning: Simulation and data visualization. *A Hand Book of Methods and Applications, Kluwer Academic Publishers*, 70:77–102, 2004.

[21] M. Hojati. Optimal political districting. *Computers Operation Research*, 23(12):1147–1161, 1996.

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

[23] M. Kuisma, öm, P. Holmstr, J. Repo, Määttä, T. Nousila-Wiik, and J. Boyd. Pre-hospital mortality in an ems system using medical priority dispatching: a community based cohort study. *Resuscitation*, 61(3):297–302, 2004.

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

[25] R.C. Larson. A hypercube queuing model for facility location and redistricting in urban emergency services. *Computers and Operations Research*, 1(1):67–75, 1974.

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

[27] R.C. Larson. Urban emergency service system: an iterative procedure for approximating performance characteristics. *Computers and Operations Research*, 23(5):845–868, 1975.

[28] S. Lee. The role of preparedness in ambulance dispatching. *Journal of the Operational Research Society*, 62:1888–1897, 2011.

[29] V. Marianov and C. ReVelle. The queuing probabilistic location set covering problem and some extensions. *Socio-Economic Planning Science*, 28(3):167–178, 1994.

[30] M.S. Maxwell, M. Restrepo, S.G. Henderson, and T. Huseyin. Approximate dynamic programming for ambulance redeployment. *INFORMS Journal on Computing*, 22(2):266–281, 2010.

[31] L.A. McLay. A maximum expected covering location model with two types of servers. *IIE Transactions*, 41(8):730–741, 2009.

[32] L.A. McLay. Emergency medical service systems that improve patient survivability. *Wiley Encyclopedia of Operations Research and Management Science*, 2010.

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

[34] L.A. McLay and M.E. Mayorga. A model for optimally dispatching ambulances to emergency calls with classification errors in patient priorities. *Under Review, Technical report, Virginia Commonwealth University, Richmond, Virginia. Available online at http://dl.dropbox.com/u/6995461/TechnicalReports/2010OptimalDispatching.pdf*, 2011.

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

[36] J.A. Paul and R. Batta. Improving hurricane disaster preparedness: Models for optimal reallocation of hospital capacity. *Int. Journal of Operational Research*, 10(2):194–213, 2011.

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

[38] C. ReVelle and K. Hoganl. The maximal availability location problem. *Transportation Science*, 23(3):192–200, 1989.

[39] C. ReVelle and K. Hoganl. The maximal reliability location problem and alpha reliable p-center problem: Derivatives of probabilistic location set covering problem. *Annals of Operations Research*, 18(1):155–174, 1989.

[40] L. P. Roppolo, A. Westfall, P.E. Pepe, L. Nobel, J. Cowan, J.J. Kay, and A.H. Idris. Dispatcher assessments for agonal breathing improve detection of cardiac arrest. *Resuscitation*, 80:769–772, 2009.

[41] L.C. Santone and N.B. Geoffrey. A computer model for the evaluation of fire station location. *National Bureau of Standards Report, U.S. Department of Commerce, Washington, DC*, 1969.

[42] E.S. Savas. Simulation and cost-effectiveness analysis of new york's emergency ambulance service. *Management Science*, 15(12):608–627, 1969.

[43] D.A. Schilling, D.J. Elzinga, J. Cohon, R.L. Church, and C. ReVelle. The team/fleet models for simultaneous facility and equipment siting. *Transportation Science*, 13(2):163–175, 1979.

[44] J.B. Schneider. Solving urban location problems; human intuition versus the computer. *Journal of the American Institute of Planners*, 37:95–99, 1971.

[45] J.B. Schneider and J.G. Symons. Locating ambulance dispatch centers in an urban region: A man-computer interactive problem solving approach. *Regional Science Research Institute, Philadelphia, Pennsylvania, RSRI Discussion Paper Series*, 49, 1971.

[46] R.D. Smith. Computer application in police manpower distribution. *Field Service Devision, International Association of Chiefs of Police, Wahington, DC*, 1961.

[47] P. Sorenson and R. Church. Integrating expected coverage and local reliability for emergency medical services location problems. *Socio-Economic Planning Sciences*, 44(1):8–18, 2010.

[48] I.G. Stiell, L. P. Nesbitt, W. Pickettl, D. Munkley, and W. Daniel. The opals major trauma study: impact of advanced life-support on survival and morbidity. *CMAJ*, 178(9):1141–1152, 2008.

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