8-2013

Joint Location and Dispatching Decisions for Emergency Medical Service Systems

Hector Toro-diaz
Clemson University, htoro@g.clemson.edu

Joint Location and Dispatching Decisions for Emergency Medical Service 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
Héctor Hernán Toro Díaz
August 2013

Accepted by:
Dr. Maria E. Mayorga, Committee Chair
Dr. Kevin Taaffe
Dr. Mary E. Kurz
Dr. Joel Greenstein

Abstract

Emergency Medical Service (EMS) systems are a service that provides acute care and transportation to a place for definitive care, to people experiencing a medical emergency. The ultimate goal of EMS systems is to save lives. The ability of EMS systems to do this effectively is impacted by several resource allocation decisions including location of servers (ambulances), districting of demand zones and dispatching rules for the servers. The location decision is strategic while the dispatching decision is operational. Those two decisions are usually made separately although both affect typical EMS performance measures. The service from an ambulance is usually time sensitive (patients generally want the ambulances to be available as soon as possible), and the demand for service is stochastic. Regulators also impose availability constraints, the most generally accepted being that 90% of high priority calls (such as those related to cardiac arrest events) should be attended to within 8 minutes and 59 seconds.

In the case of minimizing the mean response time as the only objective, previous works have shown that there are cases in which it might not be optimal to send the closest available server to achieve the minimum overall response time. Some researchers have proposed integrated models in which the two decisions are made sequentially. The main contribution of this work is precisely in developing the integration of location and dispatching decisions made simultaneously. Combining those decisions leads to complex optimization models in which even the formulation is not straightforward. In addition, given the stochastic nature of the EMS systems the models need to have a way to represent their probabilistic nature. Several researchers agree that the use of queuing theory elements in combination with location, districting and dispatching models is the best way to represent EMS systems. Often heuristic/approximate solution procedures have been

proposed and used since the use of exact methods is only suitable for small instances.

Performance indicators other than Response Time can be affected negatively when the dispatching rule is sending the closest server. For instance, there are previous works claiming that when the workload of the servers is taken into account, the nearest dispatching policy can cause workload imbalances. Therefore, researchers mentioned as a potential research direction to develop solution approaches in which location, districting and dispatching could be handled in parallel, due to the effect that all those decisions have on key performance measures for an EMS system. In this work the aim is precisely the development of an optimization framework for the joint problem of location and dispatching in the context of EMS systems. The optimization framework is based on meta heuristics. Fairness performance indicators are also considered, taking into account different points of view about the system, in addition to the standard efficiency criteria.

Initially we cover general aspects related to EMS systems, including an overall description of main characteristics being modeled as well as an initial overview of related literature. We also include an overall description and literature review with focus on solution methodologies for real instances, of two related problems: the p-median problem and the maximal covering location problem (MCLP). Those two problems provide much of the basic structure upon which the main mathematical model integrating location and dispatching decisions is built later.

Next we introduce the mathematical model (mixed-integer non-linear problem) which has embedded a queuing component describing the service nature of the system. Given the nature of the resulting model it was necessary to develop a solution algorithm. It was done based on Genetic Algorithms. We have found no benefit on using the joint approach regarding mean Response Time minimization or Expected Coverage maximization. We concluded that minimizing Response Time is a better approach than maximizing Expected Coverage, in terms of the trade-off between those two criteria.

Once the optimization framework was developed we introduced fairness ideas to the location/allocation of servers for EMS systems. Unlike the case of Response Time, we found that the joint approach finds better solutions for the fairness criteria, both from the point of view of internal and external costumers. The importance of that result lies in the fact that people not only expect

the service from ambulances to be quick, but also expect it to be fair, at least in the sense that any costumer in the system should have the same chances of receiving quick attention. From the point of view of service providers, balancing ambulance workloads is also desirable. Equity and efficiency criteria are often in conflict with each other, hence analyzing trade-offs is a first step to attempt balancing different points of view from different stakeholders.

The initial modeling and solution approach solve the problem by using a heuristic method for the overall location/allocation decisions and an exact solution to the embedded queuing model. The problem of such an approach is that the embedded queuing model increases its size exponentially with relation to the number of ambulances in the system. Thus the approach is not practical for large scale real systems, say having 10+ ambulances. Therefore we addressed the scalability problem by introducing approximation procedures to solve the embedded queuing model. The approximation procedures are faster than the exact solution method for the embedded sub-problem. Previous works mentioned that the approximated solutions are only marginally apart from the exact solution (1 to 2%). The mathematical model also changed allowing for several ambulances to be assigned to a single station, which is a typical characteristic of real world large scale EMS systems. To be able to solve bigger instances we also changed the solution procedure, using a Tabu Search based algorithm, with random initialization and dynamic size of the tabu list. The conclusions in terms of benefits of the joint approach are true for bigger systems, i.e. the joint approach allows for finding the best solutions from the point of view of several fairness criteria.

Dedication

Dedicada a mamá y papá, Fabiola Díaz Osorio y Hernando Toro Delgado.

Acknowledgments

This work would not have been possible without the support, collaboration and guidance from my adviser, Professor Maria Mayorga. My sincere gratitude goes to her. Professor Mayorga's support went beyond the academic matters into professional development and advice.

My sincere thanks to the committee members, Professors Kevin Taaffe, Mary B. Kurz and Joel Greenstein. Their suggestions helped to improve the readability and understanding of my work and also to enhance the way in which the contributions were being presented.

I am in debt to the Ambassadors Program from The Graduate School at Clemson, coordinated by Professor Bob Lippert. This program was designed to recruit international students by visiting universities in South America. My sincere thanks to the student ambassador that was in charge of helping me through the application process, Susana Lizcano, a fellow Colombian that has been my friend ever since. I became an ambassador during my last two years and enjoyed every part of it in company of Professor Lippert.

I am also grateful with all my professors at the Industrial Engineering School at Universidad del Valle (Univalle) - Colombia, from where I received both my Bachelor and Master degrees. What I learned while being a student at Univalle prepared me more than well to be able to succeed at the PhD level.

While pursuing my PhD degree I have sacrificed countless hours with my family and people I care the most. Their support, understanding and good energies made my life easier while being away, and I am grateful for that. Finally, I have to thank Chuck Horton and his family for 'adopting' me during my years at Clemson. The Horton's family at large made me feel one of their own while being at a foreign country.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Emergency Medical Service (EMS) Systems

A wise man once said "Hope for the best, but prepare for the worse". Murphy's law states that if something has the potential to go wrong, it eventually will go wrong. Emergencies happen inadvertently. We all hope to not be involved in any emergency (we hope for the best), but we also hope that if we are involved in such an incident there should be help promptly coming our way. We cannot predict when are we going to need the services of an Emergency Response System, but when we have that need we surely would prefer it to be satisfied as quickly as possible.

The aims of any EMS system are to prevent premature death, to reduce the pain and to prevent avoidable disability. In order to fulfill these objectives EMS Systems provide out-of-hospital acute care and transport to a place of definitive care, to patients with illnesses and injuries that constitute a medical emergency. Although the ultimate goal of EMS systems is to save lives, typical performance measures such as coverage and response time are used as a proxy for survivability, which is the broader objective. A demand zone is said to be covered if there is at least one facility within a predefined distance/time threshold from the demand zone. The concept of coverage is related to the availability of a satisfactory facility rather than the best possible one (Farahani et al., 2012). Li et al. (2011) pointed out that the coverage maximization approach is the most widely used by practitioners, researchers and regulators.

Researchers like Erkut et al. (2008) and McLay and Mayorga (2010) have explored the use of survivability objective functions directly, instead of using a proxy objective function. However, survivability functions are not readily available for different types of emergency calls. Most of the works analyzing the rate of survival in the context of EMS systems have focused on cardiac arrest. However, cardiac arrest accounts only for a small percentage of the total number of calls handled typically by EMS systems. Generalized survivability functions would be required to include most of the EMS demand. Another objective that has also been suggested by Lee (2011), among others, is called Preparedness. The underlying idea is to operate an EMS system not only considering the current call asking for attention, but also the expected behavior of the system based on some forecast of the upcoming calls. Or in other words, operate the system trying to be *prepared* in advance for what is coming next.

There are other objectives that an EMS system should seek to satisfy and that somehow the mathematical models usually assume implicitly. For instance, when we are dealing with a location decision we are usually worried about response time, either trying to minimize it or to set a maximum that should not be surpassed. We are implicitly assuming that the paramedics and medical personnel riding in the ambulances are well prepared to handle the emergency they are responding to (or in other words, one objective of an EMS system should be having well trained personnel); we also assume that the personnel have the adequate equipment (planners of the EMS should make it happen); we are also assuming that EMS personnel will have a professional behavior that not only can save the lives at risk, but also help the patients and their relatives to stay calm, treating them with respect. In other words, the quality of an EMS system, from the perspective of the users (patients) can not be reduced to a measure of response time, or whether or not the ambulance observed an existing standard. Those single numbers can be used for planning and design purposes, but the day to day operation of the system, in which medical personnel is arguably the biggest player, needs to be more than those single performance measures. After all, if your EMS almost always arrive pretty quickly but most of the time perform the wrong medical procedure, it is unlikely to think that the overall perception of quality is going to be positive. The mathematical models we propose to be used are just an abstraction of the system, very useful to account for

certain performance indicators. However, some other characteristics, requirements and operation rules that cannot be included in those models should be always observed in the real world.

A preliminary review of literature related to location, districting and dispatching problems in the context of Emergency Medical Services (EMS) reveals a growing trend in the number of papers published as well as in the number of characteristics from real world systems that are included in those models. The topic can be considered then as an active research area. In the context of EMS systems the location decision is related to deciding where should the idle ambulances be positioned within the geographical region where they offer their services. It is usually a discrete problem in the sense that both the number of ambulances available and the number and location of potential stations are known.

To make the problem manageable usually a certain level of aggregation is used. The standard approach is to divide the geographic region into areas (typically by imposing a grid over the map representing the region of interest). The demand rate for each cell in the grid is made up adding the calls coming from within the boundaries of the cell. For planning purposes the demand of a cell is assumed to be assigned to the center of the cell. When the EMS system is in operation and a emergency call is received, the next decision is to assign one of the idle ambulances to attend the call. This assignment process corresponds to the dispatching decision. The most common dispatching rule is assigning the idle ambulance that is closest to the place of the emergency. Once the ambulances are assigned to specific locations it is possible to rank them, in relation to each demand zone, based on the dispatching preference. For instance if the closest dispatching rule is being used, then for each demand zone the ambulance located at the closest station would be the preferred one for that demand zone; the next closest ambulance will be the second preferred, and so on and so forth. This also allows us to talk about cooperation between the ambulances (the servers), a characteristic that is distinctive of EMS systems. In some circles this is also known as backup coverage, in the sense that if the preferred station is busy, then there is a backup station (or several of them) that the EMS planners can still use to attend a particular call. The identification of demand zones also provides the basis for what is called districting. It corresponds to identifying a subset of demand zones for which a particular ambulance is going to be the preferred server. From

the point of view of creating districts the cooperation among ambulances can be seen as allowing ambulance cross-district services. That is, if a demand happens in the district of an ambulance when it is busy, then another ambulance from a different district can still be allowed to attend the demand.

The level of sophistication in the modeling of EMS systems has been growing. Several reasons can be behind this trend, such as the availability of better hardware, better software as well as novel heuristic and exact optimization algorithms and techniques. That context has also created the conditions to allow the inclusion of other objectives than the usual efficiency criteria. Issues related to fairness/equity have also emerged, both from the perspective of the final (external) users but also the internal users (medical personal and managers). From the point of view of final users it wouldn't be fair, for example, to have some costumers being attended in an amount of time that is much greater than the mean response time of the whole system. Designing a system with the least variability in response time among users would be a possible answer for that problem. In the case of internal users, an unbalanced workload can also be judged as unfair. Minimizing the variability of the workloads among the different servers can be used as an objective of the system.

1.2 Overview of literature on EMS systems planning

Although examples using criteria other than those related to efficiency can be found in EMS systems design, the commonly used objectives are minimizing the mean response time and/or maximizing coverage. Making sure that every patient has the best chances to survive can be seen as the implicit objective of any EMS. The paper by Brandeau and Chiu (1989) is an extensive survey of over 50 representative problems in location research that includes more than 200 references. According to this survey, location theory was formally introduced in 1909 but a theoretical interest is more recent, traced back to the mid-1960's. By 1989 they were able to identify a growing trend in the publication of literature related to location problems, as well as the use of location models for a much more rich range of applications, health care being one of the newest. Brotcorne et al. (2003) provided a review focused on location models and their particular application to emergency response services. They classified the location models that evolved over the past 30 years into

two main categories, deterministic and probabilistic, recognizing that the most recent models were more concerned with the representation of the stochastic nature of the systems. Location models were also distinguished in *coverage* and *median* type problems. The first class attempts to locate the servers so as to maximize the fraction of the demand that has at least one server unit within a predefined maximum distance or time. The latter type minimizes the average or total travel cost between servers and demand zones. For service systems the cost is usually measured in time.

The two seminal attempts to develop basic coverage models were the set covering location problem (SCLP) by Toregas et al. (1971) and the maximal coverage location problem (MCLP) by Church and ReVelle (1974). These basic models were followed by many extensions. TEAM and FLEET models, by Schilling et al. (1979), considered several types of servers; Marianov and ReVelle (1992) improved the MCLP model considering two types of servers required simultaneously. Multiple coverage of demands were considered in BACOP1 and BACOP2 by Hogan and Revelle (1986) and other extensions, DSM and DDSM were added by Gendreau et al. (1997, 2001). The *p-median* problem was introduced by Hakimi (1964). The use of the *p-median* model in the planning and location of facilities for EMS can be found in Calvo and Marks (1973), Carbone (1974) and Carson and Batta (1990). Daskin (1983) developed the maximum expected coverage location model (MEXCLP) including the modeling of congestion elements. Hogan and Revelle (1986) developed the maximal availability location problem (MALP I and II) and later Marianov and ReVelle (1996) improved it, endogenously calculating the availability of servers using a queuing model at each facility.

It was the work by Larson (1974) that first used queueing theory elements in facility location models by introducing the hypercube model. Larson (1975) later developed an approximation for the hypercube model due to the fact that exact calculations were prohibitive. Chiyoshi et al. (2001) pointed out, after comparing some other models that were used before, that the hypercube model was the only one with the capabilities for an accurate representation of the system. There is a variety of applications and extensions of the hypercube model to EMS system such as the works by Brandeau and Chiu (1989), Chelst and Barlach (1981), Mendonca and Morabito (2001), Atkinson et al. (2006), Atkinson et al. (2008), Iannoni and Morabito (2007), Iannoni et al.

(2008), Galvao and Morabito (2008) and Geroliminis et al. (2009), among others. In several of the mentioned works, it has also been stated that the hypercube model is a descriptive tool that allows the analysis of scenarios, but it was not designed as an optimization model. However, it is possible to embed the hypercube model into an optimization framework. Batta et al. (1989) combined MEXCLP with the hypercube into an iterative, local search algorithm. Aytug and Saydam (2002) replaced the local search by a genetic algorithm. Geroliminis et al. (2009) as well as Iannoni and Morabito (2007), Iannoni et al. (2008) and Geroliminis et al. (2011) have also embedded the hypercube model into genetic algorithms to solve the location problem.

Traditionally, location, districting and dispatching decisions have been approached separately. Iannoni et al. (2008) concluded their paper by making the suggestion of trying to develop future extensions in which location, districting and dispatching can be handled in parallel, due to the effect that all of them have on key performance measures for an EMS system. Some researchers have proposed integrated models in which the two decisions are made sequentially, like in Geroliminis et al. (2009). Chiyoshi et al. (2001) have reported that the use of queuing theory elements in combination with location, districting and dispatching models is the best way to represent these systems. For the most part heuristic/approximate solution procedures have been proposed and used since the use of exact methods is only suitable for small instances.

In the case of minimizing the mean response time as the only objective, previous works by Cuninghame-Green and Harries (1988), Carter et al. (1972) and Repede and Bernardo (1994) have shown that there are cases in which it might not be optimal to use the nearest dispatching policy to achieve the minimum average response time. Other performance indicators can be affected negatively as well when following that dispatching rule. When the workload of the servers is taken into account, as in the work by Iannoni and Morabito (2007) and Iannoni et al. (2008), they found that using the nearest dispatching policy can cause imbalances in workloads. Workloads are part of a fairness performance measurement from the point of view of internal costumers of the system. If in addition to the mean response time there are other objectives that need consideration, it is not clear how to select a dispatching policy even if the locations are fixed. Of course, if both decisions have to be made at the same time the problem is even more challenging.

1.3 p-Median and Maximal Expected Coverage Location (MEXCLP) Problems

These two problems are closely related to the problem being studied in this work. The p-Median problem corresponds to a location/allocation problem in which, from a set of candidate locations we need to select exactly p of them, assigning each demand node to an open location, with the objective of minimizing the mean distance between the open facilities and their assigned demand zones. In our case we deal with a location/allocation problem in which the ambulances available need to be located and then assigned a priority with respect to the demand zones. Our problem is more complex than the p-median because of its server-to-costumer nature. We need to consider congestion, i.e., the possibility that an allocation decision is not possible in a particular period of time, as a result of the ambulance being already busy. Furthermore, the MEXCLP problem also deals with location/allocation. In this case, for a particular location of the servers (ambulances) it is assumed that all the demand zones within a given distance/time threshold from a server are assigned to a it. The objective is to maximize the proportion of demand that is reachable within the given threshold. Both problems have been extensively studied in the literature and because of their combinatorial nature several heuristic solution procedures have been suggested. What follows is an overview of those solution procedures.

1.3.1 p-median problem

The work by Mladenovic et al. (2007) is a relatively recent survey of meta-heuristic approaches applied to the solution of the p-median problem. In turn, Reese (2005) and Varnamkhasti (2012) offer a broader overview of methods that have been reported to tackle the problem. Reese (2005) classified the solution methods into the following categories: heuristics, meta-heuristics, approximation algorithms, LP relaxations, Surrogate relaxations, IP formulations and reductions and enumeration. His work included 120 references that were selected among an even bigger number of papers (more than 200). The selection was primarily based on the consideration of other characteristics than those of the classical p-median problem, i.e. minimax or p-center problems although

related were excluded, stochastic problems were also excluded as well as multi-objective versions of the problem or multi-services or multi-commodities. Out of those 120 references, 32 were meta-heuristic related works. 42 out of the 120 references were published between 2000 and 2005, almost doubling the number of publications of the previous 5-year period (24 references, 1995-1999). The following table lists the meta-heuristics and the number of publications associated.

Table 1.1: Summary of Meta-heuristic approaches to solve the p-median problem

Meta-heuristics	Number of Papers	
	Reese (2005)	Mladenovic (2007)
Variable Neighborhood Search (VNS)	5	4
Heuristic Concentration (HC)	4	3
Genetic Algorithms (GA)	11	5
Greedy Randomized Adaptive Search Procedures	1	-
Scatter Search (SS)	1	1
Tabu Search (TS)	5	6
Simulated Annealing (SA)	3	3
Neural Networks (NN)	2	2
Ant Colony Optimization (ACO)	-	1

The work by Reese (2005) is an "annotated bibliography", and as such it does not go into details, providing only a very brief overview of the works that were considered by the authors. Both works agree in the identification of solution approaches, in particular related to the use of meta-heuristic techniques. They also agree about a real progress in the state-of-the-art related to the solution of the p-median problem, thanks to the advent of meta-heuristics. They mentioned that it is clear that earlier methods such as constructive heuristics and local search have been surpassed by the new approaches, making it possible to solve larger instances in reduced computational time and with higher quality of the solutions being obtained. From Table 1.1 it should be noted that the most common approaches have been Genetic Algorithms, Tabu Search and Variable Neighborhood Search.

1.3.2 MEXCLP problem

Li et al. (2011) presents an up-to-date review on covering models for EMS systems. They attempted a review on facility location and planning models for EMS systems, however they finally

decided to focus only on covering models. They explained that the facility location problems when applied to EMS can be classified into three categories: (1) covering models, (2) p-median models and (3) p-center models. The first category approaches the planning problem by determining an standard (maximum distance/time) to offer the attention, and making decisions so that the standard is observed. The second problem attempts to minimize the total or average distance/service time and finally, the third approach aims to minimize the maximum service time/distance for all demand points. The authors quickly realized that the first approach is the most widely used by practitioners, researchers and regulators, and therefore they focused their review on covering models. The same fact was pointed out by Roth (2005), mentioning that the focus has shifted from minimizing response time to offering a certain level of coverage. Table 1.2 is a summary of meta-heuristic approaches mentioned in the work by Li et al. (2011).

Table 1.2: Summary of Meta-heuristic approaches to solve location covering problems

Problem	Algorithm	Number of Papers
MCLP and extensions	Tabu Search	1
	Genetic Algorithms	1
MEXCLP	Genetic Algorithms	3
	Tabu Search	1
	Simulated Annealing	2
MEXCLP - Hypercube	Genetic Algorithms	3
DSM - DDSM	Tabu Search	4

In table 1.2 MCLP stands for Maximal Covering Location Problem and DSM and DDSM are the Double Standard Model and the Dynamic version of it, respectively. GAs are the most common meta-heuristic applied to the solution of MEXCLP. Additionally, in this review the only meta-heuristic identified to solve the MEXCLP that includes the hypercube model was GA. The review includes some details about the performance of the different meta-heuristics. In relation to GA the reported results are good, in comparison to exact solutions found by using an exact approach. For small cases the optimal solution is found. For larger cases, close to optimal solutions are found in reasonable computation time (smaller than the time used by the exact algorithm). An improvement that is mentioned by several authors is the use of a local search add-in to the standard GA, as well as the use of greedy initialization procedures instead of a pure random initialization of

the GA's population. Only one paper was found by this authors using TS to solve the MEXCLP problem.

The work by Rajagopalan et al. (2007) offers an experimental design framework to compare the performance of several popular meta-heuristics, such as TS (Tabu Search), GA (Genetic Algorithms), SA (Simulated Annealing) and HHC (Hybrid Hill Climbing). They solved 400 different problems, generated by using 6 different system configurations. Each configuration is a combination of the size of the grid being used to represent the demand zones, and the relative distribution of the demand across the grid (uniformly and non-uniformly). The solutions were compared to exact solutions found by using CPLEX. In general, the results suggest that the type of distribution and size of the problem, in terms of number of demand zones, has the largest effect on the quality of the solution. It is also showed that GA, in general, produces the best results in terms of quality of the solutions, however it comes at a price, higher computation times. Regarding this issue, it is worth to recall results reported by Jaramillo et al. (2002), showing that GAs can quickly generate good solutions, within 1% of the known optimal, but at the same time, after that it takes a lot of time for the GA to improve it further. This could be the same phenomenon in this case. Because of that behavior, Rajagopalan et al. (2007) suggested that a good idea might be to use GA at first to generate a pool of good solutions, and then to use another approach to improve them further.

1.4 Dissertation structure

The main results from this research are introduced in the next three chapters. Each one of these chapter adds details about available and relevant bibliographical references. Chapter 2 introduces the research problem, a combination of the location decision for ambulances (strategic level) and the dispatching decision (operational level). The main result from Chapter 2 is the joint location/allocation model for EMS systems, as well as the development of a solution strategy for the model, given the fact that it is not possible to be solved by using off-the-shelf commercial solvers. The proposed optimization framework is based on Genetic Algorithms. Congestion is modeled by using a Markovian system and an exact solution procedure is used to characterize the stochastic behavior of the system.

Chapter 3 takes advantage of the optimization framework, allowing the analysis of EMS systems from perspectives other than efficiency. The main result is the identification of fairness criteria that benefit from the joint approach. These new criteria are used to optimize the system and comparisons are made showing the different trade-offs, not only between fairness criteria but also efficiency criteria. The most commonly used approach corresponding to Expected Coverage maximization is contrasted with a Mean Response Time minimization, identifying that the latter offers a better trade-off.

Chapter 4 addresses scalability issues associated with the optimization framework. In particular, the Markovian submodel representing congestion requires a solution procedure whose number of equations increases exponentially. In addition, the mathematical model used in previous chapters allows only one ambulance to be assigned to each candidate location. In turn, real world EMS systems usually will allow for several ambulances at strategic locations. This chapter introduces changes in the model, allowing several ambulances per candidate location. The new model is also accompanied by a new way to model congestion, dropping the Markovian model and using an approximation instead, for which the assumption of exponential service times is not required. This is very important because there is literature showing that response times are not necessarily exponentially distributed. Changes in the mathematical model also required changes in the solution procedure. In this case a Tabu Search based heuristic is developed, including random initialization and dynamic tabu tenure update. Finally, in Chapter 5 a summary of contributions and open related research questions are presented.

Chapter 2

Developing the Optimization Framework

2.1 Introduction

Emergency Medical Service (EMS) systems are a public service that provides out-of-hospital acute care and transport to a place of definitive care, to patients with illnesses and injuries that constitute a medical emergency. The ultimate goal of EMS systems is to save lives. The ability of these systems to do this effectively is impacted by several resource allocation decisions including location of servers, districting of demand zones and dispatching rules for the servers. Common objectives are minimizing the mean response time and/or maximizing coverage. The relationship between minimizing response time and improving survivability has been reported by several works such as Sanchez-Mangas et al. (2010) and McLay and Mayorga (2010, 2011). A demand zone is said to be covered if there is at least one facility within a predefined distance/time threshold from the demand zone. The concept of coverage is related to the availability of a satisfactory facility rather than the best possible one (Farahani et al., 2012). Li et al. (2011) pointed out that the coverage maximization approach is the most widely used by practitioners, researchers and regulators.

Traditionally, location and dispatching decisions have been approached separately, even though various studies have shown that the servers' busy probabilities (and therefore the response

time and coverage, among other performance indicators) are sensitive to the server locations and the choice of server dispatching strategies (Batta et al., 1989; Larson and Odoni, 1981). Ambulance dispatch is the process of assigning a particular ambulance to answer an emergency call. An ambulance dispatch policy can be formed using various dispatch methods and there is no single policy that fits all systems (Li et al., 2011). The same authors emphasized that a dispatch policy has to be designed to fulfill the particular objectives and performance indicators defined by EMS providers and regulators. In our work we consider dispatch policies in which there is a single list associated with each demand zone that ranks the available servers (ambulances), or a subset of them, in order of dispatch preference. This type of list is commonly referred to as a contingency table.

The most common dispatching policy for EMS calls is rather simple in that the closest idle vehicle is usually dispatched to attend the call (Goldberg, 2004; Andersson and Varbrand, 2006). The rationale behind that policy is related to the idea of having the objective of minimizing the mean system response time. The works on allocation of distinguishable servers by Jarvis (1981) and Katehakis and Levine (1986) pointed out that under light traffic conditions using a myopic allocation policy (i.e. assigning always the closest available sever) will lead to an optimal solution, when the objective is to minimize the long run average cost (response time). For heavy traffic the same works mentioned that the optimal policy can deviate from the myopic policy. However, even in the latter case using the myopic policy still might lead to solutions that are close the the optimum (Katehakis and Levine, 1986). Related literature applied to EMS systems planning included arguments against the closest dispatching rule as a way to minimize the response time. Arguments were made originally by Carter et al. (1972) and thereafter supported by Cuninghame-Green and Harries (1988) and Repede and Bernardo (1994). In the referred works the locations of the servers are assumed to be known. We have not found references addressing the relationship between a myopic dispatching policy and expected coverage. There is usually a trade-off between response time and coverage, so that improving one of them implies a sacrifice in the other.

In this work, first we present a mathematical model that integrates the location and dispatching decisions for an EMS system. It is a non-linear mixed integer optimization model in which

even generating some of the equations is computationally intensive, therefore making it hard to solve. The Hypercube model is used providing an exact model of the stochastic queuing dynamics. The mathematical model is accompanied by the analysis of randomly generated small instances whose purpose is twofold: (i) given the small size it is possible to fully enumerate all feasible solutions hence also identifying the optimal, that can be used later for comparison purposes against faster/smarter solution strategies than enumeration; and (ii) after solving a variety of random instances it is also possible to point out some general trends observed in the optimal solutions (with respect to response time and coverage). Second, we present an optimization framework to solve the joint location and dispatching problem based on Genetic Algorithms (GAs). We present a heuristic solution procedure to solve the exact model of the system. Our work is different from previous approaches to the problem, for although we assume the general form of the dispatching policy, as a fixed preference list, we do not assume a priori any particular dispatching order (based on distance, for example). Instead, we model the location and dispatching decisions in a single mathematical model, and develop an optimization framework for its solution. In fact, since a district is the union of the demand zones assigned to a particular server, it can be said that an indirect result of our model is also a districting strategy: for each available server, all the zones having it as its first preferred server would form the server's district.

Our findings are that in fact the common dispatching rule based on the closest available server leads to the best solutions when the objective is minimizing the mean response time and locations are optimized simultaneously. Conversely, if the objective is maximizing expected coverage, then the optimal solution could deviate from the use of the closest dispatching rule. However, the best solutions based on coverage offer an increase of that indicator (with respect to the coverage attained by minimizing the mean response time) that is rather small (3.15% average increase - 95% CI: 2.75-3.55%) compared to the sacrifice in response time (65.2% average increase - 95% CI: 56.33-74.24%). Although these numbers correspond to the average results for the small instances, bigger instances showed similar behavior. The optimization procedure proposed has consistently obtained good solutions, i.e. within 1% gap compared to the best solutions obtained by full or partial enumeration procedures, which are computationally more intensive.

While our main goal was the development of the optimization framework for the solution of the joint location/dispatching problem, we discovered that little benefit can be gained from the integrated approach when using the two most commonly used criteria, namely response time and expected coverage. Thus we considered two additional criteria related to fairness, and we used one of them to illustrate the potential benefits of the joint approach. In particular we tested the variance of the individual response times as a measure of fairness from the point of view of the users of the system (demand zones). We found that in this case using a myopic policy would result in a potential deviation from the optimal policy aimed at reducing disparities, as measure by the variance of the response times. We also illustrate the trade-offs among the presented optimization criteria.

In Section 2.2 we provide the presentation of the problem as well as a review of related literature. Next, in Section 2.3 we introduce the mathematical model. Section 2.4 presents a small case study, as well as a summary of its results and implications. Section 2.5 provides a detailed description of the optimization framework based on GAs and section 2.6 introduces the case studies to which the optimization procedure is applied, as well as the results obtained. The last two sections, 2.7 and 2.8 are the discussion of the results and the conclusions, respectively. As part of the conclusions possible extensions of the present work are mentioned.

2.2 Problem presentation and related literature

In Goldberg's review of models for deployment of EMS vehicles (Goldberg, 2004), it is mentioned that little work had been done on dispatching of ambulances. Similar opinion is shared by Lee (2011), mentioning that the contributions in ambulance dispatching are sparse. In turn, Galvao and Morabito (2008) and Iannoni et al. (2011) mention as an interesting extension of their work the use of different dispatch preference lists, instead of assuming that for a given set of locations the dispatching order is based on the closest dispatching rule.

The most widely used dispatching rule under a fixed preference scheme is to send the closest unit, looking to minimize the response times (Andersson and Varbrand, 2006). The first argument against the use of such a policy was made by Carter et al. (1972). They present a case where two

units, A and B, have equally large areas of responsibility, but A's area has a significantly higher call frequency. In those conditions, the mean response time will decrease if B is allowed to respond to some of the calls for which A is the closest unit. The result was generalized for cases involving more than two units by Cuninghame-Green and Harries (1988). Repede and Bernardo (1994) also supported the argument. The works by Jarvis (1981) and Katehakis and Levine (1986) studied the optimal allocation of distinguishable servers on Markovian queuing systems, reaching a different conclusion. For a given location of the servers, so that the cost of assigning a server to a particular costumer is known (the cost in EMS planning is usually related to the amount of time that it takes for the EMS system to effectively respond to a call), these two works showed that under light traffic conditions (traffic is measured by the ratio between the mean arrival rate and the mean total service rate) the use of a myopic policy always would lead to an optimal solution, i.e. minimizing the long run average cost (response time). For heavy traffic the use of a myopic policy will deviate from the optimal, however the deviation is rather small (2 to 3%). Katehakis and Levine (1986) used 0.38 as an indicator of light traffic and 1.94 for the case of heavy traffic.

We propose a mathematical model that combines location and dispatching decisions for EMS vehicles, initially looking for optimal solutions according to maximum coverage or minimum response time. The dispatching decisions are modeled as a fixed preference scheme, meaning that there is a single list associated with each costumer that ranks the available servers (ambulances) ir order of dispatch preference. That list does not change as a result of changes in the state of the system. However, the particular unit that will be dispatched to attend each call from a demand zone is not known in advance, since the assignment depends on the availability of the servers (system's state) when the call is received. Katehakis and Levine (1986) pointed out some results from Markov Decision Theory indicating that, when the number of states of the system as well as the number of actions available to perform in every state (allocation of the servers) are finite, it suffices to consider only deterministic policies; a deterministic policy is one which, whenever the system is in particular state, the set of available actions to perform is deterministic and depends only of the actual state (in our case, which servers are busy, and which are idle).

The servers in a typical EMS system are: (i) spatially distributed in the region; (ii) share the

system workload due to cooperation among them and (iii) have different operational characteristics, such as different preferential regions (Galvao and Morabito, 2008). Those characteristics have been progressively included in different approaches used for planning EMS systems. Congestion is also a typical phenomena related to EMS systems. According to Galvao et al. (2005) the volume of calls for service may keep ambulances busy from 20 to 30% of the time.

Brotcorne et al. (2003) provided a review focused on location models and their particular application to EMS. They classified the location models that evolved over the past 30 years into two main categories, deterministic and probabilistic, recognizing that the most recent models were more concerned with the representation of the stochastic nature of the systems. Location models were also distinguished in *coverage* and *median* type problems. The first class attempts to locate the servers so as to maximize the fraction of the demand that has at least one server unit within a predefined maximum distance or time. The latter type minimizes the average or total travel time/cost between servers and demand zones.

The two seminal attempts to develop basic coverage models were the set covering location problem (SCLP) by Toregas et al. (1971) and the maximal coverage location problem (MCLP) by Church and ReVelle (1974). Extension to those basic models were developed later. TEAM and FLEET models, by Schilling et al. (1979), considered several types of servers; Marianov and ReVelle (1992) improved the MCLP model. Multiple coverage of demands were considered in BACOP1 and BACOP2 by Hogan and Revelle (1986) and other extensions, DSM and DDSM were added by Gendreau et al. (1997, 2001). The *p-median* problem was introduced by Hakimi (1964). The use of the *p-median* model in the planning and location of facilities for EMS can be found in Carbone (1974) and Carson and Batta (1990).

Basic location models are deterministic in nature and therefore do not represent the system accurately (Brotcorne et al., 2003; Jia et al., 2007a). Basic coverage models might make sense when the location of facilities are fixed, but in the case of an EMS system, as soon as a unit leaves its home base to attend a request for service, other demand points that are supposed to be covered by that unit may no longer be covered. The work by Snyder (2004) reviewed several models that address variations in the inputs, such as demands and travel times, as a way to take uncertainty into account.

The same work pointed out the importance of addressing congestion. Daskin (1983) developed the maximum expected coverage location model (MEXCLP) including the modeling of congestion elements. Hogan and Revelle (1986) developed the maximal availability location problem (MALP I and II) and later Marianov and ReVelle (1996) improved it. Farahani et al. (2012) present an extensive up to date review on covering problems in facility location. Arabani and Farahani (2012) developed a survey on facility locations dynamics.

It was the work by Larson (1974) that first used queueing theory elements in facility location models by introducing the hypercube model. Larson (1975) later developed an approximation for the hypercube model due to the fact that exact calculations were prohibitive. Chiyoshi et al. (2001) pointed out, after comparing several models, that the hypercube was the only one with the capabilities for an accurate representation of the system. There is a variety of applications and extension of the hypercube model to EMS system such as the works by Brandeau and Chiu (1989), Mendonca and Morabito (2001), Atkinson et al. (2008), Iannoni and Morabito (2007), Iannoni et al. (2008), Galvao and Morabito (2008) and Geroliminis et al. (2009), among others. It is well documented that the hypercube model is a descriptive tool allowing scenario analysis, not designed as an optimization model. However, it is possible to embed it into an optimization framework. Batta et al. (1989) combined MEXCLP with the hypercube into an iterative, local search algorithm. Aytug and Saydam (2002) replaced the local search by a genetic algorithm. Iannoni and Morabito (2007) as well as Iannoni et al. (2008) and Geroliminis et al. (2011) have embedded the hypercube model into genetic algorithms to solve the location problem. In this paper we also use the hypercube model as it exactly models the system. While hypercube approximations (Larson, 1975; Jarvis, 1985) may lead to faster solution procedures, they do not provide an exact solution. Thus, as mentioned earlier, our approach is to find a heuristic solution to an exact problem as opposed to an exact solution to an approximate problem. Future work, related to scalability issues of the proposed method is mentioned in Section 2.8.

2.3 Mathematical model

Our model is different from existing literature in that we integrate location and dispatching decisions into a single framework, whereas the mentioned references assumed the use of a priori dispatching policy, particularly based on the closest relationship.

2.3.1 Assumptions

It is assumed that the system provides service to a certain geographical region **J** that is partitioned into service regions -demand zones, cells or atoms are other terms that have been use for these partitions. A given number of servers are located at points $i \in \mathbf{I} \subset \mathbf{J}$. Demands occur solely at the center of each service region by time homogeneous Poisson requests for service and are attended at exponential service rates. Larson and Odoni (1981) have shown that reasonable deviation from this last assumption do not significantly alter the accuracy of the model.

Each service region j generates a fraction f_j of the total demand ($\sum_j f_j = 1$). The total demand is then λ and the demand of each zone is $\lambda_j \equiv \lambda f_j$. A server's primary response area (*district*) consists of those service regions to which the server would be dispatched if available. When a request for service arrives, if the primary server is available, it is dispatched immediately. The server travels to the place of the incident, spends some time at scene and then returns to its base location. If the responsible server is busy when a request for it arrives, another server will be assigned, following a fixed priority list with respect to the servers for each demand zone. The priority list can include all the servers available in the system (total backup) or only a subset of them (partial backup). If all the servers are busy, the request is considered to be lost (this typically means that it will be served by an external system). The basic model also assumes that the servers are identical and that the service time of any response unit for any call for service has an exponential distribution with mean $1/\mu$ (This assumption is reasonable if the travel times are short compared to the total service time, which is usually the case in urban areas). The service time for a call includes the set up time, the travel time from the base to the incident location, the on-scene time, a possible related follow up-time and the travel time back to the base. The response time interval is the time from when an ambulance is dispatched until it arrives at the scene.

Each server can be busy or free (idle), generating 2^N possible states for the system (where N = number of servers); the states can be mapped to the vertices of a hypercube (strictly a cube for the case of exactly 3 servers) named B_j $\left(j = 1, 2, \ldots, 2^N\right)$ of dimension N. Each vertex, or state, is denoted by an ordered set of N one digit binary numbers taking the value of 1 if the server is busy and 0 if not ($B_j \equiv \{b_1, b_2, \ldots, b_N\}$). It is assumed that only one step transitions occur, i.e. two servers cannot be assigned simultaneously. Using the convention proposed by Larson (1974), transitions are only allowed between states with Hamming distance equal to 1, where the Hamming distance d_{ij} between two vertices B_i and B_j is the number of digits by which the two vertices differ (or the 'right angle' distance between two vertices of the hypercube). The terms "upward" and "downward" Hamming distance, d_{ij}^+ and d_{ij}^-, refer to the number of binary digits switching from 0 to 1 and 1 to 0. The model of the system corresponds to a finite-state continuous time Markov process. Steady-state probabilities are determined from equations of detailed balance that express a conservation of flow between consequent states. This set of balance equations depends on both, the location of the servers and the dispatching policy.

2.3.2 Formulation

In the following formulation **J** represents the service regions; **I** are the potential location sites, $|\mathbf{I}| \leq |\mathbf{J}|$; N is the total number of response units (servers); t_{nj} is the mean response time for server n to reach region j, when available; λ is the total network-wide demand (requests/unit time); f_j is the fraction of network-wide workload generated from region $j \in \mathbf{J}$; E_{nj} is the set of states where server n is the preferred server for region j; C_N are the vertices of a N-dimensional hypercube; d_{ij}^-, d_{ij}^+ are the "downward" and "upward" Hamming distances between vertices B_i and B_j, $(d_{ij}^- + d_{ij}^+ = d_{ij})$ and λ_{ij}, μ_{ij} are the upward and downward mean rates at which transitions are made from state i to state j, corresponding to vertices B_i and B_j, given that the system is in state i. Finally, we have the decision variables:

$$x_i = \begin{cases} 1 & \text{if we locate a vehicle at potential site } i \\ 0 & \text{otherwise} \end{cases}$$

$$y_{ij}^{l} = \begin{cases} 1 & \text{if vehicle located at } i \text{ has priority } l \text{ to zone } j \\ 0 & \text{otherwise} \end{cases}$$

The following are auxiliary decision variables: ρ_{nj} is the fraction of dispatches sending unit n to region j, $n = 1, 2, \ldots, N$; $P(B_k)$ is the steady-state probability of state corresponding to vertex B_k, $k = 1, 2, \ldots, 2^N$.

The optimization problem is formulated as:

$$\text{Minimize } MRT = \sum_{n=1}^{N} \sum_{j=1}^{J} \rho_{nj} t_{nj} \tag{2.1}$$

s. t:

$$\sum_{i=1}^{I} x_i = N \tag{2.2}$$

$$x_i \in \{0, 1\}, \ i \in \mathbf{I} \tag{2.3}$$

$$\rho_{nj} = f_j \frac{\sum_{B_i \in E_{nj}} P(B_i)}{1 - P(B_{2^N})} \quad n = 1, \ldots, N; j \in \mathbf{J} \tag{2.4}$$

$$P(B_j) \left[\sum_{\substack{i \\ B_i \in C_N : d_{ij}^{+} = 1}} \lambda_{ij} + \sum_{\substack{i \\ B_i \in C_N : d_{ij}^{-} = 1}} \mu_{ij} \right] =$$

$$\sum_{\substack{i \\ B_i \in C_N : d_{ij}^{-} = 1}} \mu_{ij} P(B_i) + \sum_{\substack{i \\ B_i \in C_N : d_{ij}^{+} = 1}} \lambda_{ij} P(B_i) \quad j = 1, \ldots, 2^N \tag{2.5}$$

$$\sum_{i=0}^{2^N - 1} P(B_i) = 1 \tag{2.6}$$

$$y_{ij}^{l} \in \{0, 1\} \qquad i \in \mathbf{I}; j \in \mathbf{J}; l = 1, \ldots, N \tag{2.7}$$

$$x_i \geq y_{ij}^{l} \qquad i \in \mathbf{I}; j \in \mathbf{J}; l = 1, \ldots, N \tag{2.8}$$

$$\sum_{l=1}^{N} y_{ij}^{l} = 1 \qquad i \in \mathbf{I}; j \in \mathbf{J} \tag{2.9}$$

$$\sum_{i=1}^{I} y_{ij}^{l} = 1 \qquad j \in \mathbf{J}; l = 1, \ldots, N \tag{2.10}$$

Equation (2.1) is the objective function, Mean Response Time (MRT); constraint (2.2) determines the number of servers to be located and constraint (2.3) is the integrality constraint for the decision variable x_i. Constraint (2.4) calculates the fraction of all dispatches that send server n to region j using standard queueing theory arguments and assuming a zero-line capacity system (calls that arrive when all the servers are busy are lost); equation (2.5) represents detailed balance equations determining steady-state probabilities of the finite-state continuous time Markov process model with N servers. Note that even though it was assumed that the service rate is equal for all servers since they were identical, the general expression given by this equation allows for different service rates for different servers. For details on calculating λ_{ij} and μ_{ij}, see Geroliminis et al. (2009).

Constraint (2.6) ensures that the sum of probabilities is equal to one. Equation (2.7) is the integrality constraint for the decision variable y_{ij}^l; constraint (2.8) states the logical relationship between the location decision and the assignment of a location within the priority list of a demand zone and finally, constraints (2.9) and (2.10) assure that there is a complete priority list for each demand zone, and that within the priority list of each demand zone each server appears only once. The full model given by (2.1)-(2.10) represents the basic optimization problem in which the location of the servers and the dispatching rule for each demand zone are the decisions to be made. Also note that the steady-state probabilities are auxiliary variables that change for every full combination of location and dispatching decision.

Formally, the presented model corresponds to an NP-Hard problem (Geroliminis et al., 2009). It is a non-linear mixed integer programming model that has embedded a queuing sub-model corresponding to the finite-state continuous time Markov process. Given a particular set of locations for the servers available and a preference list for each demand zone with respect to the same servers, it is necessary to first solve the flow balance equations given by (2.5)-(2.6), before being able to calculate the value of the objective function. Although compactly written, those equations are neither easily determined nor easily solved. The number of flow balance equations equals 2^N, therefore the number of equations to solve the subproblem grows exponentially with respect to the number of servers. It has been mentioned by Galvao and Morabito (2008) that in

fact the computer time required to generate the coefficients of the linear system may be even higher than the time required to solve it. That is because of the complex relationships imposed by the combined location and dispatching decisions. The flow balance equations lead to a linear system of equations, whose exact solution requires the calculation of the inverse for the matrix of coefficients. The size of this matrix grows exponentially, therefore the time that it takes to perform a single iteration to evaluate a candidate solution can be prohibitive.

It was mentioned that maximizing coverage is the most commonly used approach to planning EMS systems. Instead of the standard coverage, we use the concept of expected coverage as presented by Ingolfsson et al. (2008), which takes into account the congestion of the system and potentially the variability in responses times. The following equations details how to calculate the expected coverage:

$$Ex.Cov = \sum_{j=1}^{J} f_j \sum_{n=1}^{N} P_{j,j(n)}(1 - P_{j(n)}) \prod_{u=1}^{n-1} P_{j(u)} \tag{2.11}$$

Where $P_{j,i}$ is the probability that station i covers node j, P_i corresponds to the busy probability of the ambulance in station i and $j(n)$ refers to the n^{th} preferred station for demand node j. Note that $P_{j,i}$ can be used as a binary variable, indicating whether or not the coverage threshold is satisfied by the available servers, but it can also be used as the probability of that coverage being possible within the given threshold, accounting for variability in travel times. In this particular case it has been used as a binary variable. Equation 2.11 replaces Equation 2.1 in the optimization model for the cases in which we are looking at maximum expected coverage.

2.4 Toy case study

In this section we introduce a case study that is small enough that we can enumerate all the possible solutions, also identifying the optimal. Because of the small size we also use the exact solution for the embedded hypercube model. For this example we use a square region on the cartesian plane and assume that we have 5 demand zones, that are also candidate locations for 3 available servers. The locations of the demand zones can be in any integer ordered pair within a

grid, starting at (0,0) and extending up to (10,10) on the plane. The demand for each zone ranges between 1 and 20 calls/period-time.

To generate different instances we use random numbers as follows. The coordinates (x, y) for each of the five demand zones are obtained by generating uniform integers between 0 and 10. For each one of the demand zones the demand is obtained by generating uniform integers between 1 and 20. The distances between demand zones correspond to right angle distances. Optimal locations are nevertheless insensitive to the choice of a distance metric (Benveniste, 1985). The service rate for the servers is assigned based on assuming a particular value for the overall utilization of the system, namely $\rho = \lambda/3\mu$. As in the works by Budge et al. (2009) and Chiyoshi et al. (2001), where ρ is varied between 0.1 and 0.9, three different scenarios of utilization are evaluated for each combination of location of demand zones and demands, by using $\rho = 0.1, 0.5, 0.9$. The server's speed is assumed to be 1.0 distance-units/time. The maximum threshold used for coverage was 7.0 distance units. 100 different set of locations are generated, and since for each one of them three scenarios are considered for the service rates, we generate 300 different instances.

One of such randomly generated problem (Table 2.1) and its respective optimal solution for MRT (Table 2.2) after enumeration is detailed next. It is worth noticing the number of possible solutions: 77,760. There are $\binom{5}{3} = 10$ possible locations. Each demand zone has an ordered list of the servers, and since there are 3 servers, each costumer can have 3! unique lists. The total number of solutions is then $10 \times 3!^5 = 77,760$. Note that the inclusion of another demand zone would increase the number of solutions to 466,560. In other words, the number of possible solutions increase by a factor of 3!. Hence, the search space for a real size problem is huge, and enumeration is no longer an alternative.

Table 2.1: Locations and demand

Locations			Demand
Index	$x - coord$	$y - coord$	
1	10	5	20
2	1	1	18
3	7	9	12
4	2	7	8
5	6	1	6

Table 2.2: Optimal(MRT) solution information

ρ	Performance indicator				Optimal
	St.Cov	Ex.Cov	MRT	P[111]	locations
0.1	1.0	0.954	2.123	0.003	1-2-3
0.5	1.0	0.721	4.340	0.134	1-2-3
0.9	1.0	0.517	5.355	0.309	1-2-4

In Table 2.2, St.Cov refers to the basic calculation of coverage, hence each demand zone is considered covered if there is at least one ambulance located at a distance of 7.0 or less distance units; St.Cov=1.0 means full coverage. However, this definition of coverage does not take into account the congestion of the system, hence we used Ex.Cov (Eq. 2.11). MRT (Eq. 2.1) is the mean system response time. The 5^{th} column corresponds to the probability of the system being busy (all the servers are attending calls), and therefore new calls would be rejected. The last column indicates the optimal locations of the servers.

At first sight results in Table 2.2 correspond to what was expected. On one hand, the increase of the overall utilization, which basically means reducing the service rates while keeping the call rates constant, causes an increase in the expected response time, as well as an increase in the busy probability. On the other hand, we can see that the standard coverage is not able to take into account the congestion phenomena. The expected coverage given by (2.11) is clearly affected by the increase of the overall utilization. The more congested the system, the lower the expected coverage.

We have solved by enumeration a total of 300 small size problems, for both minimum response time and maximum expected coverage. As expected according to the arguments expressed in section 2.2, for each one of the 300 problems, the optimal solution that minimizes MRT was the same as a solution where the locations were optimized and a dispatching list based on the closest vehicle was used. We have also observed that, when there are ties (several servers are at the exactly same minimum distance from a given demand zone), only one of the combinations leads to the optimal solution, although the use of other dispatching ranking, which would still be based on the closest rule (because of the ties), causes an increase on the objective function value that in the worse case is below 2%. Note that the change of the preference list of a single demand zone, even if

for that demand zone several servers' locations are tied, changes the overall performance indicators of the system.

Next we looked for maximum expected coverage (Ex.Cov), as given by equation (2.11). Once again, we enumerated all the possible solutions for all 300 instances to be able to identify the ones that generate the maximum coverage. In this case we have noticed that the optimal solutions do not follow the closest dispatching rule. We have also observed that there are several solutions that exhibit the same maximum coverage for a particular instance, and that the associated response time of those solutions present great variation. Since minimizing the response time is also important, in cases where there were several optimal solutions with respect to coverage we have selected the one with the minimum associated response time. Although the optimal solutions with respect to expected coverage do not follow a myopic allocation policy we have also noticed that the use of such a policy would cause a decrease in expected coverage that in all cases is below 3.8%. In fact for half of the cases it is below 1.0%.

Since both objectives are important, we compare the optimal solutions obtained with each optimization criteria. We noticed that for every instance of the problem, the maximum coverage identified is in fact greater than the coverage associated with the minimum response time solution. However, the increase in coverage is small on average, ranging from 0.18% to 17.4%, with a mean increase of 3.15% (95% CI: 2.75-3.55%). On the other hand, the increase in coverage (obtained by changing the objective function) comes as a result of worsening the response time. The increase in MRT ranges from 1.5% to 117%, averaging 65.2 (95% CI: 56.33-74.24%)%. As expected, there is a trade-off between response time and expected coverage. These results seem to indicate that focusing on minimizing the response time yields solutions that are robust with respect to the expected coverage. There were only 4 cases (out of 300) in which the proportional improvement in coverage (when maximizing coverage) was in fact higher than the corresponding increase in response time. In all the other cases the trade-off between improved coverage but increased response time is not appealing. These results are aligned with those reported by Geroliminis et al. (2009), who mentioned that the optimal locations obtained by using MCLP (a coverage maximization approach) performed up to 40% worse when the response time was evaluated by using the hypercube model.

In the next section we introduce an optimization framework that allows us to solve bigger size problems, hence allowing us to check if the observed behavior of the small instances holds for more real-world sized problems.

2.5 Genetic algorithm based optimization framework

Next we develop an optimization framework to solve the combined location and dispatching decision problem for EMS systems. The optimization is based on GAs. In his review, Goldberg (2004) suggest that the use of spatial queuing (hypercube model) or simulation procedures embedded within a heuristic search offers the greatest utility for real world EMS planning applications. Aytug and Saydam (2002) also comment on the success of GAs in solving combinatorial problems, which make them strong candidates to solve the ambulance location/allocation problem. Iannoni and Morabito (2007) as well as Iannoni et al. (2008) and Geroliminis et al. (2011) have also embedded the hypercube model into genetic algorithms to solve the location problem. As mentioned by Geroliminis et al. (2009) the objective function MRT, as a function of the location space, has many local minima, making it suitable for a global search procedure such as GAs. Jia et al. (2007b) proposed a GA to solve the problem of locating facilities to attend large scale emergencies. Shariff et al. (2012) used a GA for solving the MCLP problem applied to healthcare facility location in Malaysia.

GAs were first introduced by Holland (1975) and popularized later by Goldberg (1989). GAs are general-purpose, population based search algorithms that resemble the natural selections survival of the fittest. Particular coded schemes (solutions representations) corresponding to chromosomes represent population members. At each iteration individual solutions are evaluated and assigned a fitness value (related to the objective function being optimized). According to their fitness values, solutions are selected to construct the next generation by applying genetic operators: selection, crossover and mutation. Current members of the population are probabilistically selected based on their fitness values, where a high fitness value yields a higher chance of being selected for the next generation. After selection, current solutions may be carried to the next generation without altering (selection), or they may be crossed-over to generate the next set of solutions. Crossover

is an operator by which two solutions mutually interchange their current genes. Mutation is an operator that randomly alters the value of a gene of a selected solution. Following the work by Aytug and Saydam (2002), there are five key issues in designing a GA algorithm: (1) Selecting an appropriate solution representation, (2) an effective mutation operator, (3) an effective crossover operator, (4) a feasible initialization and (5) appropriate crossover and mutation rates as well as population size.

2.5.1 Solution representation

The present work uses the idea of a composite chromosome. That is, a chromosome that is in fact composed of several chromosomes. This representation makes sense given the nature of the problem in which there are two decisions to be made, a location decision and a dispatching decision. Therefore, those two decisions are coded in separate sub-chromosomes. Furthermore, since the dispatching decision is in fact one decision per each demand zone, that gives rise to the idea of having separate chromosomes to represent each demand zone. Figure 2.2 shows the composite chromosome for a case in which there are 3 servers to be located among 5 candidate locations to attend 5 demand zones (every demand node is a candidate to locate a server). Note that the chromosome has been divided into sub-chromosomes. The first one deals with the location decision, and therefore has size 5, with the three first components storing the location of a server. The location sub-chromosome stores more information than required, since only a subset of locations will have a server. However, it is kept that way to facilitate feasibility checking as well as the mutation operation, described later. The remaining sub-chromosomes have size 3, representing the order in which every server is ranked to attend a particular demand zone. Note that the sub-chromosome for the location decision corresponds to a permutation of the candidate locations and any sub-chromosome for the dispatching decision corresponds to a permutation of rankings.

2.5.2 Mutation operator

The standard mutation operator randomly selects a chromosome from the pool, and then goes through every one of its genes changing them randomly with a given probability. Since we are

using a composite chromosome, once a chromosome has been selected for mutation the operation should analyze every one of its sub-chromosomes. For we are working with sub-chromosomes that are a permutation, the standard mutation operator is replaced by a swapping operator. It randomly interchanges the positions of two genes within the chromosome, as depicted in Figure 2.1. Note that for the location sub-chromosome the swap is done such that the interchange occurs between an assigned location in the current solution, and a candidate location not yet selected. That is in order to avoid swaps that do not affect the solution.

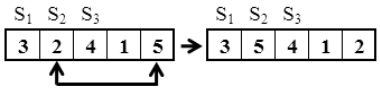


Figure 2.1: Swapping mutation operator

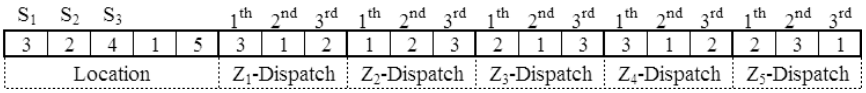


Figure 2.2: Composite chromosome

2.5.3 Cross-over operation

A single point cross-over operation is used on the implementation of the GA. The recombination of genes is done at sub-chromosomes level, which means that the the candidate cross-over points correspond to sub-chromosomes as well. To better understand the way it operates an illustrative example is given in Figure 2.3. **A** and **B** are the two parents. **O1** and **O2** represent the two offsprings that it is possible to generate. Parent **A** has been shadowed so that it is possible to trace where the genes of it are going to be after the cross-over operation. The possible crossover points are represented by vertical dashed lines.

2.5.4 Population initialization

It is usually the case that the initialization is done randomly. The existence of constraints might require us to develop initialization routines that produce feasible solutions. In this case the population of the GA can be randomly initialized, since any permutation for any sub-chromosome

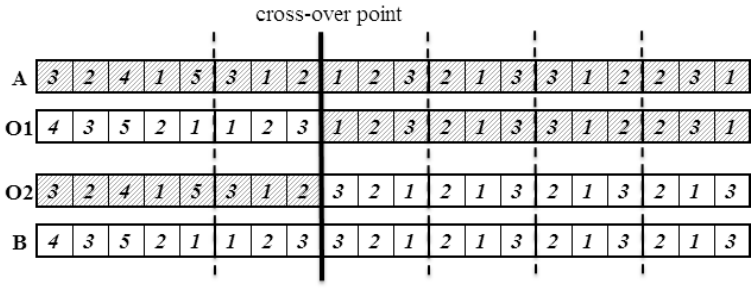


Figure 2.3: Single point cross-over for the composite chromosomes

will generate a feasible solution. However, it is also possible to use an initialization procedure to create 'good' initial solutions, using the available knowledge about the problem. Initial tests of the GA implementation were done with a randomly generated population. Based on the result from the enumeration procedure for the small case study presented in Section 2.4, a better initialization procedure is devised. Since the use of the closest dispatching rule seems to effectively helps in minimizing the response time also providing good coverage, it makes sense to use that information as part of the initialization process. In fact, when solving mid-size problems, the locations are generated randomly during the initialization, but the dispatching is based on the use of the closest servers first.

2.5.5 Cross-over and mutation rates - population size

In order to test the performance of the GA values of mutation (P_m) and cross-over (P_c) rates are required. Iannoni et al. (2008) used $P_c = 0.5$, and $P_m = 0.06$, while the population size was set to $S = 100$ individuals. In turn, Aytug and Saydam (2002) suggested $P_c = 0.6$ and $P_m = 0.03$, while the population size was set according to $S = \max(100; 0.75n)$, where n is the number of nodes in the problem being solved. The authors argued that for objective functions with potential multiple local optima, there is a trade-off between mutation and cross-over, and that large population sizes are generally favorable, at the cost of computation time. It is also mentioned that the rule of thumb $P_m = 1/L$, where L refers to the length of the chromosome could yield good results. Instead of selecting arbitrarily values for these GA parameters, in the next section we introduce an experimental design to tune-in the parameters of the GA before using it. The

implementation of the GA has been done using the Java GA framework developed by Meffert et al. (2012).

2.6 Computational results

2.6.1 Tuning the GA

A tuning procedure was carried out to find adequate values for several parameters of the GA. The purpose of any experiment is to get the maximum amount of information with the minimum expenditure of resources. A Central Composite Experimental Design (CCD) was used, which according to Montgomory (2008) is widely used because it is highly efficient and flexible. A CCD is normally used to fit a second order polynomial model of a variable of interest. In our case we are not trying to fit a polynomial model. However the combination of factors' values suggested by the CCD provide a good exploration of trade-offs between the different parameters of the GA and its general performance.

There are three parameters (factors) that need to be set up: mutation rate, cross-over rate and population size. In experimental design, for each one of these three factors it is necessary to specify a minimum and a maximum value. These values have been selected according to general recommendations of designing GAs from previous works (Iannoni et al., 2008; Aytug and Saydam, 2002). The CCD also uses the midpoint of the factor (given the minimum and maximum values), as well as the so-called axial points. Axial points correspond to values of the factors that assure that the predicted values of the fitted response surface have the same variance, if the predicted points are at the same distance from the center of the design region (Montgomory, 2008). For the case of three factors a standard CCD requires 20 runs. The first 14 runs correspond to different combinations of the factor's levels, while the last 6 runs correspond to experiments in which each factor is set to its midpoint. A standard CCD does not uses replication. We do use it (30 runs for each combination of factors), as a way to improve the statistical significance of the tests. Instead of the 6 last runs each with one replication, we have a single run setting the factors to their midpoints and we replicate it 30 times. We explore 15 combinations of factors, detailed in Table 2.3.

The minimum and maximum values for the mutation rate are 2 and 5% (the values used in the experiments are then those two, plus the center point, 3.5% and the axial points 1 and 6%). For the Cross-over the minimum and maximum values are 40 and 60% and the population size varies between 30 and 100 individuals. In all the runs of the GA to tune in the parameters, the number of evolutions is set up so that the total number of individuals being evaluated remains constant (approximately equal to 10.000). For example, if the population size is set to 30 then 334 evolutions are performed.

Table 2.3: Experimental design

Combination	P_m	P_c	Pop. Size
1	0.02	0.4	30
2	0.05	0.4	30
3	0.02	0.6	30
4	0.05	0.6	30
5	0.02	0.4	100
6	0.05	0.4	100
7	0.02	0.6	100
8	0.05	0.6	100
9	0.01	0.5	65
10	0.059	0.5	65
11	0.036	0.332	65
12	0.036	0.668	65
13	0.036	0.5	6
14	0.036	0.5	123
15	0.036	0.5	65

The results from the tuning procedure are given by the box plot graph shown in Figure 2.4. It corresponds to the tuning for MRT optimization. As it was mentioned before, for each combination of factors given in Table 2.3, the GA was run 30 times, applied to different instances and each time using a different random seed. In each case 100 evolutions of the GA were allowed. We have noticed that allowing more evolutions didn't further improve the objective value. In order to have a comparison point to tune the GA we enumerate only the location solutions for the case study ($\binom{30}{3} = 4,060$) possible location decisions). It is not possible to enumerate the dispatching decisions. It would be computationally prohibitive. For each possible location solution we use the closest rule to set the priority dispatching list of each demand zone. We then compare the

performance of the GA (GASol) against the best solution found (BestSol) following the enumeration procedure just described. The *Gap* is calculated as $(Gap = (BestSol - GASol)/BestSol)$.

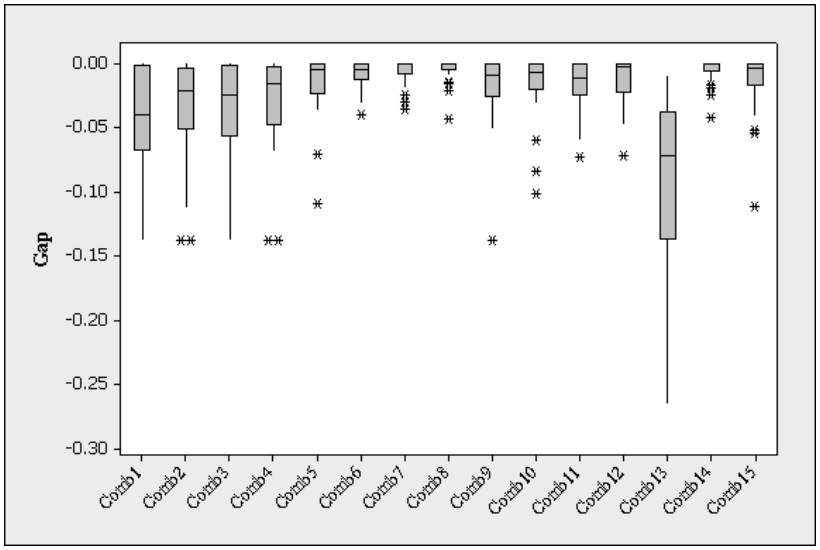


Figure 2.4: Comparative performance of the GA varying its parameters

Negative values of the Gap indicate that the GA obtained a solution with a worse objective function value than the best solution coming from the enumeration procedure. If Gap = 0 it basically means that the GA was able to find a solution with the same objective function value. Positives values of the Gap would indicate that the combination of dispatching and location decisions was useful in getting a better value for the objective function. Recall that the gap reported is the average over 30 runs. Out of the 15 combinations of factor's levels under consideration, combinations 7, 8 and 14 showed the best overall performance. All have a Gap close to zero, and exhibited low variability. We ran normality tests on the selected combinations and could not verify the normality of the data. Therefore, we performed a non-parametric test, the Wilcoxon Signed Rank Test, to obtain the Confidence Intervals (CI) for the three candidate combinations. Table 2.4 shows the results from the non-parametric test. As it can be seen, the overlapping CIs indicate that statistically there is no difference between the parameters combinations. We decided to use combination 8 which has the smaller CI.

Table 2.4: Wilcoxon Test for obtaining CIs

Combination	Median Gap (%)	Conf. Interval	
		Lower	Upper
Comb7	-0.191	-0.823	-0.058
Comb8	-0.208	-0.476	-0.049
Comb14	-0.271	-0.712	-0.109

2.6.2 Mid-size case study

We solved a bigger, mid-size problem, proposed as an instance of MCLP (Correa et al., 2007) (`http://www.lac.inpe.br/~lorena/correa/Q_MCLP_30.txt`). We analyze several scenarios, varying the number of servers to be located, considering 3 and 4 ambulances. Because of the small number of ambulances we use again an exact solution for the hypercube model. The server rates are obtained by selecting particular values for the overall utilization factor, $\rho = (\lambda/N \times \mu)$. In fact, ρ is varied between 0.1 and 0.9, with increases of 0.1. For the scenarios having 3 servers we use full backup, which means that any zone can be attended by any of the available servers. In the case of 4 servers we use partial backup, therefore each demand zone is only allowed to be served by 3 of the available servers. There are two reasons to proceed this way, that have also been suggested by Geroliminis et al. (2009): (i) from a practical perspective, allowing servers that are ranked as 4^{th} and up for a particular demand zone is not desirable, because the overall efficiency of the system would likely decrease; (ii) the calculation of transition rates for the embedded hypercube model becomes very tedious.

For each instance we ran the GA using the tuned parameters and initially minimizing the MRT. In each case we also enumerate the location solutions, and combine them with the use of the closest dispatching policy to have a full solution. That gives us a comparison point. The GA was allowed to run for 100 evolutions. We have noticed that allowing more evolutions doesn't improve the results. The performance of the GA is compared to the best solution coming from the enumeration procedure. Table 2.5 shows the results of applying the GA, in each case running it 30 times starting with different initial solutions. The experiments have been run on a PC executing Windows 7 -64 Bit, with an Intel®Core 2 Duo processor running at 2.13 GHz and 2 GB of RAM.

All the programming was done in Java. The average running time of the GA for the 3 servers scenarios was 20 seconds, while the average for the case of 4 servers was 55 seconds.

Table 2.5: Mid-size case study - MRT results

ρ	3 servers - MRT		4 servers - MRT	
	Gap(%)	CV	Gap(%)	CV
0.1	-0.26	0.0039	-0.51	0.0054
0.2	-0.07	0.0015	-0.27	0.0043
0.3	-0.08	0.0022	-0.37	0.0050
0.4	-0.02	0.0009	-0.53	0.0041
0.5	-0.02	0.0001	-0.78	0.0068
0.6	-0.03	0.0012	-0.46	0.0071
0.7	-0.14	0.0054	-1.46	0.0160
0.8	-0.04	0.0014	-0.98	0.0117
0.9	0.00	0.0000	-1.00	0.0113

As expected given the low-medium traffic (Jarvis, 1981; Katehakis and Levine, 1986), for the mid-size problem the results suggest that a policy that focuses on appropriately selecting locations in combination with dispatching the closest server minimizes mean system response time. These results serve as a validation of the general structure of the mathematical model as well as for the correctness of the optimization procedure. For the mid-size case studies we have also observed that the expected coverage associated with the solution that minimizes the MRT is smaller (on average it is 7.2% smaller, with a 95% CI: 6.14-8.31%) than the maximum observed after the enumeration procedure.

Next we approached the optimization of the system maximizing the expected coverage. A procedure similar to that described in section 2.6.1 was followed to tune the GA to be used with the new objective function, Expected Coverage. In this case the combination 7 (from table 2.3) showed the best results and therefore was selected as the values for the GA parameters. The enumeration procedure of the location decisions together with a myopic dispatching policy was used again to identify the solution with the highest expected coverage. The performance of the GA was compared against the solution from enumeration. The overall average Gap of the GA compared to the the enumeration procedure was -0.87%. The overall mean coefficient of variation of the maximum coverage was 0.0136. These performance measures of the GA show that the algorithm

was consistently able to get to the same or to a very close solution from the best found by the enumeration procedure. Compared to the solution that minimizes the response time, the average improvement in coverage is 7.9% (95% CI: 6.64-9.09%). However, this increase comes at a price, a sacrifice of MRT that on average increased by 19.2% (95% CI: 16.35-21.97%). Recall that the expected coverage of the of the solution minimizing MRT was on average 7.2% smaller than the maximum obtained with enumeration, while the increase on MRT would be on average 19.2% as a result of maximizing coverage. The joint location/allocation approach was not able to improve the solution found by combining the enumeration of locations and the closest dispatching policy.

Thus far we have introduced an optimization framework for the joint location/allocation problem, however we have noticed that for the two most common objectives the joint approach is not adding value, since the use of a myopic policy seems to suffice to get to the optimal or near optimal solution. Hence we have turned our attention to calculating other performance indicators for the system. In particular, other works have mentioned the importance of finding solutions in which the total workload is evenly distributed among the available servers, and some others have mentioned that it would be desirable to have individual response times (the mean response time for each demand zone) that do not vary too much among the demand zones. Both performance indicators are associated to the idea of fairness, either from an internal or external point of view. In Table 2.6 we present the coefficient of variation (CV) for both, mean individual workloads and mean individual response times, resulting from the solutions that optimize mean response time. In this table several instances of high CV values (for example ≥ 0.5) are observed, which implies high variability among server's workloads or demand zones' response time.

Among the several instances of the case study it is possible to notice that variability on individual response times tends to be higher (see Table 2.6), hence we attempted to improve that performance indicator by using the optimization approach already developed. Equation 2.1 gives the total average response time for the system and in doing so it includes the response time for each demand zone. We use the coefficient of variation (CV) of the individual response times as the new optimization criteria. Once again it was necessary to tune the GA with the new objective function. The GA parameters that perform the best were the same as for the case of MRT minimization. We

Table 2.6: Mid-size case study - Workloads and Individual MRT

ρ	3 servers - CV		4 servers - CV	
	Workload	Ind.MRT	Workload	Ind.MRT
0.1	0.387	0.594	0.357	0.605
0.2	0.250	0.590	0.105	0.529
0.3	0.039	0.565	0.318	0.489
0.4	0.029	0.541	0.335	0.436
0.5	0.038	0.549	0.325	0.512
0.6	0.018	0.539	0.317	0.501
0.7	0.031	0.531	0.612	0.556
0.8	0.028	0.525	0.609	0.551
0.9	0.025	0.520	0.605	0.548

used the enumeration procedure of the location decisions to get a reference point of the minimum CV for the response times and the compare those solutions with the ones obtained by the joint location/allocation approach. Table 2.7 shows the results for several instances of the problem, all of them using 3 servers.

Table 2.7: Mid-size case study - Optimizing CV Resp. Times

ρ	Min CV Ind. RT		Delta	Trade-offs (%)	
	Enum.	Loc/Disp	CV(%)	MRT	Ex. Cov
0.1	0.489	0.364	-25.577	37.853	-11.980
0.2	0.504	0.355	-29.652	32.423	-10.461
0.3	0.496	0.367	-25.948	24.485	-4.603
0.4	0.478	0.375	-21.571	19.698	-3.981
0.5	0.465	0.380	-18.285	16.345	-3.911
0.6	0.458	0.384	-16.154	13.764	-3.497
0.7	0.454	0.389	-14.300	11.954	-3.472
0.8	0.481	0.392	-18.636	11.051	-3.098
0.9	0.476	0.397	-16.534	9.612	-3.022

The second column in table 2.7 shows the minimum CV for individual response times that was obtained by using the enumeration procedure for the location decisions in combination with the closest dispatching rule. The third column shows the CV that was possible to achieve by using the proposed optimization approach while the fourth column compares the two previous, showing the relative improvement that was possible thanks to the joint approach. The last two columns show the sacrifices in MRT and Expected Coverage that come as a result of the reduction in response

times variability across demand zones. We see that there are both an increase in response time and a reduction in coverage. The size of the trade-offs depends upon the utilization (ρ) of the system. The trade-offs were calculated using the solution that minimizes response time as a reference point. For instance, for $\rho = 0.4$ there is a reduction of 21.5% in response time variability, as measured by the coefficient of variation, as well as an increase of about 20% in response time and a reduction of 4% in coverage.

2.7 Results summary and discussion

We have done extensive computational experiments using 300 small case studies (enumerating more that 70,000 solutions for each instance). We were looking for a better understanding of the potential benefits when location and dispatching decisions are made together for an EMS system. The instances have been generated randomly therefore not favoring any particular result in terms of the decisions being made. Although previous literature had suggested that the existence of demand zones with very different demand rates could lead to situations in which the dispatching based on the closest rule was not optimal, our results were in agreement with some other references showing that using a myopic policy can lead to optimal solutions. We have allowed the demand rates to vary between 1 and 20, therefore introducing differences in the demand rates. What we have found is that if the dispatching policies are designed as a fixed priority list associated to each demand zone, then focusing on finding good locations, and combining them with the use of the closest dispatching rule, yields the desired result of minimizing the mean response time.

In terms of coverage, which is also a common objective to optimize in EMS system, we have used an expected version of coverage, since previous works have made it clear that the standard coverage, which does not take into account the congestion phenomena, overestimates the real coverage. For the small instances we have found that the solutions that maximize the coverage did not use the dispatching policy based on the closest rule. However, we have also noticed that the improved coverage that comes as a result of its maximization, causes a deterioration in the mean response time. As pointed out in section 2.4, optimizing the coverage increases it less than 5% (compared to the coverage obtained by the best solution with respect to MRT), while the sacrifice

in MRT would be greater than 60%. Those results basically suggest that optimizing the MRT is a better strategy, and that in fact the results in coverage when optimizing MRT are robust, in the sense that the coverage is only 1 or 2% below its optimal value. These results about coverage were also validated with a mid-size real case study found and adapted from previous literature. For the mid-size cases the best coverage was reached when using the closest policy. The average improvement in coverage was again smaller than the average increase in response time.

Results from alternative performance indicators such as those depicted in Table 2.6, suggest some other observations. Given the values for the CVs it is not surprising that in some cases there are some demand zones with a MRT that doubles that of other zones, or one ambulance having a much heavier workload than the others. Solutions that are good from the point of view of system wide mean response time, can have other performance indicators affected negatively. Since the optimization has been done with a single objective in mind, there is no guarantee of good performance with respect to other criteria. Our results have shown that optimizing the MRT also yields good values for expected coverage. That is convenient since those two are the most common performance indicators used for planning purposes of EMS systems. We illustrated the potential benefits of the joint approach by considering a fairness performance indicator from the user point of view, namely coefficient of variation for individual response times. In this case, the joint approach was able to find better solutions than those that could be reached by using a myopic allocation policy. Of course, the improvement of a fairness objective like the one we have used has consequences, altering other performance indicators such as MRT and coverage. It would be up to the decision maker to balance those trade-offs.

2.8 Conclusions

Our main goal was to develop an optimization framework for the joint location/allocation problem for EMS systems. We combined the mathematical model and a heuristic solution procedure based on Genetic Algorithms, to be able to solve bigger instances in which enumeration is no longer an option. We were able to validate our approach. The GA has been consistently able to find the same or a pretty close solution to that obtained by full or partial enumeration procedures. In terms

of MRT minimization or Expected Coverage maximization we have noticed that the integrated approach does not offer tangible benefits. A more simpler approach considering only the location decision combined with a myopic allocation of the servers based on closest distance would be enough.

One general explanation of the observed behavior is that MRT and Expected Coverage are in fact a function of the distance (time) between servers and demand zones. Hence, locations that reduce the overall distance between servers and costumers tend to dominate the optimization procedure. Although in this case we could just have proposed an optimization procedure in which the decisions are the optimal locations, combined with the use of the closest dispatching rule, we have kept both sets of decisions as part of the optimization framework. We believe that it is important because it gives us the opportunity to attempt the optimization of other performance indicators, so that we can see the trade-offs that are being made as a result of focusing on minimizing the response time or maximizing coverage. The fact that solutions that minimize response time offer at the same time a good expected coverage is convenient, since those two criteria are the most commonly used.

We have illustrated two alternative criteria, in particular variability on individual response time, as well as variability on ambulances workloads. Those criteria can be seen as fairness performance indicators from the perspective of internal and external costumers. We used individual response time variability as a optimization criteria. As in the case of maximizing the expected coverage, when focused on reducing response time variability among demand zones it is the case that the best solutions do not follow the use of the closest dispatching rule. Furthermore, the improvements that can be made on variability are important, and not only marginal as in the case of maximizing coverage. The proposed optimization framework, already proven to work correctly, can be used to analyze the EMS system from other perspectives, gaining insight into the design of better operation strategies.

As for future research directions, we will attempt to identify other performance indicators of EMS systems for which the joint location and dispatching problem can yield substantial gains. Another potential area for future research deals with the issue of scalability. We are aware of the limitations of our approach in terms of applying the joint model and its solution procedure to real-

sized case studies, basically because the exact solution of the hypercube model will likely require extensive computation time (recall that the exact solution to the hypercube model requires solving a linear system of equations that grows exponentially in size with respect to the number of servers available in the system). However, available approximation procedures that have been suggested in the literature could be embedded in the meta-heuristic optimization framework proposed, hence reducing the computational burden and allowing the solution of bigger instances.

Chapter 3

Joint Optimization Approach and Fairness Considerations

3.1 Introduction

Emergency Medical Service (EMS) systems are a public service that provides out-of-hospital acute care and transport to a place of definitive care, to patients experiencing medical emergencies. Over the last decades there have been several developments aimed to improve the location planning of EMS systems (see reviews by Brotcorne et al. (2003), Goldberg (2004), Li et al. (2011) and Farahani et al. (2012)). The particular characteristics of EMS systems have been recognized, such as having spatial and temporal demand location (demand occurs over a given geographic area and potentially changing over time, with some periods experiencing peak demands). EMS systems are also subject to random variations in demand and response times, which leads to congestion, i.e. ambulances can be already busy attending a call when they are required to attend another call.

EMS systems are referred to as 'option goods/services' (Felder and Brinkmann, 2002); i.e. patients do not know when they are going to require the service, but when they need it they would like to have it immediately. When a person experiences an emergency and a call is made to an EMS system, there is a natural expectation of equitable service, in addition to the expectation of a quick response. However, most EMS planning literature is concerned only with efficiency. As

pointed out by Felder and Brinkmann (2002) and Bertsimas et al. (2011), efficient solutions can be unacceptable when they are achieved at the expense of some players. The same authors mentioned that there is not a general agreement on how to measure equity.

The most widely used criteria when planning EMS systems are response time and coverage. The former measures how quickly the EMS system can respond on average to the emergency calls, from the time a call is received until the time at which the ambulance reaches the site of the event. The later measures the proportion of all calls that can be reached within a predefined threshold of time/distance. Minimizing response time and maximizing coverage can both be considered efficiency design criteria. The coverage maximization approach is used by the majority of researchers, practitioners and regulators (Li et al., 2011). It is reported by Iannoni et al. (2011) that in the US the most widely used response time standard is based on National Fire Protection Association (NFPA) guidelines, where 90% of all life threatening calls are expected to be attended within 8 minutes and 59 seconds.

The location of EMS facilities occurs within a spatially distributed population, hence the costumers will be located at different distances from ambulances and are likely to experience different effects (coverage, response time, among others), potentially facing inequalities. Differences in the service provided to costumers can also be the result of congestion. This means that an ambulance assigned to a costumer as their primary option can be busy when it is required, causing a different unit to be dispatched, instead of the preferred server. The congestion of the servers depends not only of the relative location among the costumers, but also on the dispatching rules. The location and allocation of servers to costumers also affects the relative utilization of the ambulances, therefore inequalities from the point of view of the servers can also appear.

We provide a methodological approach aimed at identifying good solutions for several equity and efficiency measures, based on a joint modeling approach for the location and dispatching decisions. This section is based on the previous one, in which the optimization framework for the joint problem was introduced. Although closely related, this work is different because we focus on the optimization analysis of several fairness functions, and the trade-offs among them as well as with common efficiency measures. Conversely, the previous section's objective was specifically

the development of the joint optimization model and a solution procedure for it. Due to the complexity of the systems under study our optimization framework uses an exact model but it is solved by approximate algorithms (Genetic Algorithms). We use the optimization framework to show that the joint approach provides better solutions than using a-priori dispatching rules such as the common assignment of the closest available server, when the optimization criteria is based on equity. It is also shown that there are trade-offs among the different performance measures, leaving it up to the decision makers to balance the different criteria. We conclude that the proposed joint optimization approach can be used to gain insight about some of the implicit trade-offs between common efficiency measures and the discussed equity criteria.

Although we approach the optimization of EMS systems by using several criteria, we are not using multi-criteria optimization. Each performance measure is optimized as a single objective, and once the best solution has been obtained from that point of view, we evaluate the other performance criteria for comparison purposes. Our main contribution is to show that for the equity related criteria under consideration the use of a myopic dispatching policy would not lead to the best solutions, contrary to the case when only efficiency is considered (see Section 2). In other words, the proposed methodology that uses a joint location/allocation modeling approach adds value when fairness considerations are in place. We also identify trade-offs among several efficiency and fairness criteria. Equity is still a critical and controversial issue when allocating public resources (Stone, 2002), and there is not a general agreement as to how to measure it. Thus, we provide an overview of possible ways to measure equity pointing out important considerations related to operating EMS systems when a particular form of equity is preferred. Although fairness issues related to facility location problems have indeed been addressed by previous literature (see for example the reviews by Marsh and Schilling (1995) and Ogryczak (2000)), we contribute to the literature by analyzing systems in which not only the location plays an important role, but also the dispatching policies are critical. To the best of our knowledge, previous literature addressing fairness issues together with the location of facilities in server to costumers environments (such as EMS systems) have assumed a-priori dispatching policies, focusing specifically on the location decision. We believe that the joint approach (location and dispatching decisions), is in itself a contribution that can serve as a starting

point to better analyze EMS systems. In fact we have identified that it is a better approach, when considering the optimization of several equity criteria, since it produces better solutions than the location only approach.

In Section 3.2 we provide a review of related literature. Section 3.3 adds to the mathematical model the objective functions that are used as optimization criteria. After that we present two case studies and our computational experiments in Section 3.4, as well as a review of the results and their implications. One of the case studies is based on data collected in 2007 by the Hanover Fire/EMS Department, which is located in Hanover, VA. Finally, in Section 3.5 we offer some conclusions and future research perspectives.

3.2 Problem presentation and related literature

The servers in a typical EMS system are: (i) spatially distributed in the region; (ii) share the system workload due to cooperation among them and (iii) have different operational characteristics, such as different preferential regions (Galvao and Morabito, 2008). Those characteristics have been progressively included in different approaches used for planning EMS systems. Congestion is also a typical phenomena related to EMS systems. According to Galvao et al. (2005) the volume of calls for service may keep ambulances busy from 20 to 30% of the time. Since the demand is spatially distributed and the calls for emergency occur randomly, the servers can be out of their base station performing a service when a new call is received. In a standard location problem where congestion is not an issue, once the locations have been decided it is possible to know the distance from every costumer to the open locations. Using that distance it is possible to estimate an expected service level. For EMS systems, due to congestion and cooperation, the performance of the servers from the point of view of the costumers depends not only of the distance from the server's base station, but also on availability. Because of cooperation (usually referred to as backup, in EMS systems), even though a costumer might have a preferred server, another server might serve that costumer if the preferred server is busy when it is required. The dispatching process, i.e. determining what is the preferred server for each costumer and the relative order in which back up servers will be used, becomes then an important part of the operation of the system.

Location models for EMS systems have been extensively developed in the literature (see reviews by Brotcorne et al. (2003), Goldberg (2004), Farahani et al. (2012) and Arabani and Farahani (2012)). The dominant approach has been Coverage maximization, or Expected Coverage maximization when congestion is taken into account. Minimization of Mean Response time has also been a common optimization approach. A key development in facility location applied to EMS was the Hypercube model by Larson (1974), combining queuing theory elements to better represent the relationship between the servers and their costumers. In Goldberg's review of models for deployment of EMS vehicles (Goldberg, 2004), it is mentioned that little work had been done on dispatching of ambulances. Similar opinion is shared by Lee (2011), mentioning that the contributions in ambulance dispatching are sparse. In turn, Galvao and Morabito (2008) and Iannoni et al. (2011) mention as an interesting extension of their work the use of different dispatch preference lists, instead of assuming that for a given set of locations the dispatching order is based on the closest dispatching rule. Nonetheless, the most widely used dispatching rule under a fixed preference scheme is to send the closest unit (Andersson and Varbrand, 2006).

The works by Jarvis (1981) and Katehakis and Levine (1986) studied the optimal allocation of distinguishable servers on Markovian queuing systems. They conclude that for a given location of the servers, so that the cost of assigning a server to a particular costumer is known (the cost in EMS planning is usually related to the amount of time that it takes for the EMS system to effectively respond to a call), under light traffic conditions (traffic is measured by the ratio between the mean arrival rate and the mean total service rate) the use of a myopic policy would always lead to an optimal solution, i.e. minimizing the long run average cost (response time). For heavy traffic the use of a myopic policy will deviate from the optimal, however the deviation is rather small (2 to 3%). Katehakis and Levine (1986) used 0.38 as an indicator of light traffic and 1.94 for the case of heavy traffic. Related work on dispatching by Bandara et al. (2012) was aimed to increase patients's survivability. Locations are considered to be fixed in their approach. McLay and Mayorga (2012) presented a model for dispatching, again with fixed locations, in which efficiency and equity are balanced by introducing several fairness constraints on typical efficiency maximization models.

A mathematical model that combines location and dispatching decisions for EMS vehicles

was introduced in Section 2. The dispatching decisions are modeled as a fixed preference scheme, meaning that there is a single list associated with each costumer that ranks the available servers (ambulances) in order of dispatch preference. This list does not change as a result of changes in the state of the system. However, the particular unit that will be dispatched to attend each call from a demand zone is not known in advance, since the assignment depends on the availability of the servers (system's state) when the call is received. Katehakis and Levine (1986) pointed out some results from Markov Decision Theory indicating that, when the number of states of the system as well as the number of actions available to perform in every state (allocation of the servers) are finite, it suffices to consider only deterministic policies; a deterministic policy is one which, whenever the system is in particular state, the set of available actions to perform is deterministic and depends only of the actual state (in our case, which servers are busy, and which are idle). In Section 2 it was shown that using the closest dispatching rule leads to optimal solutions when minimizing response time. Furthermore, although in some cases the maximization of expected coverage would benefit from a dispatching rule other than sending the closest vehicle, it is also shown that the trade-off on response time is not appealing. Thus, it is preferable to minimize response time, because the associated coverage of the minimum response time solution is only marginally smaller than the optimum coverage.

Although contributions on the topic of fairness and location problems do exist, they are sparse compared to the works related to efficiency. A recent paper by Bertsimas et al. (2011) pointed this out very clearly, mentioning that "a great deal of though has been invested in understanding and axiomatically characterizing what might constitute a 'fair' allocation of resources, but beyond qualitative economic analysis, there has been little work to quantitatively characterize the trade-offs inherent in employing these notions". There are multiple interpretations of the concept of fairness and they are subjective by nature. Several principles can lead to different forms of fairness. For example, allocate resources in proportion to an existing claim (Aristotle); allocate by maximizing the sum of individual utilities (Utilitarianism); give higher priority to players that are least well off (Rawls); or allocate based on Nash's Equilibrium. Stone (2002) formulates eight definitions of equity that depend on the perspective from different stakeholders. Her definitions are aligned with

three categories: (1) who receives the service, (2) what is being allocated and (3) how resources are allocated. The conclusion is the same as offered by others: there is no single principle that is universally accepted (Felder and Brinkmann, 2002; Bertsimas et al., 2011; Leclerc et al., 2012). References on location analysis focusing on equitable service to costumers can be found in Erkut (1993), Mulligan (1991) and Ogryczak (2000).

In the reviews by Marsh and Schilling (1995) and Eiselt et al. (1995) there is a list of 20 and 19 equity measures used in location theory, respectively. Range, variance, squared coefficient of variation, variance of logarithms, absolute and relative mean deviations and the Gini coefficient based on the Lorenz curve are among the measures identified. The work by Ogryczak (2000) is also a survey on inequality measures and equitable approaches to location problems. It is a common conclusion of the mentioned works that there is no consistency in the way those measures are applied to operation research models, and that little to no consensus exists on the best way to measure equity. Very often it seems that computational tractability becomes a given reason for selecting a particular equity criteria. Furthermore, Erkut (1993) noticed that it is rather a common flaw of all relative inequality measures that while moving away from the spatial units to be serviced one gets better values of the selected measure, as the relative distances become closer to one another. Therefore care must be exercised to avoid apparently equitable solutions that are indeed highly inefficient.

Several characteristics have been suggested as important when selecting a particular equity measure. In the review by Marsh and Schilling (1995) they collected seven characteristics from previous literature, although they mention that some might only be desirable while others might be required. We briefly mention those characteristics here since we consider that they are general enough to give an idea of what to look for in an equity measure: (1) Analytic tractability, which is associated with convenience for computational purposes; (2) Appropriateness, not from a mathematical point of view but from a managerial/administrative point of view, in the sense that it should serve to represent the stakeholder's point of view in a way that is rather clear to them; (3) Impartiality, which basically calls for a measure that should depend solely on the effects of a policy and not any other ranking coming from a political point of view; (4) Principle of transfers (also

know as Pigou-Dalton efficiency), that implies that the equity measure should increase/decrease as the difference in effect between any two individuals (groups) increases or decreases, respectively; (5) Scale invariance, which applies to measures that exhibit no change in the level of equity if the effects on all groups are multiplied by the same constant; (6) Pareto optimality, that implies that, as the solution improves according to the equity measure, none of the individuals or groups being affected should be worse off, or in other words, an improved solution should cause at least one individual to be better off; and finally (7) Normalization, which is related to the principle of scale invariance.

One of the most commonly used measures of equity is the variance of the individual outcomes, since a small variance means a low dispersion of the outcomes. The variance is used as a way to provide equitable service in the works by Berman (1990), Drezner and Drezner (2007), Maimon (1986) and Drezner and Drezner (2011). The range of outcomes has also been used in the works by Drezner and Drezner (2007) and Drezner et al. (1986). The Gini coefficient, which is commonly used in Economics to account for income inequalities has also been applied to location problems by Drezner (2004), Drezner et al. (2009) and Maimon (1988). Espejo et al. (2009) introduced the envy criteria, which measures the difference between pairs of costumers (thus is a measure of equity). Since people feel no dissatisfaction when they are better off than others, only negative effects are considered. Based on Espejo et al. (2009), the minimum p-envy location model was proposed by Chanta et al. (2011a), relaxing the strict and ordinal preference assumptions made by Espejo et al. (2009), and including backup servers. Most of the equity considerations in location problems have been devoted to reducing disparities among costumers, but they are not the only players in the system. Those who provide the service are also affected by the decisions made about the system. However, as pointed out by Leclerc et al. (2012), server's equity has been overlooked in the literature. Nevertheless some works can be found looking at server equity such as Berman et al. (2009), Marn (2011) and Kalcsics et al. (2010).

There have been some attempts at using multi-objective optimization to combine efficiency and equity measures for EMS systems. However, as pointed by Ogryczak (2000) the multi-criteria framework is quite difficult to implement, even for small size problems. Ogryczak (2009) developed

an approach in which a combination of equity and efficiency functions is allowed in a bi-criteria optimization framework. They identify several inequality measures that can be combined with typical efficiency measures while preserving the consistency between the two approaches. Their approach prevents having solutions that apparently equalize the service by making decisions that in fact would deny service to all costumers. Of course if no one gets service that would be an equitable solution in a formal sense, but completely inefficient. Hooker and Williams (2012) proposed a model in which a combination of equity and utilitarianism was attempted. They used a rawlsian approach to equity, looking to improve the conditions of the least well off. However, the objective function changes to the utilitarian approach whenever the rawlsian principle takes too many resources from others to improve only marginally those that are least well off. In the work by Chanta et al. (2011b) a bi-objective model is used combining the traditional maximization of expected coverage with three other objectives (one at a time) aimed at reducing disparities between urban and rural areas.

In this work we provide a planning methodology for EMS systems, based on a modeling approach to the joint location and dispatching problem, aimed at identifying good solutions for several equity and efficiency measures. The optimization framework for the joint problem was introduced in Section 2. Although closely related, this work is different because we now focus on the analysis of several functions representing various ideas of fairness, and the different trade-offs among them as well as with common efficiency measures. Our work is also different from previous literature in that we are modeling both decisions, location and dispatching, together. This allow us to analyze the effect of both decisions over the different optimization criteria, making it possible to identify performance measures that benefit from the use of dispatching policies other than always sending the closest server available.

3.3 Mathematical model

Now we introduce the alternative optimization criteria, starting with coverage. We use the concept of expected coverage as presented by Ingolfsson et al. (2008), which takes into account the congestion of the system and potentially the variability in responses times. The standard coverage assumes that the servers are always available, and therefore overestimates the real coverage.

The following equation details how to calculate the expected coverage:

$$Ex.Cov = \sum_{j=1}^{J} f_j \sum_{n=1}^{N} P_{j,j(n)}(1 - P_{j(n)}) \prod_{u=1}^{n-1} P_{j(u)} \tag{3.1}$$

Where $P_{j,i}$ is the probability that station i covers node j, P_i corresponds to the busy probability of the ambulance in station i and $j(n)$ refers to the n^{th} preferred station for demand node j. Note that $P_{j,i}$ can be used as a binary variable, indicating whether or not the coverage threshold is satisfied by the available servers, but it can also be used as the probability of that coverage being possible within the given threshold, accounting for variability in travel times. In this particular case it has been used as a binary variable.

From the point of view of the costumers (demand zones) we focus our attention on average individual response times (IRT), which are given by:

$$IRT_j = \sum_{n=1}^{N} \rho_{nj} t_{nj} \tag{3.2}$$

The aggregated equity measures accounting for disparities among individual response times are the variance (Eq. 3.3), the squared coefficient of variation (Eq. 3.4) and the Gini index (Eq. 3.5). We use variance since it is a commonly used dispersion measure. The squared coefficient of variation and the Gini index satisfy the scale independence principle, are population size independent and also comply with the principle of transfers (Pigou-Dalton condition). The coefficient of variation is also commonly used to relate the mean and dispersion of random variables, and being a number between 0 and 1 has a rather easy interpretation. Furthermore, the Gini index is very popular in Economics.

$$V(IRT) = \frac{1}{|\mathbf{J}|} \sum_{j \in \mathbf{J}} (IRT_j - \overline{IRT})^2 \tag{3.3}$$

$$CV^2(IRT) = \frac{V(IRT)}{\overline{IRT}^2} \tag{3.4}$$

$$Gini(IRT) = \frac{1}{\overline{IRT}|\mathbf{J}|^2} \sum_{i \in \mathbf{J}} \sum_{j \in \mathbf{J}} |IRT_i - IRT_j| \tag{3.5}$$

For the case of measuring equity among servers we focused on calculating the workload of each server, and the relative differences among all server's workload. Recall that each server can be busy or free (idle), and system's states are denoted by an ordered set of N one digit binary numbers taking the value of 1 if the server is busy and 0 if not ($B_j \equiv \{b_1, b_2, \ldots, b_N\}$). The individual workloads (IWK) can then be obtained according to Equation 3.6.

$$IWK_n = \sum_{j=1:b_n=1}^{2^N} P(B_j) \tag{3.6}$$

Having the individual server's workloads, we decided to calculate the variance and the squared coefficient of variation as the equity measures for servers. Expressions to calculate those two indicators are given by Equations 3.7 and 3.8.

$$V(IWK) = \frac{1}{N} \sum_{n=1}^{N} (IWK_n - \overline{IWK})^2 \tag{3.7}$$

$$CV^2(IWK) = \frac{V(IWK)}{\overline{IWK}^2} \tag{3.8}$$

3.4 Results discussion and analysis

3.4.1 Mid-size case study

We solved a mid-size problem, proposed as an instance of the Maximal Covering Location Probem (MCLP) (`http://www.lac.inpe.br/~lorena/correa/Q_MCLP_30.txt`) (Correa et al., 2007). We analyze several scenarios locating 3 ambulances. Because of the small number of ambulances we use an exact solution for the hypercube model. The server rates are obtained by selecting particular values for the overall utilization factor, $\rho = (\lambda/N \times \mu)$. In fact, ρ is varied between 0.1 and 0.9, with increases of 0.1. We use full backup, which means that any zone can be attended by any of the available servers.

A tuning procedure has been used for every one of the optimization criteria being considered. Each combination of possible values for the GA's parameters (combinations suggested by the experimental design) was tested on randomly selected scenarios for the mid-size problem, starting the GA each time with a different random seed (therefore, a different initial population). Once the tuning was performed, the GA was run 30 times for each scenario and for each optimization criteria. The experiments have been run on a PC executing Windows 7 -64 Bit, with an Intel®Core 2 Duo processor running at 2.13 GHz and 2 GB of RAM. All the programming was done in Java. The average running time of the GA (a run is made up of 100 evolutions of a population with 100 individuals) for the 3 servers scenarios was 20 seconds.

The mean values obtained for each criteria are presented in Table 3.1, for a subset of the scenarios that were run. For each value of ρ, the optimization framework was used to get the best possible heuristic solution under each of the seven criteria. Recall that due to the complex combinatorial and non-linear nature of the problem it is not computationally attractive to get an optimal solution using a standard commercial solver applied to the mathematical model. In Table 3.1 each row corresponds to the optimization of the system under a particular criteria. Columns 3 to 9 indicate the performance of the system according to a specific criteria. In each row there is a bold number that shows the result for the criteria being optimized. The remaining numbers in each row are the result of the other criteria. For example, when $\rho = 0.1$ the minimum mean response time is 0.5877; whereas if we maximize coverage the associated response time is 0.7303. These results show trade-offs between the different criteria. We will study those trade-offs in more detail later. Note that the results coming from minimizing the variance of the individual server's workloads are the same as those when minimizing the square coefficient of variation for the workloads. This is because the total workload of the system is a constant. Different dispatching policies simply redistribute the total workload between different servers, however the average workload will remain constant. Finally, recall that it has been assumed that the system does not allow queued calls, hence if a call arrives when all the servers are busy, that call is considered to be lost (attended by an external system). In a system with full backup $Probability(loss)$ depends only on the overall ratio between total demand rate and total service rate, regardless of the particular location and

dispatching decisions. In this case, for instance, $P(loss) = 0.0033$ when the overall utilization ratio $\rho = 0.1$ and $P(loss) = 0.0501$ for $\rho = 0.3$.

The most commonly used optimization criteria for EMS systems planning are Response Time and Coverage. As suggested by previous results (see Section 2), minimizing the system response time also results in good values of coverage (compared to the maximum coverage when it is the optimization criteria); however, it is unknown how minimizing system response time affects the other criteria of interest. In Table 3.2 we take the solution resulting from MRT minimization as a base scenario, and compare the performance of the system when other criteria are used. For each value of ρ, each row in this table compares one by one the values of the first row on Table 3.1 for the same ρ, with the remaining rows (optimization criteria other than MRT). For $\rho = 0.3$, for instance, the solution that maximizes expected coverage improves the coverage by 10.51% with respect to the base scenario (going from 0.8110 to .8963), but at the same time the response time also increases by 14.32% (from 0.6788 to .7761). The variance of individual response times increases more than 100% (from 1.63E-4 to 3.42E-4) and the variance of the server's workloads increases more than 200%. The solution that minimizes the Gini coefficient of the individual response times reduces the Gini coefficient by 47.83% with respect to the base case (from 0.2329 to 0.1215). However, the response time increases by 28.52% while the expected coverage is reduced by 5.3%.

The last column in Table 3.2 shows the benefit of using the joint location and dispatching approach for this data-set. To calculate the numbers in this column each scenario was solved following an approach in which the locations were the decision variables but the dispatching rule was always sending the closest available server. Hence, the last column on Table 3.2 shows how much improvement was attained for each criteria by using the joint approach instead. On one hand, note that the expected coverage can be improved less than 1% by using the integrated approach. On the other hand, for all the fairness criteria it is possible to get improvements ranging from 15% to more than 100%. Of course, as it has been mentioned before, there is a cost (trade-off) associated to each improvement in fairness. In particular, it is easy to see that equalizing response times or workloads will worse efficiency criteria such as response time and expected coverage.

In Figure 3.1a we plot the overall results of the optimization process for the mid-size case

Table 3.1: Performance indicators - Mid size case study - Mean value

ρ	Optimz. Criteria	Performance Indicators - Mean						
		MRT	ExCov	V.IRT	V.Wkl	Gini	SCV-RT	SCV-WK
0.1	MRT	**0.5877**	0.9125	1.35E-04	1.47E-03	0.2619	0.3515	0.1481
	ExCov	0.7303	**0.9675**	4.81E-04	4.68E-03	0.3608	0.8119	0.4711
	V.IRT	0.8111	0.8022	**9.95E-05**	3.31E-03	0.1347	0.1405	0.3333
	V. Wkl	0.9352	0.7577	5.60E-04	**1.47E-07**	0.2629	0.4687	0.0000
	Gini	0.9044	0.7712	1.10E-04	2.29E-03	**0.1181**	0.1211	0.2310
	SCV-RT	0.9174	0.7625	1.09E-04	2.51E-03	0.1180	**0.1160**	0.2527
	SCV-WK	0.9409	0.7520	5.46E-04	2.07E-07	0.2626	0.4449	**0.0000**
0.3	MRT	**0.6788**	0.8110	1.63E-04	2.23E-04	0.2329	0.3188	0.0028
	ExCov	0.7761	**0.8963**	3.42E-04	5.97E-03	0.3125	0.5112	0.0735
	V.IRT	0.8423	0.7712	**1.06E-04**	7.89E-03	0.1234	0.1346	0.0972
	V. Wkl	0.8952	0.7579	3.45E-04	**3.37E-07**	0.2230	0.3266	0.0000
	Gini	0.8724	0.7702	1.12E-04	8.25E-03	**0.1215**	0.1329	0.1016
	SCV-RT	0.8961	0.7581	1.13E-04	7.36E-03	0.1251	**0.1264**	0.0907
	SCV-WK	0.8952	0.7579	3.45E-04	3.37E-07	0.2230	0.3266	**0.0000**
0.5	MRT	**0.7223**	0.7575	1.72E-04	2.26E-04	0.2205	0.2971	0.0012
	ExCov	0.8467	**0.8226**	3.95E-04	5.09E-03	0.3034	0.4949	0.0272
	V.IRT	0.8445	0.7277	**1.12E-04**	8.13E-03	0.1292	0.1413	0.0434
	V. Wkl	0.9822	0.6939	4.47E-04	**3.12E-07**	0.2185	0.3204	0.0000
	Gini	0.8651	0.7257	1.19E-04	8.90E-03	**0.1295**	0.1430	0.0475
	SCV-RT	0.8895	0.7194	1.18E-04	6.96E-03	0.1304	**0.1342**	0.0372
	SCV-WK	0.9822	0.6939	4.47E-04	3.12E-07	0.2185	0.3204	**0.0000**
0.7	MRT	**0.7490**	0.6885	1.74E-04	2.62E-04	0.2108	0.2784	0.0009
	ExCov	0.8863	**0.7390**	4.15E-04	3.48E-03	0.2952	0.4756	0.0118
	V.IRT	0.8367	0.6654	**1.19E-04**	4.36E-03	0.1366	0.1526	0.0148
	V. Wkl	0.9294	0.6481	3.10E-04	**1.71E-07**	0.2080	0.2770	0.0000
	Gini	0.8673	0.6628	1.22E-04	4.29E-03	**0.1326**	0.1467	0.0146
	SCV-RT	0.8836	0.6609	1.24E-04	5.31E-03	0.1367	**0.1429**	0.0181
	SCV-WK	0.9294	0.6481	3.10E-04	1.71E-07	0.2080	0.2770	**0.0000**
0.9	MRT	**0.7676**	0.6192	1.76E-04	1.21E-04	0.2064	0.2696	0.0003
	ExCov	0.9174	**0.6587**	4.37E-04	2.24E-03	0.2907	0.4669	0.0058
	V.IRT	0.8426	0.6010	**1.24E-04**	2.90E-03	0.1407	0.1570	0.0075
	V. Wkl	0.9189	0.5859	2.87E-04	**1.17E-07**	0.2039	0.2704	0.0000
	Gini	0.8537	0.5993	1.26E-04	1.99E-03	**0.1382**	0.1557	0.0051
	SCV-RT	0.8853	0.5971	1.29E-04	3.33E-03	0.1408	**0.1481**	0.0086
	SCV-WK	0.9189	0.5859	2.87E-04	1.17E-07	0.2039	0.2704	**0.0000**

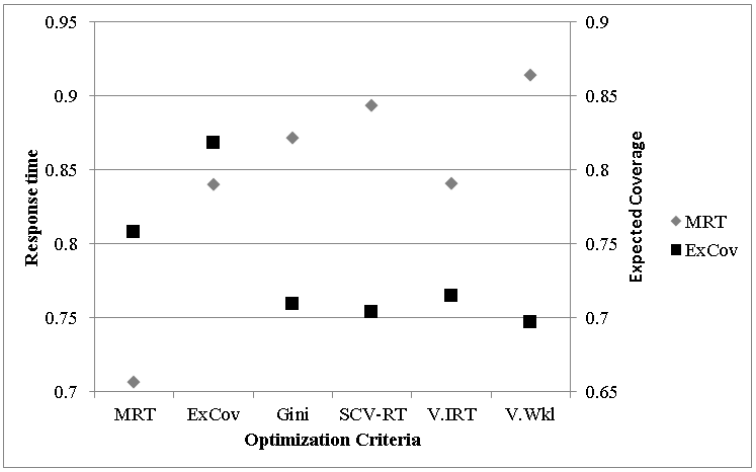
Table 3.2: Variations vs. minimizing MRT solution

ρ	Optimz. Criteria	Variations vs. MRT Solution (%)						Variation vs. Closest (%)
		MRT	ExCov	V.IRT	V.Wkl	Gini	SCV-RT	
0.1	ExCov	24.26	6.02	>200	>200	37.73	130.96	0.10
	V.IRT	38.01	-12.09	-26.27	125.11	-48.59	-60.02	-15.46
	V. Wkl	59.11	-16.97	>200	-99.99	0.38	33.35	>-200
	Gini	53.88	-15.49	-18.29	55.99	-54.91	-65.54	-71.40
	SCV-RT	56.09	-16.44	-19.22	70.69	-54.95	-67.00	-96.19
0.3	ExCov	14.32	10.51	109.59	>200	34.22	60.36	0
	V.IRT	24.07	-4.91	-35.17	>200	-47.02	-57.77	-29.89
	V. Wkl	31.87	-6.56	111.17	-99.85	-4.23	2.44	>-200
	Gini	28.52	-5.03	-31.13	>200	-47.83	-58.31	-52.57
	SCV-RT	32.00	-6.53	-31.00	>200	-46.27	-60.34	-66.71
0.5	ExCov	17.22	8.60	129.47	>200	37.59	66.58	0.70
	V.IRT	16.93	-3.93	-35.02	>200	-41.40	-52.44	-26.22
	V. Wkl	35.98	-8.39	159.70	-99.86	-0.92	7.85	>-200
	Gini	19.77	-4.19	-31.06	>200	-41.30	-51.88	-50.00
	SCV-RT	23.15	-5.03	-31.64	>200	-40.87	-54.82	-60.84
0.7	ExCov	18.34	7.32	139.27	>200	40.01	-70.84	0.30
	V.IRT	11.71	-3.35	-31.66	>200	-35.19	-45.19	-25.66
	V. Wkl	24.10	-5.86	78.86	-99.93	-1.34	-0.51	>-200
	Gini	15.79	-3.73	-29.51	>200	-37.13	-47.32	-43.58
	SCV-RT	17.98	-4.02	-28.68	>200	-35.16	-48.67	-44.03
0.9	ExCov	19.51	6.38	147.42	>200	40.84	73.22	0.60
	V.IRT	9.77	-2.94	-29.85	>200	-31.82	-41.77	-20.70
	V. Wkl	19.70	-5.38	62.38	-99.90	-1.22	0.32	>-200
	Gini	11.22	-3.22	-28.55	>200	-33.06	-42.22	-34.07
	SCV-RT	15.33	-3.57	-27.08	>200	-31.78	-45.07	-36.98

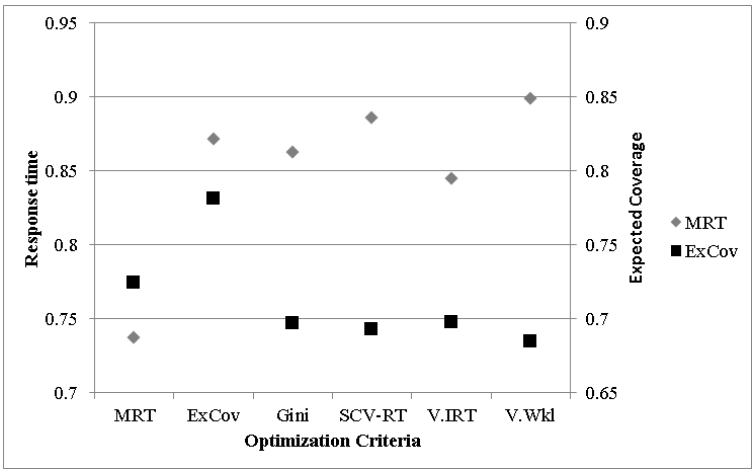
study, averaging over the different scenarios that were considered. There are two series: the response time (gray rhombus markers) and the expected coverage (black square markers). The graph shows the results on those two commonly used criteria when the optimization of the system is based on each of the different criteria (x-axis). The left vertical axis is associated with the response time, while the right vertical axis is used to plot the expected coverage. All the fairness criteria always cause a sacrifice of the efficiency criteria, response time and expected coverage. The MRT shows a minimum value when MRT is the criteria being optimized, just as expected. Note that the response time obtained under any other optimization criteria is always bigger. If we look at the expected coverage values it is possible to see that the use of the other criteria will also cause a sacrifice in coverage. The sacrifice in coverage when minimizing response time looks smaller (it is in fact, see Table 3.2) than the sacrifice in response time if the coverage is maximized. The expected coverage obtained under the optimization of the fairness criteria exhibits small variations (The ExCov values are similar for the fairness criteria). Figure 3.1b shows the results for a particular scenario (in this case $\rho = 0.6$).

As it has been mentioned, from the point of view of the final users it is important to get a quick response when an ambulance is required. At the same time it is also important to have a system that exhibits fair treatment to the users as well as to the servers. In Figure 3.2 we show the overall results of system response time along with several fairness criteria. The black markers should be read on the secondary (right) vertical axis, while the gray markers correspond to response time and are associated to the primary (left) vertical axis. The right axis is used to represent several criteria (Gini coefficient and SCV for individual response times).

Note that the solutions obtained when using the Gini coefficient, the square coefficient of variation or the variance of the individual response times, which are all fairness criteria from the point of view of the final users, are similar regarding several performance measures. Take for example the values for expected coverage, depicted in Figure 3.1b by the black squared markers. If a line is drawn connecting the three squared markers corresponding to the expected coverage when Gini, SCV-RT and V.IRT were used, that would be almost a flat line. That suggests, as mentioned, that the solutions obtained when using those three criteria (Gini coefficient, square



(a) Overall trade-offs



(b) Trade-offs for $\rho = 0.6$

Figure 3.1: Trade-offs MRT vs. Exp. Coverage mid-size case study

coefficient of variation or the variance of the individual response times) have a similar expected coverage. A similar observation can be done for the values of the Gini coefficient (black triangles markers) and the SCV-Wk (black dots) in Figure 3.2. For the case of the MRT (gray rhombus) a bigger variation is observed. Also note that regardless of the optimization criteria being used, the squared coefficient of variation of individual server's workloads is almost always below 0.1. From Table 3.2 we see that the use of the variance of individual server's workloads as optimization criteria can reduce this variance by almost 100%, with respect to the variance server's workloads associated to the solution minimizing response time. Although that seems like a big and important improvement, the fact is that the system is going from a SCV of 0.02 (which is already small) to an even smaller value of 4.03E-6 (practically equal to 0). Most managers may agree that a SCV of 0.02 is already small (which suggest that the workload is being almost evenly distributed among the servers) and no extra efforts to reduce it are required.

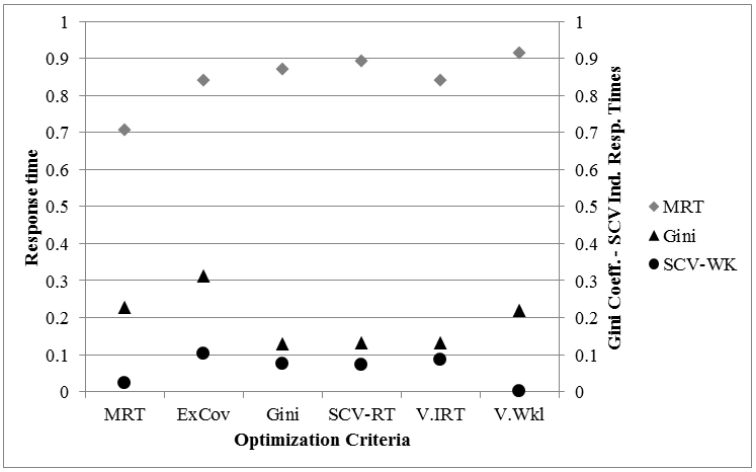


Figure 3.2: Overall trade-offs MRT vs. Fairness criteria

In Figure 3.3 we show aggregate trade-off results based on the numbers already presented in Table 3.2. The variations (percent change) shown in this graph are with respect to the solution that minimizes response time. In this figure positive variations are used to represent improvements on the different criteria. When the expected coverage maximization is applied we can see that indeed the coverage is improved, however all the other criteria deteriorate. We can also see that although the coverage is indeed higher, the change (increase) of this criteria is smaller than the

variation of any of the other criteria. In other words, increasing the coverage causes a sacrifice in all the other criteria, sacrifice that is proportionally higher than the increase in coverage.

Minimizing the Gini coefficient of the individual response times improves (reduces) not only this criteria, but also the square coefficient of variation and the variance of the individual response times. All these criteria are intended to create solutions that are fair from the point of view of the final user, by equalizing the response times. However, the coverage is reduced and the response times increases (both undesirable effects), although those two variations are smaller than the improvement in final user's fairness. The biggest sacrifice is observed in server's workloads balance. However, it is worth mentioning that the workload balance among the servers is similar under the different optimization criteria (SCV-WK almost always below 0.1). Even then, going from 0.0028 to 0.1016 (the case of $\rho = 0.3$, for instance) causes a huge relative increment, although both values are small enough in the sense of a coefficient of variation.

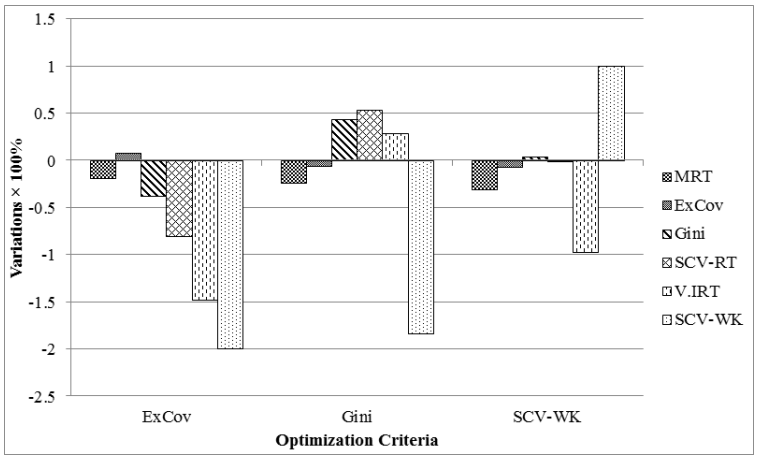


Figure 3.3: Overall trade-offs MRT vs. Fairness criteria

Figure 3.4 depicts the spatial location of the demand zones. The index assigned to every location not only identifies it but also gives an idea of its ranking according to the proportion of demand generated by each zone. The lower the index, the bigger the proportion of demand. Locations 1, 2 and 3 combined account for almost 35% of the total demand. The first 10 locations represent more than 65% of the total demand. In Table 3.3 we show the changes in the location decisions for different scenarios. Recall that every demand zone was a candidate location for an ambulance. All the fairness criteria are shown as a single category, because in all the cases the

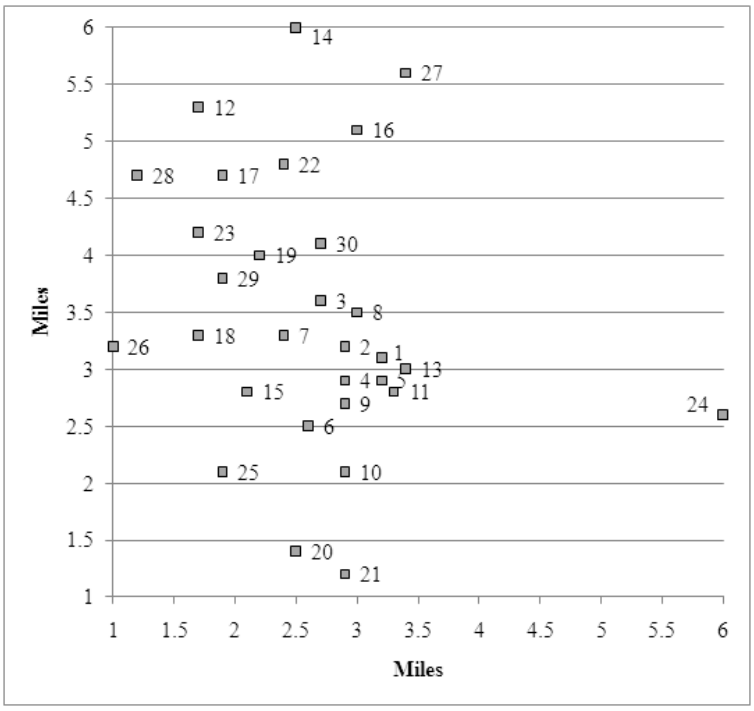


Figure 3.4: Spatial location of demand zones - Mid-size case study

locations selected when the optimization was based on those fairness criteria were the same. The differences in the performance measures are then explained by different dispatching rules. This observation is important because it shows that using the closest dispatching rule would not be enough to obtain good solutions for the equity criteria.

Table 3.3: Location decisions mid-size case study

ρ	MRT	ExCov	Fairness
0.1	2-5-17	2-3-17	1-2-3
0.3	1-3-4	3-7-15	1-2-3
0.5	2-3-4	3-7-9	1-2-3
0.7 & 0.9	2-3-4	2-3-4	1-2-3

3.4.2 Hanover County case study

We have also applied our modeling approach to a case study that uses real data from the Hanover Fire/EMS department, which is located in Hanover, VA. The county has 474 square miles

and a population nearing 100,000 individuals. The county has been divided for planning purposes in 122 demand zones. There are 16 candidate locations for 5 ambulances. The total demand rate has been estimated in 1.2 calls/hour. The average service time per call has been estimated to be 74 minutes and it is assumed to be independent of the demand zone being served. Additional details about this case study can be found in Chanta et al. (2011b). For this case study we use partial backup, allowing every demand zone to be served only by 3 out of the 5 available servers. There are several reasons to proceed this way. The first two have been suggested by Geroliminis et al. (2009): (i) from a practical perspective, allowing servers that are ranked as 4^{th} and up for a particular demand zone is not desirable, because the overall efficiency of the system would likely decrease; (ii) the calculation of transition rates for the embedded hypercube model becomes very tedious; (iii) the partial backup better represents the real system, because even when there is a server available, if it is too far away from the costumer (likely in rural areas), typically a third party will be contacted to attend that particular call. In addition to the real case study we consider two variations, increasing the demand by a factor of 1.5 and 2 respectively (which increases the overall utilization, ρ). For this case study we are also using the exact procedure to solve the hypercube model (with 5 servers the number of states is 32). The average running time of the GA for each scenario of this case study was 280 seconds.

Table 3.4 shows the mean performance indicators for the original Hanover data-set as well as for the hypothetical scenarios with increased demand. Note that for server's workloads we are showing only the squared coefficient of variation. This is because, as it was mentioned before, the variance of the workloads would lead to the same results (small differences can still be observed when running the algorithm due to rounding). Table 3.5 presents the variation (percent change) in the performance of the system when criteria other than MRT are used. The percent changes are presented only for the base case ($\rho = 0.2$). The same variations are depicted in Figure 3.5 for a subset of criteria. Positive values in Figure 3.5 represent improvements for the criteria. It is observed that improving the expected coverage sacrifices all the other criteria, with only a marginal increase in coverage. Improving equity on individual response times (by reducing the Gini coefficient) negatively affects response time and coverage, with response time increasing more than

70%. The server's workloads are also affected, increasing their relative differences, given by the square coefficient of variation. Finally, trying to equalize the server's workloads, their coefficient of variation goes from 0.37 to 0.02 but it has a major negative effect on other performance measures. For instance, response times increases over 200% and coverage decreases more than 60%.

Table 3.4: Performance indicators - Hanover case study

ρ	Optimz. Criteria	Performance Indicators - Mean						
		MRT	ExCov	V.IRT	V. Wkl	Gini	SQV-RT	SQV-Wk
0.2	MRT	**4.5270**	0.8768	0.0031	5.36E-03	0.591	2.214	0.1415
	ExCov	5.2652	**0.8928**	0.0110	6.49E-03	0.674	5.894	0.1722
	V.IRT	4.9432	0.8580	**0.0027**	5.93E-03	0.538	1.657	0.1556
	V. Wkl	18.5242	0.2699	0.1643	**8.83E-06**	0.747	6.930	0.0002
	Gini	8.3998	0.6895	0.0063	8.41E-03	**0.503**	1.301	0.2217
	SQV-RT	7.6947	0.7316	0.0052	6.11E-03	0.512	**1.173**	0.1606
	SQV-Wk	18.2834	0.2554	0.1430	9.55E-06	0.741	6.152	**0.0002**
0.3	MRT	**4.8236**	0.8305	0.0038	8.80E-03	0.600	2.417	0.1121
	ExCov	5.8109	**0.8453**	0.0106	5.89E-03	0.641	4.676	0.0748
	V.IRT	5.4450	0.8104	**0.0035**	1.15E-02	0.549	1.767	0.1480
	V. Wkl	18.6544	0.2326	0.1579	**1.32E-05**	0.736	6.371	0.0002
	Gini	8.8163	0.6388	0.0079	1.40E-02	**0.527**	1.464	0.1782
	SQV-RT	8.3356	0.6783	0.0064	9.14E-03	0.525	**1.286**	0.1151
	SQV-Wk	18.7333	0.2179	0.1690	7.09E-06	0.743	6.916	**0.0001**
0.4	MRT	**4.9672**	0.7800	0.0042	1.25E-02	0.605	2.523	0.0995
	ExCov	5.9712	**0.8118**	0.0102	4.79E-03	0.631	4.260	0.0376
	V.IRT	5.6846	0.7300	**0.0037**	2.52E-02	0.542	1.687	0.2033
	V. Wkl	18.0653	0.2297	0.1558	**1.08E-05**	0.743	6.957	0.0001
	Gini	8.8731	0.6054	0.0088	1.82E-02	**0.535**	1.586	0.1453
	SQV-RT	8.8060	0.6132	0.0079	1.31E-02	0.538	**1.397**	0.1028
	SQV-Wk	17.4731	0.2370	0.1329	1.60E-05	0.733	6.402	**0.0001**

The last column in Table 3.5 compares the quality of the solutions coming from the joint approach versus approaching the problem with only the locations as decision variables, in combination with the closest rule for the dispatching of the ambulances. Note that the improvements in criteria such as the variance of the individual response times or the Gini coefficient are smaller than those observed for the mid-size case study. This could be caused by the fact that the Hanover case study has a reduced number of candidate locations, while the mid-size case study had every demand point as a candidate for locating an ambulance. Nevertheless, the joint approach was still

able to generate better solutions, although the relative improvement was in fact smaller. Furthermore, contrary to the results for the mid-size case study, for the Hanover data set the locations selected when using different fairness criteria are different. For instance, to minimize the variance of the individual response times the locations selected are [1-6-7-8-10], while for the minimization of the Gini coefficient the locations will be [1-6-7-10-13]. 4 out of 5 locations remain the same between those two criteria. Recall that for the previous case study all the fairness criteria selected the locations with the highest proportion of demand. For the Hanover case only locations 6, 7 and 13 (selected by the two fairness criteria just mentioned) are near to zones with a high demand relative to other zones. For minimizing the variance of the server's workloads the locations will be [1-2-3-4-5], a set of locations that only shares one element with the set of locations selected for the fairness criteria based on final costumers.

Table 3.5: Variations vs. minimizing MRT solution

Optim.	Variation (%)							Variation vs.
Criteria	MRT	ExCov	V.IRT	V. Wkl	Gini	SQV-RT	SQV-Wk	Closest (%)
ExCov	16.31	1.83	>200	20.93	14.13	166.18	21.70	-1.80
V.IRT	9.19	-2.14	-10.85	10.54	-8.88	-25.18	9.96	-2.47
V. Wkl	>200	-69.22	>200	-99.84	26.49	>200	-99.84	-99.22
Gini	85.55	-21.36	107.80	56.88	-14.79	-41.25	56.65	-7.24
SQV-RT	69.97	-16.56	69.58	13.93	-13.38	-47.03	13.50	-30.95
SQV-Wk	>200	-70.87	>200	-99.82	25.56	177.81	-99.82	-99.12

For the Hanover case study there are no queued calls, therefore congestion of the system could potentially lead to 'lost' calls. There are 5 ambulances available but, as it was mentioned, partial backup is being used. In particular only 3 ambulances are allowed to serve every demand zone (the priority order for those ambulances is given as a solution by the optimization framework). Given these conditions the $Probabilty(loss)$ is not only associated with all servers being busy. In fact if the three preferred servers of a demand zone are busy when a emergency call originates from that demand zone, then the call will be lost (in reality an external EMS system is contacted to attend the call). $P(loss)$ then varies across demand zones, although it will have the same value for demand zones sharing preferred servers, regardless of the order of preference of those servers for each zone. Table 3.6 shows the values of $P(loss)$ for the Hanover case study ($\rho = 0.2$). It is possible

to see that under all the scenarios (different optimization criteria), the value of $P(loss)$ is less than 0.035. The overall $P(loss)$ for the system, which is the mean value across demand zones is almost in all cases below 0.015. These low values for $P(loss)$ are of course desirable. Given the nature of EMS systems, which are designed to respond to emergencies, it makes sense that they are provided with enough capacity to attend to those emergencies in a timely manner. In fact, as it has been mentioned before, regulators impose service level constraints so that there is a maximum expected response time threshold for emergency calls, especially those that are life threatening. This leads to low values of $P(loss)$ as the ones observed for the real data-set. Even if a call has to be attended by an external EMS system, that external service should also provide a quick response.

Table 3.6: $P(loss)$ Hanover Case Study - $\rho = 0.2$

	Min	Max	Mean	Mode
MRT	0.0054	0.0255	0.0146	0.0054
ExCov	0.0108	0.0161	0.0131	0.0115
V.IRT	0.0025	0.0343	0.0206	0.0343
Gini	0.0067	0.0193	0.0137	0.0185
SQV-RT	0.0058	0.0220	0.0123	0.0220
SQV-Wk	0.0086	0.0155	0.0133	0.0132

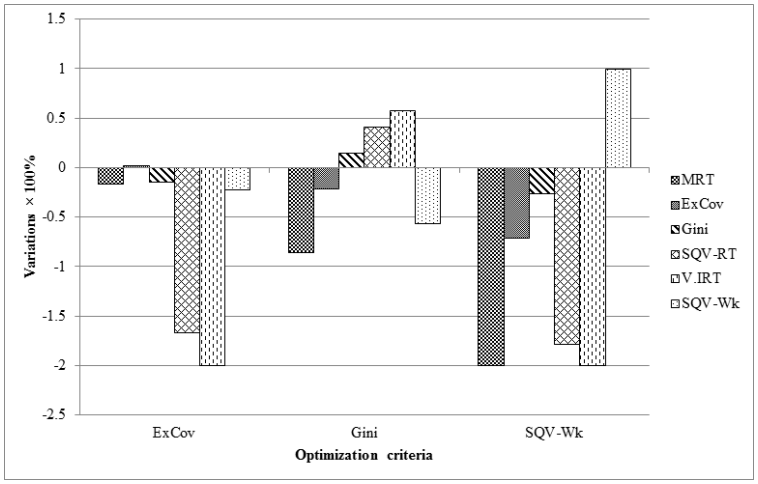


Figure 3.5: Overall trade-offs MRT vs. Fairness criteria - Hanover case study

3.5 Conclusions

According to our results, an important finding is that the joint location/allocation approach is useful to get better solutions when the optimization of the system is based on the fairness criteria under consideration. Recall that when the optimization was based on response time or expected coverage, the joint approach did not add much value, and the conclusion was then that using a myopic dispatching policy combined with the selection of good locations provided good enough solutions (about 1% from the optimal). Conversely, for the fairness criteria we have found that in fact it is necessary to model both decisions concurrently in order to get to better solutions for these criteria (improvements ranged from about 2.5% - 100% for the real Hanover data set).

As expected, focusing on different criteria causes the system to behave in different ways. The use of the different fairness criteria degraded the performance of the commonly used efficiency criteria, response time and coverage. Wether or not the sacrifice is acceptable will depend on the particular interest of the system planners. It is also possible to see that the trade-offs are different when the system is subject to a different level of congestion, as given by the parameter ρ. We have also seen that the availability of candidate facilities when considering the location of the servers can impact the results. Making it possible to evaluate EMS systems from different perspectives, adding fairness considerations to the planning process, should help decision makers and regulators to better conciliate the expectations from different stake holders. If a particular criteria or a set of them does not exhibit very good values at least it can be explained in terms of gains from some other perspective. Of course, the dominant perspective when planning a public system is a matter of political discussion and agreement between the different parties involved.

From a fairness perspective, we can see that in general the biggest trade-off occurs between equally serving the demand zones (by offering service times that do not vary that much between them) and equally distributing the total workload between the available servers. This makes sense because we are talking about different stake holders. Our approach allows for identifying the trade-offs and providing the decision makers with different perspectives for planning the system. It might be the case that additional resources are required to achieve a desired trade-off. Furthermore, we have to keep in mind that the trade-offs that have been illustrated are the result of optimizing

only one criteria at a time. Intermediate trade-offs might be also achievable and preferable in some cases. For example, instead of reducing the Gini coefficient by about 40% by sacrificing the response time by about 15%, the planners can opt for improving the Gini coefficient by about 20% only, which also means that the sacrifice in response time would likely be less than 15%. We have used three different criteria to account for variability in individual response times: the variance, the Gini coefficient and the square coefficient of variation. Although those are different criteria, as expressed by their mathematical expressions, in all cases their minimization is aimed to produce smaller differences among individual response times. While they indeed produced different solutions, the performance of those solutions is similar.

The modeling and solution approach that we introduced here could potentially be used to identify those desired trade-offs, for example by stopping the GA once you have reached a particular value of the desired criteria. The joint approach can also serve to evaluate the potential future improvements of the EMS system if more resources are added, such as new servers, or new facilities for the location of the available servers. Using some of the performance measures as constraints can be a way to control the trade-offs. It is also possible to approach the optimization problem from a different perspective, treating it as a multi-criteria optimization problem. The model that we have presented can still be used but the solution procedure needs to change, so that it is possible to generate the efficient frontier between the different criteria under optimization. Our main contribution was to identify trade-offs among several efficiency and fairness criteria, showing that for the equity related criteria under consideration the use of a myopic dispatching policy would not lead to the best solutions. To the best of our knowledge, previous literature addressing fairness issues together with the location of facilities in server to costumers environments (such as EMS systems), have assumed a-priori dispatching policies, focusing specifically on the location decision. Therefore, we believe that the joint approach (location and dispatching decisions) is in itself a contribution that can serve as a starting point for future developments, such as the multi-criteria optimization approach.

We are aware of the limitations of our approach in terms of applying the joint model and its solution procedure to bigger case studies. That is because the exact solution of the hypercube

model will likely require extensive computation time (recall that the exact solution to the hypercube model requires solving a linear system of equations that grows exponentially in size with respect to the number of servers available in the system). We are currently working on using available approximation procedures that have been suggested in the literature, and that could be embedded in the meta-heuristic optimization framework proposed, reducing the computational burden and allowing the solution of bigger instances. Nonetheless, the Hanover case study that we have solved is a real case study with more than a 120 demand zones accounting for about 100,000 individuals, and the computation time has been reasonable. Thus, even in its current development stage we believe that the proposed approach has potential applicability to a wide variety of EMS systems.

Chapter 4

Scalability of the model and solution approach

4.1 Introduction

EMS systems operate under the pressure of knowing that human lives might be directly at stake. Although not all of the emergencies attended daily by EMS systems pose serious risks to patient's lives, in the public eye there is still a natural expectation of efficient response to all calls. From a practical perspective, EMS systems typically operate by assigning priorities to the incoming calls. The higher the perceived risk for the life of a patient, the higher the priority of the call. However, it is not always possible to make the right assessment of a medical situation by phone, hence policy is designed to "play it safe" by responding to most calls as quickly as possible (McCallion, 2012).

There is abundant literature on the topic of efficient planing of EMS systems, in particular works seeking to maximize coverage (the percentage of calls attended to within a given time/distance threshold) or minimize average system-wide response time. The earliest references on the topic can be traced back to the 1970s. The planning of these systems has been approached by using different modeling techniques. There are two streams of work clearly recognized: descriptive models, which deal with characterizing the performance of the system given its characteristics,

such as queuing and simulation models; and normative models, which attempt to find the best set of decisions from a particular perspective (objective function). However, the use of exacts models is usually hindered by the fact that the resulting model's size makes them impractical, due to the amount of computation time required to obtain solutions. While it is sometimes possible to write a mathematical model in a very compact way this does not necessarily mean that it can be solved by off-the-shelf standard commercial solvers.

Objectives different from efficiency have been considered but the literature available is very sparse compared to efficiency-based works. Furthermore, most of the works including equity considerations approach the problem from the point of view of the final users. Although the patients are a very important player in the system, they are not the only stake-holders. Operating conditions that are perceived as fair by the medical personnel are also desirable. We approach the planning of EMS systems including fairness considerations and we use heuristic/approximated solution techniques allowing for the analysis of large scale real systems. Previous works combining location and dispatching decisions were only capable of analyzing mid-size systems having about five ambulances available and only one ambulance allowed per location. We use the basic ideas from the meta/heuristic Tabu Search (TS) to guide the process of identifying good solutions, as well as an approximation to the queuing behavior of the system to account for its dynamic performance (ambulance busy probabilities and dispatching probabilities). Our approach is tested on real case studies having up to 18 ambulances and 180 demand zones, running in reasonable computation time. Our TS implementation uses random initialization as well as a dynamic/reactive size of the tabu list, characteristics that are different from a classical version of the meta heuristic.

Two different real systems have been analyzed. One of them corresponds to the city of Edmonton, Canada and the other one is the city of Charlotte, in North Carolina, USA. In both cases the number of demand zones exceeds 150 and the number of servers (ambulances) used to provide the service is at least 12. The former case allows for several ambulances to be located at a single station (each candidate station has a maximum capacity that is imposed as a constraint when planing the system). The latter case allows for almost all demand zones to be candidates for locating an ambulance (only those zones located in the boundaries of the geographical region are

excluded). By using the heuristic/approximation solution approach this work addresses scalability issues mentioned in previous research.

The rest of the chapter is organized as follows: in Section 4.2 we provide a review of related literature. Section 4.3 presents the mathematical model and discusses the approximation scheme used to deal with the dynamic behavior of the system. Section 4.4 introduces the solution strategy based on Tabu Search. Then we present the case studies and computational results in Section 4.5. Finally, in Section 4.6 we offer our conclusions and future research perspectives.

4.2 Problem presentation and related literature

Over the last decades several attempts have been reported aimed to improve the location planning of EMS systems. The evolution of such attempts can be traced by consulting reviews such as Brotcorne et al. (2003); Goldberg (2004); Li et al. (2011); Farahani et al. (2012). EMS are systems with spatial and temporal demand location. Demand occurs over a given geographic area, at different rates from different zones, potentially changing over time, with some periods experiencing peak demands. Since these systems are subject to random variations in demand and response times it leads to congestion. It is very important to be able to describe the system by assigning a probability of a particular server being busy when the system is in steady state. It allows for the calculation of different performance measures so that several operating policies can be compared against each other.

Characterization of typical EMS servers includes: (i) they are spatially distributed over a given region; (ii) share the system workload following specific cooperation rules and (iii) have different operational characteristics, such as different preferential regions (Galvao and Morabito, 2008). Congestion is also a typical phenomena related to EMS systems. The volume of calls for service may keep ambulances busy from 20 to 30% of the time (Galvao et al., 2005). Since the demand is spatially distributed and calls for emergencies occur randomly, the servers can be out of their base station and in service when a new call is received. Therefore, due to congestion and cooperation, the performance of the servers from the point of view of the costumers depends not only on the distance from the server's base station (static location), but also on availability. Cooperation

among servers (usually referred to as backup in EMS systems), allows for the possibility that even though a costumer might have a preferred server, another server might attend that costumer if the preferred server is busy when it is required. The dispatching process, i.e. determining which is the preferred server for each costumer and the relative order in which back up servers will be used, becomes then an important part of the operation of the system. The planning of EMS systems over the last two decades has been heavily dominated by the coverage maximization approach, which is used by the majority of researchers, practitioners and regulators (Li et al., 2011). It has been reported by Iannoni et al. (2011) that in the US the most widely used response time standard is based on National Fire Protection Association (NFPA) and it is 8 min and 59 seconds; 90% of all life threatening calls are expected to be attended within this time threshold. Recall that the concept of coverage refers to the availability of at least one server within the given time/distance threshold. The available server(s) are then considered satisfactory but it still leaves an open question regarding to which one is the best possible alternative (Farahani et al., 2012).

Particularly important to our work is the Hypercube model proposed by Larson (1974). It was the first work that used queueing theory elements in facility location models applied to EMS systems. Larson (1975) later developed an approximation for the hypercube model due to the fact that exact calculations were prohibitive. There are a variety of applications and extension of the hypercube model to EMS systems (Brandeau and Chiu, 1989; Mendonca and Morabito, 2001; Atkinson et al., 2008; Iannoni and Morabito, 2007; Iannoni et al., 2008; Galvao and Morabito, 2008; Geroliminis et al., 2009; Toro-Díaz et al., 2013), among others. It is well documented that the hypercube model is a descriptive tool that allows the analysis of scenarios, but it was not designed as an optimization model. However, it has been embedded into optimization frameworks. Batta et al. (1989) combined MEXCLP with the hypercube into an iterative, local search algorithm. Aytug and Saydam (2002) replaced the local search by a genetic algorithm. Iannoni and Morabito (2007), Iannoni et al. (2008), Geroliminis et al. (2011) and Toro-Díaz et al. (2013) have embedded the hypercube model into genetic algorithms. In this paper we use the hypercube model and in particular an approximation to its solution. The earliest approximation was provided by Larson (1975), who developed the hypercube model. Jarvis (1985) generalized the approximation by al-

lowing general service time distributions as opposed to requiring exponential behavior. In addition, his work also considered the possibility of having different response times depending upon the unit being dispatched and the zone being attended. Goldberg and Paz (1991) extended the work by Jarvis (1985) reducing the computational effort required to get a solution. We use the more recent generalization proposed by Budge et al. (2009) which allows analyzing for analysis of systems where there might be several servers located at a single station.

The demand for EMS, in addition to having a random behavior, belongs to a category in which the users of the system claim immediate satisfaction of their needs. This is what is called 'option goods/services' (Felder and Brinkmann, 2002). Furthermore, in addition to quick response there is also a natural expectation of a fair service, meaning that all the people living in a particular area served by an EMS system should have the same chances of being attended promptly. Equalization of service is not a concern of efficiency based models. Furthermore, even if there is an agreement for having fair policies in place, the particular idea of fairness in use can be one of many different alternatives. Efficient solutions have a natural appeal in public-settings because public resources are expected to be used efficiently. However, as pointed out by Felder and Brinkmann (2002) and Bertsimas et al. (2011), efficient solutions can be unacceptable when they are achieved at the expense of some players.

In spite of the abundant literature on location of EMS facilities, the references addressing dispatching decisions are rather sparse (Goldberg, 2004; Lee, 2011). The dispatching problem was initially studied by Carter et al. (1972). The most widely used dispatching rule under a fixed preference scheme (each zone ranks the servers in order of preference, which does not change over time) is to send the closest unit (Andersson and Varbrand, 2006). Researchers such as Galvao and Morabito (2008); Iannoni et al. (2011) mention that an interesting extension of their work would be the use of different dispatch preference lists, instead of assuming that for a given set of locations the dispatching order is based on the closest dispatching rule. Jarvis (1981); Katehakis and Levine (1986) studied the optimal allocation of distinguishable servers on Markovian queuing systems. The two works pointed out some results from Markov Decision Theory indicating that, when the number of states of the system as well as the number of actions available to perform in every state (allocation

of the servers) are finite, it suffices to consider only deterministic policies, such as a fixed preference scheme. Benveniste (1985) proposed non-linear programming techniques for the solution of the combined location/districting problem assuming a continuous space for the location decisions. A formulation of the joint location/dispatching problem and a solution procedure was presented in (Toro-Díaz et al., 2013). It is shown that using the closest dispatching rule leads to optimal solutions when minimizing response time. The same authors mentioned that although in some cases the maximization of expected coverage would benefit from a dispatching rule other than sending the closest vehicle, the trade-off on response time is not acceptable. Related work on dispatching was aimed to increase patients's survivability (Bandara et al., 2012); locations are considered to be fixed in their approach. In McLay and Mayorga (2012) the authors presented a model for dispatching, again with fixed locations, in which efficiency and equity are balanced by introducing several fairness constraints on typical efficiency oriented models. Baptista and Oliveira (2012) presented a case study for Lisbon EMS systems management in which several dispatching policies are proposed and compared. The model is not normative however, and therefore it is not possible to derive optimal dispatching policies using their approach.

Contributions on the topic of fairness related to location problems are also sparse (Bertsimas et al., 2011). There are multiple interpretations of the concept of fairness and they are subjective by nature. For instance, allocate resources in proportion to an existing claim; allocate by maximizing the sum of individual utilities; give higher priority to those who are least well off; or allocate based on Nash's Equilibrium. Stone (2002) formulates eight definitions of equity that depend on the perspective from different stakeholders. Her definitions are aligned with three categories: (1) who receives the service, (2) what is being allocated and (3) how resources are allocated. The general agreement is that there is no single principle that is universally accepted (Felder and Brinkmann, 2002; Bertsimas et al., 2011; Leclerc et al., 2012). References on location analysis focusing on equitable service to costumers can be found in Erkut (1993); Mulligan (1991); Ogryczak (2000). The reviews by Marsh and Schilling (1995) and Eiselt et al. (1995) list several equity measures used in location theory. Range, variance, squared coefficient of variation, variance of logarithms, absolute and relative mean deviations and the Gini coefficient based on the Lorenz curve are among the measures

identified. The work by Ogryczak (2000) is also a survey on inequality measures and equitable approaches to location problems. It is rather a common flaw of all relative inequality measures that by moving away from the spatial units to be serviced one gets better values of equality, as the relative distances become closer to one another (Erkut, 1993). Our goal is to present a modeling and solution approach to the planing of EMS systems in which location and dispatching decisions are made simultaneously to balance efficiency and fairness. We overcome deficiencies observed in previous literature related to the size of the systems being analyzed. We also perform an analysis of the trade-offs between performance criteria.

4.3 Modeling approach

Our model extends the results introduced in Chapter 2, in particular addressing scalability issues introducing the possibility of having more than one server per candidate location. The dynamic behavior of the system is modeled by using an approximation procedure to solve the underlying hypercube model. Location and dispatching decisions are integrated into a single framework instead of assuming the use of an a priori dispatching policy, particularly based on the closest distance.

4.3.1 Assumptions

It is assumed that the system provides service to a certain geographical region $\mathbf{J}$ that is partitioned into service regions. A given number of servers may be located at points $i \in \mathbf{I} \subset \mathbf{J}$. Demands occur solely at the center of each service region by time homogeneous Poisson requests for service and are attended at service rates exhibiting a general distribution and that can be different depending upon the pair (server, costumer) being considered. Note that by lifting the assumption of exponential service times the spatial queuing system is no longer Markovian. Previous works have noticed that service times are better modeled by using a lognormal distribution (Budge et al., 2009; Rajagopalan et al., 2008).

Each service region j generates a fraction f_j of the total demand ($\sum_j f_j = 1$). The total demand rate is λ and the demand of each zone is $\lambda_j \equiv \lambda f_j$. When a request for service arrives, if

the primary responsible base station has at least one server available, it is dispatched immediately. The server travels to the place of the incident, spends some time at scene and then returns to its base location before being assigned to the next request. If the primary responsible station does not have a server available a server from another base will be assigned, following a fixed priority list with respect to the base stations for each demand zone. If all the servers are busy the request is considered to be lost (this typically means that it will be referred to an external system). The model assumes that the servers are identical. The service time of any response unit for any call for service has a general distribution with mean depending upon the server location and the demand zone being served. The service time for a call includes the set up time, the travel time from the base to the incident location, the on-scene time, a possible related follow up-time and the travel time back to the base. The response time interval is the time from when an ambulance is dispatched until it arrives at the scene.

Steady-state probabilities for the underlying spatial queuing system are determined by using the approximated hypercube model from Budge et al. (2009). In their work, instead of assigning busy probabilities to each server they focused on finding busy probabilities associated to base stations. The assumption is that once a call has been assigned to a base station, the particular server that will be sent is selected randomly among those available at that particular location. The approximation algorithm depends on the locations having servers, how many servers there are per location and also on the dispatching policy.

4.3.2 Formulation

$\mathbf{J}$ represents the set of service regions; $\mathbf{I}$ is the set of potential location sites, $|\mathbf{I}| \leq |\mathbf{J}|$; N is the total number of response units (servers); M_i is the maximum number of ambulances allowed at station base i; t_{ij} is the mean response time for a server from station i to reach region j, when available; τ_{ij} is the mean service time for any server from base i when attending region j; λ is the total network-wide demand (requests/unit time); f_j is the fraction of network-wide demand generated from region $j \in \mathbf{J}$. The decision variables are as follows:

$$x_i = \begin{cases} 1 & \text{if potential site } i \text{ is open} \\ 0 & \text{otherwise} \end{cases}$$

$$y_{ij}^l = \begin{cases} 1 & \text{if station } i \text{ has priority } l \text{ to zone } j \\ 0 & \text{otherwise} \end{cases}$$

w_i = number of servers assigned to potential site i

S = total number of base stations that are open (auxiliary variable)

ρ_{ij} = fraction of dispatches sending a unit from site i to region j (auxiliary variable)

The optimization problem, using mean response time (MRT) as the objective function, is formulated as:

$$\text{Minimize } MRT = \sum_{i=1}^{I} \sum_{j=1}^{J} f_j \rho_{ij} t_{ij} \tag{4.1}$$

s. t:

$$\sum_{i=1}^{I} w_i = N \tag{4.2}$$

$$w_i \leq M_i x_i \qquad i \in \mathbf{I} \tag{4.3}$$

$$\sum_{i=1}^{I} x_i = S \tag{4.4}$$

$$x_i \geq y_{ij}^l \qquad i \in \mathbf{I}; j \in \mathbf{J}; l = 1, \ldots, S \tag{4.5}$$

$$\sum_{l=1}^{S} y_{ij}^l = 1 \qquad i \in \mathbf{I}; j \in \mathbf{J} \tag{4.6}$$

$$\sum_{i=1}^{I} y_{ij}^l = 1 \qquad j \in \mathbf{J}; l = 1, \ldots, S \tag{4.7}$$

$$x_i \in \{0, 1\} \qquad i \in \mathbf{I} \tag{4.8}$$

$$y_{ij}^l \in \{0, 1\} \qquad i \in \mathbf{I}; j \in \mathbf{J}; l = 1, \ldots, S \tag{4.9}$$

Equation (4.1) is the objective function; constraint (4.2) determines the number of servers to be located; constraint (4.3) restricts the assignment of ambulances only to the open sites; equation (4.4) defines the auxiliary variable accounting for the total number of stations open. Constraint

(4.5) states the logical relationship between the location decision and the assignment of a location within the priority list of a demand zone and finally, constraints (4.6) and (4.7) assure that there is a complete priority list for each demand zone, and that within the priority list of each demand zone each server appears only once. Constraint (4.8) is the integrality constraint for the decision variable x_i and (4.9) is the integrality constraint for the decision variable y_{ij}^l. The model given by (4.1)-(4.9) represents the basic optimization problem in which the location of the servers and the dispatching rule for each demand zone are the decisions to be made. Recall that the auxiliary variable ρ_{ij} needs to be calculated by analyzing the queuing behavior of the system. The dispatching probabilities change whenever any of the location or dispatching decisions change.

The resulting formulation corresponds to a non-linear mixed integer programming model that has embedded a queuing sub-model corresponding to a finite-state continuous time stochastic process. It is an NP-Hard problem (Geroliminis et al., 2009). Given a particular set of locations for the available servers and a preference list for each demand zone with respect to the servers' locations, it is necessary to solve the embedded queuing model before being able to calculate the value of the objective function. There is a complex relationship imposed by the combined location and dispatching decisions. The stochastic sub-system can be analyzed by writing a set of flow balance equations that in turn will lead to a linear system of equations, whose exact solution requires the calculation of the inverse for the matrix of coefficients. The size of this matrix grows exponentially (with respect to the number of servers), therefore the time that it takes to perform a single iteration to evaluate a candidate solution can be computationally prohibitive.

4.3.3 Solving the embedded queuing model

The dispatching probabilities ρ_{ij} depend on the particular configuration of the system: where the servers are and how they will be dispatched to attend incoming emergency calls. The location and dispatching decisions are given by the decision variables x_i, y_{ij}^l and w_i. In what follows we detail the approximation procedure to obtain the dispatching probabilities. It is based in the work by Budge et al. (2009). We detail the procedure in terms of our notation for convenience of the reader. As in the seminal works by Larson (1975) and Jarvis (1985) the approximation

procedure is derived by initially assuming that the servers operate independently, and then developing correction factors to account for server cooperation. The algorithm starts by calculating the following:

b_{kj} = kth preferred station for node j (given by variables y_{ij}^l)

$s_{(k)j}$ = number of vehicles at the kth preferred station for node j (given by y_{ij}^l and w_i)

$z_{(k)j} = s_{(1)j} + s_{(2)j} + \ldots + s_{(k)j}$

$\tau_{(k)j}$ = average response time from kth preferred base station to costumer j

Step 0: Set iteration counter $h = 0$. Initialize busy fraction (r) and system-wide average service time (τ), under the assumption of the EMS operating as an Erlang Loss System M/G/N/0 (superscripts indicate iteration counters). r_i corresponds to the busy fraction of each open station. P_0 and P_N correspond to the probability of the system being idle (all ambulances are available) and the probability of all servers being busy, respectively.

$$\tau^0 = \frac{1}{\lambda N} \sum_{i=1}^{I} w_i \sum_{j=1}^{J} \lambda_j \tau_{ij}$$
$$r_i^0 = r^0 = \lambda \tau^0 (1 - P_N^0)/N$$

Step 1: Calculate P_0^h and P_N^h

Step 2: Calculate V_i^h as follows

$$V_i^h = \sum_{j=1}^{J} \lambda_j \tau_{ij} Q_j(\{S_{(k)j}\}, \rho^{h-1}, a_{ij}) \prod_{l=1}^{a_{ij}} (r_{(l)j}^{h-1})^{s_{(l)j}}$$

$\{S_{(k)j}\}$ is an ordered set containing the number of servers available at each kth preferred location for costumer j; r_i represents the busy fraction of each station i while r stands for the system-wide average server utilization, $r = \rho(1 - P_N)$; ρ is the overall system utilization ($\rho = \lambda\tau/N$); a_{ij} is the order of preference of station i in the priority list of demand zone j. Finally, the correction factors Q_j are defined as follows:

$$Q_j(\{S_{(k)j}\}, \rho^{h-1}, a_{ij}) = \frac{P_0 \sum_{n=z_{(k-1)j}}^{N-1} \frac{(\rho N)^n}{n!} \left[\prod_{u=0}^{z_{(k-1)j}-1} \frac{n-u}{s-u} - \prod_{u=0}^{z_{(k)j}-1} \frac{n-u}{s-u} \right]}{r^{z_{(k-1)j}} \left(1 - r^{s_{(k)j}}\right)}$$

After obtaining the values for V_i^h the busy fractions need to be updated. If $r^{h-1} \leq 0.5$ then

the update should be done using the following:

$$r_i^h = \frac{V_i^h}{w_i + (r_i^{h-1})^{w_i - 1} V_i^h}$$

For relatively high average server utilization, i.e. $r^{h-1} > 0.5$, the updating should be done using the following expression:

$$r_i^h = \left(\frac{V_i^h}{V_i^h + w_i / (r_i^{h-1})^{w_i - 1}} \right)^{1/w_i}$$

Step 3: Calculate the dispatching probabilities using the following

$$\rho_{ij} \approx Q_j(\{S_{(k)j}\}, \rho, k) \prod_{l=1}^{k-1} r_{(l)j}^{s_{(l)j}} (1 - r_i^{w_i})$$

and then normalize them using $\rho_{ij}^h \leftarrow \rho_{ij}^h (1 - P_N^h) / \sum_{i=1}^{S} \rho_{ij}^h$

update τ^h, ρ^h, r^h as follows:

$$\tau^h = \frac{1}{\lambda(1 - P - N)} \sum_{j=1}^{J} \lambda_j \sum_{i=1}^{S} \rho_{ij}^h \tau_{ij}, \ \rho^h = \lambda \tau^h / N, \ r^h = \frac{1}{N} \sum_{i=1}^{S} w_i r_i^h$$

Step 4: If $|r_i^h - r_i^{h-1}| < \xi$ for all i, then stop. Otherwise, set $h = h + 1$ and go to *Step* 1.

The value of ξ controls the convergence of the algorithm. The developers of the algorithm mentioned that there is not a theoretical guarantee for the convergence of the algorithm, but that it did converge in all their test cases, which covered a wide range of operational conditions for real EMS systems. For further discussion and details the reader is advised to consult the original work by Budge et al. (2009).

4.3.4 Alternative performance criteria

The model introduced in section 4.3.2 has Mean Response Time as the objective function. However we have performed experiments with other objective functions, including some that are related to fairness. The Expected Coverage, which is an efficiency criteria widely used in real EMS settings is defined next.

$$Ex.Cov = \sum_{j=1}^{J} \sum_{i=1}^{I} f_j \rho_{ij} P_{i,j} \tag{4.10}$$

where $P_{i,j}$ is the probability that station i covers node j. Note that $P_{i,j}$ can be used as a binary variable, assuming deterministic response times, indicating whether or not the coverage threshold is satisfied by the available servers, but it can also be used as the probability of that coverage is possible within the given threshold, accounting for variability in travel times.

From the point of view of the costumers (demand zones) we calculate average individual response times (IRT), which are given by equation (4.11).

$$IRT_j = \sum_{i=1}^{I} \rho_{ij} t_{ij} \tag{4.11}$$

Once the individual response times are obtained then it is possible to calculate the following criteria, Squared Coefficient of Variation (equation (4.12)) and Gini Coefficient (equation (4.13)), both of which measure dispersion among the individual values. The Gini coefficient has a value between 0 and 1, with 0 meaning that all the individual values are identical. A value of 0 for the squared coefficient of variation also means identical values, or standard deviation equal to 0. The squared coefficient of variation and the Gini index satisfy the scale independence principle, are population size independent and also comply with the principle of transfers (Pigou-Dalton condition), which are desirable characteristics of fairness criteria (Marsh and Schilling, 1995).

$$CV^2(IRT_j) = \frac{\sigma^2(IRT)}{\overline{IRT}^2} \tag{4.12}$$

$$Gini(IRT) = \frac{1}{\overline{IRT}|\mathbf{J}|^2} \sum_{i \in \mathbf{J}} \sum_{j \in \mathbf{J}} |IRT_i - IRT_j| \tag{4.13}$$

To measure equity among servers initially we obtain the workload of each station, and then the relative differences among each station's workload. The approximation procedure detailed in Section 4.3.3 includes the calculation of individual station's busy fractions r_i. Note that in this case we are attempting to equalize the workload assigned to each open station, some of which may have more than one server assigned. The underlying assumption is that the total workload assigned to a particular station would be divided proportionally among its servers.

We use the squared coefficient of variation for server's workload (Equation 4.14). IWK stands for Individual Station Workload.

$$CV^2(IWK) = \frac{\sigma^2(IWK)}{\overline{IWK}^2} \tag{4.14}$$

4.4 Solution approach

Tabu Search is a path-based metaheuristic based on local search. Local search procedures used to solve optimization problems start with a given solution (usually easy to generate and if possible feasible), and then define a set of neighboring solutions that are evaluated looking for an improvement in the value of the optimization criteria. However, at each iteration of a TS algorithm it moves to the best neighbor, even if the best neighbor has a worse objective value. Furthermore, TS uses the idea of memory, keeping track of the last solutions visited (or their characteristics) to avoid cycling. Solutions stored in memory are considered Tabu for a number of iterations (tabu tenure), and hence the algorithm considers it illegal to move to those solutions. The use of memory and moving to the best neighbor (even at the cost of a deterioration in the objective function value) are ways to escape local optima during the search procedure. Seminal ideas of TS were introduced in Glover (1986), with further developments and refinements in Hansen (1986); Glover (1989). Our approach is not a pure TS implementation since we are using random components in the search procedure, including random initialization and a tabu list with dynamic (reactive) size.

Reviews by Reese (2005) and Mladenovic et al. (2007) related to solving the p-median problem found TS and Genetic Algorithms (GA) to be the most common metaheuristics used. The mathematical structure of the p-median problem is close to the structure of the problem we are approaching. The p-median is a location/allocation problem although it does not include cooperation among servers, which is a distinctive characteristic of EMS systems. The review by Li et al. (2011) on covering models applied to EMS also identified TS and GAs as the predominant solution approaches. Although typical covering models assume a-priori dispatching rules, the mathematical structure of covering problems in particular as applied to EMS system is closely related to that of our work.

4.4.1 Solution representation

The main decision variables in our problem are where to locate ambulances, how many ambulances to locate at each open site and the priority of open stations regarding each demand zone. For the location decision we use an array with as many components as the number of ambulances available to locate. Each component of the vector can take an index value representing the candidate locations. This representation allows for controlling the maximum number of servers in the system as well as checking if the capacity of each station base is being observed. For the dispatching decision we concentrate on identifying only the preferred server, hence we have a representation based on an array with as many components as demand zones. The values that are valid for each component of the dispatching array depend on the location decision, and in particular on how many open stations are being considered. Note that it would be invalid to use an index including all candidate locations, since several of them are not going be active in a particular solution. Focusing only on the preferred server to each demand zone has the advantage of allowing a compact representation of the dispatching decision. Furthermore, by analyzing the dynamic queuing behavior of the EMS systems under consideration we have noticed that the fraction of dispatches from the preferred base station is usually the highest, which is a desirable characteristic of an EMS system. The preferred server is expected to offer the best possible service whereas the backup servers are only acceptable when the preferred one is busy. The solution representation provides only the preferred base station (which can have several ambulances assigned). The remaining open stations are organized in order of preference according to the distance from the demand zone, with the highest priority given to the closest base station.

4.4.2 Neighborhood structure

For a given solution we define its neighbor considering either the location portion of the decision or the dispatching portion of the decision. During each iteration a decision is made randomly regarding whether a neighborhood is generated based on the location decision or the dispatching decision. Each neighborhood has a 50% chance of being used. In the case of a neighborhood based on the location decision each possible interchange is considered, as follows. For each location having

at least one ambulance assigned in the current solution we consider interchanging one ambulance with a different location that still has available capacity. If the neighborhood has to be generated based on the current dispatching portion of the solution, then for every demand zone we consider interchanging its current preferred base station with any of the other open stations. Recall that any change in the solution, either on location or dispatching, requires the evaluation of the objective function, which means the hypercube approximation procedure has the be executed. If required, in order to reduce the computational burden of the algorithm, it is possible to reduce the size of the neighborhood. This can be done for instance by using an acceptance probability for each possible move, although it also reduces the power of the local search component of the TS algorithm.

4.4.3 Tabu tenure

A critical component of any TS implementation is the size of the tabu list, i.e. for how long (number of iterations) the algorithm will penalize the solutions (or specific characteristics of their generation process) already visited. Instead of experimenting with several sizes of the tabu list we are using a reactive tabu list tenure. The idea of having a memory attached to a local search is to prevent the search procedure from getting trapped at a local optima. However, if the application of repeated local searches is yielding good results it would be beneficial to keep the algorithm progressing in the direction suggested by the local search. In this case, the tabu list will likely get in the way and therefore it makes sense to shorten the list, allowing the algorithm to keep inspecting a potentially good area of the solution space. On the other hand, if the application of several local searches is not allowing the optimization criteria to improve it may be a sign of the algorithm getting trapped at a local optima. In this case it makes sense to use the tabu list to force the algorithm to move to a different exploration area (by increasing the tabu tenure), even at the cost of accepting a deterioration in the optimization criteria, at least temporarily. In Erdogan et al. (2010) the authors solved what they termed MEXCLP+PR+SSBP problem (Maximal Expected Coverage Location Problem + Probabilistic Response Time + Station Specific Busy Probabilities). They suggested a TS heuristic for the solution of the problem in which the tabu tenure was set to 10. Gendreau et al. (1997) suggested a tenure in the range 10-30. As mentioned before we use a

dynamic size of the tabu list based on the progress of the optimization search. As lower and upper bound for the size we use the range 5-30.

4.5 Computational results

In our experiments we have used the following configuration of the algorithm. The TS was stopped after 20 iterations in which no improvement was made to the objective function, or else a maximum number of iterations was set to 50. In many cases we have observed convergence of the algorithm in less than the maximum number of iterations allowed. We ran experiments allowing as many as a 100 iterations, however the average number of iterations for convergence was 43. The algorithm started from a random feasible solution and for each criteria and under each scenario 30 different runs were made. In all the cases the algorithm converged to solution values for which the coefficient of variation was below 2%. The tabu list starts with a size of 5, which is the lower bound used in our dynamic tabu tenure approach. The tabu list size is updated if after 5 iterations of the algorithm the best solution remains the same. The update is done according to $tenure_{new} = max\{1.5tenure_{current}, tenure_{max}\}$. If the algorithm is progressing through improved solutions from one iteration to the next then the tabu tenure is updated according to $tenure_{new} = min\{0.9tenure_{current}, tenure_{min}\}$. The proposed updating scheme increases or decreases the size at small steps, and worked well in our experiments for quick convergence of the algorithm.

Two case studies are analyzed next. All the required programming was done in Java (v1.5), using the Open Tabu Search package (Harder, 2013). Experiments were executed on a PC running Windows®7 -64 Bit, with an Intel®Core 2 Duo processor running at 2.13 GHz and 2 GB of RAM. The first case study corresponds to the city of Edmonton, Canada. The data has been previously used for the computational experiments reported in Erkut et al. (2008); Erdogan et al. (2010). Edmonton data is available from `http://apps.business.ualberta.ca/aingolfsson/data/`. It corresponds to a city of about 800,000 people. For the analysis the city has been divided into 180 demand zones. There are 16 candidate locations for ambulances across the city, 7 of which have capacity for more than one ambulance. The average call rate used was 3.5 calls/hour, which for the

base case (with 12 ambulances) gives an overall utilization of 0.3 (ratio between total demand and total service rates). Additional details about this case study can be found in Erkut et al. (2008); Erdogan et al. (2010).

The second case study corresponds to Mecklenburg County (Greater Charlotte), North Carolina. The data for this case study corresponds to the year 2004. This region had a population of about 800,000 (in 2004). In that year approximately 62,092 calls were classified as emergency calls. For the analysis, the region was divided into 168 demand zones, by imposing a grid of two mile by two mile squares. The demand zones are at the same time candidate locations for the ambulances, excluding the areas that correspond to the boundary of the region. After excluding those zones, there are 121 candidate locations, each one having capacity for only one ambulance. The case study includes 18 ambulances, with an overall utilization of about 0.55. Additional details about the case study can be found in Rajagopalan et al. (2008).

4.5.1 Edmonton case study

Table 4.1 shows the results after applying the solution procedure to the data from Edmonton. Bold values correspond to the best result under each optimization criteria. We are assuming partial backup, allowing only 3 of the open stations to respond to calls for every demand zone. Recall that in this case there may be locations with more than 1 ambulance assigned, hence it is still possible that more than 3 ambulances are allowed to serve a demand zone. According to Geroliminis et al. (2009), from a practical perspective, being served from locations that are ranked as 4^{th} and up for a particular demand zone is not desirable, because the overall efficiency of the system would likely decrease. In addition, the partial backup tends to better represent real systems, because even when there is a server available, if it is too far away from the costumer, typically a third party will be contacted to attend that particular call. The average running time of the algorithm for the Edmonton scenarios ranges from 14 minutes for the base case with 12 ambulances to 28 minutes for the case of 18 ambulances. Recall that the solution space has a combinatorial nature. Furthermore, it is important to note that the increase in solution time is not exponential, as it would be if an exact solution for the queuing submodel were used.

Table 4.1: Edmonton Overall Results

Num. Ambulances	Optimization Criteria	Performance Indicator				
		MRT	Exp.Cov	Gini	SCV-RT	SCV-Wk
12	MRT	**6.283**	0.792	0.617	2.695	0.071
	ExCov	6.283	**0.792**	0.617	2.695	0.071
	Gini	10.927	0.476	**0.545**	1.287	0.115
	SCV-RT	12.129	0.415	0.543	**1.191**	0.128
	SCV-Wk	9.580	0.536	0.636	2.245	**0.002**
15	MRT	**5.964**	0.832	0.626	2.918	0.110
	ExCov	5.965	**0.833**	0.630	2.933	0.113
	Gini	10.004	0.553	**0.549**	1.420	0.144
	SCV-RT	10.665	0.371	0.556	**1.244**	0.085
	SCV-Wk	8.806	0.577	0.678	3.334	**0.011**
18	MRT	**5.807**	0.853	0.635	3.071	0.206
	ExCov	5.807	**0.853**	0.635	3.071	0.206
	Gini	10.185	0.543	**0.562**	1.446	0.236
	SCV-RT	10.350	0.401	0.549	**1.195**	0.308
	SCV-Wk	8.709	0.536	0.656	2.551	**0.005**

MRT is given in minutes

The Edmonton data includes the modeling of response times as random variables (Log-normal distribution). For each pair (potential location, demand zone) the coverage probability is given, i.e., we know how likely the response time is to be less or equal than the coverage threshold (8 minutes in this case). Details on how the response times were modeled can be found elsewhere (Ingolfsson et al., 2008). This modeling choice is a closer representation to the real behavior of the system; if instead of using this approach the coverage were calculated using the expected response times as deterministic values, then for all the scenarios considered the Expected Coverage would be above 90%. Note that adding servers improves the efficiency criteria, response time and coverage. However, going from 15 to 18 servers (an increase of 20% in the number of ambulances) improves coverage and response time only by about 5%. In other words we have a decreasing marginal impact of each additional vehicle, a result that was also mentioned in Goldberg and Paz (1991).

Applying the efficiency criteria (coverage and response time) result in the same or very close solutions. Both efficiency criteria are distance-based, and the use of coverage as a probability as opposed to a pure binary variable leads to a expected coverage function that is smoother, which

may explain why the solutions to Mean Response Time and Expected Coverage problems are the same. From Table 4.1 we can also see that there are trade-offs between the efficiency and fairness criteria. Figure 4.1 illustrates the trade-offs, comparing the solutions obtained when applying fairness criteria against the solution from Expected Coverage/Response time, for the base case (12 ambulances). Positive variations represent improvements. Categories in x-axes correspond to the different optimization criteria being used.

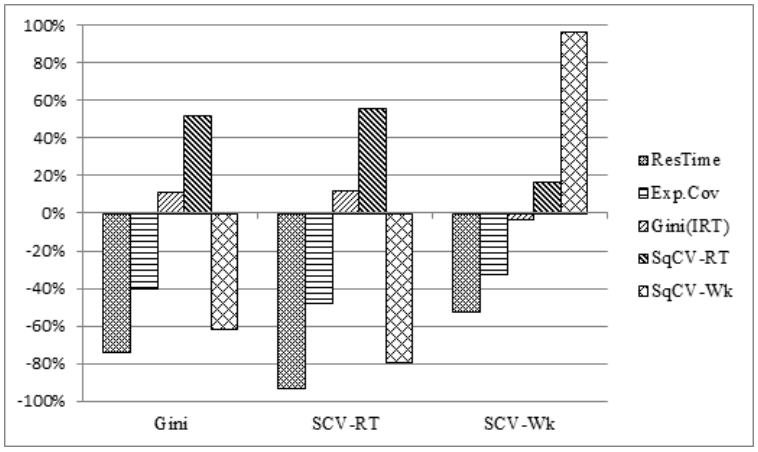


Figure 4.1: Edmonton trade-offs Efficiency vs. Fairness criteria

Improvements in equity from the point of view of the costumers (demand zones) are in general accompanied by a sacrifice in efficiency that is proportionally bigger. From the point of view of server's workloads it possible to see that improving workload balance causes a decrease in efficiency that is proportionally smaller. However, going back to Table 4.1 we can see that the variation in workloads under different criteria, as measured by the squared coefficient of variation (SQV-Wk), is in general only a small fraction of the variation on individual response times (SQV-RT). For the base case, considering all the optimization criteria being used, the maximum SQV-Wk is 0.128, whereas the minimum SQV-RT is 1.191. Given the fact that there is bigger disparity in response times than in workloads we focus the following analysis on individual's response time variation. In particular we further analyze the trade-offs between coverage and squared coefficient of variation of individual response times.

So far we have used a single objective approach, and once we get the best heuristic solution from a particular criteria then we obtain the value of the remaining performance criteria. To

better analyze the trade-offs between Exp.Cov and SQV-RT of individual response times we use a bi-objective optimization approach, based on the ideas of the $\xi-$constraint method. The basic idea is to use one of the criteria as optimization objective while the other one acts as a constraint. Initially the two objectives are used individually to identify the best possible performance of the system from their perspective. This is what has been done so far. Next we set one of them as a constraint, and optimize the system based on the other. For instance, for the base case (with 12 ambulances) the maximum Exp.Cov was 0.792, and the solution reaching this value has a SQV-RT equal to 2.695. If we set a constraint on Exp.Cov so that it has to be a value strictly less than 0.792 and optimize the system based on SQV-RT, then it is possible to identify different trade-offs. Formally, the $\xi-$constraints method requires to perform several optimization scenarios, by choosing an specific value for ξ. We used expected coverage as the constraint and optimize the system based on SQV-RT. The lower and upper bounds for the expected coverage were 0.415 and 0.792. We divided that interval into 10 sections, hence our ξ value is 0.037. The results can be seen in Figure 4.2. The dots in Figure 4.2 correspond to a heuristic approximation to the Pareto front (or efficient frontier). Solutions that do not belong to the Pareto front are known as dominated solutions (for the same value of one of the objective functions it is possible to get a different solution that strictly improves the other objective). The Pareto front is an important decision tool that would help decision makers to understand how much sacrifice in one criteria is reasonable in order to get improvements in the other. Solutions that do not belong to the Pareto front should be avoided.

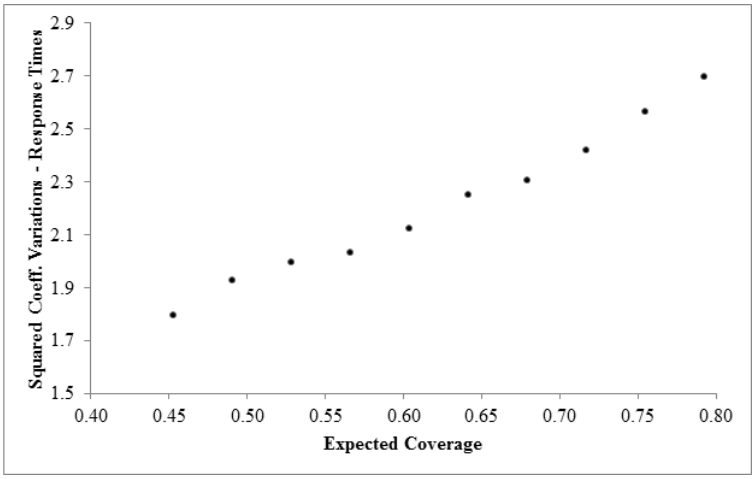


Figure 4.2: Edmonton trade-offs Coverage vs. SQV-RT

4.5.2 Charlotte case study

Because of the assumption that every demand zone (except those at the boundaries) is a candidate location for an ambulance we change the neighborhood structure used within the optimization procedure. This is done in order to reduce the computational effort of the algorithm and to better match a local search strategy to the data structure. Originally the local search was designed based on interchanging an ambulance between its current location and every alternative location available. With 121 potential locations available it would generate too many potential moves for the local search step of the algorithm. Instead, we use the proposed grid (2 miles by 2 miles squares) to generate the neighborhood of a current location: interchanges are only made between the current location and those potential locations that surround it on the grid. The neighborhood for the dispatching decision is the same that was used for the Edmonton case study. Partial backup is again used, allowing only up to three servers to be in the list of potential units to attend each demand zone.

Table 4.2 shows the overall results for the Charlotte case study using 18 ambulances. The data available for Charlotte includes call volume for each demand zone and different periods during each day of the week. It has been shown that call volume differs across the week and at different hours during each day (Rajagopalan et al., 2008). The results shown next correspond to demand observed Mondays from 4:00PM to 6:00PM, which corresponds to a peak of demand during the day. The overall utilization is 0.55. The same stopping criteria explained before was used, as well as the same dynamic updating procedure for the tabu list size. Initial solutions are generated randomly and 30 runs are used for each criteria. For each criteria the coefficient of variation of the 30 different runs was below 3%. The average running time of the algorithm was 18 minutes.

Note that for the Charlotte case study the coverage is based on deterministic response times, therefore the coverage probability would be either 0 or 1 for each demand zone depending on the relative location of demand zones and open stations. Despite the coverage probability being 0 or 1 it is worth noticing that the availability of a server (overall probability of being busy) also plays a role in determining the expected coverage of each demand zone. We do not have enough data to model response times as a probability distribution, which would allow for smoothing the expected coverage

Table 4.2: Charlotte Overall Results

Optimization Criteria	Performance Indicator				
	MRT	ExCov	Gini	SCV-RT	SCV-Wk
MRT	**3.544**	0.963	0.457	0.790	0.094
ExCov	5.219	**0.987**	0.603	2.236	0.214
Gini	5.956	0.836	**0.386**	0.487	0.224
SCV-RT	6.435	0.803	0.391	**0.479**	0.222
SCV-Wk	8.953	0.630	0.584	2.112	**0.007**

MRT is given in minutes

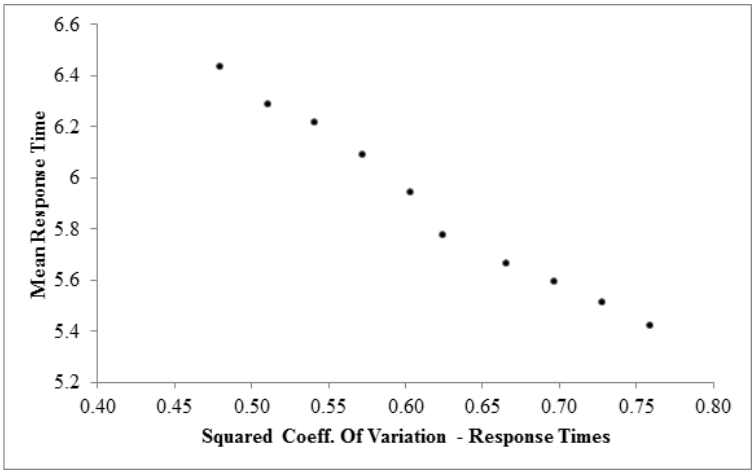


Figure 4.3: Charlotte trade-offs MRT vs. SQV-RT

function. This change in the modeling approach should explain obtaining different solutions for the efficiency criteria. For the previous case study using response time or expected coverage led to the same or similar solutions. We can see that the solution minimizing mean response time is overall better than the one maximizing coverage: the sacrifice in coverage when reducing response time is only about 2.5%, and the solution that minimizes response time also offers better values for each of the fairness criteria.

Analyzing the fairness criteria it is possible to see that for this case study the variations on individual response times are also proportionally bigger than variations on individual server's workloads. We focus then on analyzing the trade-offs between response time and squared coefficient of variation of individual's response times. The minimum response time solution has a SQV-RT of 0.790, while the minimum SQV-RT found was 0.479. Figure 4.3 shows the trade-offs between

the two criteria, optimizing the response time subject to a constraint on SQV-RT. We use the $\xi-$constraint method, varying SQV-RT between 0.479 and 0.79 with increments of 0.031, resulting in 10 points. This is again a bi-criteria approach that provides decision makers with a better look at balancing these two competing criteria. It may be the case that in a particular setting one criteria is more important than the other, according to stakeholders, and therefore one criteria would be preferred as the optimization objective. However, a decision tool such as the Pareto front helps to identify tradeoffs that are unacceptable, by showing that it is possible to get a different solution with the same value on one criteria and strictly a better value on the other criteria.

4.6 Conclusions

Although the analysis and planing of EMS systems is usually based on performance criteria related to efficiency, it is clear that there are other perspectives that also play an important role on the overall quality perception of these systems. In addition to the natural expectation of quick response to emergency calls, in the public eye there is also a concern for equity considerations, since EMS systems are there to assist people when they are more vulnerable. Hence, there is also a natural expectation that anyone experiencing an emergency should have the same chances of receiving quick response. However, there is not a unique and universally accepted way to consider equity, a conclusion shared by several authors. As there is not an agreement about what equity should entail for a particular system, it is difficult to include fairness considerations during the planning of the system. In addition, several groups of stakeholders will usually have different objectives, and improving a single perspective can lead to bad solutions from other's points of view.

We approached the planning of EMS systems taking into consideration not only the typical efficiency criteria but also fairness considerations, from the point of view of both final users and internal providers. We used a fairness perspective in which equalizing the performance of the system with respect to individual stake holders is a desirable characteristic. Hence, we looked to reduce disparities in the mean response time of different demand zones, as well as disparities in the workloads of the servers (ambulances). Previous literature had suggested that the combination

of location and allocation decisions was required in order to optimize the system from a fairness perspective. However, the joint problem leads to a formulation of exponential size that was impractical to use for large scale real systems. With an exponential number of equations to be solved, the problem quickly saturates the processing capacity and the memory of a computer. We contribute to the literature by developing a new mathematical model that addresses the large scale limitations, allowing the location of several ambulances at a single station. Furthermore, our solution procedure is based on a heuristic optimization technique (Tabu Search), and uses an approximation algorithm to evaluate the stochastic performance of the system. We were able to solve instances with as many as 18 ambulances and 180 demand zones in under 20 minutes of computation time on a regular desktop PC. For an exact solution approach, with 18 ambulances the transition matrix of the queuing submodel could potentially have 6.8E10 entries, and reserving computer memory to store such a big amount of intermediate values would likely be infeasible.

In addition to addressing the scalability issues we also add a multiobjective (bi-criteria) perspective, identifying efficiency and fairness criteria that are in conflict with each other and studying the tradeoffs so that a Pareto front can be derived. Although it is true that efficiency considerations are usually considered to be the most important we believe that having the opportunity to identify the performance from other perspectives should add value to the decision making process. It may be the case that a reasonable sacrifice on efficiency can led to a more equitable solution, and then it may make sense to change the configuration of the system accordingly.

Although our results are promising, as a way to achieve a more holistic planning process for EMS systems we are also aware of potential limitations of the approach. Real EMS systems are subject to a lot of pressure and often times decisions are made that do not comply with the original design of the system. The modeling approach does not include situations such as possible relocation of ambulances, for instance in response to temporal/spatial variations in demand, nor does the model take into consideration the priority of emergency calls, a practice that is very common in the real world. We studied the demand for high priority calls, and based our results on that demand, making the assumption that if the system is reasonably well designed for the calls with the highest priority, it should be able to respond adequately to less time-sensitive calls.

Furthermore, although the solution times are small enough so that several configurations of the system can be evaluated, they are still quite large if real-time decisions were to be made based on our solution approach.

Chapter 5

Final remarks

5.1 Summary of contributions

We have approached the study of EMS systems combining decisions from two different levels: the location of ambulances, which is a strategic decision, and the dispatching rules, which correspond to the operational level. Our joint approach is a contribution to the literature on EMS systems planning, where we did not find previous references combining the two decisions on a single framework. The work was inspired by previous work mentioning that the use of the most commonly accepted dispatching rule, which is to send the closest ambulance available to attend any emergency call, could lead to suboptimal solutions from the point of view of the efficiency of the system, in particular the mean system response time.

Initially we developed the mathematical model for the joint problem. The resulting problem is NP Hard and has two main components: the first dealing with the combinatorial nature of the location/allocation decisions, and the second dealing with the evaluation of the system, which is subject to congestion. Demand for service is indeed random as well as service times; therefore ambulances can be busy when they are required to attend the next emergency call. Obtaining the busy probabilities of the ambulances is a critical step in order to be able to estimate performance indicators for the system.

The complexity of the mathematical model required the development of solution strategies

based on heuristics. Initially a solution based on Genetic Algorithms (GAs) was proposed and validated. The GA is used to guide the search of good solutions. The evaluation of each solution requires solving the congestion submodel, which is a queuing model. The GA solution framework used an exact solution for the queuing model.

The GA optimization framework was used to analyze several case studies, starting with very small cases that were possible to solve to optimality by using full enumeration. A mid size case study taken from the literature was also solved, as well as a real case study of an EMS system. In all cases we found that the minimization of Mean Response Time can be done by using the closest dispatching rule to determine the dispatching preference of the servers with respect to each demand zone. We also analyze Expected Coverage, concluding that in most cases the closest dispatching rule also produces the best solution for coverage (or solutions that are only marginally worse, as much as 2%). We also found that the solution that minimizes Mean Response Time also offers a competitive value of coverage, whereas the sacrifice in Response Time to maximize coverage is not appealing.

The optimization framework was then used to analyze EMS systems from other perspectives, in particular taking into account fairness criteria. We focused on equalizing individual response times and equalizing servers workload, treating each objective separately. Several performance measures were used to represent the idea of fairness, such as variance and Gini coefficient. In general we found that the joint approach is required in order to get the best solutions from the perspective of the fairness criteria we used. In all cases our joint approach was able to find better solutions from the perspective of fairness than an approach in which the closest dispatching rule is used.

The potential improvements on fairness criteria come at a price: sacrifices on efficiency. Our solution approach allowed for identifying different trade-offs between the several criteria that we used. Providing these trade-offs is also a main contribution from our work since previous literature did not address directly the measure of criteria other than efficiency based for EMS systems. We believe that it is important for decision makers to know how their systems are performing from different points of view. Even if there is a dominant perspective it adds value

knowing the effect that the selected optimization strategy has on alternative views of the system. This is particularly important when we are talking about public systems such as EMS, because there are several stakeholders involved and each one surely has some expectations. Realizing how the system is performing as a result of selecting particular criteria to optimize is a starting point. If the resulting optimal or best policy causes some stakeholders to be affected more than others, then knowing the trade-offs can serve to identify what sacrifices are acceptable to regain balance.

Finally we studied scalability issues related to our proposed joint model and its solution approach. It turned out that using en exact model for the congestion/queuing component is not a good idea if the approach is going to be used on large scale real world EMS systems. Initially we used the exact approach as a way to understand the benefits of the joint model. After realizing that there were benefits related to optimize fairness objectives we turned our attention to make the solution process scalable. In order to do so we developed a new mathematical model, allowing for several ambulances to be located at a single station since this is typical on large scale EMS systems. We also changed our solution approach from a population based heuristic to a path based heuristic, in this case Tabu search (TS). Due to the dynamic nature of the dispatching decision, related to the fact that several ambulances can be located at a single station, and therefore we do not know beforehand how many open stations we are going to have, the solution representation that was natural for GA was no longer valid.

The new solution procedure is based on Tabu search but also has characteristics that are not part of a classic implementation. We used random initialization, also a random component determining the type of neighborhood that is used at each iteration (related to either the location or dispatching decision), and dynamic update of the size of the Tabu list. By using the new optimization framework we were able to analyze two different real world large scale EMS systems with as much as 18 ambulances and 180 demand zones, covering a population of about 800,000 people. The solution approach allows for identifying trade-offs between the different criteria in reasonable computation time, usually under 20 minutes on a regular desktop based PC.

5.2 Managerial insights

The planning of EMS systems is heavily dominated by the use of efficiency criteria. We have seen this trend being mentioned not only in several references cited throughout the document, but also on opinions and comments that were received after presenting our results at several academic conferences. Among the efficiency criteria the most commonly used is Coverage (or Expected Coverage). In fact, we have found that not only in the US but also in other countries (such as Canada, England, Germany, Netherlands) the minimum standard required for EMS systems is given as a function of coverage.

We have found that selecting coverage maximization as the optimization criteria can have undesirable effects on other criteria, including response time, which is also a key performance measure for EMS systems. In fact, for critical emergencies such as cardiac arrest and similar time-sensitive incidents, reducing the response time increases the chances of saving lives. In all of our case studies the coverage maximization approach causes a sacrifice in response time that is proportionally bigger than the additional coverage obtained. We believe this is a very important result for practitioners. If the overall coverage resulting from a response time minimization approach is good enough to meet the regulations that are in place, then minimizing response time is a better approach. Some EMS planners will not realize this; since they are being measured by the coverage that they are providing it is only natural that their planning approach would be based on coverage maximization.

The public nature of EMS systems as well as their key role in saving human lives makes them particularly subject to close scrutiny. Although efficiency is a *must*, equity considerations are also becoming part of the discussion. We believe that our results are important for practitioners because we show explicitly the trade-offs between several criteria, including some criteria that look for equitable solutions from several perspectives. Even if a particular solution and operating conditions are in agreement with the regulation that applies (in particular regarding minimum coverage), the solution can also cause big differences between individual stakeholders. Our modeling framework and solution methodology allows for analizing/optimizing several performance indicators, making it possible to compare different planning perspectives.

A key component of our modeling and solution approach is the integrated location/allocation framework. We have shown that the use of a-priori dispatching rules, and in particular to send the closest available ambulance, generates good (optimal in most cases) solutions only if we are using efficiency criteria. If fairness considerations are included then the joint approach generates better solutions. This is also an important result for practitioners looking at improving the planning of their EMS systems including perspectives other than efficiency. Although sending the closest ambulance available is the most commonly used dispatching rule, we have shown that deviating from that policy can have benefits, in the form of more equitable solutions. The trade-offs provide elements for negotiating the right balance between efficiency and fairness.

5.3 Limitations and future work

We conclude our work by mentioning some of the limitations that we have identified and that at the same time are areas for potential extension of our work. There are several characteristics of real world EMS systems that we have left out from our modeling and solution approach. For instance, it is a typical practice to use a priority classification system for the calls that each EMS receives. We have assumed that there is only one type of call (or that objectives are the same for all types of calls), in fact focusing on high priority calls (or life threatening situations), assuming that if our solutions perform well for these critical calls, the performance for less critical incidents should be also reasonable good. The inclusion of several call types as part of the joint model and subsequently solution approach can lead to a closer representation of the system, although the resulting model will be also harder to solve.

EMS systems can also have different types of ambulances/personnel, such as basic units that are used only for transportation, and more advanced units that can be used to perform medical procedures on site or while in transit to a hospital. The use of different types of ambulances can be combined to identifying the priority of each call, so that different priority also means using one or another type of ambulance. We have assumed that all the ambulances available were of the same type, which is a simplification of the system nonetheless commonly used for planning purposes.

A common practice in real world EMS systems is the relocation of ambulances, or dynamic

allocation of ambulances. This means that instead of determining a unique location for an ambulance, it can be changed over time depending upon the behavior of the system. The relocation can be the result of identifying that the demand is changing over time, exhibiting particular patterns. In that case one approach could be running a location model for each period, and then figuring out a way to move the ambulances between locations (if required) according to the results for each period. A more extreme approach to relocation is to move the ambulances in real time, so that whenever an ambulance leaves its current base to perform a service, the remaining idle ambulances can be relocated. In the latter case a very fast algorithm is required to determine the best relocation possible. Our approach could not be used in such a scenario because the solution times would be too long.

In order to better understand some of the trade-offs we have performed bi-objective optimization experiments with our last solution framework. A potential extension to our work would be to use a pure multi-objective optimization approach, allowing for the generation of the whole Pareto front, potentially including more than two objectives. There are specialized multi-objective optimization algorithms, some of them based on meta-heuristics, hence this looks like a promising area to extend our work.

Bibliography

T. Andersson and P. Varbrand. Decision support tools for ambulance dispatch and relocation. *Journal of the Operational Research Society*, 58(2):195–201, 2006.

Alireza Boloori Arabani and Reza Zanjirani Farahani. Facility location dynamics: An overview of classifications and applications. *Computers and Industrial Engineering*, 62(1):408 – 420, 2012.

J.B. Atkinson, I.N. Kovalenko, N. Kuznetsov, and K.V. Mikhalevich. Heuristic methods for the analysis of a queueing system describing emergency medical service deployed along a highway. *Cybern. Syst. Anal.*, 42(3):379–391, 2006.

J.B. Atkinson, I.N. Kovalenko, N. Kuznetsov, and K.V. Mykhalevych. A hypercube queueing loss model with customer-dependent service rates. *European Journal of Operational Research*, 191 (1):223–239, 2008.

Haldun Aytug and Cem Saydam. Solving large-scale maximum expected covering location problems by genetic algorithms: A comparative study. *European Journal of Operational Research*, 141(3): 480–494, 2002.

Damitha Bandara, Maria E. Mayorga, and Laura A. McLay. Optimal dispatching strategies for emergency vehicles to increase patient survivability. *International Journal of Operational Research*, 15(2):195–214, 2012.

Susana Baptista and RuiCarvalho Oliveira. A case study on the application of an approximated hypercube model to emergency medical systems management. *Central European Journal of Operations Research*, 20(4):559–581, 2012.

Rajan Batta, June M. Dolan, and Nirup N. Krishnamurthy. The maximal expected covering location problem: Revisited. *Transportation Science*, 23(4):277–287, 1989.

Regina Benveniste. Solving the combined zoning and location problem for several emergency units. *The Journal of the Operational Research Society*, 36(5):433–450, 1985.

Oded Berman. Mean-variance location problems. *Transportation Science*, 24(4):287–293, 1990.

Oded Berman, Zvi Drezner, Arie Tamir, and George Wesolowsky. Optimal location with equitable loads. *Annals of Operations Research*, 167(1):307–325, 2009.

Dimitris Bertsimas, Vivek F. Farias, and Nikolaos Trichakis. The price of fairness. *Operation Research*, 59(1):17–31, January 2011.

Margaret L. Brandeau and Samuel S. Chiu. An overview of representative problems in location research. *Management Science*, 35(6):645–674, 1989.

Luce Brotcorne, Gilbert Laporte, and Fredric Semet. Ambulance location and relocation models. *European Journal of Operational Research*, 147(3):451–463, 2003.

Susan Budge, Armann Ingolfsson, and Erhan Erkut. Technical note: 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.

Alberto B. Calvo and David H. Marks. Location of health care facilities: An analytical approach. *Socio-Economic Planning Sciences*, 7(5):407–422, 1973.

Robert Carbone. Public facility location under stochastic demand. *INFOR*, 12(3):261–270, 1974.

Yolanda M. Carson and Rajan Batta. Locating an ambulance on the amherst campus of the state university of new york at buffalo. *Interfaces*, 20(5):43–49, 1990.

Grace M. Carter, Jan M. Chaiken, and Edward Ignall. Response areas for two emergency units. *Operations Research*, 20(3):571–594, 1972.

Sunarin Chanta, Maria E. Mayorga, Mary E. Kurz, and Laura A. McLay. The minimum p-envy location problem: a new model for equitable distribution of emergency resources. *IIE Transactions on Healthcare Systems Engineering*, 1(2):101–115, 2011a.

Sunarin Chanta, Maria E. Mayorga, and Laura A. McLay. Improving emergency service in rural areas: a bi-objective covering location model for ems systems. *Annals of Operations Research*, pages 1–27, 2011b.

Kenneth R. Chelst and Ziv Barlach. Multiple unit dispatches in emergency services: Models to estimate system performance. *Management Science*, 27(12):1390–1409, 1981.

Fernando Chiyoshi, Roberto D. Galvao, and Reinaldo Morabito. Modelo hipercubo: Analise e resultados para o caso de servidores nao-homogeneos. *Pesquisa Operacional*, 21:199–218, 2001.

Richard Church and Charles R. ReVelle. The maximal covering location problem. *Papers in Regional Science*, 32(1):101–118, 1974.

Francisco Correa, Antonio-Augusto Chaves, and Luiz Antonio Nogueira Lorena. Hybrid heuristics for the probabilistic maximal covering location-allocation problem. *Operational Research*, 7: 323–343, 2007. ISSN 1109-2858.

R. A. Cuninghame-Green and G. Harries. *Nearest-neighbour rules for emergency services*. National Emergency Training Center, Emmitsburg, MD, 1988.

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

Tammy Drezner. Location of casualty collection points. *Environment and Planning C: Government and Policy*, 22(6):899–912, 2004.

Tammy Drezner and Zvi Drezner. Equity models in planar location. *Computational Management Science*, 4(1):1–16, 2007.

Tammy Drezner and Zvi Drezner. A note on equity across groups in facility location. *Naval Research Logistics (NRL)*, 58(7):705–711, 2011.

Tammy Drezner, Zvi Drezner, and Jeffery Guyse. Equitable service by a facility: Minimizing the gini coefficient. *Comput. Oper. Res.*, 36(12):3240–3246, December 2009.

Z. Drezner, J. F. Thisse, and G. O. Wesolowsky. The minimax-min location problem. *Journal of Regional Science*, 26(1):87–101, 1986.

Horst A. Eiselt, Gilbert Laporte, and Administration University of New Brunswick. Faculty of. *Objectives in location problems*. Faculty of Administration, University of New Brunswick, Fredericton, NB, 1995.

Gnes Erdogan, Erhan Erkut, Armann Ingolfsson, and Gilbert Laporte. Scheduling ambulance crews for maximum coverage. *JORS*, 61(4):543–550, 2010.

Erhan Erkut. Inequality measures for location problems. *Location Science*, 1(3):199–217, 1993.

Erhan Erkut, Armann Ingolfsson, and G. Erdogan. Ambulance location for maximum survival. *Naval Research Logistics*, 55(1):42–58, 2008.

Inmaculada Espejo, Alfredo Marin, Justo Puerto, and Antonio M. Rodriguez. A comparison of formulations and solution methods for the minimum-envy location problem. *Comput. Oper. Res.*, 36(6):1966–1981, 2009.

Reza Zanjirani Farahani, Nasrin Asgari, Nooshin Heidari, Mahtab Hosseininia, and Mark Goh. Covering problems in facility location: A review. *Computers and Industrial Engineering*, 62(1): 368 – 407, 2012.

Stefan Felder and Henrik Brinkmann. Spatial allocation of emergency medical services: minimising the death rate or providing equal access? *Regional Science and Urban Economics*, 32(1):27–45, 2002.

Roberto D. Galvao and Reinaldo Morabito. Emergency service systems: The use of the hypercube queueing model in the solution of probabilistic location problems. *International Transactions in Operational Research*, 15(5):525–549, 2008.

Roberto D. Galvao, Fernando Y. Chiyoshi, and Reinaldo Morabito. Towards unified formulations and extensions of two classical probabilistic location models. *Computers and Operations Research*, 32(1):15–33, 2005.

Michel Gendreau, Gilbert Laporte, and Fredric Semet. Solving an ambulance location model by tabu search. *Location Science*, 5(2):75–88, 1997.

Michel Gendreau, Gilbert Laporte, Fredric Semet, Universit De Montral, Succursale Centre-ville, Montral Hc J, and Montral Hc J. A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Computing*, 27:1641–1653, 2001.

Nikolas Geroliminis, Matthew G. Karlaftis, and Alexander Skabardonis. A spatial queuing model for the emergency vehicle districting and location problem. *Transportation Research Part B: Methodological*, 43(7):798–811, 2009.

Nikolas Geroliminis, Konstantinos Kepaptsoglou, and Matthew G. Karlaftis. A hybrid hypercube and genetic algorithm approach for deploying many emergency response mobile units in an urban network. *European Journal of Operational Research*, 210(2):287–300, 2011.

Fred Glover. Future paths for integer programming and links to artificial intelligence. *Comput. Oper. Res.*, 13(5):533–549, May 1986.

Fred Glover. Tabu search part i. *ORSA Journal on Computing*, 1(3):190–206, 1989.

David Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional, 1989.

Jeffrey Goldberg. Operations research models for the deployment of emergency services vehicles. *EMS Management Journal*, 1(1):20–39, 2004.

Jeffrey Goldberg and Luis Paz. Locating emergency vehicle bases when service time depends on call location. *Transportation Science*, 25(4):264–280, 1991.

S. L. Hakimi. Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations Research*, 12(3):450–459, 1964.

P. Hansen. The Steepest Ascent Mildest Descent Heuristic for Combinatorial Programming. In *Proceedings of the Congress on Numerical Methods in Combinatorial Optimization*, Capri, Italy, 1986.

Robert Harder. Open ts - java tabu search package. `http://www.coin-or.org/Ots`, 2013.

Kathleen Hogan and Charles Revelle. Concepts and applications of backup coverage. *Management Science*, 32(11):1434–1444, 1986.

J. H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975.

J. N. Hooker and H. P. Williams. Combining equity and utilitarianism in a mathematical programming model. *Management Science*, 2012.

Ana Paula Iannoni and Reinaldo Morabito. A multiple dispatch and partial backup hypercube queuing model to analyze emergency medical systems on highways. *Transportation Research Part E: Logistics and Transportation Review*, 43(6):755–771, 2007.

Ana Paula Iannoni, Reinaldo Morabito, and Cem Saydam. A hypercube queueing model embedded into a genetic algorithm for ambulance deployment on highways. *Ann Op Res*, 157(1):207–224, 2008.

Ana Paula Iannoni, Reinaldo Morabito, and Cem Saydam. Optimizing large-scale emergency medical system operations on highways using the hypercube queuing model. *Socio-Economic Planning Sciences*, 45(3):105–117, 2011.

Armann Ingolfsson, Susan Budge, and Erhan Erkut. Optimal ambulance location with random delays and travel times. *Health Care Management Science*, 11:262–274, 2008. ISSN 1386-9620.

Jorge H. Jaramillo, Joy Bhadury, and Rajan Batta. On the use of genetic algorithms to solve location problems. *Computers and Operations Research*, 29(6):761–779, 2002.

James P. Jarvis. Optimal assignments in a markovian queueing system. *Computers and Operations Research*, 8(1):17–23, 1981.

James P. Jarvis. Approximating the equilibrium behavior of multi-server loss systems. *Management Science*, 31(2):235–239, 1985.

Hongzhong Jia, Fernando Ordonez, and Maged Dessouky. A modeling framework for facility location of medical services for large-scale emergencies. *IIE Transactions*, 39(1):41–55, 2007a.

Hongzhong Jia, Fernando Ordonez, and Maged Dessouky. Solution approaches for facility location of medical supplies for large-scale emergencies. *Computers and Industrial Engineering*, 52(2):257 – 276, 2007b. ISSN 0360-8352.

Jrg Kalcsics, Stefan Nickel, Justo Puerto, and Antonio Rodriguez-Chia. The ordered capacitated facility location problem. *TOP*, 18(1):203–222, 2010.

Michael N. Katehakis and Alan Levine. Allocation of distinguishable servers. *Computers and Operations Research*, 13(1):85–93, 1986.

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

Richard C. Larson. Approximating the performance of urban emergency service systems. *Operations Research*, 23(5):845–868, 1975.

Richard C. Larson and Amadeo R. Odoni. *Urban Operations Research*. Prentice Hall, http://web.mit.edu/urban_or_book/www/book/, 1981.

Philip D. Leclerc, Laura A. McLay, and Maria E. Mayorga. *Modeling Equity for Allocating Public Resources Community-Based Operations Research*, volume 167 of *International Series in Operations Research and Management Science*, pages 97–118. Springer New York, 2012.

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

Xueping Li, Zhaoxia Zhao, Xiaoyan Zhu, and Tami Wyatt. Covering models and optimization techniques for emergency response facility location and planning: a review. *Mathematical Methods of Operations Research*, pages 1–30, July 2011.

Oded Maimon. The variance equity measure in locational decision theory. *Annals of Operations Research*, 6(5):147–160, 1986.

Oded Maimon. An algorithm for the lorenz measure in locational decisions on trees. *Journal of Algorithms*, 9(4):583–596, 1988.

Vladimir Marianov and Charles ReVelle. The capacitated standard response fire protection siting problem: Deterministic and probabilistic models. *Annals of Operations Research*, 40(1):303–322, 1992.

Vladimir Marianov and Charles ReVelle. The queueing maximal availability location problem: A model for the siting of emergency vehicles. *European Journal of Operational Research*, 93(1): 110–120, 1996.

Alfredo Marn. The discrete facility location problem with balanced allocation of customers. *European Journal of Operational Research*, 210(1):27–38, 2011.

Michael T. Marsh and David A. Schilling. Equity measurement in facility location analysis: A review and framework. *Location Science*, 3(1):61, 1995.

T. McCallion. The great ambulance response time debate, February 2012. EMS Insider, Online Journal.

Laura A. McLay and Maria E. Mayorga. Evaluating emergency medical service performance measures. *Health Care Management Science*, 13(2):124–136, 2010.

Laura A. McLay and Maria E. Mayorga. Evaluating the impact of performance goals on dispatching decisions in emergency medical service. *IIE Transactions on Healthcare Systems Engineering*, 1 (3):185–196, 2011.

Laura A. McLay and Maria E. Mayorga. A dispatching model for server-to-customer systems that balances efficiency and equity. *Manufacturing and Service Operations Management*, 2012.

Klaus Meffert, Javier Meseguer, Enrique DMart, Vos Jerry, and Neil Rotstan. Jgap - java genetic algorithms and genetic programming package. `http://jgap.sf.net`, 2012.

F. C. Mendonca and R. Morabito. Analysing emergency medical service ambulance deployment on a brazilian highway using the hypercube model. *The Journal of the Operational Research Society*, 52(3):261–270, 2001.

Nenad Mladenovic, Jack Brimberg, Pierre Hansen, and Jos A. Moreno-Prez. The p-median problem: A survey of metaheuristic approaches. *European Journal of Operational Research*, 179(3):927–939, 2007.

Douglas Montgomory. *Design and analysis of experiments*. Wiley, Hoboken, NJ, 2008.

G. F. Mulligan. Equity measures and facility location. *Papers on regional science*, 70:345–365, 1991.

Wodzimierz Ogryczak. Inequality measures and equitable approaches to location problems. *European Journal of Operational Research*, 122(2):374–391, 2000.

Wodzimierz Ogryczak. Inequality measures and equitable locations. *Annals of Operations Research*, 167(1):61–86, 2009.

Hari K. Rajagopalan, F. Elizabeth Vergara, Cem Saydam, and Jing Xiao. Developing effective meta-heuristics for a probabilistic location model via experimental design. *European Journal of Operational Research*, 177(1):83–101, 2007.

Hari K. Rajagopalan, Cem Saydam, and Jing Xiao. A multiperiod set covering location model for dynamic redeployment of ambulances. *Computers and Operations Research*, 35(3):814 – 826, 2008.

J. Reese. Methods for solving the p-median problem: An annotated bibliography. Technical report, Trinity University, 2005. Technical report No. 96 - Math Technical Reports.

John F. Repede and John J. Bernardo. Developing and validating a decision support system for locating emergency medical vehicles in louisville, kentucky. *European Journal of Operational Research*, 75(3):567–581, 1994.

Ron Roth. EMS quality improvement made ridiculously easy, 2005. Department of Emergency Medicine. Pitssburg.

Roco Sanchez-Mangas, Antonio Garca-Ferrrer, Aranzazu de Juan, and Antonio Martn Arroyo. The probability of death in road traffic accidents. how important is a quick medical response? *Accident Analysis and Prevention*, 42(4):1048–1056, 2010.

David Schilling, D. Jack Elzinga, Jared Cohon, Richard Church, and Charles ReVelle. The team/fleet models for simultaneous facility and equipment siting. *Transportation Science*, 13 (2):163–175, 1979.

S.S. Radiah Shariff, Noor Hasnah Moin, and Mohd Omar. Location allocation modeling for health-care facility planning in malaysia. *Computers and Industrial Engineering*, 62(4):1000 – 1010, 2012.

Lawrence V. Snyder. Facility location under uncertainty: A review. *IIE Transactions*, 38:547–564, 2004.

Deborah A. Stone. *Policy paradox : the art of political decision making*. Norton, New York, 2002.

Constantine Toregas, Ralph Swain, Charles ReVelle, and Lawrence Bergman. The location of emergency service facilities. *Operations Research*, 19(6):1363–1373, 1971.

Héctor Toro-Díaz, Maria E. Mayorga, Chanta Sunarin, and Laura A. McLay. Joint location and dispatching decisions for emergency medical services. *Computers and Industrial Engineering*, 64: 917–928, 2013.

M. Jalali Varnamkhasti. Overview of the algorithms for solving the p-median facility location problems. *Advanced Studies in Biology*, 4(2):49 55, 2012.